Numerical Matrix Analysis

Numerical Matrix Analysis

Linear Systems *and* Least Squares

Ilse C. F. Ipsen

North Carolina State University
Raleigh, North Carolina

10 9 8 7 6 5 4 3 2 1

Library of Congress Cataloging-in-Publication Data

Ipsen, Ilse C. F.
Numerical matrix analysis : linear systems and least squares / Ilse C. F. Ipsen.
p. cm.
Includes index.
ISBN 978-0-898716-76-4
1. Least squares. 2. Linear systems. 3. Matrices. I. Title.
QA275.I67 2009
511'.42--dc22

2009008514

To Stan

Contents

Preface

This book was written for a first-semester graduate course in matrix theory at North Carolina State University. The students come from applied and pure mathematics, all areas of engineering, and operations research. The book is self-contained. The main topics covered in detail are linear system solution, least squares problems, and singular value decomposition.

My objective was to present matrix analysis in the context of numerical computation, with numerical conditioning of problems, and numerical stability of algorithms at the forefront. I tried to present the material at a basic level, but in a mathematically rigorous fashion.

Main Features. This book differs in several regards from other numerical linear algebra textbooks.

- *Systematic development of numerical conditioning.*
  Perturbation theory is used to determine sensitivity of problems as well as numerical stability of algorithms, and the perturbation results built on each other.
  For instance, a condition number for matrix multiplication is used to derive a residual bound for linear system solution (Fact 3.5), as well as a least squares bound for perturbations on the right-hand side (Fact 5.11).
- *No floating point arithmetic.*
  There is hardly any mention of floating point arithmetic, for three main reasons. First, sensitivity of numerical problems is, in general, not caused by arithmetic in finite precision. Second, many special-purpose devices in engineering applications perform fixed point arithmetic. Third, sensitivity is an issue even in symbolic computation, when input data are not known exactly.
- *Numerical stability in exact arithmetic.*
  A simplified concept of numerical stability is introduced to give quantitative intuition, while avoiding tedious roundoff error analyses. The message is that unstable algorithms come about if one decomposes a problem into ill-conditioned subproblems.
  Two bounds for this simpler type of stability are presented for general direct solvers (Facts 3.14 and 3.17). These bounds imply, in turn, stability bounds for solvers based on the following factorizations: LU (Corollary 3.22), Cholesky (Corollary 3.31), and QR (Corollary 3.33).

- *Simple derivations.*
  The existence of a QR factorization for nonsingular matrices is deduced very simply from the existence of a Cholesky factorization (Fact 3.32), without any commitment to a particular algorithm such as Householder or Gram–Schmidt.
  A new intuitive proof is given for the optimality of the singular value decomposition (Fact 4.13), based on the distance of a matrix from singularity.
  I derive many relative perturbation bounds with regard to the perturbed solution, rather than the exact solution. Such bounds have several advantages: They are computable; they give rise to intermediate absolute bounds (which are useful in the context of fixed point arithmetic); and they are easy to derive.
  Especially for full rank least squares problems (Fact 5.14), such a perturbation bound can be derived fast, because it avoids the Moore–Penrose inverse of the perturbed matrix.
- *High-level view of algorithms.*
  Due to widely available high-quality mathematical software for small dense matrices, I believe that it is not necessary anymore to present detailed implementations of direct methods in an introductory graduate text. This frees up time for analyzing the accuracy of the output.
- *Complex arithmetic.*
  Results are presented for complex rather than real matrices, because engineering problems can give rise to complex matrices. Moreover, limiting one's focus to real matrices makes it difficult to switch to complex matrices later on. Many properties that are often taken for granted in the real case no longer hold in the complex case.
- *Exercises.*
  The exercises contain many useful facts. A separate category of easier exercises, labeled with roman numerals, is appropriate for use in class.

Acknowledgments. I thank Nick Higham, Rizwana Rehman, Megan Sawyer, and Teresa Selee for providing helpful suggestions and all MA523 students for giving me the opportunity to develop the material in this book. It has been a pleasure working with the SIAM publishing staff, in particular with Sara Murphy, who made possible the publication of this book, and Lou Primus, who patiently and competently dealt with all my typesetting requests.

Ilse Ipsen
Raleigh, NC, USA
December 2008

References

Gene H. Golub and Charles F. Van Loan: *Matrix Computations*, Third Edition, Johns Hopkins Press, 1996

Nicholas J. Higham: *Accuracy and Stability of Numerical Algorithms*, Second Edition, SIAM, 2002

Roger A. Horn and Charles A. Johnson: *Matrix Analysis*, Cambridge University Press, 1985

Peter Lancaster and Miron Tismenetsky: *The Theory of Matrices*, Second Edition, Academic Press, 1985

Carl D. Meyer: *Matrix Analysis and Applied Linear Algebra*, SIAM, 2000

Gilbert Strang: *Linear Algebra and Its Applications*, Third Edition, Harcourt Brace Jovanovich, 1988

Introduction

The goal of this book is to help you understand the *sensitivity* of matrix computations to errors in the input data. There are two important reasons for such errors.

(i) *Input data may not be known exactly.*
For instance, your weight on the scale tends to be 125 pounds, but may change to 126 or 124 depending where you stand on the scale. So, you are sure that the leading digits are 12, but you are not sure about the third digit. Therefore the third digit is considered to be in error.

(ii) *Arithmetic operations can produce errors.*
Arithmetic operations may not give the exact result when they are carried out in finite precision, e.g., in floating point arithmetic or in fixed point arithmetic. This happens, for instance, when $1/3$ is computed as .33333333.

There are matrix computations that are sensitive to errors in the input. Consider the system of linear equations

$$\begin{aligned}\frac{1}{3}x_1 + \frac{1}{3}x_2 &= 1,\\ \frac{1}{3}x_1 + .3x_2 &= 0,\end{aligned}$$

which has the solution $x_1 = -27$ and $x_2 = 30$. Suppose we make a small change in the second equation and change the coefficient from .3 to $\frac{1}{3}$. The resulting linear system

$$\begin{aligned}\frac{1}{3}x_1 + \frac{1}{3}x_2 &= 1,\\ \frac{1}{3}x_1 + \frac{1}{3}x_2 &= 0\end{aligned}$$

has no solution. A small change in the input causes a drastic change in the output, i.e., the total loss of the solution. Why did this happen? How can we predict that something like this can happen? That is the topic of this book.

1. Matrices

We review the basic matrix operations.

1.1 What Is a Matrix?

An array of numbers

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{pmatrix}$$

with m rows and n columns is an $m \times n$ matrix. Element a_{ij} is located in position (i,j). The elements a_{ij} are *scalars*, namely, real or complex numbers. The set of real numbers is $\mathbb{R}$, and the set of complex numbers is $\mathbb{C}$.

We write $A \in \mathbb{R}^{m\times n}$ if A is an $m \times n$ matrix whose elements are real numbers, and $A \in \mathbb{C}^{m\times n}$ if A is an $m \times n$ matrix whose elements are complex numbers. Of course, $\mathbb{R}^{m\times n} \subset \mathbb{C}^{m\times n}$. If $m = n$, then we say that A is a *square matrix* of order n.

For instance,

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}$$

is a 2×4 matrix with elements $a_{13} = 3$ and $a_{24} = 8$.

Vectors. A *row vector* $y = \begin{pmatrix} y_1 & \dots & y_m \end{pmatrix}$ is a $1 \times m$ matrix, i.e., $y \in \mathbb{C}^{1\times m}$. A *column vector*

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

is an $n \times 1$ matrix, i.e., $x \in \mathbb{C}^{n\times 1}$ or shorter, $x \in \mathbb{C}^n$. If the elements of x are real, then $x \in \mathbb{R}^n$.

Submatrices. Sometimes we need only those elements of a matrix that are situated in particular rows and columns.

Definition 1.1. *Let $A \in \mathbb{C}^{m\times n}$ have elements a_{ij}. If $1 \le i_1 < i_2 < \cdots < i_k \le m$ and $1 \le j_1 < j_2 < \cdots < j_l \le n$, then the $k \times l$ matrix*

$$\begin{pmatrix} a_{i_1,j_1} & a_{i_1,j_2} & \dots & a_{i_1,j_l} \\ a_{i_2,j_1} & a_{i_2,j_2} & \dots & a_{i_2,j_l} \\ \vdots & \vdots & & \vdots \\ a_{i_k,j_1} & a_{i_k,j_2} & \dots & a_{i_k,j_l} \end{pmatrix}$$

is called a submatrix *of A. The submatrix is a* principal submatrix *if it is square and its diagonal elements are diagonal elements of A, that is, $k = l$ and $i_1 = j_1$, $i_2 = j_2, \dots, i_k = j_k$.*

Example. If

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix},$$

then the following are submatrices of A:

$$\begin{pmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 4 & 6 \end{pmatrix}, \qquad \begin{pmatrix} a_{21} & a_{23} \end{pmatrix} = \begin{pmatrix} 4 & 6 \end{pmatrix}, \qquad \begin{pmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \\ a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 2 & 3 \\ 5 & 6 \\ 8 & 9 \end{pmatrix}.$$

The submatrix

$$\begin{pmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{pmatrix} = \begin{pmatrix} 1 & 3 \\ 7 & 9 \end{pmatrix}$$

is a principal matrix of A, as are the diagonal elements a_{11}, a_{22}, a_{33}, and A itself. ∎

Notation. Most of the time we will use the following notation:

- Matrices: uppercase Roman or Greek letters, e.g., A, Λ.
- Vectors: lowercase Roman letters, e.g., x, y.
- Scalars: lowercase Greek letters, e.g., α;
  or lowercase Roman with subscripts, e.g., x_i, a_{ij}.
- Running variables: i, j, k, l, m, and n.

The elements of the matrix A are called a_{ij} or A_{ij}, and the elements of the vector x are called x_i.

Zero Matrices. The *zero matrix* $0_{m\times n}$ is the $m \times n$ matrix all of whose elements are zero. When m and n are clear from the context, we also write 0. We say $A = 0$

if all elements of the matrix A are equal to zero. The matrix A is nonzero, $A \neq 0$, if at least one element of A is nonzero.

Identity Matrices. The *identity matrix* of order n is the real square matrix

$$I_n = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}$$

with ones on the diagonal and zeros everywhere else (instead of writing many zeros, we often write blanks). In particular, $I_1 = 1$. When n is clear from the context, we also write I.

The columns of the identity matrix are also called *canonical vectors* e_i. That is, $I_n = \begin{pmatrix} e_1 & e_2 & \ldots & e_n \end{pmatrix}$, where

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \qquad e_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \qquad \ldots, \qquad e_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}.$$

Exercises

(i) Hilbert Matrix.
A square matrix of order n whose element in position (i,j) is $\frac{1}{i+j-1}$, $1 \leq i,j \leq n$, is called a *Hilbert* matrix.
Write down a Hilbert matrix for $n = 5$.

(ii) Toeplitz Matrix.
Given $2n-1$ scalars α_k, $-n+1 \leq k \leq n-1$, a matrix of order n whose element in position (i,j) is α_{j-i}, $1 \leq i,j \leq n$, is called a *Toeplitz* matrix.
Write down the Toeplitz matrix of order 3 when $\alpha_i = i$, $-2 \leq i \leq 2$.

(iii) Hankel Matrix.
Given $2n-1$ scalars α_k, $0 \leq k \leq 2n-2$, a matrix of order n whose element in position (i,j) is α_{i+j-2}, $1 \leq i,j \leq n$, is called a *Hankel* matrix.
Write down the Hankel matrix of order 4 for $\alpha_i = i$, $0 \leq i \leq 6$.

(iv) Vandermonde Matrix.
Given n scalars α_i, $1 \leq i \leq n$, a matrix of order n whose element in position (i,j) is α_i^{j-1}, $1 \leq i,j \leq n$, is called a *Vandermonde* matrix. Here we interpret $\alpha_i^0 = 1$ even for $\alpha_i = 0$. The numbers α_i are also called *nodes* of the Vandermonde matrix.
Write down the Vandermonde matrix of order 4 when $\alpha_i = i$, $1 \leq i \leq 3$, and $\alpha_4 = 0$.

(v) Is a square zero matrix a Hilbert, Toeplitz, Hankel, or Vandermonde matrix?

(vi) Is the identity matrix a Hilbert, Toeplitz, Hankel, or Vandermonde matrix?

(vii) Is a Hilbert matrix a Hankel matrix or a Toeplitz matrix?

1.2 Scalar Matrix Multiplication

Each element of the matrix is multiplied by a scalar. If $A \in \mathbb{C}^{m\times n}$ and λ a scalar, then the elements of the *scalar matrix product* $\lambda A \in \mathbb{C}^{m\times n}$ are

$$(\lambda A)_{ij} \equiv \lambda a_{ij}.$$

Multiplying the matrix $A \in \mathbb{C}^{m\times n}$ by the scalar zero produces a zero matrix,

$$0\,A = 0_{m\times n},$$

where the first zero is a scalar, while the second zero is a matrix with the same number of rows and columns as A. Scalar matrix multiplication is associative,

$$(\lambda\mu)\,A = \lambda\,(\mu A).$$

Scalar matrix multiplication by -1 corresponds to negation,

$$-A \equiv (-1)\,A.$$

Exercise

(i) Let $x \in \mathbb{C}^n$ and $\alpha \in \mathbb{C}$. Prove: $\alpha x = 0$ if and only if $\alpha = 0$ or $x = 0$.

1.3 Matrix Addition

Corresponding elements of two matrices are added. The matrices must have the same number of rows and the same number of columns. If A and $B \in \mathbb{C}^{m\times n}$, then the elements of the *sum* $A+B \in \mathbb{C}^{m\times n}$ are

$$(A+B)_{ij} \equiv a_{ij} + b_{ij}.$$

Properties of Matrix Addition.

- Adding the zero matrix does not change anything. That is, for any $m \times n$ matrix A,
$$0_{m\times n} + A = A + 0_{m\times n} = A.$$
- Matrix addition is commutative,
$$A+B = B+A.$$
- Matrix addition is associative,
$$(A+B)+C = A+(B+C).$$
- Matrix addition and scalar multiplication are distributive,
$$\lambda\,(A+B) = \lambda A + \lambda B, \qquad (\lambda+\mu)\,A = \lambda A + \mu A.$$

One can use the above properties to save computations. For instance, computing $\lambda A + \lambda B$ requires twice as many operations as computing $\lambda(A+B)$. In the special case $B = -C$, computing $(A+B)+C$ requires two matrix additions, while $A+(B+C) = A+0 = A$ requires no work.

A special type of addition is the sum of scalar vector products.

Definition 1.2. *A* linear combination *of m column (or row) vectors $v_1, \ldots, v_m$, $m \geq 1$, is*

$$\alpha_1 v_1 + \cdots + \alpha_m v_m,$$

where the scalars $\alpha_1, \ldots, \alpha_m$ are the coefficients.

Example. Any vector in $\mathbb{R}^n$ or $\mathbb{C}^n$ can be represented as a linear combination of canonical vectors,

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n. \qquad \blacksquare$$

1.4 Inner Product (Dot Product)

The product of a row vector times an equally long column vector produces a single number. If

$$x = \begin{pmatrix} x_1 & \ldots & x_n \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

then the *inner product* of x and y is the scalar

$$xy = x_1 y_1 + \cdots + x_n y_n.$$

Example. A sum of n scalars a_i, $1 \leq i \leq n$, can be represented as an inner product of two vectors with n elements each,

$$\sum_{j=1}^{n} a_j = \begin{pmatrix} a_1 & a_2 & \ldots & a_n \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}. \qquad \blacksquare$$

Example. A polynomial $p(\alpha) = \sum_{j=0}^{n} \lambda_j \alpha^j$ of degree n can be represented as an inner product of two vectors with $n+1$ elements each,

$$p(\alpha) = \begin{pmatrix} 1 & \alpha & \ldots & \alpha^n \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} \lambda_0 & \lambda_1 & \ldots & \lambda_n \end{pmatrix} \begin{pmatrix} 1 \\ \alpha \\ \vdots \\ \alpha^n \end{pmatrix}. \qquad \blacksquare$$

Exercise

(i) Let $n \geq 1$ be an integer. Represent $n(n+1)/2$ as an inner product of two vectors with n elements each.

1.5 Matrix Vector Multiplication

The product of a matrix and a vector is again a vector. There are two types of matrix vector multiplications: matrix times column vector and row vector times matrix.

Matrix Times Column Vector. The product of matrix times column vector is again a column vector. We present two ways to describe the operations that are involved in a matrix vector product. Let $A \in \mathbb{C}^{m\times n}$ with rows r_j and columns c_j, and let $x \in \mathbb{C}^n$ with elements x_j,

$$A = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix} = \begin{pmatrix} c_1 & \dots & c_n \end{pmatrix}, \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

View 1: Ax is a column vector of inner products, so that element j of Ax is the inner product of row r_j with x,

$$Ax = \begin{pmatrix} r_1 x \\ \vdots \\ r_m x \end{pmatrix}.$$

View 2: Ax is a linear combination of columns

$$Ax = c_1 x_1 + \cdots + c_n x_n.$$

The vectors in the linear combination are the columns c_j of A, and the coefficients are the elements x_j of x.

Example. Let

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

The first view shows that Ae_2 is equal to column 2 of A. That is,

$$Ae_2 = 0 \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + 1 \cdot \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix} + 0 \cdot \begin{pmatrix} 0 \\ 0 \\ 3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}.$$

The second view shows that the first and second elements of Ae_2 are equal to zero. That is,

$$Ax = \begin{pmatrix} \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} e_2 \\ \begin{pmatrix} 0 & 0 & 0 \end{pmatrix} e_2 \\ \begin{pmatrix} 1 & 2 & 3 \end{pmatrix} e_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}. \qquad \blacksquare$$

Example. Let A be the Toeplitz matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}.$$

The first view shows that the last element of Ax is equal to zero. That is,

$$Ax = \begin{pmatrix} \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 1 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix} \end{pmatrix} x = \begin{pmatrix} x_2 \\ x_3 \\ x_4 \\ 0 \end{pmatrix}. \qquad \blacksquare$$

Row Vector Times Matrix. The product of a row vector times a matrix is a row vector. There are again two ways to think about this operation. Let $A \in \mathbb{C}^{m \times n}$ with rows r_j and columns c_j, and let $y \in \mathbb{C}^{1 \times m}$ with elements y_j,

$$A = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix} = \begin{pmatrix} c_1 & \dots & c_n \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 & \dots & y_m \end{pmatrix}.$$

View 1: yA is a row vector of inner products, where element j of yA is an inner product of y with the column c_j,

$$yA = \begin{pmatrix} yc_1 & \dots & yc_n \end{pmatrix}.$$

View 2: yA is a linear combination of rows of A,

$$yA = y_1 r_1 + \cdots + y_m r_m.$$

The vectors in the linear combination are the rows r_j of A, and the coefficients are the elements y_j of y.

Exercises

(i) Show that Ae_j is the jth column of the matrix A.

(ii) Let A be an $m \times n$ matrix and e the $n \times 1$ vector of all ones. What does Ae do?

(iii) Let $\alpha_1 v_1 + \cdots + \alpha_m v_m = 0$ be a linear combination of vectors $v_1, \ldots, v_m$. Prove: If one of the coefficients α_j is nonzero, then one of the vectors can be represented as a linear combination of the other vectors.

1. Let $A, B \in \mathbb{C}^{m\times n}$. Prove: $A = B$ if and only if $Ax = Bx$ for all $x \in \mathbb{C}^n$.

1.6 Outer Product

The product of a column vector times a row vector gives a matrix (this is not to be confused with an inner product which produces a single number). If

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 & \ldots & y_n \end{pmatrix},$$

then the *outer product* of x and y is the $m \times n$ matrix

$$xy = \begin{pmatrix} x_1 y_1 & \ldots & x_1 y_n \\ \vdots & & \vdots \\ x_m y_1 & \ldots & x_m y_n \end{pmatrix}.$$

The vectors in an outer product are allowed to have different lengths. The columns of xy are multiples of each other, and so are the rows. That is, each column of xy is a multiple of x,

$$xy = \begin{pmatrix} xy_1 & \ldots & xy_n \end{pmatrix},$$

and each row of xy is a multiple of y,

$$xy = \begin{pmatrix} x_1 y \\ \vdots \\ x_m y \end{pmatrix}.$$

Example. A Vandermonde matrix of order n all of whose nodes are the same, e.g., equal to α, can be represented as the outer product

$$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \begin{pmatrix} 1 & \alpha & \ldots & \alpha^{n-1} \end{pmatrix}.$$ ■

Exercise

(i) Write the matrix below as an outer product:

$$\begin{pmatrix} 4 & 5 \\ 8 & 10 \\ 12 & 15 \end{pmatrix}.$$

1.7 Matrix Multiplication

The product of two matrices A and B is defined if the number of columns in A is equal to the number of rows in B. Specifically, if $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times p}$, then $AB \in \mathbb{C}^{m\times p}$. We can describe matrix multiplication in four different ways. Let $A \in \mathbb{C}^{m\times n}$ with rows a_j, and let $B \in \mathbb{C}^{n\times p}$ with columns b_j:

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix}, \qquad B = \begin{pmatrix} b_1 & \dots & b_p \end{pmatrix}.$$

View 1: AB is a block row vector of matrix vector products. The columns of AB are matrix vector products of A with columns of B,

$$AB = \begin{pmatrix} Ab_1 & \dots & Ab_p \end{pmatrix}.$$

View 2: AB is a block column vector of matrix vector products, where the rows of AB are matrix vector products of the rows of A with B,

$$AB = \begin{pmatrix} a_1 B \\ \vdots \\ a_m B \end{pmatrix}.$$

View 3: The elements of AB are inner products, where element (i,j) of AB is an inner product of row i of A with column j of B,

$$(AB)_{ij} = a_i b_j, \qquad 1 \le i \le m, \quad 1 \le j \le p.$$

View 4: If we denote by c_i the columns of A and r_i the rows of B,

$$A = \begin{pmatrix} c_1 & \dots & c_n \end{pmatrix}, \qquad B = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix},$$

then AB is a sum of outer products, $AB = c_1 r_1 + \cdots + c_n r_n$.

Properties of Matrix Multiplication.

- Multiplying by the identity matrix does not change anything. That is, for an $m \times n$ matrix A,
$$I_m\, A = A\, I_n = A.$$
- Matrix multiplication is associative,
$$A\,(BC) = (AB)\,C.$$

- Matrix multiplication and addition are distributive,
$$A\,(B+C) = AB + AC, \qquad (A+B)\,C = AC + BC.$$
- Matrix multiplication is *not* commutative.
For instance, if
$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},$$
then
$$AB = \begin{pmatrix} 0 & 2 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = BA.$$

Example. Associativity can save work. If
$$A = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}, \qquad C = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix},$$
then computing the product
$$(AB)\,C = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \\ 4 & 8 & 12 \\ 5 & 10 & 15 \end{pmatrix} \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}$$
requires more operations than
$$A\,(BC) = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix} \cdot 10. \qquad \blacksquare$$

Warning. Don't misuse associativity. For instance, if
$$A = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 4 \\ 5 & 5 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 2 & 3 \end{pmatrix}, \qquad C = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix},$$
it looks as if we could compute
$$A\,(BC) = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ 4 & 4 \\ 5 & 5 \end{pmatrix} \cdot 10.$$

However, the product ABC is not defined because AB is not defined (here we have to view BC as a 1×1 matrix rather than just a scalar). In a product ABC, all adjacent products AB and BC have to be defined. Hence the above option $A\,(BC)$ is not defined either.

Matrix Powers. A special case of matrix multiplication is the repeated multiplication of a square matrix by itself. If A is a nonzero square matrix, we define $A^0 \equiv I$, and for any integer $k > 0$,

$$A^k = \overbrace{A \ldots A}^{k \text{ times}} = A^{k-1}\, A = A\, A^{k-1}.$$

Definition 1.3. *A square matrix is*

- *involutory if $A^2 = I$,*
- *idempotent (or a projector) if $A^2 = A$,*
- *nilpotent if $A^k = 0$ for some integer $k > 0$.*

Example. For any scalar α,

$$\begin{pmatrix} 1 & \alpha \\ 0 & -1 \end{pmatrix} \quad \text{is involutory,}$$

$$\begin{pmatrix} 1 & \alpha \\ 0 & 0 \end{pmatrix} \quad \text{is idempotent,}$$

and

$$\begin{pmatrix} 0 & \alpha \\ 0 & 0 \end{pmatrix} \quad \text{is nilpotent.}$$

■

Exercises

(i) Which is the only matrix that is both idempotent and involutory?

(ii) Which is the only matrix that is both idempotent and nilpotent?

(iii) Let $x \in \mathbb{C}^{n\times 1}$, $y \in \mathbb{C}^{1\times n}$. When is xy idempotent? When is it nilpotent?

(iv) Prove: If A is idempotent, then $I - A$ is also idempotent.

(v) Prove: If A and B are idempotent and $AB = BA$, then AB is also idempotent.

(vi) Prove: A is involutory if and only if $(I - A)(I + A) = 0$.

(vii) Prove: If A is involutory and $B = \frac{1}{2}(I + A)$, then B is idempotent.

(viii) Let $x \in \mathbb{C}^{n\times 1}$, $y \in \mathbb{C}^{1\times n}$. Compute $(xy)^3 x$ using only inner products and scalar multiplication.

1. Fast Matrix Multiplication.
   One can multiply two complex numbers with only three real multiplications instead of four. Let $\alpha = \alpha_1 + \imath\alpha_2$ and $\beta = \beta_1 + \imath\beta_2$ be two complex numbers,

where $\imath^2 = -1$ and $\alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathbb{R}$. Writing

$$\alpha\beta = \alpha_1\beta_1 - \alpha_2\beta_2 + \imath\,[(\alpha_1+\alpha_2)(\beta_1+\beta_2) - \alpha_1\beta_1 - \alpha_2\beta_2]$$

shows that the complex product $\alpha\beta$ can be computed with three real multiplications: $\alpha_1\beta_1$, $\alpha_2\beta_2$, and $(\alpha_1+\beta_1)(\alpha_2+\beta_2)$.
Show that this approach can be extended to the multiplication AB of two complex matrices $A = A_1 + \imath A_2$ and $B = B_1 + \imath B_2$, where $A_1, A_2 \in \mathbb{R}^{m\times n}$ and $B_1, B_2 \in \mathbb{R}^{n\times p}$. In particular, show that no commutativity laws are violated.

1.8 Transpose and Conjugate Transpose

Transposing a matrix amounts to turning rows into columns and vice versa. If

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix},$$

then its transpose $A^T \in \mathbb{C}^{n\times m}$ is obtained by converting rows to columns,

$$A^T = \begin{pmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \vdots & \vdots & & \vdots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{pmatrix}.$$

There is a second type of transposition that requires more work when the matrix elements are complex numbers. A complex number α is written $\alpha = \alpha_1 + \imath\alpha_2$, where $\imath^2 = -1$ and $\alpha_1, \alpha_2 \in \mathbb{R}$. The complex conjugate of the scalar α is $\overline{\alpha} = \alpha_1 - \imath\alpha_2$.

If $A \in \mathbb{C}^{m\times n}$ is a matrix, its conjugate transpose $A^* \in \mathbb{C}^{n\times m}$ is obtained by converting rows to columns and, in addition, taking the complex conjugates of the elements,

$$A^* = \begin{pmatrix} \overline{a}_{11} & \overline{a}_{21} & \dots & \overline{a}_{m1} \\ \overline{a}_{12} & \overline{a}_{22} & \dots & \overline{a}_{m2} \\ \vdots & \vdots & & \vdots \\ \overline{a}_{1n} & \overline{a}_{2n} & \dots & \overline{a}_{mn} \end{pmatrix}.$$

Example. If

$$A = \begin{pmatrix} 1+2\imath & 5 \\ 3-\imath & 6 \end{pmatrix},$$

then

$$A^T = \begin{pmatrix} 1+2\imath & 3-\imath \\ 5 & 6 \end{pmatrix}, \qquad A^* = \begin{pmatrix} 1-2\imath & 3+\imath \\ 5 & 6 \end{pmatrix}.$$ ■

Example. We can express the rows of the identity matrix in terms of canonical vectors,

$$I_n = \begin{pmatrix} e_1^T \\ \vdots \\ e_n^T \end{pmatrix} = \begin{pmatrix} e_1^* \\ \vdots \\ e_n^* \end{pmatrix}. \qquad \blacksquare$$

Fact 1.4 (Properties of Transposition).

- For real matrices, the conjugate transpose and the transpose are identical. That is, if $A \in \mathbb{R}^{m\times n}$, then $A^* = A^T$.
- Transposing a matrix twice gives back the original,

$$(A^T)^T = A, \qquad (A^*)^* = A.$$

- Transposition does not affect a scalar, while conjugate transposition conjugates the scalar,

$$(\lambda A)^T = \lambda A^T, \qquad (\lambda A)^* = \bar{\lambda} A^*.$$

- The transpose of a sum is the sum of the transposes,

$$(A+B)^T = A^T + B^T, \qquad (A+B)^* = A^* + B^*.$$

- The transpose of a product is the product of the transposes with the factors in reverse order,

$$(AB)^T = B^T A^T, \qquad (AB)^* = B^* A^*.$$

Example. Why do we have to reverse the order of the factors when the transpose is pulled inside the product AB? Why isn't $(AB)^T = A^T B^T$? One of the reasons is that one of the products may not be defined. If

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

then

$$(AB)^T = \begin{pmatrix} 2 & 2 \end{pmatrix},$$

while the product $A^T B^T$ is not be defined. $\blacksquare$

Exercise

(i) Let A be an $n \times n$ matrix, and let Z be the matrix with $z_{j,j+1} = 1$, $1 \le j \le n-1$, and all other elements zero. What does ZAZ^T do?

1.9 Inner and Outer Products, Again

Transposition comes in handy for the representation of inner and outer products. If

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

then

$$x^* y = \bar{x}_1 y_1 + \cdots + \bar{x}_n y_n, \qquad y^* x = \bar{y}_1 x_1 + \cdots + \bar{y}_n x_n.$$

Example. Let $\alpha = \alpha_1 + \imath \alpha_2$ be a complex number, where $\imath^2 = -1$ and $\alpha_1, \alpha_2 \in \mathbb{R}$. With

$$x = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}$$

the absolute value of α can be represented as the inner product, $|\alpha| = \sqrt{x^* x}$. ■

Fact 1.5 (Properties of Inner Products). Let $x, y \in \mathbb{C}^n$.

1. $y^* x$ is the complex conjugate of $x^* y$, i.e., $y^* x = \overline{(x^* y)}$.
2. $y^T x = x^T y$.
3. $x^* x = 0$ if and only if $x = 0$.
4. If x is real, then $x^T x = 0$ if and only if $x = 0$.

Proof. Let $x = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix}^T$ and $y = \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix}^T$. For the first equality write $y^* x = \sum_{j=1}^n \bar{y}_j x_j = \sum_{j=1}^n x_j \bar{y}_j$. Since complex conjugating twice gives back the original, we get $\sum_{j=1}^n x_j \bar{y}_j = \sum_{j=1}^n \bar{\bar{x}}_j \bar{y}_j = \sum_{j=1}^n \overline{\bar{x}_j y_j} = \overline{\sum_{j=1}^n \bar{x}_j y_j} = \overline{x^* y}$, where the long overbar denotes complex conjugation over the whole sum.

As for the third statement, $0 = x^* x = \sum_{j=1}^n \bar{x}_j x_j = \sum_{j=1}^n |x_j|^2$ if and only if $x_j = 0$, $1 \le j \le n$, if and only if $x = 0$. □

Example. The identity matrix can be represented as the outer product

$$I_n = e_1 e_1^T + e_2 e_2^T + \cdots + e_n e_n^T. \qquad ■$$

Exercises

(i) Let x be a column vector. Give an example to show that $x^T x = 0$ can happen for $x \neq 0$.

(ii) Let $x \in \mathbb{C}^n$ and $x^* x = 1$. Show that $I_n - 2xx^*$ is involutory.

(iii) Let A be a square matrix with $a_{j,j+1} = 1$ and all other elements zero. Represent A as a sum of outer products.

1.10 Symmetric and Hermitian Matrices

We look at matrices that remain unchanged by transposition.

Definition 1.6. *A matrix $A \in \mathbb{C}^{n \times n}$ is*

- *symmetric if $A^T = A$,*
- *Hermitian if $A^* = A$,*
- *skew-symmetric if $A^T = -A$,*
- *skew-Hermitian if $A^* = -A$.*

The identity matrix I_n is symmetric and Hermitian. The square zero matrix $0_{n \times n}$ is symmetric, skew-symmetric, Hermitian, and skew-Hermitian.

Example. Let $\imath^2 = -1$.

$$\begin{pmatrix} 1\imath & 2\imath \\ 2\imath & 4 \end{pmatrix} \text{ is symmetric,} \qquad \begin{pmatrix} 1 & 2\imath \\ -2\imath & 4 \end{pmatrix} \text{ is Hermitian,}$$

$$\begin{pmatrix} 0 & 2\imath \\ -2\imath & 0 \end{pmatrix} \text{ is skew-symmetric,} \qquad \begin{pmatrix} 1\imath & 2\imath \\ 2\imath & 4\imath \end{pmatrix} \text{ is skew-Hermitian.}$$ ■

Example. Let $\imath^2 = -1$.

$$\begin{pmatrix} 0 & \imath \\ \imath & 0 \end{pmatrix}$$

is symmetric and skew-Hermitian, while

$$\begin{pmatrix} 0 & -\imath \\ \imath & 0 \end{pmatrix}$$

is Hermitian and skew-symmetric. ■

Fact 1.7. If $A \in \mathbb{C}^{m \times n}$, then AA^T and A^TA are symmetric, while AA^* and A^*A are Hermitian.

If $A \in \mathbb{C}^{n \times n}$, then $A + A^T$ is symmetric, and $A + A^*$ is Hermitian.

Exercises

(i) Is a Hankel matrix symmetric, Hermitian, skew-symmetric, or skew-Hermitian?

(ii) Which matrix is both symmetric and skew-symmetric?

(iii) Prove: If A is a square matrix, then $A - A^T$ is skew-symmetric and $A - A^*$ is skew-Hermitian.

(iv) Which elements of a Hermitian matrix cannot be complex?

(v) What can you say about the diagonal elements of a skew-symmetric matrix?

(vi) What can you say about the diagonal elements of a skew-Hermitian matrix?

(vii) If A is symmetric and λ is a scalar, does this imply that λA is symmetric? If yes, give a proof. If no, give an example.

(viii) If A is Hermitian and λ is a scalar, does this imply that λA is Hermitian? If yes, give a proof. If no, give an example.

(ix) Prove: If A is skew-symmetric and λ is a scalar, then λA is skew-symmetric.

(x) Prove: If A is skew-Hermitian and λ is a scalar, then λA is, in general, not skew-Hermitian.

(xi) Prove: If A is Hermitian, then $\imath A$ is skew-Hermitian, where $\imath^2 = -1$.

(xii) Prove: If A is skew-Hermitian, then $\imath A$ is Hermitian, where $\imath^2 = -1$.

(xiii) Prove: If A is a square matrix, then $\imath(A - A^*)$ is Hermitian, where $\imath^2 = -1$.

1. Prove: Every square matrix A can be written $A = A_1 + A_2$, where A_1 is Hermitian and A_2 is skew-Hermitian.
2. Prove: Every square matrix A can be written $A = A_1 + \imath A_2$, where A_1 and A_2 are Hermitian and $\imath^2 = -1$.

1.11 Inverse

We want to determine an inverse with respect to matrix multiplication. Inversion of matrices is more complicated than inversion of scalars. There is only one scalar that does not have an inverse: 0. But there are many matrices without inverses.

Definition 1.8. *A matrix $A \in \mathbb{C}^{n\times n}$ is* nonsingular *(or* invertible*) if A has an* inverse*, that is, if there is a matrix A^{-1} so that $AA^{-1} = I = A^{-1}A$. If A does not have an inverse, it is* singular.

Example.

- A 1×1 matrix is invertible if it is nonzero.
- An involutory matrix is its own inverse: $A^2 = I$. ∎

Fact 1.9. The inverse is unique.

Proof. Let $A \in \mathbb{C}^{n\times n}$, and let $AB = BA = I_n$ and $AC = CA = I_n$ for matrices $B, C \in \mathbb{C}^{n\times n}$. Then

$$B = BI_n = B(AC) = (BA)C = I_nC = C. \qquad \square$$

It is often easier to determine that a matrix is singular than it is to determine that a matrix is nonsingular. The fact below illustrates this.

Fact 1.10. Let $A \in \mathbb{C}^{n\times n}$ and $x, b \in \mathbb{C}^n$.

- If $x \neq 0$ and $Ax = 0$, then A is singular.
- If $x \neq 0$ and A is nonsingular, then $Ax \neq 0$.
- If $Ax = b$, where A is nonsingular and $b \neq 0$, then $x \neq 0$.

Proof. To prove the first statement, assume to the contrary that A is nonsingular and has an inverse A^{-1}. Then $0 = Ax$ implies $0 = A^{-1}Ax = I_n x = x$, hence $x = 0$, which contradicts the assumption $x \neq 0$. Therefore A must be singular.

The proofs for the other two statements are similar. □

Fact 1.11. An idempotent matrix is either the identity or else is singular.

Proof. If A is idempotent, then $A^2 = A$. Hence $0 = A^2 - A = A(A - I)$. Either $I - A = 0$, in which case A is the identity, or else $I - A \neq 0$, in which case it has a nonzero column and Fact 1.10 implies that A is singular. □

Now we show that inversion and transposition can be exchanged.

Fact 1.12. If A is invertible, then A^T and A^* are also invertible, and

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

Proof. Show that $(A^{-1})^*$ fulfills the conditions for an inverse of A^*:

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

and

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

The proof for A^T is similar. □

Because inverse and transpose can be exchanged, we can simply write A^{-*} and A^{-T}.

The expression below is useful because it can break apart the inverse of a sum.

Fact 1.13 (Sherman–Morrison Formula). If $A \in \mathbb{C}^{n\times n}$ is nonsingular, and $V \in \mathbb{C}^{m\times n}$, $U \in \mathbb{C}^{n\times m}$ are such that $I + VA^{-1}U$ is nonsingular, then

$$(A + UV)^{-1} = A^{-1} - A^{-1}U\left(I + VA^{-1}U\right)^{-1} VA^{-1}.$$

Here is an explicit expression for the inverse of a partitioned matrix.

Fact 1.14. Let $A \in \mathbb{C}^{n\times n}$ and

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

If A_{11} and A_{22} are nonsingular, then

$$A^{-1} = \begin{pmatrix} S_1^{-1} & -A_{11}^{-1}A_{12}S_2^{-1} \\ -A_{22}^{-1}A_{21}S_1^{-1} & S_2^{-1} \end{pmatrix},$$

where $S_1 = A_{11} - A_{12}A_{22}^{-1}A_{21}$ and $S_2 = A_{22} - A_{21}A_{11}^{-1}A_{12}$.

Matrices of the form S_1 and S_2 are called *Schur complements*.

Exercises

(i) Prove: If A and B are invertible, then $(AB)^{-1} = B^{-1}A^{-1}$.

(ii) Prove: If $A, B \in \mathbb{C}^{n\times n}$ are nonsingular, then $B^{-1} = A^{-1} - B^{-1}(B-A)A^{-1}$.

(iii) Let $A \in \mathbb{C}^{m\times n}$, $B \in \mathbb{C}^{n\times m}$ be such that $I + BA$ is invertible. Show that $(I+BA)^{-1} = I - B(I+AB)^{-1}A$.

(iv) Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $u \in \mathbb{C}^{n\times 1}$, $v \in \mathbb{C}^{1\times n}$, and $vA^{-1}u \neq -1$. Show that

$$(A+uv)^{-1} = A^{-1} - \frac{A^{-1}uvA^{-1}}{1+vA^{-1}u}.$$

(v) The following expression for the partitioned inverse requires only A_{11} to be nonsingular but not A_{22}.
Let $A \in \mathbb{C}^{n\times n}$ and

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Show: If A_{11} is nonsingular and $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$, then

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{pmatrix}.$$

(vi) Let $x \in \mathbb{C}^{1\times n}$ and $A \in \mathbb{C}^{n\times n}$. Prove: If $x \neq 0$ and $xA = 0$, then A is singular.

(vii) Prove: The inverse, if it exists, of a Hermitian (symmetric) matrix is also Hermitian (symmetric).

(viii) Prove: If A is involutory, then $I - A$ or $I + A$ must be singular.

(ix) Let A be a square matrix so that $A + A^2 = I$. Prove: A is invertible.

(x) Prove: A nilpotent matrix is always singular.

1. Let $S \in \mathbb{R}^{n\times n}$. Show: If S is skew-symmetric, then $I - S$ is nonsingular. Give an example to illustrate that $I - S$ can be singular if $S \in \mathbb{C}^{n\times n}$.
2. Let x be a nonzero column vector. Determine a row vector y so that $yx = 1$.
3. Let A be a square matrix and let α_j be scalars, at least two of which are nonzero, such that $\sum_{j=0}^{k} \alpha_j A^j = 0$. Prove: If $\alpha_0 \neq 0$, then A is nonsingular.
4. Prove: If $(I-A)^{-1} = \sum_{j=0}^{k} A^j$ for some integer $k \geq 0$, then A is nilpotent.
5. Let $A, B \in \mathbb{C}^{n\times n}$. Prove: If $I + BA$ is invertible, then $I + AB$ is also invertible.

1.12 Unitary and Orthogonal Matrices

These are matrices whose inverse is a transpose.

Definition 1.15. *A matrix $A \in \mathbb{C}^{n \times n}$ is*

- *unitary if $AA^* = A^*A = I$,*
- *orthogonal if $AA^T = A^TA = I$.*

The identity matrix is orthogonal as well as unitary.

Example 1.16. Let c and s be scalars with $|c|^2 + |s|^2 = 1$. The matrices

$$\begin{pmatrix} c & s \\ -\bar{s} & \bar{c} \end{pmatrix}, \qquad \begin{pmatrix} c & s \\ \bar{s} & -\bar{c} \end{pmatrix}$$

are unitary. ∎

The first matrix above gets its own name.

Definition 1.17. *If $c, s \in \mathbb{C}$ so that $|c|^2 + |s|^2 = 1$, then the unitary 2×2 matrix*

$$\begin{pmatrix} c & s \\ -\bar{s} & \bar{c} \end{pmatrix}$$

is called a Givens rotation. *If c and s are also real, then the Givens rotation $\begin{pmatrix} c & s \\ -s & c \end{pmatrix}$ is orthogonal.*

When a Givens rotation is real, then both diagonal elements are the same. When a Givens rotation is complex, then the diagonal elements are complex conjugates of each other. A unitary matrix of the form

$$\begin{pmatrix} -c & s \\ \bar{s} & \bar{c} \end{pmatrix},$$

where the real parts of the diagonal elements have different signs, is a *reflection*; it is *not* a Givens rotation.

An orthogonal matrix that can reorder the rows or columns of a matrix is called a *permutation matrix.* It is an identity matrix whose rows have been reordered (permuted). One can also think of a permutation matrix as an identity matrix whose columns have been reordered. Here is the official definition.

Definition 1.18 (Permutation Matrix). *A square matrix is a* permutation matrix *if it contains a single one in each column and in each row, and zeros everywhere else.*

Example. The following are permutation matrices.

- The identity matrix I.

- The exchange matrix

$$J = \begin{pmatrix} & & 1 \\ & \iddots & \\ 1 & & \end{pmatrix}, \qquad J\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_n \\ \vdots \\ x_2 \\ x_1 \end{pmatrix}.$$

- The upper circular shift matrix

$$Z = \begin{pmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & \ddots & 1 \\ 1 & & & & 0 \end{pmatrix}, \qquad Z\begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} x_n \\ x_1 \\ \vdots \\ x_{n-1} \end{pmatrix}. \qquad \blacksquare$$

Fact 1.19 (Properties of Permutation Matrices).

1. Permutation matrices are orthogonal and unitary.
   That is, if P is a permutation matrix, then $PP^T = P^TP = PP^* = P^*P = I$.
2. The product of permutation matrices is again a permutation matrix.

Exercises

(i) Prove: If A is unitary, then A^*, A^T, and $\overline{A}$ are unitary.

(ii) What can you say about an involutory matrix that is also unitary (orthogonal)?

(iii) Which idempotent matrix is unitary and orthogonal?

(iv) Prove: If A is unitary, so is $\imath A$, where $\imath^2 = -1$.

(v) Prove: The product of unitary matrices is unitary.

(vi) Partitioned Unitary Matrices.
Let $A \in \mathbb{C}^{n\times n}$ be unitary and partition $A = \begin{pmatrix} A_1 & A_2 \end{pmatrix}$, where A_1 has k columns, and A_2 has $n-k$ columns. Show that $A_1^*A_1 = I_k$, $A_2^*A_2 = I_{n-k}$, and $A_1^*A_2 = 0$.

(vii) Let $x \in \mathbb{C}^n$ and $x^*x = 1$. Prove: $I_n - 2xx^*$ is Hermitian and unitary. Conclude that $I_n - 2xx^*$ is involutory.

(viii) Show: If P is a permutation matrix, then P^T and P^* are also permutation matrices.

(ix) Show: If $\begin{pmatrix} P_1 & P_2 \end{pmatrix}$ is a permutation matrix, then $\begin{pmatrix} P_2 & P_1 \end{pmatrix}$ is also a permutation matrix.

1.13 Triangular Matrices

Triangular matrices occur frequently during the solution of systems of linear equations, because linear systems with triangular matrices are easy to solve.

Definition 1.20. *A matrix $A \in \mathbb{C}^{n\times n}$ is* upper triangular *if $a_{ij} = 0$ for $i > j$. That is,*

$$A = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ & \ddots & \vdots \\ & & a_{nn} \end{pmatrix}.$$

A matrix $A \in \mathbb{C}^{n\times n}$ is lower triangular *if A^T is upper triangular.*

Fact 1.21. Let A and B be upper triangular, with diagonal elements a_{jj} and b_{jj}, respectively.

- $A + B$ and AB are upper triangular.
- The diagonal elements of AB are $a_{ii}b_{ii}$.
- If $a_{jj} \neq 0$ for all j, then A is invertible, and the diagonal elements of A^{-1} are $1/a_{jj}$.

Definition 1.22. *A triangular matrix A is* unit triangular *if it has ones on the diagonal, and* strictly triangular *if it has zeros on the diagonal.*

Example. The identity matrix is unit upper triangular and unit lower triangular. The square zero matrix is strictly lower triangular and strictly upper triangular. ■

Exercises

(i) What does an idempotent triangular matrix look like? What does an involutory triangular matrix look like?

1. Prove: If A is unit triangular, then A is invertible, and A^{-1} is unit triangular. If A and B are unit triangular, then so is the product AB.
2. Show that a strictly triangular matrix is nilpotent.
3. Explain why the matrix $I - \alpha e_i e_j^T$ is triangular. When does it have an inverse? Determine the inverse in those cases where it exists.
4. Prove:

$$\begin{pmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^n \\ & 1 & \alpha & \ddots & \vdots \\ & & \ddots & \ddots & \alpha^2 \\ & & & 1 & \alpha \\ & & & & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & -\alpha & & & \\ & 1 & -\alpha & & \\ & & \ddots & \ddots & \\ & & & 1 & -\alpha \\ & & & & 1 \end{pmatrix}.$$

5. Uniqueness of LU Factorization.
   Let L_1, L_2 be unit lower triangular, and U_1, U_2 nonsingular upper triangular. Prove: If $L_1U_1 = L_2U_2$, then $L_1 = L_2$ and $U_1 = U_2$.

6. Uniqueness of QR Factorization.
   Let Q_1, Q_2 be unitary (or orthogonal), and R_1, R_2 upper triangular with positive diagonal elements. Prove: If $Q_1 R_1 = Q_2 R_2$, then $Q_1 = Q_2$ and $R_1 = R_2$.

1.14 Diagonal Matrices

Diagonal matrices are special cases of triangular matrices; they are upper and lower triangular at the same time.

Definition 1.23. *A matrix $A \in \mathbb{C}^{n\times n}$ is* diagonal *if $a_{ij} = 0$ for $i \neq j$. That is,*

$$A = \begin{pmatrix} a_{11} & & \\ & \ddots & \\ & & a_{nn} \end{pmatrix}.$$

The identity matrix and the square zero matrix are diagonal.

Exercises

(i) Prove: Diagonal matrices are symmetric. Are they also Hermitian?

(ii) Diagonal matrices commute.
Prove: If A and B are diagonal, then AB is diagonal, and $AB = BA$.

(iii) Represent a diagonal matrix as a sum of outer products.

(iv) Which diagonal matrices are involutory, idempotent, or nilpotent?

(v) Prove: If a matrix is unitary and triangular, it must be diagonal. What are its diagonal elements?

1. Let D be a diagonal matrix. Prove: If $D = (I + A)^{-1} A$, then A is diagonal.

2. Sensitivity, Errors, and Norms

Two difficulties arise when we solve systems of linear equations or perform other matrix computations.

(i) Errors in matrix elements.
Matrix elements may be contaminated with errors from measurements or previous computations, or they may simply not be known exactly. Merely inputting numbers into a computer or calculator can cause errors (e.g., when $1/3$ is stored as .33333333). To account for all these situations, we say that the matrix elements are afflicted with *uncertainties* or are *perturbed*. In general, perturbations of the inputs cause difficulties when the outputs are "sensitive" to changes in the inputs.

(ii) Errors in algorithms.
Algorithms may not compute an exact solution, because computing the exact solution may not be necessary, may take too long, may require too much storage, or may not be practical. Furthermore, arithmetic operations in finite precision may not be performed exactly.

In this book, we focus on perturbations of inputs, and how these perturbations affect the outputs.

2.1 Sensitivity and Conditioning

In real life, *sensitive* means[1] "acutely affected by external stimuli," "easily offended or emotionally hurt," or "responsive to slight changes." A sensitive person can be easily upset by small events, such as having to wait in line for a few minutes. Hardware can be sensitive: A very slight turn of a faucet may change the water from freezing cold to scalding hot. The slightest turn of the steering wheel when driving on an icy surface can send the car careening into a spin. Organs can be

[1] The Concise Oxford English Dictionary

sensitive: Healthy skin may not even feel the prick of a needle, while it may cause extreme pain on burnt skin.

It is no different in mathematics. Steep functions, for instance, can be sensitive to small perturbations in the input.

Example. Let $f(x) = 9^x$ and consider the effect of a small perturbation to the input of $f(50) = 9^{50}$, such as

$$f(50.5) = \sqrt{9}\,9^{50} = 3f(50).$$

Here a 1 percent change in the input causes a 300 percent change of the output. ■

Systems of linear equations are sensitive when a small modification in the matrix or the right-hand side causes a large change in the solution.

Example 2.1. The linear system $Ax = b$ with

$$A = \begin{pmatrix} 1/3 & 1/3 \\ 1/3 & .3 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

has the solution

$$x = \begin{pmatrix} -27 \\ 30 \end{pmatrix}.$$

However, a small change of the $(2,2)$ element from .3 to $1/3$ results in the total loss of the solution, because the system $\tilde{A}x = b$ with

$$\tilde{A} = \begin{pmatrix} 1/3 & 1/3 \\ 1/3 & 1/3 \end{pmatrix}$$

has no solution. ■

A linear system like the one above whose solution is sensitive to small perturbations in the matrix is called *ill-conditioned*. Here is another example of ill-conditioning.

Example. The linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1+\epsilon \end{pmatrix}, \qquad b = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \qquad 0 < \epsilon \ll 1,$$

has the solution

$$x = \frac{1}{\epsilon}\begin{pmatrix} -2-\epsilon \\ 2 \end{pmatrix}.$$

But changing the $(2,2)$ element of A from $1+\epsilon$ to 1 results in the loss of the solution, because the linear system $\tilde{A}x = b$ with

$$\tilde{A} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

has no solution. This happens regardless of how small ϵ is. ■

An ill-conditioned linear system can also be sensitive to small perturbations in the right-hand side, as the next example shows.

Example 2.2. The linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1+\epsilon \end{pmatrix}, \qquad b = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \qquad 0 < \epsilon \ll 1,$$

has the solution $x = \begin{pmatrix} 2 & 0 \end{pmatrix}^T$. Changing the leading element in the right-hand side from 2 to $2+\epsilon$ alters the solution radically. That is, the system $A\tilde{x} = \tilde{b}$ with

$$\tilde{b} = \begin{pmatrix} 2 \\ 2+\epsilon \end{pmatrix}$$

has the solution $\tilde{x} = \begin{pmatrix} 1 & 1 \end{pmatrix}^T$, which is completely different from x. ■

Important. Ill-conditioning of a linear system has nothing to do with how we compute the solution. Ill-conditioning is a property of the linear system. Hence there is, in general, nothing you can do about ill-conditioning.

In an ill-conditioned linear system, errors in the matrix or in the right-hand side can be amplified so that the errors in the solution are much larger. Our aim is to determine which properties of a linear system are responsible for ill-conditioning, and how one can quantify ill-conditioning.

2.2 Absolute and Relative Errors

To quantify ill-conditioning, we need to assess the size of errors.

Example. Suppose you have $y = 10$ dollars in your bank account. But the bank makes a mistake and subtracts 20 dollars from your account, so that your account now has a negative balance of $\tilde{y} = -10$ dollars. The account is overdrawn, and all kinds of bad consequences ensue.

Now imagine this happens to Bill Gatez. He has $g = 10^{11}$ dollars in his account, and if the bank subtracts by mistake 20 dollars from his balance, he still has $\tilde{g} = 10^{11} - 20$ dollars.

In both cases, the bank makes the same error,

$$y - \tilde{y} = g - \tilde{g} = 20.$$

But you are much worse off than Bill Gatez. You are now in debt, while Bill Gatez has so much money, he may not even notice the error. In your case, the error is larger than your credit; while in Bill Gatez's case, the error is only a tiny part of his fortune.

How can we express mathematically that the bank's error is much worse for you than for Bill Gatez? We can compare the error to the balance in your account: $\frac{y-\tilde{y}}{y} = 2$. This shows that the error is twice as large as your original balance.

For Bill Gatez we obtain $\frac{g-\tilde{g}}{g} = 2 \cdot 10^{-10}$, so that the error is only a tiny fraction of his balance. Now it's clear that the bank's error is much more serious for you than it is for Bill Gatez. ∎

A difference like $y - \tilde{y}$ measures an *absolute error*, while $\frac{y-\tilde{y}}{y}$ and $\frac{y-\tilde{y}}{\tilde{y}}$ measure *relative errors*. We use relative errors if we want to know how large the error is when compared to the original quantity. Often we are not interested in the signs of the errors, so we consider the absolute values $\frac{|y-\tilde{y}|}{|y|}$ and $\frac{|y-\tilde{y}|}{|\tilde{y}|}$.

Definition 2.3. *If the scalar $\tilde{x}$ is an approximation to the scalar x, then we call $|x-\tilde{x}|$ an* absolute error. *If $x \neq 0$, then we call $\frac{|x-\tilde{x}|}{|x|}$ a* relative error. *If $\tilde{x} \neq 0$, then $\frac{|x-\tilde{x}|}{|\tilde{x}|}$ is also a relative error.*

A relative error close to or larger than 1 means that an approximation is totally inaccurate. To see this, suppose that $\frac{|x-\tilde{x}|}{|x|} \geq 1$. Then $|x-\tilde{x}| \geq |x|$, which means that the absolute error is larger than the quantity we are trying to compute. If we approximate $x = 0$ by $\tilde{x} \neq 0$, however small, then the relative error is always $\frac{|0-\tilde{x}|}{|\tilde{x}|} = 1$. Thus, the only approximation to 0 that has a small relative error is 0 itself.

In contrast to an absolute error, a relative error can give information about how many digits two numbers have in common. As a rule of thumb, if

$$\frac{|x-\tilde{x}|}{|x|} \leq 5 \cdot 10^{-d},$$

then we say that the numbers x and $\tilde{x}$ agree to d decimal digits.

Example. If $x = 1$ and $\tilde{x} = 1.003$, then $\frac{|x-\tilde{x}|}{|x|} = 3 \cdot 10^{-3} \leq 5 \cdot 10^{-3}$, so that x and $\tilde{x}$ agree to three decimal digits.

According to the above definition, the numbers $x = 1$ and $\hat{x} = .997$ also agree to three decimal digits because $\frac{|x-\hat{x}|}{|x|} = 3 \cdot 10^{-3} \leq 5 \cdot 10^{-3}$. ∎

2.3 Floating Point Arithmetic

Many computations in science and engineering are carried out in floating point arithmetic, where all real numbers are represented by a finite set of floating point numbers. All floating point numbers are stored in the same, fixed number of bits regardless of how small or how large they are. Many computers are based on IEEE double precision arithmetic where a floating point number is stored in 64 bits.

The floating point representation $\hat{x}$ of a real number x differs from x by a factor close to one, and satisfies[2]

$$\hat{x} = x(1+\epsilon_x), \qquad \text{where} \quad |\epsilon_x| \leq \mathsf{u}.$$

[2] We assume that x lies in the range of normalized floating point numbers, so that no underflow or overflow occurs.

Here u is the "unit roundoff" that specifies the accuracy of floating point arithmetic. In IEEE double precision arithmetic $\mathsf{u} = 2^{-53} \approx 10^{-16}$. If $x \neq 0$, then

$$\frac{|x - \hat{x}|}{|x|} \leq |\epsilon_x|.$$

This means that conversion to floating point representation causes relative errors. We say that a floating point number $\hat{x}$ is a *relative perturbation* of the exact number x.

Since floating point arithmetic causes relative perturbations in the inputs, it makes sense to determine relative—rather than absolute—errors in the output. As a consequence, we will pay more attention to relative errors than to absolute errors.

The question now is how elementary arithmetic operations are affected when they are performed on numbers contaminated with small relative perturbations, such as floating point numbers. We start with subtraction.

2.4 Conditioning of Subtraction

Subtraction is the only elementary operation that is sensitive to relative perturbations. The analogy below of the captain and the battleship can help us understand why.

Example. To find out how much he weighs, the captain first weighs the battleship with himself on it, and then he steps off the battleship and weighs it without himself on it. At the end he subtracts the two weights. Intuitively we have a vague feeling for why this should not give an accurate estimate for the captain's weight. Below we explain why.

Let $\tilde{x}$ represent the weight of the battleship plus captain, and $\tilde{y}$ the weight of the battleship without the captain, where

$$\tilde{x} = 112233\underline{9}, \qquad \tilde{y} = 112233\underline{7}.$$

Due to the limited precision of the scale, the underlined digits are uncertain and may be in error. The captain computes as his weight $\tilde{x} - \tilde{y} = \underline{2}$. This difference is totally inaccurate because it is derived from uncertainties, while all the accurate digits have cancelled out. This is an example of "catastrophic cancellation." ■

Catastrophic cancellation occurs when we subtract two numbers that are uncertain, and when the difference between these two numbers is as small as the uncertainties. We will now show that catastrophic cancellation occurs when subtraction is ill-conditioned with regard to relative errors.

Let $\tilde{x}$ be a perturbation of the scalar x and $\tilde{y}$ a perturbation of the scalar y. We bound the error in $\tilde{x} - \tilde{y}$ in terms of the errors in $\tilde{x}$ and $\tilde{y}$.

Absolute Error. From

$$|(\tilde{x} - \tilde{y}) - (x - y)| \leq |\tilde{x} - x| + |\tilde{y} - y|,$$

we see that the absolute error in the difference is bounded by the absolute errors in the inputs. Therefore we say that subtraction is *well-conditioned in the absolute sense*. In the above example, the last digit of $\tilde{x}$ and $\tilde{y}$ is uncertain, so that $|\tilde{x}-x| \leq 9$ and $|\tilde{y}-y| \leq 9$, and the absolute error is bounded by $|(\tilde{x}-\tilde{y})-(x-y)| \leq 18$.

Relative Error. However, the relative error in the difference can be much larger than the relative error in the inputs. In the above example we can estimate the relative error from

$$\frac{|(\tilde{x}-\tilde{y})-(x-y)|}{|\tilde{x}-\tilde{y}|} \leq \frac{18}{2} = 9,$$

which suggests that the computed difference $\tilde{x}-\tilde{y}$ is completely inaccurate.

In general, this severe loss of accuracy can occur when we subtract two nearly equal numbers that are in error. The bound in Fact 2.4 below shows that subtraction can be *ill-conditioned in the relative sense* if the difference is much smaller in magnitude than the inputs.

Fact 2.4 (Relative Conditioning of Subtraction). Let x, y, $\tilde{x}$, and $\tilde{y}$ be scalars. If $x \neq 0$, $y \neq 0$, and $x \neq y$, then

$$\underbrace{\frac{|(\tilde{x}-\tilde{y})-(x-y)|}{|x-y|}}_{\text{relative error in output}} \leq \kappa \underbrace{\max\left\{\frac{|\tilde{x}-x|}{|x|}, \frac{|\tilde{y}-y|}{|y|}\right\}}_{\text{relative error in input}},$$

where

$$\kappa = \frac{|x|+|y|}{|x-y|}.$$

The positive number κ is a *relative condition number for subtraction*, because it quantifies how relative errors in the input can be amplified, and how sensitive subtraction can be to relative errors in the input. When $\kappa \gg 1$, subtraction is ill-conditioned in the relative sense and is called catastrophic cancellation.

If we do not know x, y, or $x-y$, but want an estimate of the condition number, we can use instead the bound

$$\frac{|(\tilde{x}-\tilde{y})-(x-y)|}{|\tilde{x}-\tilde{y}|} \leq \tilde{\kappa} \max\left\{\frac{|\tilde{x}-x|}{|\tilde{x}|}, \frac{|\tilde{y}-y|}{|\tilde{y}|}\right\}, \qquad \tilde{\kappa} = \frac{|\tilde{x}|+|\tilde{y}|}{|\tilde{x}-\tilde{y}|},$$

provided $\tilde{x} \neq 0$, $\tilde{y} \neq 0$, and $\tilde{x} \neq \tilde{y}$,

Remark 2.5. *Catastrophic cancellation does not occur when we subtract two numbers that are* exact.

Catastrophic cancellation can only occur when we subtract two numbers that have relative errors. It is the amplification of these relative errors that leads to catastrophe.

Exercises

1. Relative Conditioning of Multiplication.
   Let x, y, $\tilde{x}$, $\tilde{y}$ be nonzero scalars. Show:
   $$\left|\frac{xy-\tilde{x}\tilde{y}}{xy}\right| \leq (2+\epsilon)\epsilon, \qquad \text{where} \quad \epsilon = \max\left\{\left|\frac{x-\tilde{x}}{x}\right|, \left|\frac{y-\tilde{y}}{y}\right|\right\},$$
   and if $\epsilon \leq 1$, then
   $$\left|\frac{xy-\tilde{x}\tilde{y}}{xy}\right| \leq 3\epsilon.$$
   Therefore, if the relative error in the inputs is not too large, then the condition number of multiplication is at most 3. We can conclude that multiplication is well-conditioned in the relative sense, provided the inputs have small relative perturbations.
2. Relative Conditioning of Division.
   Let x, y, $\tilde{x}$, $\tilde{y}$ be nonzero scalars, and let
   $$\epsilon = \max\left\{\left|\frac{x-\tilde{x}}{x}\right|, \left|\frac{y-\tilde{y}}{y}\right|\right\}.$$
   Show: If $\epsilon < 1$, then
   $$\left|\frac{x/y-\tilde{x}/\tilde{y}}{x/y}\right| \leq \frac{2\epsilon}{1-\epsilon},$$
   and if $\epsilon < 1/2$, then
   $$\left|\frac{x/y-\tilde{x}/\tilde{y}}{x/y}\right| \leq 4\epsilon.$$
   Therefore, if the relative error in the operands is not too large, then the condition number of division is at most 4. We can conclude that division is well-conditioned in the relative sense, provided the inputs have small relative perturbations.

2.5 Vector Norms

In the context of linear system solution, the error in the solution constitutes a vector. If we do not want to pay attention to individual components of the error, perhaps because there are too many components, then we can combine all errors into a single number. This is akin to a grade point average which combines all grades into a single number. Mathematically, this "combining" is accomplished by norms. We start with vector norms, which measure the length of a vector.

Definition 2.6. *A vector norm $\|\cdot\|$ is a function from $\mathbb{C}^n$ to $\mathbb{R}$ with three properties:*

N1: $\|x\| \geq 0$ *for all* $x \in \mathbb{C}^n$, *and* $\|x\| = 0$ *if and only if* $x = 0$.

N2: $\|x+y\| \leq \|x\|+\|y\|$ *for all* $x,\ y \in \mathbb{C}^n$ *(triangle inequality).*

N3: $\|\alpha x\| = |\alpha|\,\|x\|$ *for all* $\alpha \in \mathbb{C},\ x \in \mathbb{C}^n$.

The vector p-norms below are useful for computational purposes, as well as analysis.

Fact 2.7 (Vector *p*-Norms). Let $x \in \mathbb{C}^n$ with elements $x = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix}^T$. The *p-norm*

$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p\right)^{1/p}, \qquad p \geq 1,$$

is a vector norm.

Example.

- If e_j is a canonical vector, then $\|e_j\|_p = 1$ for $p \geq 1$.
- If $e = \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^T \in \mathbb{R}^n$, then

$$\|e\|_1 = n, \qquad \|e\|_\infty = 1, \qquad \|e\|_p = n^{1/p}, \quad 1 < p < \infty. \qquad \blacksquare$$

The three p-norms below are the most popular, because they are easy to compute.

- One norm: $\|x\|_1 = \sum_{j=1}^n |x_j|$.
- Two (or Euclidean) norm: $\|x\|_2 = \sqrt{\sum_{j=1}^n |x_j|^2} = \sqrt{x^* x}$.
- Infinity (or maximum) norm: $\|x\|_\infty = \max_{1\leq j\leq n} |x_j|$.

Example. If $x = \begin{pmatrix} 1 & 2 & \cdots & n \end{pmatrix}^T \in \mathbb{R}^n$, then

$$\|x\|_1 = \frac{1}{2}n(n+1), \qquad \|x\|_2 = \sqrt{\frac{1}{6}n(n+1)(2n+1)}, \qquad \|x\|_\infty = n. \qquad \blacksquare$$

The inequalities below bound inner products in terms of norms.

Fact 2.8. Let $x, y \in \mathbb{C}^n$. Then

Hölder inequality: $|x^* y| \leq \|x\|_1\, \|y\|_\infty$
Cauchy–Schwarz inequality: $|x^* y| \leq \|x\|_2\, \|y\|_2$.

Moreover, $|x^* y| = \|x\|_2 \|y\|_2$ if and only if x and y are multiples of each other.

Example. Let $x \in \mathbb{C}^n$ with elements $x = \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix}^T$. The Hölder inequality and Cauchy–Schwarz inequality imply, respectively,

$$\left|\sum_{i=1}^n x_i\right| \leq n \max_{1\leq i\leq n} |x_i|, \qquad \left|\sum_{i=1}^n x_i\right| \leq \sqrt{n}\, \|x\|_2. \qquad \blacksquare$$

Definition 2.9. *A nonzero vector $x \in \mathbb{C}^n$ is called* unit-norm *vector in the $\|\cdot\|$ norm if $\|x\| = 1$. The vector $x/\|x\|$ has unit norm.*

Example. Let e be the $n \times 1$ vector of all ones. Then

$$1 = \|e\|_\infty = \left\| \frac{1}{n} e \right\|_1 = \left\| \frac{1}{\sqrt{n}} e \right\|_2.$$

The canonical vectors e_i have unit norm in any p-norm. ∎

Normwise Errors. We determine how much information the norm of an error gives about individual, componentwise errors.

Definition 2.10. *If $\tilde{x}$ is an approximation to a vector $x \in \mathbb{C}^n$, then $\|x - \tilde{x}\|$ is a* normwise absolute error. *If $x \neq 0$ or $\tilde{x} \neq 0$, then $\frac{\|x-\tilde{x}\|}{\|x\|}$ and $\frac{\|x-\tilde{x}\|}{\|\tilde{x}\|}$ are* normwise relative errors.

How much do we lose when we replace componentwise errors by normwise errors? For vectors $x, \tilde{x} \in \mathbb{C}^n$, the infinity norm is equal to the largest absolute error,

$$\|x - \tilde{x}\|_\infty = \max_{1 \le j \le n} |x_j - \tilde{x}_j|.$$

For the one and two norms we have

$$\max_{1 \le j \le n} |x_j - \tilde{x}_j| \le \|x - \tilde{x}\|_1 \le n \max_{1 \le j \le n} |x_j - \tilde{x}_j|$$

and

$$\max_{1 \le j \le n} |x_j - \tilde{x}_j| \le \|x - \tilde{x}\|_2 \le \sqrt{n} \max_{1 \le j \le n} |x_j - \tilde{x}_j|.$$

Hence absolute errors in the one and two norms can overestimate the worst componentwise error by a factor that depends on the vector length n.

Unfortunately, normwise relative errors give much less information about componentwise relative errors.

Example. Let $\tilde{x}$ be an approximation to a vector x where

$$x = \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}, \quad 0 < \epsilon \ll 1, \qquad \tilde{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The normwise relative error $\frac{\|x-\tilde{x}\|_\infty}{\|x\|_\infty} = \epsilon$ is small. However, the componentwise relative error in the second component, $\frac{|x_2-\tilde{x}_2|}{|x_2|} = 1$, shows that $\tilde{x}_2$ is a totally inaccurate approximation to x_2 in the relative sense. ∎

The preceding example illustrates that a normwise relative error can be small, even if individual vector elements have a large relative error. In the infinity norm, for example, the normwise relative error only bounds the relative error corresponding to a component of x with the largest magnitude. To see this, let $|x_k| = \|x\|_\infty$. Then

$$\frac{\|x-\tilde{x}\|_\infty}{\|x\|_\infty} = \frac{\max_{1\le j\le n}|x_j-\tilde{x}_j|}{|x_k|} \ge \frac{|x_k-\tilde{x}_k|}{|x_k|}.$$

For the normwise relative errors in the one and two norms we incur additional factors that depend on the vector length n,

$$\frac{\|x-\tilde{x}\|_1}{\|x\|_1} \ge \frac{1}{n}\frac{|x_k-\tilde{x}_k|}{|x_k|}, \qquad \frac{\|x-\tilde{x}\|_2}{\|x\|_2} \ge \frac{1}{\sqrt{n}}\frac{|x_k-\tilde{x}_k|}{|x_k|}.$$

Therefore, normwise relative errors give no information about relative errors in components of smaller magnitude. If relative errors in individual vector components are important, then do not use normwise errors.

Remark 2.11. *When measuring the normwise relative error of an approximation $\tilde{x}$ to x, the question is which error to measure, $\frac{\|x-\tilde{x}\|}{\|x\|}$ or $\frac{\|x-\tilde{x}\|}{\|\tilde{x}\|}$? If $\|\tilde{x}\| \approx \|x\|$, then the two errors are about the same. In general, the two errors are related as follows. Let $x \neq 0$, $\tilde{x} \neq 0$, and*

$$\epsilon = \frac{\|x-\tilde{x}\|}{\|x\|}, \qquad \tilde{\epsilon} = \frac{\|x-\tilde{x}\|}{\|\tilde{x}\|}.$$

If $\epsilon < 1$, then

$$\frac{\epsilon}{1+\epsilon} \le \tilde{\epsilon} \le \frac{\epsilon}{1-\epsilon}.$$

This follows from $\tilde{\epsilon} = \epsilon\|x\|/\|\tilde{x}\|$ and $1-\tilde{\epsilon} \le \|x\|/\|\tilde{x}\| \le 1+\tilde{\epsilon}$.

Exercises

(i) Let $x \in \mathbb{C}^n$. Prove: $\|x\|_2 \le \sqrt{\|x\|_1\|x\|_\infty}$.

(ii) For each equality below, determine a class of vectors that satisfy the equality:

$$\|x\|_1 = \|x\|_\infty, \qquad \|x\|_1 = n\|x\|_\infty, \qquad \|x\|_2 = \|x\|_\infty, \qquad \|x\|_2 = \sqrt{n}\|x\|_\infty.$$

(iii) Give examples of vectors $x, y \in \mathbb{C}^n$ with $x^*y \neq 0$ for which $|x^*y| = \|x\|_1\|y\|_\infty$. Also find examples for $|x^*y| = \|x\|_2\|y\|_2$.

(iv) The p-norm of a vector does not change when the vector is permuted. Prove: If P is a permutation matrix, then $\|Px\|_p = \|x\|_p$.

(v) The two norm of a vector does not change when the vector is multiplied by a unitary matrix. Prove: If the matrix $V \in \mathbb{C}^{n\times n}$ is unitary, then $\|Vx\|_2 = \|x\|_2$ for any vector $x \in \mathbb{C}^n$.

(vi) Prove: If $Q \in \mathbb{C}^{n\times n}$ is unitary and $x \in \mathbb{C}^n$ is a nonzero vector with $Qx = \lambda x$, where λ is a scalar, then $|\lambda| = 1$.

1. Verify that the vector p-norms do indeed satisfy the three properties of a vector norm in Definition 2.6.

2. Reverse Triangle Inequality.
   Let $x, y \in \mathbb{C}^n$ and let $\|\cdot\|$ be a vector norm. Prove: $\big| \|x\| - \|y\| \big| \le \|x-y\|$.
3. Theorem of Pythagoras.
   Prove: If $x, y \in \mathbb{C}^n$ and $x^*y = 0$, then $\|x \pm y\|_2^2 = \|x\|_2^2 + \|y\|_2^2$.
4. Parallelogram Equality.
   Let $x, y \in \mathbb{C}^n$. Prove: $\|x+y\|_2^2 + \|x-y\|_2^2 = 2(\|x\|_2^2 + \|y\|_2^2)$.
5. Polarization Identity.
   Let $x, y \in \mathbb{C}^n$. Prove: $\Re(x^*y) = \frac{1}{4}(\|x+y\|_2^2 - \|x-y\|_2^2)$, where $\Re(\alpha)$ is the real part of a complex number α.
6. Let $x \in \mathbb{C}^n$. Prove:
$$\begin{aligned} \|x\|_2 &\le \|x\|_1 \le \sqrt{n}\|x\|_2, \\ \|x\|_\infty &\le \|x\|_2 \le \sqrt{n}\|x\|_\infty, \\ \|x\|_\infty &\le \|x\|_1 \le n\|x\|_\infty. \end{aligned}$$
7. Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. Show that $\|x\|_A = \|Ax\|_p$ is a vector norm.

2.6 Matrix Norms

We need to separate matrices from vectors inside the norms. To see this, let $Ax = b$ be a nonsingular linear system, and let $A\tilde{x} = \tilde{b}$ be a perturbed system. The normwise absolute error is $\|x - \tilde{x}\| = \|A^{-1}(b - \tilde{b})\|$. In order to isolate the perturbation and derive a bound of the form $\|A^{-1}\|\,\|b - \tilde{b}\|$, we have to define a norm for matrices.

Definition 2.12. *A matrix norm $\|\cdot\|$ is a function from $\mathbb{C}^{m\times n}$ to $\mathbb{R}$ with three properties:*

N1: $\|A\| \ge 0$ *for all $A \in \mathbb{C}^{n\times m}$, and $\|A\| = 0$ if and only if $A = 0$.*

N2: $\|A + B\| \le \|A\| + \|B\|$ *for all A, $B \in \mathbb{C}^{m\times n}$ (triangle inequality).*

N3: $\|\alpha A\| = |\alpha|\,\|A\|$ *for all $\alpha \in \mathbb{C}$, $A \in \mathbb{C}^{m\times n}$.*

Because of the triangle inequality, matrix norms are well-conditioned, in the absolute sense and in the relative sense.

Fact 2.13. If $A, E \in \mathbb{C}^{m\times n}$, then $\big| \|A+E\| - \|A\| \big| \le \|E\|$.

Proof. The triangle inequality implies $\|A+E\| \le \|A\| + \|E\|$, hence $\|A+E\| - \|A\| \le \|E\|$. Similarly $\|A\| = \|(A+E) - E\| \le \|A+E\| + \|E\|$, so that $-\|E\| \le \|A+E\| - \|A\|$. The result follows from
$$-\|E\| \le \|A+E\| - \|A\| \le \|E\|. \qquad \square$$

The matrix p-norms below are based on the vector p-norms and measure how much a matrix can stretch a unit-norm vector.

Fact 2.14 (Matrix p-Norms). Let $A \in \mathbb{C}^{n\times m}$. The *p-norm*

$$\|A\|_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p=1} \|Ax\|_p$$

is a matrix norm.

Remark 2.15. *The matrix p-norms are extremely useful because they satisfy the following* submultiplicative inequality.

Let $A \in \mathbb{C}^{m\times n}$ and $y \in \mathbb{C}^n$. Then

$$\|Ay\|_p \leq \|A\|_p \|y\|_p.$$

This is clearly true for $y = 0$, and for $y \neq 0$ it follows from

$$\|A\|_p = \max_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p} \geq \frac{\|Ay\|_p}{\|y\|_p}.$$

The matrix one norm is equal to the maximal absolute column sum.

Fact 2.16 (One Norm). Let $A \in \mathbb{C}^{m\times n}$. Then

$$\|A\|_1 = \max_{1\leq j\leq n} \|Ae_j\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|.$$

Proof.

- The definition of p-norms implies

  $$\|A\|_1 = \max_{\|x\|_1=1} \|Ax\|_1 \geq \|Ae_j\|_1, \qquad 1 \leq j \leq n.$$

  Hence $\|A\|_1 \geq \max_{1\leq j\leq n} \|Ae_j\|_1$.

- Let $y = \begin{pmatrix} y_1 & \dots & y_n \end{pmatrix}^T$ be a vector with $\|A\|_1 = \|Ay\|_1$ and $\|y\|_1 = 1$. Viewing the matrix vector product Ay as a linear combination of columns of A, see Section 1.5, and applying the triangle inequality for vector norms gives

  $$\begin{aligned}\|A\|_1 = \|Ay\|_1 = \|y_1 Ae_1 + \cdots + y_n Ae_n\|_1 &\leq |y_1|\|Ae_1\|_1 + \cdots + |y_n|\|Ae_n\|_1 \\ &\leq (|y_1| + \cdots + |y_n|) \max_{1\leq j\leq n} \|Ae_j\|_1.\end{aligned}$$

  From $|y_1| + \cdots + |y_n| = \|y\|_1 = 1$ follows $\|A\|_1 \leq \max_{1\leq j\leq n} \|Ae_j\|_1$. □

The matrix infinity norm is equal to the maximal absolute row sum.

Fact 2.17 (Infinity Norm). Let $A \in \mathbb{C}^{m\times n}$. Then

$$\|A\|_\infty = \max_{1\leq i\leq m} \|A^* e_i\|_1 = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|.$$

Proof. Denote the rows of A by $r_i^* = e_i^* A$, and let r_k have the largest one norm, $\|r_k\|_1 = \max_{1\le i\le m} \|r_i\|_1$.

- Let y be a vector with $\|A\|_\infty = \|Ay\|_\infty$ and $\|y\|_\infty = 1$. Then

$$\|A\|_\infty = \|Ay\|_\infty = \max_{1\le i\le m} |r_i^* y| \le \max_{1\le i\le m} \|r_i\|_1 \|y\|_\infty = \|r_k\|_1,$$

  where the inequality follows from Fact 2.8. Hence $\|A\|_\infty \le \max_{1\le i\le} \|r_i\|_1$.
- For any vector y with $\|y\|_\infty = 1$ we have $\|A\|_\infty \ge \|Ay\|_\infty \ge |r_k^* y|$. Now we show how to choose the elements of y such that $\|r_k^* y\| = \|r_k\|_1$. Let $r_k^* = \begin{pmatrix}\rho_1 & \dots & \rho_n\end{pmatrix}$ be the elements of r_k^*. Choose the elements of y such that $\rho_j y_j = |\rho_j|$. That is, if $\rho_j = 0$, then $y_j = 0$, and otherwise $y_j = |\rho_j|/\rho_j$. Then $\|y\|_\infty = 1$ and $|r_k^* y| = \sum_{j=1}^n \rho_j y_j = \sum_{j=1}^n |\rho_j| = \|r_k\|_1$. Hence

$$\|A\|_\infty \ge |r_k^* y| = \|r_k\|_1 = \max_{1\le i\le m} \|r_i\|_1. \qquad \square$$

The p-norms satisfy the following *submultiplicative inequality*.

Fact 2.18 (Norm of a Product). If $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times p}$, then

$$\|AB\|_p \le \|A\|_p \|B\|_p.$$

Proof. Let $x \in \mathbb{C}^p$ such that $\|AB\|_p = \|ABx\|_p$ and $\|x\|_p = 1$. Applying Remark 2.15 twice gives

$$\|AB\|_p = \|ABx\|_p \le \|A\|_p \|Bx\|_p \le \|A\|_p \|B\|_p \|x\|_p = \|A\|_p \|B\|_p. \qquad \square$$

Since the computation of the two norm is more involved, we postpone it until later. However, even without knowing how to compute it, we can still derive several useful properties of the two norm. If x is a column vector, then $\|x\|_2^2 = x^* x$. We show below that an analogous property holds for matrices. We also show that a matrix and its transpose have the same two norm.

Fact 2.19 (Two Norm). Let $A \in \mathbb{C}^{m\times n}$. Then

$$\|A^*\|_2 = \|A\|_2, \qquad \|A^* A\|_2 = \|A\|_2^2.$$

Proof. The definition of the two norm implies that for some $x \in \mathbb{C}^n$ with $\|x\|_2 = 1$ we have $\|A\|_2 = \|Ax\|_2$. The definition of the *vector* two norm implies

$$\|A\|_2^2 = \|Ax\|_2^2 = x^* A^* A x \le \|x\|_2 \|A^* A x\|_2 \le \|A^* A\|_2,$$

where the first inequality follows from the Cauchy–Schwarz inequality in Fact 2.8 and the second inequality from the two norm of $A^* A$. Hence $\|A\|_2^2 \le \|A^* A\|_2$. Fact 2.18 implies $\|A^* A\|_2 \le \|A^*\|_2 \|A\|_2$. As a consequence,

$$\|A\|_2^2 \le \|A^* A\|_2 \le \|A^*\|_2 \|A\|_2, \qquad \|A\|_2 \le \|A^*\|_2.$$

The same reasoning applied to AA^* gives

$$\|A^*\|_2^2 \leq \|AA^*\|_2 \leq \|A\|_2\|A^*\|_2, \qquad \|A^*\|_2 \leq \|A\|_2.$$

Therefore $\|A^*\|_2 = \|A\|_2$ and $\|A^*A\|_2 = \|A\|_2^2$. □

If we omit a piece of a matrix, the norm does not increase but it can decrease.

Fact 2.20 (Norm of a Submatrix). Let $A \in \mathbb{C}^{m\times n}$. If B is a submatrix of A, then $\|B\|_p \leq \|A\|_p$.

Exercises

(i) Let $D \in \mathbb{C}^{n\times n}$ be a diagonal matrix with diagonal elements d_{jj}. Show that $\|D\|_p = \max_{1\leq j\leq n} |d_{jj}|$.

(ii) Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. Show: $\|A\|_p\|A^{-1}\|_p \geq 1$.

(iii) Show: If P is a permutation matrix, then $\|P\|_p = 1$.

(iv) Let $P \in \mathbb{R}^{m\times m}$, $Q \in \mathbb{R}^{n\times n}$ be permutation matrices and let $A \in \mathbb{C}^{m\times n}$. Show: $\|PAQ\|_p = \|A\|_p$.

(v) Let $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ be unitary. Show: $\|U\|_2 = \|V\|_2 = 1$, and $\|UBV\|_2 = \|B\|_2$ for any $B \in \mathbb{C}^{m\times n}$.

(vi) Let $x \in \mathbb{C}^n$. Show: $\|x^*\|_2 = \|x\|_2$ without using Fact 2.19.

(vii) Let $x \in \mathbb{C}^n$. Is $\|x\|_1 = \|x^*\|_1$, and $\|x\|_\infty = \|x^*\|_\infty$? Why or why not?

(viii) Let $x \in \mathbb{C}^n$ be the vector of all ones. Determine

$$\|x\|_1, \quad \|x^*\|_1, \quad \|x\|_\infty, \quad \|x^*\|_\infty, \quad \|x\|_2, \quad \|x^*\|_2.$$

(ix) For each of the two equalities, determine a class of matrices A that satisfy the equality $\|A\|_\infty = \|A\|_1$, and $\|A\|_\infty = \|A\|_1 = \|A\|_2$.

(x) Let $A \in \mathbb{C}^{m\times n}$. Then $\|A\|_\infty = \|A^*\|_1$.

1. Verify that the matrix p-norms do indeed satisfy the three properties of a matrix norm in Definition 2.12.
2. Let $A \in \mathbb{C}^{m\times n}$. Prove:

$$\max_{i,j} |a_{ij}| \leq \|A\|_2 \leq \sqrt{mn}\max_{i,j} |a_{ij}|,$$

$$\frac{1}{\sqrt{n}}\|A\|_\infty \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty,$$

$$\frac{1}{\sqrt{m}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{n}\|A\|_1.$$

3. Norms of Outer Products.
Let $x \in \mathbb{C}^m$ and $y \in \mathbb{C}^n$. Show:

$$\|xy^*\|_2 = \|x\|_2\|y\|_2, \qquad \|xy^*\|_\infty = \|x\|_\infty\|y\|_1.$$

4. Given an approximate solution z, here is the matrix perturbation of smallest two norm that "realizes" z, in the sense that the perturbed system has z as a solution.
   Let $A \in \mathbb{C}^{n \times n}$, $Ax = b$, and $z \neq 0$. Show: Among all matrices E with $(A+E)z = b$ the matrix $E_0 = (b - Az)z^\dagger$ has the smallest two norm, where $z^\dagger = (z^*z)^{-1}z^*$.
5. Norms of Idempotent Matrices.
   Show: If $A \neq 0$ is idempotent, then $\|A\|_p \geq 1$. If A is also Hermitian, then $\|A\|_2 = 1$.
6. Let $A \in \mathbb{C}^{n \times n}$. Show: Among all Hermitian matrices, $\frac{1}{2}(A + A^*)$ is the matrix that is closest to A in the two norm.

2.7 Conditioning of Matrix Addition and Multiplication

We derive normwise relative bounds for matrix addition and subtraction, as well as for matrix multiplication.

Fact 2.21 (Matrix Addition). Let $U, V, \tilde{U}, \tilde{V} \in \mathbb{C}^{m \times n}$ such that $U, V, U+V \neq 0$. Then

$$\frac{\|\tilde{U} + \tilde{V} - (U+V)\|_p}{\|U+V\|_p} \leq \frac{\|U\|_p + \|V\|_p}{\|U+V\|_p} \max\{\epsilon_U, \epsilon_V\},$$

where

$$\epsilon_U = \frac{\|\tilde{U} - U\|_p}{\|U\|_p}, \qquad \epsilon_V = \frac{\|\tilde{V} - V\|_p}{\|V\|_p}.$$

Proof. The triangle inequality implies

$$\begin{aligned}\|\tilde{U} + \tilde{V} - (U+V)\|_p &\leq \|\tilde{U} - U\|_p + \|\tilde{V} - V\|_p = \|U\|_p\epsilon_U + \|V\|_p\epsilon_V \\ &\leq (\|U\|_p + \|V\|_p)\max\{\epsilon_U, \epsilon_V\}.\end{aligned}$$ □

The condition number for adding, or subtracting, the matrices U and V is $(\|U\|_p + \|V\|_p)/\|U+V\|_p$. It is analogous to the condition number for scalar subtraction in Fact 2.4. If $\|U\|_p + \|V\|_p \approx \|U+V\|_p$, then the matrix addition $U+V$ is well-conditioned in the normwise relative sense. But if $\|U\|_p + \|V\|_p \gg \|U+V\|_p$, then the matrix addition $U+V$ is ill-conditioned in the normwise relative sense.

Fact 2.22 (Matrix Multiplication). Let $U, \tilde{U} \in \mathbb{C}^{m \times n}$ and $V, \tilde{V} \in \mathbb{C}^{n \times p}$ such that $U, V, UV \neq 0$. Then

$$\frac{\|\tilde{U}\tilde{V} - UV\|_p}{\|UV\|_p} \leq \frac{\|U\|_p\|V\|_p}{\|UV\|_p}(\epsilon_U + \epsilon_V + \epsilon_U\epsilon_V),$$

where

$$\epsilon_U = \frac{\|\tilde{U} - U\|_p}{\|U\|_p}, \qquad \epsilon_V = \frac{\|\tilde{V} - V\|_p}{\|V\|_p}.$$

Proof. If $\tilde{U} = U + E$ and $\tilde{V} = V + F$, then

$$\tilde{U}\tilde{V} - UV = (U+E)(V+F) - UV = UF + EV + EF.$$

Now take norms, apply the triangle inequality, and divide by $\|UV\|_p$. □

Fact 2.22 shows that the normwise relative condition number for multiplying matrices U and V is $\|U\|_p\|V\|_p/\|UV\|_p$. If $\|U\|_p\|V\|_p \approx \|UV\|_p$, then the matrix multiplication UV is well-conditioned in the normwise relative sense. However, if $\|U\|_p\|V\|_p \gg \|UV\|_p$, then the matrix multiplication UV is ill-conditioned in the normwise relative sense.

Exercises

(i) What is the two-norm condition number of a product where one of the matrices is unitary?

(ii) Normwise absolute condition number for matrix multiplication when one of the matrices is perturbed.
Let $U, V \in \mathbb{C}^{n\times n}$, and U be nonsingular. Show:

$$\frac{\|F\|_p}{\|U^{-1}\|_p} \le \|U(V+F) - UV\|_p \le \|U\|_p\|F\|_p.$$

(iii) Here is a bound on the normwise relative error for matrix multiplication with regard to the perturbed product.
Let $U \in \mathbb{C}^{m\times n}$ and $V \in \mathbb{C}^{n\times m}$. Show: If $(U+E)(V+F) \ne 0$, then

$$\frac{\|(U+E)(V+F) - UV\|_p}{\|(U+E)(V+F)\|_p} \le \frac{\|U+E\|_p\|V+F\|_p}{\|(U+E)(V+F)\|_p}\left(\epsilon_U + \epsilon_V + \epsilon_U\epsilon_V\right),$$

where

$$\epsilon_U = \frac{\|E\|_p}{\|U+E\|_p}, \qquad \epsilon_V = \frac{\|F\|_p}{\|V+F\|_p}.$$

2.8 Conditioning of Matrix Inversion

We determine the sensitivity of the inverse to perturbations in the matrix.

We start by bounding the inverse of a perturbed identity matrix. If the norm of the perturbation is sufficiently small, then the perturbed identity matrix is nonsingular.

Fact 2.23 (Inverse of Perturbed Identity). If $A \in \mathbb{C}^{n\times n}$ and $\|A\|_p < 1$, then $I + A$ is nonsingular and

$$\frac{1}{1+\|A\|_p} \le \|(I+A)^{-1}\|_p \le \frac{1}{1-\|A\|_p}.$$

If also $\|A\|_p \le 1/2$, then $\|(I+A)^{-1}\|_p \le 2$.

Proof. Suppose, to the contrary, that $\|A\|_p < 1$ and $I+A$ is singular. Then there is a vector $x \neq 0$ such that $(I+A)x = 0$. Hence $\|x\|_p = \|Ax\|_p \leq \|A\|_p\|x\|_p$ implies $\|A\|_p \geq 1$, a contradiction.

- Lower bound: $I = (I+A)(I+A)^{-1}$ implies
$$1 = \|I\|_p \leq \|I+A\|_p\|(I+A)^{-1}\|_p \leq (1+\|A\|_p)\|(I+A)^{-1}\|_p.$$
- Upper bound: From
$$I = (I+A)(I+A)^{-1} = (I+A)^{-1} + A(I+A)^{-1}$$
follows
$$1 = \|I\|_p \geq \|(I+A)^{-1}\|_p - \|A(I+A)^{-1}\|_p \geq (1-\|A\|_p)\|(I+A)^{-1}\|_p.$$
If $\|A\|_p \leq 1/2$, then $1/(1-\|A\|_p) \leq 2$. □

Below is the corresponding result for inverses of general matrices.

Corollary 2.24 (Inverse of Perturbed Matrix). *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and $\|A^{-1}E\|_p < 1$. Then $A+E$ is nonsingular and*
$$\|(A+E)^{-1}\|_p \leq \frac{\|A^{-1}\|_p}{1-\|A^{-1}E\|_p}.$$
If also $\|A^{-1}\|_p\|E\|_p \leq 1/2$, then $\|(A+E)^{-1}\|_p \leq 2\|A^{-1}\|_p$.

Proof. Since A is nonsingular, we can write $A+E = A(I+A^{-1}E)$. From $\|A^{-1}E\|_p < 1$ follows with Fact 2.23 that $I+A^{-1}E$ is nonsingular. Hence $A+E$ is nonsingular. Its inverse can be written as $(A+E)^{-1} = (I+A^{-1}E)^{-1}A^{-1}$. Now take norms and apply Fact 2.23.

The second assertion follows from $\|A^{-1}E\|_p \leq \|A^{-1}\|_p\|E\|_p \leq 1/2$. □

Corollary 2.24 implies that if the perturbation E is sufficiently small, then $\|(A+E)^{-1}\|_p$ exceeds $\|A^{-1}\|_p$ by a factor of at most two.

We use the above bounds to derive normwise condition numbers for the inverses of general nonsingular matrices. A perturbation of a nonsingular matrix remains nonsingular if the perturbation is small enough in the normwise relative sense.

Fact 2.25. If $A \in \mathbb{C}^{n\times n}$ is nonsingular and $\|A^{-1}E\|_p < 1$, then
$$\|(A+E)^{-1} - A^{-1}\|_p \leq \|A^{-1}\|_p \frac{\|A^{-1}E\|_p}{1-\|A^{-1}E\|_p}.$$
If also $\|A^{-1}\|_p\|E\|_p \leq 1/2$, then
$$\frac{\|(A+E)^{-1} - A^{-1}\|_p}{\|A^{-1}\|_p} \leq 2\kappa_p(A)\frac{\|E\|_p}{\|A\|_p},$$
where $\kappa_p(A) = \|A\|_p\|A^{-1}\|_p \geq 1$.

Proof. Corollary 2.24 implies that $A+E$ is nonsingular. Abbreviating $F = A^{-1}E$, we obtain for the absolute difference

$$\begin{aligned}(A+E)^{-1} - A^{-1} &= (I+F)^{-1}A^{-1} - A^{-1}\\ &= \Big((I+F)^{-1} - I\Big)A^{-1} = -(I+F)^{-1}FA^{-1},\end{aligned}$$

where the last equation follows from $(I+F)^{-1}(I+F) = I$. Taking norms and applying the first bound in Fact 2.23 yields

$$\|(A+E)^{-1} - A^{-1}\|_p \le \|(I+F)^{-1}\|_p \|F\|_p \|A^{-1}\|_p \le \|A^{-1}\|_p \, \frac{\|F\|}{1-\|F\|_p}.$$

If $\|A^{-1}\|_p \|E\|_p \le 1/2$, then the second bound in Fact 2.23 implies

$$\|(A+E)^{-1} - A^{-1}\|_p \le 2\|F\|_p \|A^{-1}\|_p,$$

where

$$\|F\|_p \le \|A^{-1}\|_p \|E\|_p = \|A\|_p \|A^{-1}\|_p \frac{\|E\|_p}{\|A\|_p} = \kappa_p(A)\frac{\|E\|_p}{\|A\|_p}.$$

The lower bound for $\kappa_p(A)$ follows from

$$1 = \|I\|_p = \|AA^{-1}\|_p \le \|A\|_p \|A^{-1}\|_p = \kappa_p(A).$$ □

Remark 2.26. *We can conclude the following from Fact 2.25*:

- *The inverse of A is well-conditioned in the absolute sense if its norm is "small." In particular, the perturbed matrix is nonsingular if the perturbation has small enough norm.*
- *The inverse of A is well-conditioned in the relative sense if $\kappa_p(A)$ is "close to"* 1. *Note that $\kappa_p(A) \ge 1$.*

Definition 2.27. *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. The number $\kappa_p(A) = \|A\|_p \|A^{-1}\|_p$ is a* normwise relative condition number of A with respect to inversion.

According to Fact 2.25, a perturbed matrix $A+E$ is nonsingular if $\|A^{-1}E\|_p < 1$. Is this bound pessimistic, or is it tight? Does it imply that if $\|A^{-1}E\|_p = 1$, then $A+E$ can be singular? The answer is "yes." We illustrate this now for the two norm.

Example 2.28. Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. We show how to construct an outer product E such that $\|A^{-1}E\|_2 = 1$ and $A+E$ is singular.

Set $E = -yx^*/\|x\|_2^2$, where $x \neq 0$ and $y \neq 0$ are vectors we still need to choose. Since E is an outer product, Exercise 3 in Section 2.6 implies

$$\|A^{-1}E\|_2 = \frac{\|(A^{-1}y)x^*\|_2}{\|x\|_2^2} = \frac{\|A^{-1}y\|_2\|x\|_2}{\|x\|_2^2} = \frac{\|A^{-1}y\|_2}{\|x\|_2}.$$

Choosing $x = A^{-1}y$ gives $\|A^{-1}E\|_2 = 1$ and $(A+E)x = Ax + Ex = Ax - y = 0$. Since $(A+E)x = 0$ for $x \neq 0$, the matrix $A+E$ must be singular.

Therefore, if A is nonsingular, $y \neq 0$ is any vector, $x = A^{-1}y$, and $E = yx^*/\|x\|_2^2$, then $\|A^{-1}E\|_2 = 1$ and $A+E$ is singular. ∎

Exercise 3 in Section 2.6 implies that the two norm of the perturbation in Example 2.28 is $\|E\|_2 = \|y\|_2/\|x\|_2 = \|y\|_2/\|A^{-1}y\|_2$. What is the smallest two norm a matrix E can have that makes $A+E$ singular? We show that the smallest norm such an E can have is equal to $1/\|A^{-1}\|_2$.

Fact 2.29 (Absolute Distance to Singularity). Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. Then

$$\min\{\|E\|_2 : A+E \text{ is singular}\} = \frac{1}{\|A^{-1}\|_2}.$$

Proof. Let $E \in \mathbb{C}^{n\times n}$ be any matrix such that $A+E$ is singular. Then there is a vector $x \neq 0$ so that $(A+E)x = 0$. Hence $\|x\|_2 = \|A^{-1}Ex\|_2 \leq \|A^{-1}\|_2\|E\|_2\|x\|_2$ implies $\|E\|_2 \geq 1/\|A^{-1}\|_2$. Since this is true for any E that makes $A+E$ singular, $1/\|A^{-1}\|_2$ is a lower bound for the absolute distance of A to singularity.

Now we show that there is a matrix E_0 that achieves equality. Construct E_0 as in Example 2.28, and choose the vector y such that $\|A^{-1}\|_2 = \|A^{-1}y\|_2$ and $\|y\|_2 = 1$. Then $\|E_0\|_2 = \|y\|_2\|A^{-1}y\|_2 = 1/\|A^{-1}\|_2$. □

Corollary 2.30 (Relative Distance to Singularity). *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. Then*

$$\min\left\{\frac{\|E\|_2}{\|A\|_2} : A+E \textit{ is singular}\right\} = \frac{1}{\kappa_2(A)},$$

where $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2$.

Therefore, matrices that are ill-conditioned with respect to inversion are close to singular, and vice versa. In other words, matrices that are close to being singular have sensitive inverses.

The example below illustrates that absolute and relative distance to singularity are not the same.

Example. Just because a matrix is close to singularity in the absolute sense does not imply that it is also close to singularity in the relative sense. To see this, let

$$A = \begin{pmatrix} \epsilon & \epsilon \\ 0 & \epsilon \end{pmatrix}, \qquad 0 < \epsilon \ll 1, \qquad A^{-1} = \begin{pmatrix} \frac{1}{\epsilon} & \frac{1}{\epsilon} \\ 0 & \frac{1}{\epsilon} \end{pmatrix}.$$

Exercise 2 in Section 2.6 implies for an $n\times n$ matrix B that $\|B\|_2 \leq n\max_{ij}|b_{ij}|$. Hence $\epsilon \leq \|A\|_2 \leq 2\epsilon$ and $\frac{1}{\epsilon} \leq \|A^{-1}\|_2 \leq \frac{2}{\epsilon}$. Therefore,

$$\frac{\epsilon}{2} \leq \frac{1}{\|A^{-1}\|_2} \leq \epsilon, \qquad \frac{1}{4} \leq \frac{1}{\kappa_2(A)} \leq 1,$$

so that A is close to singularity in the absolute sense, but far from singularity in the relative sense. ∎

Exercises

(i) Let $A \in \mathbb{C}^{n\times n}$ be unitary. Show: $\kappa_2(A) = 1$.

(ii) Let $A, B \in \mathbb{C}^{n\times n}$ be nonsingular. Show: $\kappa_p(AB) \leq \kappa_p(A)\kappa_p(B)$.

(iii) Residuals for Matrix Inversion.
Let $A, A+E \in \mathbb{C}^{n\times n}$ be nonsingular, and let $Z = (A+E)^{-1}$. Show:

$$\|AZ - I_n\|_p \leq \|E\|_p\|Z\|_p, \qquad \|ZA - I_n\|_p \leq \|E\|_p\|Z\|_p.$$

1. For small enough perturbations, the identity matrix is well-conditioned with respect to inversion, in the normwise absolute and relative sense.
Show: If $A \in \mathbb{C}^{n\times n}$ and $\|A\|_p < 1$, then

$$\|(I+A)^{-1} - I\|_p \leq \frac{\|A\|_p}{1-\|A\|_p},$$

and if $\|A\|_p \leq 1/2$, then

$$\|(I+A)^{-1} - I\|_p \leq 2\|A\|_p.$$

2. If the norm of A is small enough, then $(I+A)^{-1} \approx I - A$.
Let $A \in \mathbb{C}^{n\times n}$ and $\|A\|_p \leq 1/2$. Show:

$$\|(I-A) - (I+A)^{-1}\|_p \leq 2\|A\|_p^2.$$

3. One can also bound the relative error with regard to $(A+E)^{-1}$.
Let A and $A+E$ be nonsingular. Show:

$$\frac{\|(A+E)^{-1} - A^{-1}\|_p}{\|(A+E)^{-1}\|_p} \leq \kappa_p(A)\frac{\|E\|_p}{\|A\|_p}.$$

4. A matrix $A \in \mathbb{C}^{n\times n}$ is called *strictly column diagonally dominant* if

$$\sum_{i=1, i\neq j}^{n} |a_{ij}| < |a_{jj}|, \qquad 1 \leq j \leq n.$$

Show: A strictly column diagonally dominant matrix is nonsingular.

5. Let $A \in \mathbb{C}^{n\times n}$ be nonsingular. Show: $\kappa_p(A) \geq \|A\|_p/\|A-B\|_p$ for any singular matrix $B \in \mathbb{C}^{n\times n}$.

3. Linear Systems

We present algorithms for solving systems of linear equations whose coefficient matrix is nonsingular, and we discuss the accuracy of these algorithms.

3.1 The Meaning of $Ax = b$

First we examine when a linear system has a solution.

Fact 3.1 (Two Views of a Linear System). Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^{m\times 1}$.

1. The linear system $Ax = b$ has a solution if and only if there is a vector x that solves the m equations

$$r_1x = b_1, \qquad \dots, \qquad r_mx = b_m,$$

where

$$A = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix}.$$

2. The linear system $Ax = b$ has a solution if and only if b is a linear combination of the columns of A,

$$b = a_1x_1 + \cdots + a_nx_n,$$

where

$$A = \begin{pmatrix} a_1 & \dots & a_n \end{pmatrix}, \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

When the matrix is nonsingular, the linear system has a solution for any right-hand side, and the solution can be represented in terms of the inverse of A.

Corollary 3.2 (Existence and Uniqueness). *If $A \in \mathbb{C}^{n\times n}$ is nonsingular, then $Ax = b$ has the unique solution $x = A^{-1}b$ for every $b \in \mathbb{C}^n$.*

Before we discuss algorithms for solving linear systems we need to take into account, as discussed in Chapter 2, that the matrix and right-hand side may be contaminated by uncertainties. This means, instead of solving $Ax = b$, we solve a perturbed system $(A+E)z = b+f$. We want to determine how sensitive the solution is to the perturbations f and E.

Even if we don't know the perturbations E and f, we can estimate them from the approximate solution z. To this end, define the residual $r = Az - b$. We can view z as the solution to a system with perturbed right-hand side, $Az = b+r$. If $z \neq 0$, then we can also view z as the solution to a system with perturbed matrix,

$$(A+E)z = b, \qquad \text{where} \quad E = -\frac{rz^*}{\|z\|_2^2},$$

see Exercise 1 below.

Exercises

(i) Determine the solution to $Ax = b$ when A is unitary (orthogonal).

(ii) Determine the solution to $Ax = b$ when A is involutory.

(iii) Let A consist of several columns of a unitary matrix, and let b be such that the linear system $Ax = b$ has a solution. Determine a solution to $Ax = b$.

(iv) Let A be idempotent. When does the linear system $Ax = b$ have a solution for every b?

(v) Let A be a triangular matrix. When does the linear system $Ax = b$ have a solution for *any* right-hand side b?

(vi) Let $A = uv^*$ be an outer product, where u and v are column vectors. For which b does the linear system $Ax = b$ have a solution?

(vii) Determine a solution to the linear system $\begin{pmatrix} A & B \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$ when A is nonsingular. Is the solution unique?

1. Matrix Perturbations from Residuals.
   This problem shows how to construct a matrix perturbation from the residual. Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $Ax = b$, and $z \in \mathbb{C}^n$ a nonzero approximation to x. Show that $(A+E_0)z = b$, where $E_0 = (b - Az)z^\dagger$ and $z^\dagger = (z^*z)^{-1}z^*$; and that $(A+E)z = b$, where $E = E_0 + G(I - zz^\dagger)$ and $G \in \mathbb{C}^{n\times n}$ is any matrix.
2. In Problem 1 above show that, among all matrices F that satisfy$(A+F)z = b$, the matrix E_0 is one with smallest two norm, i.e., $\|E_0\|_2 \leq \|F\|_2$.

3.2 Conditioning of Linear Systems

We derive normwise bounds for the conditioning of linear systems. The following two examples demonstrate that it is not obvious how to estimate the accuracy of an approximate solution z for a linear system $Ax = b$. In particular, they illustrate that the residual $r = Az - b$ may give misleading information about how close z is to x.

Example 3.3. We illustrate that a totally wrong approximate solution can have a small residual norm.

Consider the linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1+\epsilon \end{pmatrix}, \qquad b = \begin{pmatrix} 2 \\ 2+\epsilon \end{pmatrix}, \qquad 0 < \epsilon \ll 1, \qquad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

whose solution x is approximated by $z = \begin{pmatrix} 2 & 0 \end{pmatrix}^T$. The residual

$$r = Az - b = \begin{pmatrix} 0 \\ -\epsilon \end{pmatrix}$$

has a small norm, $\|r\|_p = \epsilon$, because ϵ is small. This appears to suggest that z does a good job of solving the linear system. However, comparing z to the exact solution,

$$z - x = \begin{pmatrix} -1 \\ 1 \end{pmatrix},$$

shows that z is a bad approximation to x. Therefore, a small residual norm does not imply that z is close to x. ■

The same thing can happen even for triangular matrices, as the next example shows.

Example 3.4. For the linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 10^8 \\ 0 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1+10^8 \\ 1 \end{pmatrix}, \qquad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

consider the approximate solution

$$z = \begin{pmatrix} 0 \\ 1+10^{-8} \end{pmatrix}, \qquad r = Az - b = \begin{pmatrix} 0 \\ 10^{-8} \end{pmatrix}.$$

As in the previous example, the residual has small norm, i.e., $\|r\|_p = 10^{-8}$, but z is totally inaccurate,

$$z - x = \begin{pmatrix} -1 \\ 10^{-8} \end{pmatrix}.$$

Again, the residual norm is deceptive. It is small even though z is a bad approximation to x. ■

The bound below explains why inaccurate approximations can have residuals with small norm.

Fact 3.5 (Residual Bound). Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $Ax = b$, and $b \neq 0$. If $r = Az - b$, then

$$\frac{\|z-x\|_p}{\|x\|_p} \le \kappa_p(A)\, \frac{\|r\|_p}{\|A\|_p\|x\|_p}.$$

Proof. If $b \neq 0$ and A is nonsingular, then $x \neq 0$; see Fact 1.10. The desired bound follows immediately from the perturbation bound for matrix multiplication: Apply Fact 2.22 to $U = \tilde{U} = A^{-1}$, $V = b$, $\tilde{V} = b + r$, $\epsilon_U = 0$, and $\epsilon_V = \|r\|_p/\|b\|_p$ to obtain

$$\frac{\|z-x\|_p}{\|x\|_p} \le \frac{\|A^{-1}\|_p\|b\|_p}{\|A^{-1}b\|_p}\,\frac{\|r\|_p}{\|b\|_p} = \|A\|_p\|A^{-1}\|_p\,\frac{\|r\|_p}{\|A\|_p\|x\|_p}. \qquad \square$$

The quantity $\kappa_p(A)$ is the normwise relative condition number of A with respect to inversion; see Definition 2.27. The bound in Fact 3.5 implies that the linear system $Ax = b$ is well-conditioned if $\kappa_p(A)$ is small. In particular, if $\kappa_p(A)$ is small and the relative residual norm $\frac{\|r\|_p}{\|A\|_p\|x\|_p}$ is also small, then the approximate solution z has a small error (in the normwise relative sense). However, if $\kappa_p(A)$ is large, then the linear system is ill-conditioned. We return to Examples 3.3 and 3.4 to illustrate the bound in Fact 3.5.

Example. The linear system $Ax = b$ in Example 3.3 is

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1+\epsilon \end{pmatrix}, \qquad b = \begin{pmatrix} 2 \\ 2+\epsilon \end{pmatrix}, \qquad 0 < \epsilon \ll 1, \qquad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and has an approximate solution $z = \begin{pmatrix} 2 & 0 \end{pmatrix}^T$ with residual

$$r = Az - b = \begin{pmatrix} 0 \\ -\epsilon \end{pmatrix}.$$

The relative error in the infinity norm is $\|z-x\|_\infty/\|x\|_\infty = 1$, indicating that z has no accuracy whatsoever. To see what the bound in Fact 3.5 predicts, we determine the inverse

$$A^{-1} = \frac{1}{\epsilon}\begin{pmatrix} 1+\epsilon & -1 \\ -1 & 1 \end{pmatrix},$$

the matrix norms

$$\|A\|_\infty = 2+\epsilon, \qquad \|A^{-1}\|_\infty = \frac{2+\epsilon}{\epsilon}, \qquad \kappa_\infty(A) = \frac{(2+\epsilon)^2}{\epsilon},$$

as well as the ingredients for the relative residual norm

$$\|r\|_\infty = \epsilon, \qquad \|x\|_\infty = 1, \qquad \frac{\|r\|_\infty}{\|A\|_\infty\|x\|_\infty} = \frac{\epsilon}{2+\epsilon}.$$

Since $\kappa_\infty(A) \approx 4/\epsilon$, the system $Ax = b$ is ill-conditioned. The bound in Fact 3.5 equals

$$\frac{\|z-x\|_\infty}{\|x\|_\infty} \le \kappa_\infty(A)\,\frac{\|r\|_\infty}{\|A\|_\infty\|x\|_\infty} = 2+\epsilon,$$

and so it correctly predicts the total inaccuracy of z. The small relative residual norm of about $\epsilon/2$ here is deceptive because the linear system is ill-conditioned. ■

Even triangular systems are not immune from ill-conditioning.

Example 3.6. The linear system $Ax = b$ in Example 3.4 is

$$A = \begin{pmatrix} 1 & 10^8 \\ 0 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1+10^8 \\ 1 \end{pmatrix}, \qquad x = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

and has an approximate solution $z = \begin{pmatrix} 0 & 1+10^{-8} \end{pmatrix}^T$ with residual

$$r = Az - b = \begin{pmatrix} 0 \\ 10^{-8} \end{pmatrix}.$$

The normwise relative error in the infinity norm is $\|z-x\|_\infty/\|x\|_\infty = 1$ and indicates that z has no accuracy. From

$$A^{-1} = \begin{pmatrix} 1 & -10^8 \\ 0 & 1 \end{pmatrix}$$

we determine the condition number for $Ax = b$ as $\kappa_\infty(A) = (1+10^8)^2 \approx 10^{16}$. Note that conditioning of triangular systems cannot be detected by merely looking at the diagonal elements; the diagonal elements of A are equal to 1 and far from zero, but nevertheless A is ill-conditioned with respect to inversion.

The relative residual norm is

$$\frac{\|r\|_\infty}{\|A\|_\infty\|x\|_\infty} = \frac{10^{-8}}{1+10^8} \approx 10^{-16}.$$

As a consequence, the bound in Fact 3.5 equals

$$\frac{\|z-x\|_\infty}{\|x\|_\infty} \le \kappa_\infty(A)\,\frac{\|r\|_\infty}{\|A\|_\infty\|x\|_\infty} = (1+10^8)10^{-8} \approx 1,$$

and it correctly predicts that z has no accuracy at all. ■

The residual bound below does not require knowledge of the exact solution. The bound is analogous to the one in Fact 3.5 but bounds the relative error with regard to the perturbed solution.

Fact 3.7 (Computable Residual Bound). Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and $Ax = b$. If $z \neq 0$ and $r = Az - b$, then

$$\frac{\|z-x\|_p}{\|z\|_p} \le \kappa_p(A)\,\frac{\|r\|_p}{\|A\|_p\|z\|_p}.$$

We will now derive bounds that separate the perturbations in the matrix from those in the right-hand side. We first present a bound with regard to the relative error in the perturbed solution because it is easier to derive.

Fact 3.8 (Matrix and Right-Hand Side Perturbation). Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and let $Ax = b$. If $(A+E)z = b+f$ with $z \neq 0$, then

$$\frac{\|z-x\|_p}{\|z\|_p} \leq \kappa_p(A)\left(\epsilon_A + \epsilon_f\right),$$

where

$$\epsilon_A = \frac{\|E\|_p}{\|A\|_p}, \qquad \epsilon_f = \frac{\|f\|_p}{\|A\|_p\|z\|_p}.$$

Proof. In the bound in Fact 3.7, the residual r accounts for both perturbations, because if $(A+E)z = b+f$, then $r = Az-b = f-Ez$. Replacing $\|r\|_p \leq \|E\|_p\|z\|_p + \|f\|_p$ in Fact 3.7 gives the desired bound. □

Below is an analogous bound for the error with regard to the exact solution. In contrast to Fact 3.8, the bound below requires the perturbed matrix to be nonsingular.

Fact 3.9 (Matrix and Right-Hand Side Perturbation). Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, and let $Ax = b$ with $b \neq 0$. If $(A+E)z = b+f$ with $\|A^{-1}\|_p\|E\|_p \leq 1/2$, then

$$\frac{\|z-x\|_p}{\|x\|_p} \leq 2\kappa_p(A)\left(\epsilon_A + \epsilon_f\right),$$

where

$$\epsilon_A = \frac{\|E\|_p}{\|A\|_p}, \qquad \epsilon_f = \frac{\|f\|_p}{\|A\|_p\|x\|_p}.$$

Proof. We could derive the desired bound from the perturbation bound for matrix multiplication in Fact 2.22 and matrix inversion in Fact 2.25. However, the resulting bound would not be tight, because it does not exploit any relation between matrix and right-hand side. This is why we start from scratch.

Subtracting $(A+E)x = b+Ex$ from $(A+E)z = b+f$ gives $(A+E)(z-x) = f-Ex$. Corollary 2.24 implies that $A+E$ is nonsingular. Hence we can write $z-x = (A+E)^{-1}(-Ex+f)$. Taking norms and applying Corollary 2.24 yields

$$\|z-x\|_p \leq 2\|A^{-1}\|_p(\|E\|_p\|x\|_p + \|f\|_p) = 2\kappa_p(A)(\epsilon_A+\epsilon_f)\,\|x\|_p. \qquad \square$$

We can simplify the bound in Fact 3.9 and obtain a weaker version.

Corollary 3.10. *Let $Ax = b$ with $A \in \mathbb{C}^{n\times n}$ nonsingular and $b \neq 0$. If $(A+E)z = b+f$ with $\|A^{-1}\|_p\|E\|_p < 1/2$, then*

$$\frac{\|z-x\|_p}{\|x\|_p} \leq 2\kappa_p(A)\left(\epsilon_A+\epsilon_b\right), \qquad \textit{where} \quad \epsilon_A = \frac{\|E\|_p}{\|A\|_p}, \quad \epsilon_b = \frac{\|f\|_p}{\|b\|_p}.$$

Proof. In Fact 3.9 bound $\|b\|_p \leq \|A\|_p \|x\|_p$. □

Effect of the Right-Hand Side. So far we have focused almost exclusively on the effect that the matrix has on the conditioning of the linear system, and we have ignored the right-hand side. The advantage of this approach is that the resulting perturbation bounds hold for all right-hand sides. However, the bounds can be too pessimistic for some right-hand sides, as the following example demonstrates.

Example 3.11. We illustrate that a favorable right-hand side can improve the conditioning of a linear system. Let's change the right-hand side in Example 3.6 and consider the linear system $Ax = b$ with

$$A = \begin{pmatrix} 1 & 10^8 \\ 0 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad x = \begin{pmatrix} 1-10^8 \\ 1 \end{pmatrix}$$

and the approximate solution

$$z = \begin{pmatrix} -10^8 - 9 \\ 1 + 10^{-7} \end{pmatrix}, \qquad r = Az - b = \begin{pmatrix} 0 \\ 10^{-7} \end{pmatrix}.$$

Although $\kappa_\infty(A) \approx 10^{16}$ implies that A is ill-conditioned with respect to inversion, the relative error in z is surprisingly small,

$$\frac{\|z-x\|_\infty}{\|x\|_\infty} = \frac{10}{1-10^8} \approx 10^{-7}.$$

The bound in Fact 3.5 recognizes this, too. From

$$\kappa_\infty(A) = (1+10^8)^2, \qquad \frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} = \frac{10^{-7}}{(10^8-1)(10^8+1)},$$

we obtain

$$\frac{\|z-x\|_\infty}{\|x\|_\infty} \leq \kappa_\infty(A)\,\frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} = \frac{10^8+1}{10^8-1}\,10^{-7} \approx 10^{-7}.$$

So, what is happening here? Observe that the relative residual norm is extremely small, $\frac{\|r\|_\infty}{\|A\|_\infty \|x\|_\infty} \approx 10^{-23}$, and that the norms of the matrix and solution are large compared to the norm of the right-hand side; i.e., $\|A\|_\infty \|x\|_\infty \approx 10^{16} \gg \|b\|_\infty = 1$. We can represent this situation by writing the bound in Fact 3.5 as

$$\frac{\|z-x\|_\infty}{\|x\|_\infty} \leq \frac{\|A^{-1}\|_\infty \|b\|_\infty}{\|A^{-1}b\|_\infty}\,\frac{\|r\|_\infty}{\|b\|_\infty}.$$

Because $\|A^{-1}\|_\infty \|b\|_\infty / \|A^{-1}b\|_\infty \approx 1$, the matrix multiplication of A^{-1} with b is well-conditioned with regard to changes in b. Hence the linear system $Ax = b$ is well-conditioned for this very particular right-hand side b. ■

Exercises

(i) Absolute Residual Bounds.
Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $Ax = b$, and $r = Az - b$ for some $z \in \mathbb{C}^n$. Show:

$$\frac{\|r\|_p}{\|A\|_p} \leq \|z-x\|_p \leq \|A^{-1}\|_p\|r\|_p.$$

(ii) Lower Bounds for Normwise Relative Error.
Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $Ax = b$, $b \neq 0$, and $r = Az - b$ for some $z \in \mathbb{C}^n$. Show:

$$\frac{\|r\|_p}{\|A\|_p\|x\|_p} \leq \frac{\|z-x\|_p}{\|x\|_p}, \qquad \frac{1}{\kappa_p(A)}\,\frac{\|r\|_p}{\|b\|_p} \leq \frac{\|z-x\|_p}{\|x\|_p}.$$

(iii) Relation between Relative Residual Norms.
Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $Ax = b$, $b \neq 0$, and $r = Az - b$ for some $z \in \mathbb{C}^n$. Show:

$$\frac{\|r\|_p}{\|A\|_p\|x\|_p} \leq \frac{\|r\|_p}{\|b\|_p} \leq \kappa_p(A)\frac{\|r\|_p}{\|A\|_p\|x\|_p}.$$

(iv) If a linear system is well-conditioned, and the relative residual norm is small, then the approximation has about the same norm as the solution.
Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and $b \neq 0$. Prove: If

$$\rho\kappa < 1, \qquad \text{where} \quad \kappa = \kappa_p(A), \qquad \rho = \frac{\|b - Az\|_p}{\|b\|_p},$$

then

$$1 - \kappa\rho \leq \frac{\|z\|_p}{\|x\|_p} \leq 1 + \kappa\rho.$$

(v) For this special right-hand side, the linear system is well-conditioned with regard to changes in the right-hand side.
Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $Ax = b$, and $Az = b + f$. Show: If $\|A^{-1}\|_p = \|A^{-1}b\|_p/\|b\|_p$, then

$$\frac{\|z-x\|_p}{\|x\|_p} \leq \frac{\|f\|_p}{\|b\|_p}.$$

1. Let $A \in \mathbb{C}^{n\times n}$ be the bidiagonal matrix

$$A = \begin{pmatrix} 1 & -\alpha & & & \\ & 1 & -\alpha & & \\ & & \ddots & \ddots & \\ & & & 1 & -\alpha \\ & & & & 1 \end{pmatrix}.$$

(a) Show:

$$\kappa_\infty(A) = \begin{cases} \frac{|\alpha|+1}{|\alpha|-1}\,(|\alpha|^n - 1) & \text{if } |\alpha| \neq 1, \\ 2n & \text{if } |\alpha| = 1. \end{cases}$$

Hint: See Exercise 4 in Section 1.13.

(b) Suppose we want to compute an approximation to the solution of $Ax = e_n$ when $\alpha = 2$ and $n = 100$. How small, approximately, must the residual norm be so that the normwise relative error bound is less than .1?

2. Componentwise Condition Numbers.
Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $b \neq 0$, and $Ax = b$. Prove: If $x_j \neq 0$, then

$$\frac{|z_j - x_j|}{|x_j|} \leq \kappa_j \frac{\|b - Az\|_p}{\|b\|_p}, \qquad \text{where} \quad \kappa_j = \frac{\|x\|_p}{|x|_j}\,\|e_j^* A^{-1}\|_p \|A\|_p.$$

We can interpret κ_j as the condition number for x_j. Which components of x would you expect to be sensitive to perturbations?

3. Condition Estimation.
Let A be nonsingular. Show how to determine a lower bound for $\kappa_p(A)$ with one linear system solution involving A.

3.3 Solution of Triangular Systems

Linear systems with triangular matrices are easy to solve. In the algorithm below we use the symbol "≡" to represent an assignment of a value.

ALGORITHM 3.1. Upper Triangular System Solution.

Input: Nonsingular, upper triangular matrix $A \in \mathbb{C}^{n\times n}$, vector $b \in \mathbb{C}^n$
Output: $x = A^{-1}b$

1. If $n = 1$, then $x \equiv b/A$.
2. If $n > 1$, partition

$$A = \begin{array}{c} \\ n-1 \\ 1 \end{array}\!\!\begin{array}{c} \begin{array}{cc} n-1 & 1 \end{array} \\ \begin{pmatrix} \hat{A} & a \\ 0 & a_{nn} \end{pmatrix} \end{array}, \qquad x = \begin{array}{c} n-1 \\ 1 \end{array}\begin{pmatrix} \hat{x} \\ x_n \end{pmatrix}, \qquad b = \begin{array}{c} n-1 \\ 1 \end{array}\begin{pmatrix} \hat{b} \\ b_n \end{pmatrix}.$$

(i) Set $x_n \equiv b_n/a_{nn}$.

(ii) Repeat the process on the smaller system $\hat{A}\hat{x} = \hat{b} - x_n a$.

The process of solving an upper triangular system is also called *backsubstitution*, and the process of solving a lower triangular system is called *forward elimination*.

Exercises

(i) Describe an algorithm to solve a nonsingular lower triangular system.

(ii) Solution of Block Upper Triangular Systems.
Even if A is not triangular, it may have a coarser triangular structure of which one can take advantage. For instance, let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

where A_{11} and A_{22} are nonsingular. Show how to solve $Ax = b$ by solving two smaller systems.

(iii) Conditioning of Triangular Systems.
This problem illustrates that a nonsingular triangular matrix is ill-conditioned if a diagonal element is small in magnitude compared to the other nonzero matrix elements.
Let $A \in \mathbb{C}^{n\times n}$ be upper triangular and nonsingular. Show:

$$\kappa_\infty(A) \geq \frac{\|A\|_\infty}{\min_{1\leq j\leq n} |a_{jj}|}.$$

3.4 Stability of Direct Methods

We do *not* solve general nonsingular systems $Ax = b$ by first forming A^{-1} and then multiplying by b (likewise, you would not compute $2/4$ by first forming $1/4$ and then multiplying by 2). It is too expensive and numerically less accurate; see Exercise 4 below.

A more efficient approach factors A into a product of simpler matrices and then solves a sequence of simpler linear systems. Examples of such factorizations include:

- LU factorization: $A = LU$ (if it exists), where L is lower triangular, and U is upper triangular.
- Cholesky factorization: $A = LL^*$ (if it exists), where L is lower triangular.
- QR factorization: $A = QR$, where Q is unitary and R is upper triangular. If A is real, then Q is real orthogonal.

Methods that solve linear systems by first factoring a matrix are called *direct methods*. In general, a direct method factors $A = S_1S_2$ (where "S" stands for "simpler matrix") and then computes the solution $x = A^{-1}b = S_2^{-1}S_1^{-1}b$ by solving two linear systems.

ALGORITHM 3.2. Direct Method.

Input: Nonsingular matrix $A \in \mathbb{C}^{n\times n}$, vector $b \in \mathbb{C}^n$
Output: Solution of $Ax = b$

```
1. Factor $A = S_1S_2$.
```

2. Solve the system $S_1 y = b$.
3. Solve the system $S_2 x = y$.

Each step of the above algorithm is itself a computational problem that may be sensitive to perturbations. We need to make sure that the algorithm does not introduce additional sensitivity by containing unnecessary ill-conditioned steps. For a direct method, this means that the factors S_1 and S_2 should be well-conditioned with respect to inversion. The example below illustrates that this cannot be taken for granted. That is, even if A is well-conditioned with respect to inversion, S_1 or S_2 can be ill-conditioned.

Example 3.12. The linear system $Ax = b$ with

$$A = \begin{pmatrix} \epsilon & 1 \\ 1 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} 1+\epsilon \\ 1 \end{pmatrix}, \qquad 0 < \epsilon \le 1/2,$$

has the solution $x = \begin{pmatrix} 1 & 1 \end{pmatrix}^T$. The linear system is well-conditioned because

$$A^{-1} = \begin{pmatrix} 0 & 1 \\ 1 & -\epsilon \end{pmatrix}, \qquad \kappa_\infty(A) = (1+\epsilon)^2 \le 9/4.$$

We can factor $A = S_1 S_2$ where

$$S_1 = \begin{pmatrix} 1 & 0 \\ \frac{1}{\epsilon} & 1 \end{pmatrix}, \qquad S_2 = \begin{pmatrix} \epsilon & 1 \\ 0 & -\frac{1}{\epsilon} \end{pmatrix}$$

and then solve the triangular systems $S_1 y = b$ and $S_2 x = y$. Suppose that we compute the factorization and the first linear system solution exactly, i.e.,

$$A = S_1 S_2, \qquad S_1 y = b, \qquad y = \begin{pmatrix} 1+\epsilon \\ -\frac{1}{\epsilon} \end{pmatrix},$$

and that we make errors only in the solution of the second system, i.e.,

$$S_2 z = y + r_2 = \begin{pmatrix} 1 \\ -\frac{1}{\epsilon} \end{pmatrix}, \qquad r_2 = \begin{pmatrix} -\epsilon \\ 0 \end{pmatrix}.$$

Then the computed solution satisfies

$$z = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \qquad \frac{\|z - x\|_\infty}{\|x\|_\infty} = 1.$$

The relative error is large because the leading component of z is completely wrong—although A is very well-conditioned. What happened? The triangular matrices S_1 and S_2 contain elements that are much larger in magnitude than the elements of A,

$$\|A\|_\infty = 1+\epsilon, \qquad \|S_1\|_\infty = \frac{1+\epsilon}{\epsilon}, \qquad \|S_2\|_\infty = \frac{1}{\epsilon},$$

and the same is true for the inverses,

$$\|A^{-1}\|_\infty = 1+\epsilon, \qquad \|S_1^{-1}\|_\infty = \|S_2^{-1}\|_\infty = \frac{1+\epsilon}{\epsilon}.$$

The condition numbers for S_1 and S_2 are

$$\kappa_\infty(S_1) = \left(\frac{1+\epsilon}{\epsilon}\right)^2 \approx \frac{1}{\epsilon^2}, \qquad \kappa_\infty(S_2) = \frac{1+\epsilon}{\epsilon^2} \approx \frac{1}{\epsilon^2}.$$

As a consequence, S_1 and S_2 are ill-conditioned with respect to inversion. Although the original linear system $Ax = b$ is well-conditioned, the algorithm contains steps that are ill-conditioned, namely, the solution of the linear systems $S_1 y = b$ and $S_2 x = y$. ∎

We want to avoid methods, like the one above, that factor a well-conditioned matrix into two ill-conditioned matrices. Such methods are called *numerically unstable*.

Definition 3.13. *An algorithm is (very informally)* numerically stable in exact arithmetic *if each step in the algorithm is not much worse conditioned than the original problem.*

If an algorithm contains steps that are much worse conditioned than the original problem, the algorithm is called numerically unstable.

The above definition talks about "stability in exact arithmetic," because in this book we do not take into account errors caused by floating arithmetic operations (analyses that estimate such errors can be rather tedious). However, if a problem is numerically unstable in exact arithmetic, then it is also numerically unstable in finite precision arithmetic, so that a distinction is not necessary in this case.

Below we analyze how the conditioning of the factors S_1 and S_2 affects the stability of Algorithm 3.2. The bounds are expressed in terms of relative residual norms from the linear systems.

Fact 3.14 (Stability in Exact Arithmetic of Direct Methods). Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $Ax = b$, $b \neq 0$, and

$$A + E = S_1 S_2, \qquad \epsilon_A = \frac{\|E\|_p}{\|A\|_p},$$

$$S_1 y = b + r_1, \qquad \epsilon_1 = \frac{\|r_1\|_p}{\|b\|_p},$$

$$S_2 z = y + r_2, \qquad \epsilon_2 = \frac{\|r_2\|_p}{\|y\|_p}.$$

If $\|A^{-1}\|_p \|E\|_p \leq 1/2$, then

$$\frac{\|z - x\|_p}{\|x\|_p} \leq \underbrace{2\kappa_p(A)}_{\text{condition}} (\epsilon_A + \epsilon_1 + \epsilon),$$

where

$$\epsilon = \underbrace{\frac{\|S_2^{-1}\|_p \|S_1^{-1}\|_p}{\|(A+E)^{-1}\|_p}}_{\text{stability}} \epsilon_2(1+\epsilon_1).$$

Proof. Expanding the right-hand side gives

$$(A+E)z = S_1 S_2 z = S_1(y+r_2) = S_1 y + S_1 r_2 = b + r_1 + S_1 r_2.$$

The obvious approach would be to apply Fact 3.9 to the perturbed linear system $(A+E)z = b + r_1 + S_1 r_2$. However, the resulting bound would be too pessimistic, because we did not exploit the relation between the matrix and the right-hand side. Instead, we can exploit this relation by subtracting $(A+E)x = b + Ex$ to obtain

$$(A+E)(z-x) = -Ex + r_1 + S_1 r_2.$$

Corollary 2.24 implies that $A+E$ is nonsingular, so that

$$z - x = (A+E)^{-1}(-Ex + r_1) + S_2^{-1} r_2.$$

Taking norms gives

$$\|z-x\|_p \le \|(A+E)^{-1}\|_p (\|E\|_p \|x\|_p + \|r_1\|_p) + \|S_2^{-1}\|_p \|r_2\|_p.$$

Substituting $\|r_1\|_p = \epsilon_1 \|b\|_p \le \epsilon_1 \|A\|_p \|x\|_p$ gives

$$\|z-x\|_p \le \|(A+E)^{-1}\|_p \|A\|_p (\epsilon_A + \epsilon_1) \|x\|_p + \|S_2^{-1}\|_p \|r_2\|_p.$$

It remains to bound $\|r_2\|_p$. From $\|r_2\|_p = \epsilon_2 \|y\|_p$ and $y = S_1^{-1}(b + r_1)$ follows

$$\|r_2\|_p = \epsilon_2 \|y\|_p \le \|S_1^{-1}\|_p (\|b\|_p + \|r_1\|_p).$$

Bounding $\|r_1\|_p$ as above yields

$$\|r_2\|_p \le \|S_1^{-1}\|_p \|A\|_p \|x\|_p \epsilon_2 (1+\epsilon_1).$$

We substitute this bound for $\|r_2\|_p$ into the above bound for $\|z-x\|_p$,

$$\|z-x\|_p \le \|A\|_p \|x\|_p \left(\|(A+E)^{-1}\|_p (\epsilon_A + \epsilon_1) + \|S_2^{-1}\|_p \|S_1^{-1}\|_p \epsilon_2 (1+\epsilon_1) \right).$$

Factoring out $\|(A+E)^{-1}\|_p$ and applying Corollary 2.24 gives the desired bound. □

Remark 3.15.

- *The numerical stability in exact arithmetic of a direct method can be represented by the condition number for multiplying the two matrices S_2^{-1} and S_1^{-1}, see Fact* 2.22, *since*

$$\frac{\|S_2^{-1}\|_p \|S_1^{-1}\|_p}{\|(A+E)^{-1}\|_p} = \frac{\|S_2^{-1}\|_p \|S_1^{-1}\|_p}{\|S_2^{-1} S_1^{-1}\|_p}.$$

- *If* $\|S_2^{-1}\|_p\|S_1^{-1}\|_p \approx \|(A+E)^{-1}\|_p$, *then the matrix multiplication* $S_2^{-1}S_1^{-1}$ *is well-conditioned. In this case the bound in Fact* 3.14 *is approximately* $2\kappa_p(A)(\epsilon_A+\epsilon_1+\epsilon_2(1+\epsilon_1))$, *and Algorithm* 3.2 *is numerically stable in exact arithmetic.*
- *If* $\|S_2^{-1}\|_p\|S_1^{-1}\|_p \gg \|(A+E)^{-1}\|_p$, *then Algorithm* 3.2 *is unstable.*

Example 3.16. Returning to Example 3.12 we see that

$$\kappa_\infty(A)=(1+\epsilon)^2, \qquad \frac{\|S_1^{-1}\|_\infty\|S_2^{-1}\|_\infty}{\|A^{-1}\|_\infty}=\frac{1+\epsilon}{\epsilon^2}, \qquad \frac{\|r_2\|_\infty}{\|y\|_\infty}=\epsilon^2.$$

Hence the bound in Fact 3.14 equals $2(1+\epsilon)^3$, and it correctly indicates the inaccuracy of z. ∎

The following bound is similar to the one in Fact 3.14, but it bounds the relative error with regard to the computed solution.

Fact 3.17 (A Second Stability Bound). Let $A\in\mathbb{C}^{n\times n}$ be nonsingular, $Ax=b$, and

$$A+E=S_1S_2, \qquad \epsilon_A=\frac{\|E\|_p}{\|A\|_p},$$

$$S_1y=b+r_1, \qquad \epsilon_1=\frac{\|r_1\|_p}{\|S_1\|_p\|y\|_p},$$

$$S_2z=y+r_2, \qquad \epsilon_2=\frac{\|r_2\|_p}{\|S_2\|_p\|z\|_p},$$

where $y\neq 0$ and $z\neq 0$. Then

$$\frac{\|z-x\|_p}{\|z\|_p}\leq \underbrace{\kappa_p(A)}_{\text{condition}}\,(\epsilon_A+\epsilon),$$

where

$$\epsilon=\underbrace{\frac{\|S_1\|_p\|S_2\|_p}{\|A\|_p}}_{\text{stability}}(\epsilon_2+\epsilon_1(1+\epsilon_2)).$$

Proof. As in the proof of Fact 3.14 we start by expanding the right-hand side,

$$(A+E)z=S_1S_2z=S_1(y+r_2)=S_1y+S_1r_2=b+r_1+S_1r_2.$$

The residual is $r=Az-b=-Ez+S_1y+S_1r_2=b+r_1+S_1r_2$. Take norms and substitute the expressions for $\|r_1\|_p$ and $\|r_2\|_p$ to obtain

$$\|r\|_p\leq\|E\|_p\|z\|_p+\epsilon_1\|S_1\|_p\|y\|_p+\epsilon_2\|S_1\|_p\|S_2\|_p\|z\|_p.$$

To bound $\|y\|_p$ write $y=S_2z-r_2$, take norms, and replace $\|r_2\|_p=\epsilon_2\|S_2\|_p\|z\|_p$ to get

$$\|y\|_p\leq\|S_2\|_p\|y\|_p+\|r_2\|_p=\|S_2\|_p\|z\|_p(1+\epsilon_2).$$

Substituting this into the bound for $\|r\|_p$ gives

$$\begin{aligned}\|r\|_p &\leq \|z\|_p \left(\|E\|_p + \|S_1\|_p\|S_2\|_p\epsilon_1(1+\epsilon_2) + \|S_1\|_p\|S_2\|_p\epsilon_2\right)\\ &= \|A\|_p\|z\|_p(\epsilon_A+\epsilon).\end{aligned}$$

The relative error bound now follows from Fact 3.7. □

In Fact 3.17, the numerical stability is represented by the factor $\|S_1\|_p\|S_2\|_p/\|A\|_p$. If $\|S_1\|_p\|S_2\| \gg \|A\|_p$, then Algorithm 3.2 is unstable.

Exercises

1. The following bound is slightly tighter than the one in Fact 3.14.
Under the conditions of Fact 3.14 show that

$$\frac{\|z-x\|_p}{\|x\|_p} \leq 2\kappa_p(A)\left[\epsilon_A + \rho_p(A,b)\,\epsilon\right],$$

where

$$\rho_p(A,b) = \frac{\|b\|_p}{\|A\|_p\|x\|_p}, \qquad \epsilon = \frac{\|S_2^{-1}\|_p\|S_1^{-1}\|_p}{\|(A+E)^{-1}\|_p}\,\epsilon_2(1+\epsilon_1)+\epsilon_1.$$

2. The following bound suggests that Algorithm 3.2 is unstable if the first factor is ill-conditioned with respect to inversion.
Under the conditions of Fact 3.14 show that

$$\frac{\|z-x\|_p}{\|x\|_p} \leq 2\kappa_p(A)\left[\epsilon_A + \epsilon_1 + \kappa_p(S_1)\,\epsilon_2(1+\epsilon_1)\right].$$

3. The following bound suggests that Algorithm 3.2 is unstable if the second factor is ill-conditioned with respect to inversion.
Let $Ax = b$ where A is nonsingular. Also let

$$A = S_1S_2, \qquad S_1y = b, \qquad S_2z = y + r_2, \qquad \text{where} \quad \epsilon_2 = \frac{\|r_2\|_p}{\|S_2\|_p\|z\|_p}$$

and $z \neq 0$. Show that

$$\frac{\|z-x\|_p}{\|z\|_p} \leq \kappa_p(S_2)\,\epsilon_2.$$

4. How Not to Solve Linear Systems.
One could solve a linear system $Ax = b$ by forming A^{-1}, and then multiplying A^{-1} by b. The bound below suggests that this approach is likely to be numerically less accurate than a direct solver.

Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and $Ax = b$ with $b \neq 0$. Let $A+E \in \mathbb{C}^{n\times n}$ with $\|A^{-1}\|_p \|E\| p \leq 1/2$. Compute $Z = (A+E)^{-1}$ and $z = Z(b+f)$. Show that

$$\frac{\|z-x\|_p}{\|x\|_p} \leq \kappa_p(A) \left(2\frac{\|A^{-1}\|_p \|b\|_p}{\|A^{-1}b\|_p} \epsilon_A + \epsilon_f \right),$$

where

$$\epsilon_A = \frac{\|E\|_p}{\|A\|_p}, \qquad \epsilon_f = \frac{\|f\|_p}{\|A\|_p \|x\|_p},$$

and compare this to the bound in Fact 3.9.
Hint: Use the perturbation bounds for matrix multiplication and matrix inversion in Facts 2.22 and 2.25.

3.5 LU Factorization

The LU factorization of a matrix is the basis for Gaussian elimination.

Definition 3.18. *Let $A \in \mathbb{C}^{n\times n}$. A factorization $A = LU$, where L is unit lower triangular and U is upper triangular, is called an* LU factorization *of A.*

The LU factorization of a nonsingular matrix, if it exists, is unique; see Exercise 5 in Section 1.13. Unfortunately, there are matrices that do not have an LU factorization, as the example below illustrates.

Example 3.19. The nonsingular matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

cannot be factored into $A = LU$, where L is lower triangular and U is upper triangular. Suppose to the contrary that it could. Then

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ l & 1 \end{pmatrix} \begin{pmatrix} u_1 & u_1 \\ 0 & u_3 \end{pmatrix}.$$

The first column of the equality implies that $u_1 = 0$, and $lu_1 = 1$ so $u_1 \neq 0$, a contradiction. ∎

Example 3.12 illustrates that a matrix A that is well-conditioned with respect to inversion can have LU factors that are ill-conditioned with respect to inversion. Algorithm 3.3 below shows how to permute the rows of a nonsingular matrix so that the permuted matrix has an LU factorization. Permuting the rows of A is called *partial pivoting*—as opposed to *complete pivoting* where both rows and columns are permuted. In order to prevent the factors from being too ill-conditioned, Algorithm 3.3 chooses a permutation matrix so that the elements of L are bounded.

ALGORITHM 3.3. LU Factorization with Partial Pivoting.

Input: Nonsingular matrix $A \in \mathbb{C}^{n\times n}$
Output: Permutation matrix P, unit lower triangular matrix L,
upper triangular matrix U such that $PA = LU$

1. If $n = 1$, then $P \equiv 1$, $L \equiv 1$, and $U \equiv A$.
2. If $n > 1$, then choose a permutation matrix P_n such that

$$P_n A = \begin{array}{c c} & \begin{array}{cc} 1 & n-1 \end{array} \\ \begin{array}{c} 1 \\ n-1 \end{array} & \begin{pmatrix} \alpha & a \\ d & A_{n-1} \end{pmatrix} \end{array},$$

where α has the largest magnitude among all elements in the leading column, i.e., $|\alpha| \geq \|d\|_\infty$, and factor

$$P_n A = \begin{pmatrix} 1 & 0 \\ l & I_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & a \\ 0 & S \end{pmatrix},$$

where $l \equiv d\alpha^{-1}$ and $S \equiv A_{n-1} - la$.

3. Compute $P_{n-1}S = L_{n-1}U_{n-1}$, where P_{n-1} is a permutation matrix, L_{n-1} is unit lower triangular, and U_{n-1} is upper triangular.
4. Then

$$P \equiv \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1} \end{pmatrix} P_n, \qquad L \equiv \begin{pmatrix} 1 & 0 \\ P_{n-1}l & L_{n-1} \end{pmatrix}, \qquad U \equiv \begin{pmatrix} \alpha & a \\ 0 & U_{n-1} \end{pmatrix}.$$

Remark 3.20.

- *Each iteration of step* 2 *in Algorithm* 3.3 *determines one column of L and one row of U.*
- *Partial pivoting ensures that the magnitude of the* multipliers *is bounded by one; i.e., $\|l\|_\infty \leq 1$ in step* 2 *of Algorithm* 3.3. *Therefore, all elements of L have magnitude less than or equal to one.*
- *The scalar α is called a* pivot, *and the matrix $S = A_{n-1} - d\alpha^{-1}a$ is a* Schur complement. *We already encountered Schur complements in Fact* 1.14, *as part of the inverse of a partitioned matrix. In this particular Schur complement S the matrix $d\alpha^{-1}a$ is an outer product.*
- *The multipliers can be easily recovered from L, because they are elements of L. Step* 4 *of Algorithm* 3.3 *shows that the first column of L contains the multipliers $P_{n-1}l$ that zero out elements in the first column. Similarly, column i of L contains the multipliers that zero out elements in column i. However, the multipliers cannot be easily recovered from L^{-1}.*
- *Step* 4 *of Algorithm* 3.3 *follows from $S = P_{n-1}^T L_{n-1}U_{n-1}$, extracting the permutation matrix,*

$$P_n A = \begin{pmatrix} 1 & 0 \\ 0 & P_{n-1}^T \end{pmatrix} \begin{pmatrix} 1 & 0 \\ P_{n-1}l & I_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & a \\ 0 & L_{n-1}U_{n-1} \end{pmatrix}$$

and separating lower and upper triangular parts

$$\begin{pmatrix} 1 & 0 \\ P_{n-1}l & I_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & a \\ 0 & L_{n-1}U_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ P_{n-1}l & L_{n-1} \end{pmatrix} \begin{pmatrix} \alpha & a \\ 0 & U_{n-1} \end{pmatrix}.$$

- *In the vector $P_{n-1}l$, the permutation P_{n-1} reorders the multipliers l, but does not change their values. To combine all permutations into a single permutation matrix P, we have to pull all permutation matrices in front of the lower triangular matrix. This, in turn, requires reordering the multipliers in earlier steps.*

Fact 3.21 (LU Factorization with Partial Pivoting). Every nonsingular matrix A has a factorization $PA = LU$, where P is a permutation matrix, L is unit lower triangular, and U is nonsingular upper triangular.

Proof. Perform an induction proof based on Algorithm 3.3. □

A factorization $PA = LU$ is, in general, not unique because there are many choices for the permutation matrix.

With a factorization $PA = LU$, the rows of the linear system $Ax = b$ are rearranged, and the system to be solved is $PAx = Pb$. The process of solving this linear system is called *Gaussian elimination with partial pivoting*.

ALGORITHM 3.4. Gaussian Elimination with Partial Pivoting.

Input: Nonsingular matrix $A \in \mathbb{C}^{n\times n}$, vector $b \in \mathbb{C}^n$
Output: Solution of $Ax = b$

1. Factor $PA = LU$ with Algorithm 3.3.
2. Solve the system $Ly = Pb$.
3. Solve the system $Ux = y$.

The next bound implies that Gaussian elimination with partial pivoting is stable in exact arithmetic if the elements of U are not much larger in magnitude than those of A.

Corollary 3.22 (Stability in Exact Arithmetic of Gaussian Elimination with Partial Pivoting). *If $A \in \mathbb{C}^{n\times n}$ is nonsingular, $Ax = b$, and*

$$P(A+E) = LU, \qquad \epsilon_A = \frac{\|E\|_\infty}{\|A\|_\infty},$$

$$Ly = Pb + r_L, \qquad \epsilon_L = \frac{\|r_L\|_\infty}{\|L\|_\infty \|y\|_\infty},$$

$$Uz = y + r_U, \qquad \epsilon_U = \frac{\|r_U\|_\infty}{\|U\|_\infty \|z\|_\infty},$$

where $y \neq 0$ and $z \neq 0$, then

$$\frac{\|z-x\|_\infty}{\|z\|_\infty} \le \kappa_\infty(A)\,(\epsilon_A + \epsilon), \qquad \text{where} \quad \epsilon = n\frac{\|U\|_\infty}{\|A\|_\infty}\big(\epsilon_U + \epsilon_L(1+\epsilon_U)\big).$$

Proof. Apply Fact 3.17 to $A+E=S_1S_2$, where $S_1=P^TL$ and $S_2=U$. Permutation matrices do not change p-norms, see Exercise (iv) in Section 2.6, so that $\|P^TL\|_\infty=\|L\|_\infty$. Because the multipliers are the elements of L, and $|l_{ij}|\le 1$ with partial pivoting, we get $\|L\|_\infty\le n$. □

The ratio $\|U\|_\infty/\|A\|_\infty$ represents the element growth during Gaussian elimination. In practice, $\|U\|_\infty/\|A\|_\infty$ tends to be small, but there are $n\times n$ matrices for which $\|U\|_\infty/\|A\|_\infty=2^{n-1}/n$ is possible; see Exercise 2 below. If $\|U\|_\infty\gg\|A\|_\infty$, then Gaussian elimination is unstable.

Exercises

(i) Determine the LU factorization of a nonsingular lower triangular matrix A. Express the elements of L and U in terms of the elements of A.

(ii) Determine a factorization $A=LU$ when A is upper triangular.

(iii) For

$$A=\begin{pmatrix}0 & 0\\ A_1 & 0\end{pmatrix},$$

with A_1 nonsingular, determine a factorization $PA=LU$ where L is unit lower triangular and U is upper triangular.

(iv) LDU Factorization.
One can make an LU factorization more symmetric by requiring that both triangular matrices have ones on the diagonal and factoring $A=LD\tilde{U}$, where L is unit lower triangular, D is diagonal, and $\tilde{U}$ is unit upper triangular.
Given an LU factorization $A=LU$, express the diagonal elements d_{ii} of D and the elements $\tilde{u}_{ij}$ in terms of elements of U.

(v) Block LU Factorization.
Suppose we can partition the invertible matrix A as

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

where A_{11} is invertible. Verify that A has the block factorization $A=LU$ where

$$L=\begin{pmatrix}I & 0\\ A_{21}A_{11}^{-1} & I\end{pmatrix},\qquad U=\begin{pmatrix}A_{11} & A_{12}\\ 0 & S\end{pmatrix},$$

and $S\equiv A_{22}-A_{21}A_{11}^{-1}A_{12}$ is a *Schur complement*. Note that L is unit lower triangular. However, U is only block upper triangular, because A_{11} and S are in general not triangular. Hence a block LU factorization is not the same as an LU factorization.
Determine a block LDU factorization $A=LDU$, where L is unit lower triangular, U is unit upper triangular, and D is block diagonal.

(vi) The matrix

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

does not have an LU factorization. However, it does have a block LU factorization $A = LU$ with

$$A_{11} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Determine L and U.

(vii) UL Factorization.
Analogous to Algorithm 3.3, present an algorithm that factors any square matrix A into $PA = UL$, where P is a permutation matrix, U is unit upper triangular, and L is lower triangular.

1. Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and P a permutation matrix such that

$$PA = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

with A_{11} nonsingular. Show: If all elements of $A_{21}A_{11}^{-1}$ are less than one in magnitude, then

$$\kappa_\infty\left(A_{22} - A_{21}A_{11}^{-1}A_{12}\right) \le n^2\kappa_\infty(A).$$

2. Compute the LU factorization of the $n \times n$ matrix

$$A = \begin{pmatrix} 1 & & & & 1 \\ -1 & 1 & & & 1 \\ -1 & -1 & 1 & & 1 \\ \vdots & & \ddots & \ddots & \vdots \\ -1 & \dots & \dots & -1 & 1 \end{pmatrix}.$$

Show that pivoting is not necessary. Determine the one norms of A and U.

3. Let $A \in \mathbb{C}^{n\times n}$ and $A + uv^*$ be nonsingular, where $u, v \in \mathbb{C}^n$. Show how to solve $(A + uv^*)x = b$ using two linear system solves with A, two inner products, one scalar vector multiplication, and one vector addition.

4. This problem shows that if Gaussian elimination with partial pivoting encounters a small pivot, then A must be ill-conditioned.
Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and $PA = LU$, where P is a permutation matrix, L is unit triangular with elements $|l_{ij}| \le 1$, and U is upper triangular with elements u_{ij}. Show that $\kappa_\infty(A) \ge \|A\|_\infty / \min_j |u_{jj}|$.

5. The following matrices G are generalizations of the lower triangular matrices in the LU factorization. The purpose of G is to transform all elements of a column vector into zeros, except for the kth element.

Let $G = I_n - ge_k^T$, where $g \in \mathbb{C}^n$ and $1 \leq k \leq n$. Which conditions do the elements of g have to satisfy so that G is invertible? Determine G^{-1} when it exists.
Given an index k and a vector $x \in \mathbb{C}^n$, which conditions do the elements of x have to satisfy so that $Gx = e_k$? Determine the vector g when it exists.

3.6 Cholesky Factorization

It would seem natural that a Hermitian matrix should have a factorization that reflects the symmetry of the matrix. For an $n \times n$ Hermitian matrix, we need to store only $n(n+1)/2$ elements, and it would be efficient if the same were true for the factorization. Unfortunately, this is not possible in general. For instance, the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

is nonsingular and Hermitian. But it cannot be factored into a lower times upper triangular matrix, as illustrated in Example 3.19. Fortunately, a certain class of matrices, so-called *Hermitian positive definite* matrices, do admit a symmetric factorization.

Definition 3.23. *A Hermitian matrix $A \in \mathbb{C}^{n\times n}$ is* positive definite *if $x^*Ax > 0$ for all $x \in \mathbb{C}^n$ with $x \neq 0$.*

A Hermitian matrix $A \in \mathbb{C}^{n\times n}$ is positive semidefinite *if $x^*Ax \geq 0$ for all $x \in \mathbb{C}^n$.*

A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite *if $x^TAx > 0$ for all $x \in \mathbb{R}^n$ with $x \neq 0$, and* positive semidefinite *if $x^TAx \geq 0$ for all $x \in \mathbb{R}^n$.*

A positive semidefinite matrix A can have $x^*Ax = 0$ for $x \neq 0$.

Example. The 2×2 Hermitian matrix

$$A = \begin{pmatrix} 1 & \beta \\ \bar{\beta} & 1 \end{pmatrix}$$

is positive definite if $|\beta| < 1$, and positive semidefinite if $|\beta|^2 = 1$. ■

We derive several properties of Hermitian positive definite matrices. We start by showing that all Hermitian positive definite matrices are nonsingular.

Fact 3.24. If $A \in \mathbb{C}^{n\times n}$ is Hermitian positive definite, then A is nonsingular.

Proof. Suppose to the contrary that A were singular. Then $Ax = 0$ for some $x \neq 0$, implying $x^*Ax = 0$ for some $x \neq 0$, which contradicts the positive definiteness of A; i.e., $x^*Ax > 0$ for all $x \neq 0$. □

Hermitian positive definite matrices have positive diagonal elements.

Fact 3.25. If $A \in \mathbb{C}^{n\times n}$ is Hermitian positive definite, then its diagonal elements are positive.

Proof. Since A is positive definite, we have $x^*Ax > 0$ for any $x \neq 0$, and in particular $0 < e_j^*Ae_j = a_{jj}$, $1 \leq j \leq n$. □

Below is a transformation that preserves Hermitian positive definiteness.

Fact 3.26. If $A \in \mathbb{C}^{n\times n}$ is Hermitian positive definite and $B \in \mathbb{C}^{n\times n}$ is nonsingular, then B^*AB is also Hermitian positive definite.

Proof. The matrix B^*AB is Hermitian because A is Hermitian. Since B is nonsingular, $y = Bx \neq 0$ if and only if $x \neq 0$. Hence

$$x^*B^*ABx = (Bx)^*\,A\,(Bx) = y^*Ay > 0$$

for any vector $y \neq 0$, so that B^*AB is positive definite. □

At last we show that principal submatrices and Schur complements inherit Hermitian positive definiteness.

Fact 3.27. If $A \in \mathbb{C}^{n\times n}$ is Hermitian positive definite, then its leading principal submatrices and Schur complements are also Hermitian positive definite.

Proof. Let B be a $k\times k$ principal submatrix of A, for some $1 \leq k \leq n-1$. The submatrix B is Hermitian because it is a principal submatrix of a Hermitian matrix. To keep the notation simple, we permute the rows and columns of A so that the submatrix B occupies the leading rows and columns. That is, let P be a permutation matrix, and partition

$$\hat{A} = P^TAP = \begin{pmatrix} B & A_{12} \\ A_{12}^* & A_{22} \end{pmatrix}.$$

Fact 3.26 implies that $\hat{A}$ is also Hermitian positive definite. Thus $x^*\hat{A}x > 0$ for any vector $x \neq 0$. In particular, let $x = \begin{pmatrix} y \\ 0 \end{pmatrix}$ for $y \in \mathbb{C}^k$. Then for any $y \neq 0$ we have

$$0 < x^*\hat{A}x = \begin{pmatrix} y^* & 0 \end{pmatrix} \begin{pmatrix} B & A_{12} \\ A_{12}^* & A_{22} \end{pmatrix} \begin{pmatrix} y \\ 0 \end{pmatrix} = y^*By.$$

This means $y^*By > 0$ for $y \neq 0$, so that B is positive definite. Since the submatrix B is a principal submatrix of a Hermitian matrix, B is also Hermitian. Therefore, any principal submatrix B of A is Hermitian positive definite.

Now we prove Hermitian positive definiteness for Schur complements. Fact 3.24 implies that B is nonsingular. Hence we can set

$$L = \begin{pmatrix} I_k & 0 \\ -A_{12}^*B^{-1} & I_{n-k} \end{pmatrix},$$

so that

$$L\,\hat{A}\,L^* = \begin{pmatrix} B & 0 \\ 0 & S \end{pmatrix}, \qquad \text{where} \quad S = A_{22} - A_{12}^* B^{-1} A_{12}.$$

Since L is unit lower triangular, it is nonsingular. From Fact 3.26 follows then that $L\hat{A}L^*$ is Hermitian positive definite. Earlier in this proof we showed that principal submatrices of Hermitian positive definite matrices are Hermitian positive definite, thus the Schur complement S must be Hermitian positive definite. □

Now we have all the tools we need to factor Hermitian positive definite matrices. The following algorithm produces a symmetric factorization $A = LL^*$ for a Hermitian positive definite matrix A. The algorithm exploits the fact that the diagonal elements of A are positive and the Schur complements are Hermitian positive definite.

Definition 3.28. *Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. A factorization $A = LL^*$, where L is (lower or upper) triangular with positive diagonal elements, is called a* Cholesky factorization *of A.*

Below we compute a lower-upper Cholesky factorization $A = LL^*$ where L is a lower triangular matrix.

ALGORITHM 3.5. Cholesky Factorization.

Input: Hermitian positive definite matrix $A \in \mathbb{C}^{n\times n}$
Output: Lower triangular matrix L with positive diagonal elements
such that $A = LL^*$

1. If $n = 1$, then $L \equiv \sqrt{A}$.
2. If $n > 1$, partition and factor

$$A = \begin{array}{cc} & \begin{array}{cc} 1 & n-1 \end{array} \\ \begin{array}{c} 1 \\ n-1 \end{array} & \begin{pmatrix} \alpha & a^* \\ a & A_{n-1} \end{pmatrix} \end{array} = \begin{pmatrix} \alpha^{1/2} & 0 \\ a\alpha^{-1/2} & I_{n-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} \alpha^{1/2} & \alpha^{-1/2}a^* \\ 0 & I_{n-1} \end{pmatrix},$$

where $S \equiv A_{n-1} - a\alpha^{-1}a^*$.
3. Compute $S = L_{n-1}L_{n-1}^*$, where L_{n-1} is lower triangular with positive diagonal elements.
4. Then

$$L \equiv \begin{pmatrix} \alpha^{1/2} & 0 \\ a\alpha^{-1/2} & L_{n-1} \end{pmatrix}.$$

A Cholesky factorization of a positive matrix is unique.

Fact 3.29 (Uniqueness of Cholesky factorization). Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. If $A = LL^*$ where L is lower triangular with positive diagonal, then L is unique. Similarly, if $A = LL^*$ where L is upper triangular with positive diagonal elements, then L is unique.

Proof. This can be shown in the same way as the uniqueness of the LU factorization. □

The following result shows that one can use a Cholesky factorization to determine whether a Hermitian matrix is positive definite.

Fact 3.30. Let $A \in \mathbb{C}^{n\times n}$ be Hermitian. A is positive definite if and only if $A = LL^*$ where L is triangular with positive diagonal elements.

Proof. Algorithm 3.5 shows that if A is positive definite, then $A = LL^*$. Now assume that $A = LL^*$. Since L is triangular with positive diagonal elements, it is nonsingular. Therefore, $Lx \neq 0$ for $x \neq 0$, and $x^*Ax = \|L^*x\|_2^2 > 0$. □

The next bound shows that a Cholesky solver is numerically stable in exact arithmetic.

Corollary 3.31 (Stability of Cholesky Solver). *Let $A \in \mathbb{C}^{n\times n}$ and let $A+E$ be Hermitian positive definite matrices, $Ax = b$, $b \neq 0$, and*

$$A+E = LL^*, \qquad \epsilon_A = \frac{\|E\|_2}{\|A\|_2},$$
$$Ly = b + r_1, \qquad \epsilon_1 = \frac{\|r_1\|_2}{\|b\|_2},$$
$$L^*z = y + r_2, \qquad \epsilon_2 = \frac{\|r_2\|_2}{\|y\|_2}.$$

If $\|A^{-1}\|_2\|E\|_2 \leq 1/2$, then

$$\frac{\|z-x\|_2}{\|x\|_2} \leq 2\kappa_2(A)\,(\epsilon_A + \epsilon_1 + \epsilon_2(1+\epsilon_1)).$$

Proof. Apply Fact 3.14 to $A+E$, where $S_1 = L$ and $S_2 = L^*$. The stability factor is $\|L^{-*}\|_2\|L^{-1}\|_2/\|(A+E)^{-1}\|_2 = 1$ because Fact 2.19 implies

$$\|(A+E)^{-1}\|_2 = \|L^{-*}L^{-1}\|_2 = \|L^{-1}\|_2^2 = \|L^{-*}\|_2\|L^{-1}\|_2. \qquad \square$$

Exercises

(i) The magnitude of an off-diagonal element of a Hermitian positive definite matrix is bounded by the geometric mean of the corresponding diagonal elements.
Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. Show: $|a_{ij}| < \sqrt{a_{ii}a_{jj}}$ for $i \neq j$.
Hint: Use the positive definiteness of the Schur complement.

(ii) The magnitude of an off-diagonal element of a Hermitian positive definite matrix is bounded by the arithmetic mean of the corresponding diagonal elements.

Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. Show: $|a_{ij}| \leq (a_{ii}+a_{jj})/2$ for $i \neq j$.
Hint: Use the relation between arithmetic and geometric mean.

(iii) The largest element in magnitude of a Hermitian positive definite matrix is on the diagonal.
Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. Show: $\max_{1\leq i,j\leq n} |a_{ij}| = \max_{1\leq i\leq n} a_{ii}$.

(iv) Let $A \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. Show: A^{-1} is also positive definite.

(v) Modify Algorithm 3.5 so that it computes a factorization $A = LDL^*$ for a Hermitian positive definite matrix A, where D is diagonal and L is unit lower triangular.

(vi) Upper-Lower Cholesky Factorization.
Modify Algorithm 3.5 so that it computes a factorization $A = L^*L$ for a Hermitian positive definite matrix A, where L is lower triangular with positive diagonal elements.

(vii) Block Cholesky Factorization.
Partition the Hermitian positive definite matrix A as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Analogous to the block LU factorization in Exercise (v) of Section 3.5 determine a factorization $A = LL^*$, where L is block lower triangular. That is, L is of the form

$$L = \begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix},$$

where L_{11} and L_{22} are in general not lower triangular.

(viii) Let

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

be Hermitian positive definite. Show:

$$\|A_{22} - A_{21}A_{11}^{-1}A_{12}\|_2 \leq \|A\|_2$$

and

$$\kappa_2(A_{22} - A_{21}A_{11}^{-1}A_{12}) \leq \kappa_2(A).$$

(ix) Prove: $A = MM^*$ for some nonsingular matrix M if and only if A is Hermitian positive definite.

(x) Generalized Cholesky Factorization.
Let $M \in \mathbb{C}^{n\times n}$ be Hermitian positive definite. Prove: If $M = M_1^*M_1 = M_2^*M_2$, for square matrices M_1 and M_2, then there exists a unitary matrix Q such that $M_2 = QM_1$.

(xi) Let $M = A + \imath B$ be Hermitian positive definite, where $\imath^2 = -1$, and A and B are real square matrices. Show that the matrix

$$C = \begin{pmatrix} A & -B \\ B & A \end{pmatrix}$$

is real symmetric positive definite.

3.7 QR Factorization

The QR factorization is a matrix factorization where one of the factors is unitary and the other one is triangular. We derive the existence of a QR factorization from the Cholesky factorization.

Fact 3.32. Every nonsingular matrix $A \in \mathbb{C}^{n\times n}$ has a unique factorization $A = QR$, where Q is unitary and R is upper triangular with positive diagonal elements.

Proof. Since A is nonsingular, $Ax \neq 0$ for $x \neq 0$, and $x^*A^*Ax = \|Ax\|_2^2 > 0$, which implies that $M = A^*A$ is Hermitian positive definite. Let $M = LL^*$ be a Cholesky factorization of M, where L is lower triangular with positive diagonal elements. Then $M = A^*A = LL^*$. Multiplying by A^{-*} on the left gives $A = QR$, where $Q = A^{-*}L$, and where $R = L^*$ is upper triangular with positive diagonal elements. Exercise (ix) in Section 3.6 shows that Q is unitary.

The uniqueness of the QR factorization follows from the uniqueness of the Cholesky factorization, as well as from Exercise 6 in Section 1.13. □

The bound below shows that a QR solver is numerically stable in exact arithmetic.

Corollary 3.33 (Stability of QR Solver). *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular, $Ax = b$, $b \neq 0$, and*

$$A + E = QR, \qquad \epsilon_A = \frac{\|E\|_2}{\|A\|_2},$$

$$Qy = b + r_1, \qquad \epsilon_1 = \frac{\|r_1\|_2}{\|b\|_2},$$

$$Rz = y + r_2, \qquad \epsilon_2 = \frac{\|r_2\|_2}{\|y\|_2}.$$

If $\|A^{-1}\|_2\|E\|_2 \leq 1/2$, then

$$\frac{\|z - x\|_2}{\|x\|_2} \leq 2\kappa_2(A)\left(\epsilon_A + \epsilon_1 + \epsilon_2(1 + \epsilon_1)\right).$$

Proof. Apply Fact 3.14 to $A + E$, where $S_1 = Q$ and $S_2 = R$. The stability factor is $\|R^{-1}\|_2\|Q^*\|_2/\|(A+E)^{-1}\|_2 = 1$, because Exercise (v) in Section 2.6 implies $\|Q^*\|_2 = 1$ and $\|(A+E)^{-1}\|_2 = \|R^{-1}\|_2$. □

There are many ways to compute a QR factorization. Here we present an algorithm that is based on *Givens rotations*; see Definition 1.17. Givens rotations are unitary, see Example 1.16, and they are often used to introduce zeros into matrices. Let's start by using a Givens rotation to introduce a single zero into a vector.

Example. Let $x, y \in \mathbb{C}$.

$$\begin{pmatrix} c & s \\ -\overline{s} & \overline{c} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} d \\ 0 \end{pmatrix}, \qquad \text{where} \quad d = \sqrt{|x|^2 + |y|^2}.$$

If $x = y = 0$, then $c = 1$ and $s = 0$; otherwise $c = \overline{x}/d$ and $s = \overline{y}/d$. That is, if both components of the vector are zero, then there is nothing to do and the unitary matrix is the identity. Note that $d \geq 0$ and $|c|^2 + |s|^2 = 1$. ∎

When introducing zeros into a longer vector, we embed each Givens rotation in an identity matrix.

Example. Suppose we want to zero out elements 2, 3, and 4 in a 4×1 vector with a unitary matrix. We can apply three Givens rotations in the following order.

1. Apply a Givens rotation to rows 3 and 4 to zero out element 4,

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_4 & s_4 \\ 0 & 0 & -\overline{s}_4 & \overline{c}_4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ y_3 \\ 0 \end{pmatrix},$$

where $y_3 = \sqrt{|x_3|^2 + |x_4|^2} \geq 0$. If $x_4 = x_3 = 0$, then $c_4 = 1$ and $s_4 = 0$; otherwise $c_4 = \overline{x}_3/y_3$ and $s_4 = \overline{x}_4/y_3$.

2. Apply a Givens rotation to rows 2 and 3 to zero out element 3,

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & c_3 & s_3 & 0 \\ 0 & -\overline{s}_3 & \overline{c}_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ y_3 \\ 0 \end{pmatrix} = \begin{pmatrix} x_1 \\ y_2 \\ 0 \\ 0 \end{pmatrix},$$

where $y_2 = \sqrt{|x_2|^2 + |y_3|^2} \geq 0$. If $y_3 = x_2 = 0$, then $c_3 = 1$ and $s_3 = 0$; otherwise $c_3 = \overline{x}_2/y_2$ and $s_3 = \overline{y}_3/y_2$.

3. Apply a Givens rotation to rows 1 and 2 to zero out element 2,

$$\begin{pmatrix} c_2 & s_2 & 0 & 0 \\ -\overline{s}_2 & \overline{c}_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ y_2 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} y_1 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

where $y_1 = \sqrt{|x_1|^2 + |y_2|^2} \geq 0$. If $y_2 = x_1 = 0$, then $c_2 = 1$ and $s_2 = 0$; otherwise $c_2 = \overline{x}_1/y_1$ and $s_2 = \overline{y}_2/y_1$.

Therefore $Qx = y_1 e_1$, where $y_1 = \|Qx\|_2$ and

$$Q = \begin{pmatrix} c_2 & s_2 & 0 & 0 \\ -\overline{s}_2 & \overline{c}_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & c_3 & s_3 & 0 \\ 0 & -\overline{s}_3 & \overline{c}_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_4 & s_4 \\ 0 & 0 & -\overline{s}_4 & \overline{c}_4 \end{pmatrix}. \qquad \blacksquare$$

There are many possible orders in which to apply Givens rotations, and Givens rotations don't have to operate on adjacent rows either. The example below illustrates this.

Example. Here is another way to to zero out elements 2, 3, and 4 in a 4 × 1 vector. We can apply three Givens rotations that all involve the leading row.

1. Apply a Givens rotation to rows 1 and 4 to zero out element 4,

$$\begin{pmatrix} c_4 & 0 & 0 & s_4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\overline{s}_4 & 0 & 0 & \overline{c}_4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} y_1 \\ x_2 \\ x_3 \\ 0 \end{pmatrix},$$

where $y_1 = \sqrt{|x_1|^2 + |x_4|^2} \geq 0$. If $x_4 = x_1 = 0$, then $c_4 = 1$ and $s_4 = 0$; otherwise $c_4 = \overline{x}_1 / y_1$ and $s_4 = \overline{x}_4 / y_1$.

2. Apply a Givens rotation to rows 1 and 3 to zero out element 3,

$$\begin{pmatrix} c_3 & 0 & s_3 & 0 \\ 0 & 1 & 0 & 0 \\ -\overline{s}_3 & 0 & \overline{c}_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ x_2 \\ x_3 \\ 0 \end{pmatrix} = \begin{pmatrix} z_1 \\ x_2 \\ 0 \\ 0 \end{pmatrix},$$

where $z_1 = \sqrt{|y_1|^2 + |x_3|^2} \geq 0$. If $x_3 = y_1 = 0$, then $c_3 = 1$ and $s_3 = 0$; otherwise $c_3 = \overline{y}_1 / z_1$ and $s_3 = \overline{x}_3 / z_1$.

3. Apply a Givens rotation to rows 1 and 2 to zero out element 2,

$$\begin{pmatrix} c_2 & s_2 & 0 & 0 \\ -\overline{s}_2 & \overline{c}_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ y_2 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} u_1 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

where $u_1 = \sqrt{|z_1|^2 + |x_2|^2} \geq 0$. If $x_2 = z_1 = 0$, then $c_2 = 1$ and $s_2 = 0$; otherwise $c_2 = \overline{z}_1 / u_1$ and $s_2 = \overline{x}_2 / u_1$.

Therefore $Qx = u_1 e_1$, where $u_1 = \|Qx\|_2$ and

$$Q = \begin{pmatrix} c_2 & s_2 & 0 & 0 \\ -\overline{s}_2 & \overline{c}_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_3 & 0 & s_3 & 0 \\ 0 & 1 & 0 & 0 \\ -\overline{s}_3 & 0 & \overline{c}_3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c_4 & 0 & 0 & s_4 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\overline{s}_4 & 0 & 0 & \overline{c}_4 \end{pmatrix}. \qquad \blacksquare$$

The preceding examples demonstrate that if a Givens rotation operates on rows i and j, then the c and s elements occupy positions (i,i), (i,j), (j,i), and (j,j).

At last here is a sketch of how one can reduce a square matrix to upper triangular form by means of Givens rotations.

Example. We introduce zeros one column at a time, from left to right, and within a column from bottom to top. The Givens rotations operate on adjacent rows. Elements that can be nonzero are represented by $*$. Elements that were affected by the ith Givens rotation have the label i. We start by introducing zeros into column 1,

$$\begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \xrightarrow{1} \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \xrightarrow{2} \begin{pmatrix} * & * & * & * \\ 2 & 2 & 2 & 2 \\ 0 & 2 & 2 & 2 \\ 0 & 1 & 1 & 1 \end{pmatrix} \xrightarrow{3} \begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 \\ 0 & 2 & 2 & 2 \\ 0 & 1 & 1 & 1 \end{pmatrix}.$$

Now we introduce zeros into column 2, and then into column 3,

$$\begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 \\ 0 & 2 & 2 & 2 \\ 0 & 1 & 1 & 1 \end{pmatrix} \xrightarrow{4} \begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 \\ 0 & 4 & 4 & 4 \\ 0 & 0 & 4 & 4 \end{pmatrix} \xrightarrow{5} \begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 5 & 5 & 5 \\ 0 & 0 & 5 & 5 \\ 0 & 0 & 4 & 4 \end{pmatrix} \xrightarrow{6} \begin{pmatrix} 3 & 3 & 3 & 3 \\ 0 & 5 & 5 & 5 \\ 0 & 0 & 6 & 6 \\ 0 & 0 & 0 & 6 \end{pmatrix}. \quad \blacksquare$$

Below is the general algorithm.

ALGORITHM 3.6. QR Factorization for Nonsingular Matrices.

Input: Nonsingular matrix $A \in \mathbb{C}^{n\times n}$
Output: Unitary matrix $Q \in \mathbb{C}^{n\times n}$ and upper triangular matrix $R \in \mathbb{C}^{n\times n}$ with positive diagonal elements such that $A = QR$

1. If $n = 1$, then $Q \equiv A/|A|$ and $R \equiv |A|$.
2. If $n > 1$, zero out elements $n, n-1, \ldots, 2$ in column 1 of A as follows.
   (i) Set $\begin{pmatrix} b_{n1} & b_{n2} & \ldots & b_{nn} \end{pmatrix} = \begin{pmatrix} a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix}$.
   (ii) For $i = n, n-1, \ldots, 2$
   Zero out element $(i,1)$ by applying a rotation to rows i and $i-1$,

$$\begin{pmatrix} c_i & s_i \\ -\bar{s}_i & \bar{c}_i \end{pmatrix} \begin{pmatrix} a_{i-1,1} & a_{i-1,2} & \ldots & a_{i-1,n} \\ b_{i1} & b_{i2} & \ldots & b_{in} \end{pmatrix} = \begin{pmatrix} b_{i-1,1} & b_{i-1,2} & \ldots & b_{i-1,n} \\ 0 & \hat{a}_{i2} & \ldots & \hat{a}_{in} \end{pmatrix},$$

   where $b_{i-1,1} \equiv \sqrt{|b_{i1}|^2 + |a_{i-1,1}|^2}$. If $b_{i1} = a_{i-1,1} = 0$, then $c_i \equiv 1$ and $s_i \equiv 0$; otherwise $c_i \equiv \bar{a}_{i-1,1}/b_{i-1,1}$ and $s_i \equiv \bar{b}_{i1}/b_{i-1,1}$.
   (iii) Multiply all $n-1$ rotations,

$$Q_n^* \equiv \begin{pmatrix} c_2 & s_2 & 0 \\ -\bar{s}_2 & \bar{c}_2 & 0 \\ 0 & 0 & I_{n-2} \end{pmatrix} \cdots \begin{pmatrix} I_{n-2} & 0 & 0 \\ 0 & c_n & s_n \\ 0 & -\bar{s}_n & \bar{c}_n \end{pmatrix}.$$

(iv) Partition the transformed matrix,

$$Q_n^* A = \begin{pmatrix} r_{11} & r^* \\ 0 & \hat{A} \end{pmatrix}, \qquad \text{where} \quad \hat{A} \equiv \begin{pmatrix} \hat{a}_{22} & \dots & \hat{a}_{2n} \\ \vdots & & \vdots \\ \hat{a}_{n2} & \dots & \hat{a}_{nn} \end{pmatrix},$$

$r^* \equiv \begin{pmatrix} b_{12} & \dots & b_{1n} \end{pmatrix}$, and $r_{11} \equiv b_{11} > 0$.

3. Compute $\hat{A} = Q_{n-1} R_{n-1}$, where Q_{n-1} is unitary and R_{n-1} is upper triangular with positive diagonal elements.
4. Then

$$Q \equiv Q_n \begin{pmatrix} 1 & 0 \\ 0 & Q_{n-1} \end{pmatrix}, \qquad R \equiv \begin{pmatrix} r_{11} & r^* \\ 0 & R_{n-1} \end{pmatrix}.$$

Exercises

(i) Determine the QR factorization of a real upper triangular matrix.

(ii) QR Factorization of Outer Product.
Let $x, y \in \mathbb{C}^n$, and apply Algorithm 3.6 to xy^*. How many Givens rotations do you have to apply at the most? What does the upper triangular matrix R look like?

(iii) Let $A \in \mathbb{C}^{n \times n}$ be a tridiagonal matrix, that is, only elements a_{ii}, $a_{i+1,i}$, and $a_{i,i+1}$ can be nonzero; all other elements are zero. We want to compute a QR factorization $A = QR$ with $n-1$ Givens rotations. In which order do the elements have to be zeroed out, on which rows do the rotations act, and which elements of R can be nonzero?

(iv) QL Factorization.
Show: Every nonsingular matrix $A \in \mathbb{C}^{n \times n}$ has a unique factorization $A = QL$, where Q is unitary and L is lower triangular with positive diagonal elements.

(v) Computation of QL Factorization.
Suppose we want to compute the QL factorization of a nonsingular matrix $A \in \mathbb{C}^{n \times n}$ with Givens rotations. In which order do the elements have to be zeroed out, and on which rows do the rotations act?

(vi) The elements in a Givens rotation

$$G = \begin{pmatrix} c & s \\ -\overline{s} & \overline{c} \end{pmatrix}$$

are named to invoke an association with sine and cosine, because $|c|^2 + |s|^2 = 1$. One can also express the elements in terms of tangents or cotangents. Let

$$G \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} d \\ 0 \end{pmatrix}, \qquad \text{where} \quad d = \sqrt{|x|^2 + |y|^2}.$$

Show the following: If $|y| > |x|$, then

$$\tau = \frac{x}{y}, \qquad s = \frac{\overline{y}}{|y|} \frac{1}{\sqrt{1 + |\tau|^2}}, \qquad c = \overline{\tau} s,$$

and if $|x| > |y|$, then

$$\tau = \frac{y}{x}, \qquad c = \frac{\overline{x}}{|x|}\frac{1}{\sqrt{1+|\tau|^2}}, \qquad s = \overline{\tau}c.$$

(vii) Householder Reflections.
Here is another way to introduce zeros into a vector without changing its two norm. Let $x \in \mathbb{C}^n$ and $x_1 \neq 0$. Define $Q = I - 2vv^*/v^*v$, where $v = x + \alpha\|x\|_2 e_1$ and $\alpha = x_1/|x_1|$. Show that Q is unitary and that $Qx = -\alpha\|x\|_2 e_1$. The matrix Q is called a *Householder reflection.*

(viii) Householder Reflections for Real Vectors.
Let $x, y \in \mathbb{R}^n$ with $\|x\|_2 = \|y\|_2$. Show how to choose a vector v in the Householder reflection so that $Qx = y$.

3.8 QR Factorization of Tall and Skinny Matrices

We look at rectangular matrices $A \in \mathbb{C}^{m\times n}$ with at least as many rows as columns, i.e., $m \geq n$. If A is involved in a linear system $Ax = b$, then we must have $b \in \mathbb{C}^m$ and $x \in \mathbb{C}^n$. Such linear systems do not always have a solution; and if they do happen to have a solution, then the solution may not be unique.

Example. If

$$A = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

then the linear system $Ax = b$ has a solution only for those b all of whose elements are the same, i.e., $\beta = b_1 = b_2 = b_3$. In this case the solution is $x = \beta$. ■

Fortunately, there is one right-hand side for which a linear system $Ax = b$ always has a solution, namely, $b = 0$. That is, $Ax = 0$ always has the solution $x = 0$. However, $x = 0$ may not be the only solution.

Example. If

$$A = \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 1 & -1 \end{pmatrix},$$

then $Ax = 0$ has infinitely many solutions $x = \begin{pmatrix} x_1 & x_2 \end{pmatrix}^T$ with $x_1 = x_2$. ■

We distinguish matrices A where $x = 0$ is the unique solution for $Ax = 0$.

Definition 3.34. *Let $A \in \mathbb{C}^{m\times n}$. The columns of A are* linearly independent *if $Ax = 0$ implies $x = 0$. If $Ax = 0$ has infinitely many solutions, then the columns of A are* linearly dependent.

Example.

- The columns of a nonsingular matrix A are linearly independent.
- If A is nonsingular, then the matrix $\begin{pmatrix} A \\ 0 \end{pmatrix}$ has linearly independent columns.

- Let $x \in \mathbb{C}^n$. If $x \neq 0$, then x consists of a single, linearly independent column. If $x = 0$, then x is linearly dependent.
- If $A \in \mathbb{C}^{m\times n}$ with $A^*A = I_n$, then A has linearly independent columns. This is because multiplying $Ax = 0$ on the left by A^* implies $x = 0$.
- If the linear system $Ax = b$ has a solution x, then the matrix $B = \begin{pmatrix} A & b \end{pmatrix}$ has linearly dependent columns. That is because $B\begin{pmatrix} x \\ -1 \end{pmatrix} = 0$. ∎

How can we tell whether a tall and skinny matrix has linearly independent columns? We can use a QR factorization.

ALGORITHM 3.7. QR Factorization for Tall and Skinny Matrices.

Input: Matrix $A \in \mathbb{C}^{m\times n}$ with $m \geq n$
Output: Unitary matrix $Q \in \mathbb{C}^{m\times m}$ and upper triangular matrix $R \in \mathbb{C}^{n\times n}$ with nonnegative diagonal elements such that $A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$

1. If $n = 1$, then Q is a unitary matrix that zeros out elements $2, \ldots, m$ of A, and $R \equiv \|A\|_2$.
2. If $n > 1$, then, as in Algorithm 3.6, determine a unitary matrix $Q_m \in \mathbb{C}^{m\times m}$ to zero out elements $2, \ldots, m$ in column 1 of A, so that
$$Q_m^* A = \begin{pmatrix} r_{11} & r^* \\ 0 & \hat{A} \end{pmatrix},$$
where $r_{11} \geq 0$ and $\hat{A} \in \mathbb{C}^{(m-1)\times(n-1)}$.
3. Compute $\hat{A} = Q_{m-1}\begin{pmatrix} R_{n-1} \\ 0 \end{pmatrix}$, where $Q_{m-1} \in \mathbb{C}^{(m-1)\times(m-1)}$ is unitary, and $R_{n-1} \in \mathbb{C}^{(n-1)\times(n-1)}$ is upper triangular with nonnegative diagonal elements.
4. Then
$$Q \equiv Q_m \begin{pmatrix} 1 & 0 \\ 0 & Q_{m-1} \end{pmatrix}, \qquad R \equiv \begin{pmatrix} r_{11} & r^* \\ 0 & R_{n-1} \end{pmatrix}.$$

Fact 3.35. Let $A \in \mathbb{C}^{m\times n}$ with $m \geq n$, and $A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$ where $Q \in \mathbb{C}^{m\times m}$ is unitary, and $R \in \mathbb{C}^{n\times n}$ is upper triangular. Then A has linearly independent columns if and only if R has nonzero diagonal elements.

Proof. Since Q is nonsingular, $Ax = Q\begin{pmatrix} R \\ 0 \end{pmatrix} 0 \Rightarrow x = 0$ if and only if $Rx = 0 \Rightarrow x = 0$. This is the case if and only if R is nonsingular and has nonzero diagonal elements. □

One can make a QR factorization more economical by reducing the storage and omitting part of the unitary matrix.

Fact 3.36 (Thin QR Factorization). If $A \in \mathbb{C}^{m\times n}$ with $m \geq n$, then there exists a matrix $Q_1 \in \mathbb{C}^{m\times n}$ with $Q_1^* Q_1 = I_n$, and an upper triangular matrix $R \in \mathbb{C}^{n\times n}$ with nonnegative diagonal elements so that $A = Q_1 R$.

Proof. Let $A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$ be a QR factorization as in Fact 3.35. Partition $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$, where Q_1 has n columns. Then $A = Q_1 R$. □

Definition 3.37. *If* $A \in \mathbb{C}^{m\times n}$ *and* $A^*A = I_n$, *then the columns of* A *are* orthonormal.

For a square matrix the thin QR decomposition is identical to the full QR decomposition.

Example 3.38. The columns of a unitary or an orthogonal matrix $A \in \mathbb{C}^{n\times n}$ are orthonormal because $A^*A = I_n$, and so are the rows because $AA^* = I_n$. This means, a square matrix with orthonormal columns must be a unitary matrix. A real square matrix with orthonormal columns is an orthogonal matrix. ■

Exercises

(i) Let $A \in \mathbb{C}^{m\times n}$, $m \geq n$, with thin QR factorization $A = QR$. Show: $\|A\|_2 = \|R\|_2$.

(ii) Uniqueness of Thin QR Factorization.
Let $A \in \mathbb{C}^{m\times n}$ have linearly independent columns. Show: If $A = QR$, where $Q \in \mathbb{C}^{m\times n}$ satisfies $Q^*Q = I_n$ and R is upper triangular with positive diagonal elements, then Q and R are unique.

(iii) Generalization of Fact 3.35.
Let $A \in \mathbb{C}^{m\times n}$, $m \geq n$, and $A = B\begin{pmatrix} C \\ 0 \end{pmatrix}$, where $B \in \mathbb{C}^{m\times n}$ has linearly independent columns, and $C \in \mathbb{C}^{n\times n}$. Show: A has linearly independent columns if and only if C is nonsingular.

(iv) Let $A \in \mathbb{C}^{m\times n}$ where $m > n$. Show: There exists a matrix $Z \in \mathbb{C}^{m\times(m-n)}$ such that $Z^*A = 0$.

(v) Let $A \in \mathbb{C}^{m\times n}$, $m \geq n$, have a thin QR factorization $A = QR$. Express the kth column of A as a linear combination of columns of Q and elements of R. How many columns of Q are involved?

(vi) Let $A \in \mathbb{C}^{m\times n}$, $m \geq n$, have a thin QR factorization $A = QR$. Determine a QR factorization of $A - Qe_1e_1^*R$ from the QR factorization of A.

(vii) Let $A = \begin{pmatrix} a_1 & \dots & a_n \end{pmatrix}$ have linearly independent columns a_j, $1 \leq j \leq n$. Let $A = QR$ be a thin QR factorization where $Q = \begin{pmatrix} q_1 & \dots & q_n \end{pmatrix}$ and R is upper triangular with positive diagonal elements. Express the elements of R in terms of the columns a_j of A and the columns q_j of Q.

(viii) Let A be a matrix with linearly independent columns. Show how to compute the lower-upper Cholesky factorization of A^*A without forming the product A^*A.

(ix) Bessel's Inequality.
Let $V \in \mathbb{C}^{m \times n}$ with $V = \begin{pmatrix} v_1 & \dots & v_n \end{pmatrix}$ have orthonormal columns, and let $x \in \mathbb{C}^m$. Show:

$$\sum_{j=1}^{n} |v_j^* x|^2 \le x^* x.$$

1. QR Factorization with Column Pivoting.
This problem presents a method to compute QR factorizations of arbitrary matrices. Let $A \in \mathbb{C}^{m \times n}$ with $\operatorname{rank}(A) = r$. Then there exists a permutation matrix P so that

$$AP = Q \begin{pmatrix} R_1 & R_2 \\ 0 & 0 \end{pmatrix},$$

where R_1 is an upper triangular nonsingular matrix.

(a) Show how to modify Algorithm 3.7 so that it computes such a factorization. In the first step, choose a permutation matrix P_n that brings the column with largest two norm to the front; i.e.,

$$\|AP_n e_1\|_2 = \max_{1 \le j \le n} \|AP_n e_j\|_2.$$

(b) Show that the diagonal elements of R_1 have decreasing magnitudes; i.e., $(R_1)_{11} \ge (R_1)_{22} \ge \cdots \ge (R_1)_{rr}$.

4. Singular Value Decomposition

In order to solve linear systems with a general rectangular coefficient matrix, we introduce the singular value decomposition. It is one of the most important tools in numerical linear algebra, because it contains a lot of information about a matrix, including rank, distance to singularity, column space, row space, and null spaces.

Definition 4.1 (SVD). *Let $A \in \mathbb{C}^{m\times n}$. If $m \geq n$, then a* singular value decomposition (SVD) *of A is a decomposition*

$$A = U\begin{pmatrix}\Sigma\\ 0\end{pmatrix}V^*, \qquad \textit{where} \quad \Sigma = \begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_n\end{pmatrix}, \quad \sigma_1 \geq \cdots \geq \sigma_n \geq 0,$$

and $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary.

If $m \leq n$, then an SVD of A is

$$A = U\begin{pmatrix}\Sigma & 0\end{pmatrix}V^*, \qquad \textit{where} \quad \Sigma = \begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_m\end{pmatrix}, \quad \sigma_1 \geq \cdots \geq \sigma_m \geq 0,$$

and $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary.

The matrix U is called a left singular vector matrix, *V is called a* right singular vector matrix, *and the scalars σ_j are called* singular values.

Remark 4.2.

- *An $m\times n$ matrix has* $\min\{m,n\}$ *singular values.*
- *The singular values are unique, but the singular vector matrices are not. Although an SVD is not unique, one often says "the SVD" instead of "an SVD."*

- *Let $A \in \mathbb{C}^{m\times n}$ with $m \geq n$. If $A = U\begin{pmatrix}\Sigma\\0\end{pmatrix}V^*$ is an SVD of A, then $A^* = V\begin{pmatrix}\Sigma & 0\end{pmatrix}U^*$ is an SVD of A^*. Therefore, A and A^* have the same singular values.*
- *$A \in \mathbb{C}^{n\times n}$ is nonsingular if and only if all singular values are nonzero, i.e., $\sigma_j > 0$, $1 \leq j \leq n$.*
  If $A = U\Sigma V^*$ is an SVD of A, then $A^{-1} = V\Sigma^{-1}U^*$ is an SVD of A^{-1}.

Example 4.3. The 2×2 matrix

$$A = \begin{pmatrix}1 & \alpha\\0 & 1\end{pmatrix}$$

has a smallest singular value equal to

$$\sigma_2 = \left(\frac{2}{2+|\alpha|^2+|\alpha|\sqrt{4+|\alpha|^2}}\right)^{1/2}.$$

As $|\alpha| \to \infty$, the smallest singular value approaches zero, $\sigma_2 \to 0$, so that the absolute distance of A to singularity decreases. ■

Exercises

(i) Let $A \in \mathbb{C}^{n\times n}$. Show: All singular values of A are the same if and only if A is a multiple of a unitary matrix.

(ii) Show that the singular values of a Hermitian idempotent matrix are 0 and 1.

(iii) Show: $A \in \mathbb{C}^{n\times n}$ is Hermitian positive definite if and only if it has an SVD $A = V\Sigma V^*$ where Σ is nonsingular.

(iv) Let $A, B \in \mathbb{C}^{m\times n}$. Show: A and B have the same singular values if and only if there exist unitary matrices $Q \in \mathbb{C}^{n\times n}$ and $P \in \mathbb{C}^{m\times m}$ such that $B = PAQ$.

(v) Let $A \in \mathbb{C}^{m\times n}$, $m \geq n$, with QR decomposition $A = Q\begin{pmatrix}R\\0\end{pmatrix}$, where $Q \in \mathbb{C}^{m\times m}$ is unitary and $R \in \mathbb{C}^{n\times n}$. Determine an SVD of A from an SVD of R.

(vi) Determine an SVD of a column vector, and an SVD of a row vector.

(vii) Let $A \in \mathbb{C}^{m\times n}$ with $m \geq n$. Show: The singular values of A^*A are the squares of the singular values of A.

1. Show: If $A \in \mathbb{C}^{n\times n}$ is Hermitian positive definite and $\alpha > -\sigma_n$, then $A+\alpha I_n$ is also Hermitian positive definite with singular values $\sigma_j+\alpha$.
2. Let $A \in \mathbb{C}^{m\times n}$ and $\alpha > 0$. Express the singular values of $(A^*A+\alpha I)^{-1}A^*$ in terms of α and the singular values of A.
3. Let $A \in \mathbb{C}^{m\times n}$ with $m \geq n$. Show: The singular values of $\begin{pmatrix}I_n\\A\end{pmatrix}$ are equal to $\sqrt{1+\sigma_j^2}$, $1 \leq j \leq n$.

4.1 Extreme Singular Values

The smallest and largest singular values of a matrix provide information about the two norm of the matrix, the distance to singularity, and the two norm of the inverse.

Fact 4.4 (Extreme Singular Values). If $A \in \mathbb{C}^{m\times n}$ has singular values $\sigma_1 \geq \cdots \geq \sigma_p$, where $p = \min\{m,n\}$, then

$$\|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_1, \qquad \min_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sigma_p.$$

Proof. The two norm of A does not change when A is multiplied by unitary matrices; see Exercise (iv) in Section 2.6. Hence $\|A\|_2 = \|\Sigma\|_2$. Since Σ is a diagonal matrix, Exercise (i) in Section 2.6 implies $\|\Sigma\|_2 = \max_j |\sigma_j| = \sigma_1$.

To show the expression for σ_p, assume that $m \geq n$, so $p = n$. Then A has an SVD $A = U\begin{pmatrix}\Sigma\\0\end{pmatrix}V^*$. Let z be a vector so that $\|z\|_2 = 1$ and $\|Az\|_2 = \min_{\|x\|_2=1}\|Ax\|_2$. With $y = V^*z$ we get

$$\min_{\|x\|_2=1}\|Ax\|_2 = \|Az\|_2 = \|\Sigma V^*z\|_2 = \|\Sigma y\|_2 = \left(\sum_{i=1}^n \sigma_i^2|y_i|^2\right)^{1/2} \geq \sigma_n\|y\|_2 = \sigma_n.$$

Thus, $\sigma_n \leq \min_{\|x\|_2=1}\|Ax\|_2$. As for the reverse inequality,

$$\sigma_n = \|\Sigma e_n\|_2 = \|U^*AVe_n\|_2 = \|A(Ve_n)\|_2 \geq \min_{\|x\|_2=1}\|Ax\|_2.$$

The proof for $m < n$ is analogous. □

The extreme singular values are useful because they provide information about the two-norm condition number with respect to inversion, and about the distance to singularity.

The expressions below show that the largest singular value determines how much a matrix can stretch a unit-norm vector and the smallest singular value determines how much a matrix can shrink a unit-norm vector.

Fact 4.5. If $A \in \mathbb{C}^{n\times n}$ is nonsingular with singular values $\sigma_1 \geq \cdots \geq \sigma_n > 0$, then

$$\|A^{-1}\|_2 = \frac{1}{\sigma_n}, \qquad \kappa_2(A) = \frac{\sigma_1}{\sigma_n}.$$

The absolute distance of A to singularity is

$$\sigma_n = \min\{\|E\|_2 : \ A+E \text{ is singular}\}$$

and the relative distance is

$$\frac{\sigma_n}{\sigma_1} = \min\left\{\frac{\|E\|_2}{\|A\|_2} : \ A+E \text{ is singular}\right\}.$$

Proof. Remark 4.2 implies that $1/\sigma_j$ are the singular values of A^{-1}, so that $\|A^{-1}\|_2 = \max_j 1/|\sigma_j| = 1/\sigma_n$. The expressions for the distance to singularity follow from Fact 2.29 and Corollary 2.30. □

Fact 4.5 implies that a nonsingular matrix is almost singular in the absolute sense if its smallest singular value is close to zero. If the smallest and largest singular values are far apart, i.e., if $\sigma_1 \gg \sigma_n$, then the matrix is ill-conditioned with respect to inversion in the normwise relative sense, and it is almost singular in the relative sense.

The singular values themselves are well-conditioned in the normwise absolute sense. We show this below for the extreme singular values.

Fact 4.6. Let $A, A+E \in \mathbb{C}^{m\times n}$, $p = \min\{m,n\}$, and let $\sigma_1 \geq \cdots \geq \sigma_p$ be the singular values of A and $\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_p$ the singular values of $A+E$. Then

$$|\tilde{\sigma}_1 - \sigma_1| \leq \|E\|_2, \qquad |\tilde{\sigma}_p - \sigma_p| \leq \|E\|_2.$$

Proof. The inequality for σ_1 follows from $\sigma_1 = \|A\|_2$ and Fact 2.13, which states that norms are well-conditioned.

Regarding the bound for σ_p, let y be a vector so that $\sigma_p = \|Ay\|_2$ and $\|y\|_2 = 1$. Then the triangle inequality implies

$$\begin{aligned}\tilde{\sigma}_p &= \min_{\|x\|_2=1} \|(A+E)x\|_2 \leq \|(A+E)y\|_2 \leq \|Ay\|_2 + \|Ey\|_2 \\ &= \sigma_p + \|Ey\|_2 \leq \sigma_p + \|E\|_2.\end{aligned}$$

Hence $\tilde{\sigma}_p - \sigma_p \leq \|E\|_2$. To show that $-\|E\|_2 \leq \tilde{\sigma}_p - \sigma_p$, let y be a vector so that $\tilde{\sigma}_p = \|(A+E)y\|_2$ and $\|y\|_2 = 1$. Then the triangle inequality yields

$$\begin{aligned}\sigma_p &= \min_{\|x\|_2=1} \|Ax\|_2 \leq \|Ay\|_2 = \|(A+E)y - Ey\|_2 \leq \|(A+E)y\|_2 + \|Ey\|_2 \\ &= \tilde{\sigma}_p + \|Ey\|_2 \leq \tilde{\sigma}_p + \|E\|_2.\end{aligned}$$ □

Exercises

1. Extreme Singular Values of a Product.
   Let $A \in \mathbb{C}^{k\times m}$, $B \in \mathbb{C}^{m\times n}$, $q = \min\{k,n\}$, and $p = \min\{m,n\}$. Show:
   $$\sigma_1(AB) \leq \sigma_1(A)\sigma_1(B), \qquad \sigma_q(AB) \leq \sigma_1(A)\sigma_p(B).$$
2. Appending a column to a tall and skinny matrix does not increase the smallest singular value but can decrease it, because the new column may depend linearly on the old ones. The largest singular value does not decrease but it can increase, because more “mass” is added to the matrix.
   Let $A \in \mathbb{C}^{m\times n}$ with $m > n$, $z \in \mathbb{C}^m$, and $B = \begin{pmatrix} A & z \end{pmatrix}$. Show: $\sigma_{n+1}(B) \leq \sigma_n(A)$ and $\sigma_1(B) \geq \sigma_1(A)$.
3. Appending a row to a tall and skinny matrix does not decrease the smallest singular value but can increase it. Intuitively, this is because the columns

become longer which gives them an opportunity to become more linearly independent. The largest singular value does not decrease but can increase, because more "mass" is added to the matrix.

Let $A \in \mathbb{C}^{m\times n}$ with $m \geq n$, $z \in \mathbb{C}^n$, and $B = \begin{pmatrix} A \\ z^* \end{pmatrix}$. Show that

$$\sigma_n(B) \geq \sigma_n(A), \qquad \sigma_1(A) \leq \sigma_1(B) \leq \sqrt{\sigma_1(A)^2 + \|z\|_2^2}.$$

4.2 Rank

For a nonsingular matrix, all singular values are nonzero. For a general matrix, the number of nonzero singular values measures how much "information" is contained in a matrix, while the number of zero singular values indicates the amount of "redundancy."

Definition 4.7 (Rank). *The number of nonzero singular values of a matrix $A \in \mathbb{C}^{m\times n}$ is called the* rank *of A. An $m \times n$ zero matrix has rank* 0.

Example 4.8.

- If $A \in \mathbb{C}^{m\times n}$, then $\operatorname{rank}(A) \leq \min\{m,n\}$.
  This follows from Remark 4.2.
- If $A \in \mathbb{C}^{n\times n}$ is nonsingular, then $\operatorname{rank}(A) = n = \operatorname{rank}(A^{-1})$.
  A nonsingular matrix A contains the maximum amount of information, because it can reproduce any vector $b \in \mathbb{C}^n$ by means of $b = Ax$.
- For any $m \times n$ zero matrix 0, $\operatorname{rank}(0) = 0$.
  The zero matrix contains no information. It can only reproduce the zero vector, because $0x = 0$ for any vector x.
- If $A \in \mathbb{C}^{m\times n}$ has $\operatorname{rank}(A) = n$, then A has an SVD $A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^*$, where Σ is nonsingular. This means, all singular values of A are nonzero.
- If $A \in \mathbb{C}^{m\times n}$ has $\operatorname{rank}(A) = m$, then A has an SVD $A = U \begin{pmatrix} \Sigma & 0 \end{pmatrix} V^*$, where Σ is nonsingular. This means, all singular values of A are nonzero. ∎

A nonzero outer product uv^* contains little information: because $uv^*x = (v^*x)u$, the outer product uv^* can produce only multiples of the vector u.

Remark 4.9 (Outer Product). *If $u \in \mathbb{C}^m$ and $v \in \mathbb{C}^n$ with $u \neq 0$ and $v \neq 0$, then* $\operatorname{rank}(uv^*) = 1$.

To see this, determine an SVD of uv^*. Let $U \in \mathbb{C}^{m\times m}$ be a unitary matrix so that $U^*u = \|u\|_2 e_1$, and let $V \in \mathbb{C}^{n\times n}$ be a unitary matrix so that $V^*v = \|v\|_2 e_1$. Substituting these expressions into uv^* shows that $uv^* = U\Sigma V^*$ is an SVD, where $\Sigma \in \mathbb{R}^{m\times n}$ and $\Sigma = \|u\|_2\|v\|_2 e_1 e_1^*$. Therefore, the singular values of uv^* are $\|u\|_2\|v\|_2$, and $(\min\{m,n\} - 1)$ *zeros. In particular, $\|uv^*\|_2 = \|u\|_2\|v\|_2$.*

The above example demonstrates that a nonzero outer product has rank one. Now we show that a matrix of rank r can be represented as a sum of r outer products. To this end we distinguish the columns of the left and right singular vector matrices.

Definition 4.10 (Singular Vectors). *Let $A \in \mathbb{C}^{m\times n}$, with SVD $A = U\begin{pmatrix}\Sigma\\ 0\end{pmatrix}V^*$ if $m \geq n$, and SVD $A = U\begin{pmatrix}\Sigma & 0\end{pmatrix}V^*$ if $m \leq n$. Set $p = \min\{m,n\}$ and partition*

$$U = \begin{pmatrix}u_1 & \dots & u_m\end{pmatrix}, \qquad V = \begin{pmatrix}v_1 & \dots & v_n\end{pmatrix}, \qquad \Sigma = \begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_p\end{pmatrix},$$

where $\sigma_1 \geq \cdots \geq \sigma_p \geq 0$.

We call σ_j the jth singular value, u_j the jth left singular vector, *and v_j the jth* right singular vector.

Corresponding left and right singular vectors are related to each other.

Remark 4.11. *Let A have an SVD as in Definition* 4.10. *Then*

$$Av_i = \sigma_i u_i, \qquad A^* u_i = \sigma_i v_i, \qquad 1 \leq i \leq p.$$

This follows from the fact that U and V are unitary, and Σ is Hermitian.

Now we are ready to derive an economical representation for a matrix, where the size of the representation is proportional to the rank of the matrix. Fact 4.12 below shows that a matrix of rank r can be expressed in terms of r outer products. These outer products involve the singular vectors associated with the nonzero singular values.

Fact 4.12 (Reduced SVD). Let $A \in \mathbb{C}^{m\times n}$ have an SVD as in Definition 4.10. If $\operatorname{rank}(A) = r$, then

$$A = \sum_{j=1}^{r} \sigma_j u_j v_j^*.$$

Proof. From $\operatorname{rank}(A) = r$ follows $\sigma_1 \geq \cdots \geq \sigma_r > 0$. Confine the nonzero singular values to the matrix Σ_r, so that

$$\Sigma_r = \begin{pmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_r\end{pmatrix}, \qquad \text{and} \qquad A = U\begin{pmatrix}\Sigma_r & 0\\ 0 & 0\end{pmatrix}V^*$$

is an SVD of A. Partitioning the singular vectors conformally with the nonzero singular values,

$$U = \begin{array}{cc} r & m-r \\ \Big(U_r & U_{m-r}\Big)\end{array}, \qquad V = \begin{array}{cc} r & n-r \\ \Big(V_r & V_{n-r}\Big)\end{array},$$

yields $A = U_r \Sigma_r V_r^*$. Using $U_r = \begin{pmatrix}u_1 & \dots & u_r\end{pmatrix}$ and $V_r = \begin{pmatrix}v_1 & \dots & v_r\end{pmatrix}$, and viewing matrix multiplication as an outer product, as in View 4 of Section 1.7, shows

$$A = U_r \Sigma_r V_r^* = \begin{pmatrix}\sigma_1 u_1 & \dots & \sigma_r u_r\end{pmatrix}\begin{pmatrix}v_1^*\\ \vdots\\ v_r^*\end{pmatrix} = \sum_{j=1}^{r} \sigma_j u_j v_j^*. \qquad \square$$

For a nonsingular matrix, the reduced SVD is equal to the ordinary SVD.

Based on the above outer product representation of a matrix, we will now show that the singular vectors associated with the k largest singular values of A determine the rank k matrix that is closest to A in the two norm. Moreover, the $(k+1)$st singular value of A is the absolute distance of A, in the two norm, to the set of rank k matrices.

Fact 4.13 (Optimality of the SVD). Let $A \in \mathbb{C}^{m\times n}$ have an SVD as in Definition 4.9. If $k < \operatorname{rank}(A)$, then the absolute distance of A to the set of rank k matrices is

$$\sigma_{k+1} = \min_{B\in\mathbb{C}^{m\times n}, \operatorname{rank}(B)=k} \|A-B\|_2 = \|A-A_k\|_2,$$

where $A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^*$.

Proof. Write the SVD as

$$A = U\begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} V^*, \qquad \text{where} \quad \Sigma_1 = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_{k+1} \end{pmatrix}$$

and $\sigma_1 \geq \cdots \geq \sigma_{k+1} > 0$, so that Σ_1 is nonsingular. The idea is to show that the distance of Σ_1 to the set of singular matrices, which is σ_{k+1}, is a lower bound for the distance of A to the set of all rank k matrices.

Let $C \in \mathbb{C}^{m\times n}$ be a matrix with $\operatorname{rank}(C) = k$, and partition

$$U^*CV = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

where C_{11} is $(k+1)\times(k+1)$. From $\operatorname{rank}(C) = k$ follows $\operatorname{rank}(C_{11}) \leq k$ (although it is intuitively clear, it is proved rigorously in Fact 6.19), so that C_{11} is singular. Since the two norm is invariant under multiplication by unitary matrices, we obtain

$$\begin{aligned} \|A-C\|_2 &= \left\| \begin{pmatrix} \Sigma_1 & \\ & \Sigma_2 \end{pmatrix} - U^*CV \right\|_2 \\ &= \left\| \begin{pmatrix} \Sigma_1 - C_{11} & -C_{12} \\ -C_{21} & \Sigma_2 - C_{22} \end{pmatrix} \right\|_2 \geq \|\Sigma_1 - C_{11}\|_2. \end{aligned}$$

Since Σ_1 is nonsingular and C_{11} is singular, Facts 2.29 and 4.5 imply that $\|\Sigma_1 - C_{11}\|_2$ is bounded below by the distance of Σ_1 from singularity, and

$$\|\Sigma_1 - C_{11}\|_2 \geq \min\{\|\Sigma_1 - B_{11}\|_2 : \; B_{11} \text{ is singular}\} = \sigma_{k+1}.$$

A matrix C for which $\|A-C\|_2 = \sigma_{k+1}$ is $C = A_k$. This is because

$$C_{11} = \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \\ & & \sigma_k & \\ & & & 0 \end{pmatrix}, \quad C_{12} = 0, \quad C_{21} = 0, \quad C_{22} = 0.$$

Since the $\Sigma_1 - C_{11}$ has k diagonal elements equal to zero, and the diagonal elements of Σ_2 are less than or equal to σ_{k+1}, we obtain

$$\|A-C\|_2 = \left\| \begin{pmatrix} \Sigma_1 - C_{11} & 0 \\ 0 & \Sigma_2 \end{pmatrix} \right\|_2 = \left\| \begin{pmatrix} \sigma_{k+1} & \\ & \Sigma_2 \end{pmatrix} \right\|_2 = \sigma_{k+1}. \qquad \square$$

The singular values also help us to relate the rank of A to the rank of A^*A and AA^*. This will be important later on for the solution of least squares problems.

Fact 4.14. For any matrix $A \in \mathbb{C}^{m\times n}$,

1. $\operatorname{rank}(A) = \operatorname{rank}(A^*)$,
2. $\operatorname{rank}(A) = \operatorname{rank}(A^*A) = \operatorname{rank}(AA^*)$,
3. $\operatorname{rank}(A) = n$ if and only if A^*A is nonsingular,
4. $\operatorname{rank}(A) = m$ if and only if AA^* is nonsingular.

Proof.

1. This follows from Remark 4.2, because A and A^* have the same singular values.
2. If $m \geq n$, then A has an SVD $A = U \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^*$, and $A^*A = V\Sigma^2 V^*$ is an SVD of A^*A. Since Σ and Σ^2 have the same number of nonzero diagonal elements, $\operatorname{rank}(A) = \operatorname{rank}(A^*A)$. Also, $AA^* = U \begin{pmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{pmatrix} U^*$ is an SVD of AA^*. As before, $\operatorname{rank}(A) = \operatorname{rank}(AA^*)$ because Σ and Σ^2 have the same number of nonzero diagonal elements.
   A similar argument applies when $m < n$.
3. Since A^*A is $n \times n$, A^*A is nonsingular if and only if $n = \operatorname{rank}(A^*A) = \operatorname{rank}(A)$, where the second equality follows from item 2.
4. The proof is similar to that of item 3. □

In item 3 above the matrix A has linearly independent columns, and in item 4 it has linearly independent rows. Below we give another name to such matrices.

Definition 4.15 (Full Rank). *A matrix $A \in \mathbb{C}^{m\times n}$ has* full column rank *if* $\operatorname{rank}(A) = n$, *and* full row rank *if* $\operatorname{rank}(A) = m$.

A matrix $A \in \mathbb{C}^{m\times n}$ has full rank *if A has full column rank or full row rank. A matrix that does not have full rank is* rank deficient.

Example.

- A nonsingular matrix has full row rank and full column rank.
- A nonzero column vector has full column rank, and a nonzero row vector has full row rank.
- If $A \in \mathbb{C}^{n\times n}$ is nonsingular, then $\begin{pmatrix} A & B \end{pmatrix}$ has full row rank for any matrix $B \in \mathbb{C}^{n\times m}$, and $\begin{pmatrix} A \\ C \end{pmatrix}$ has full column rank, for any matrix $C \in \mathbb{C}^{m\times n}$.
- A singular square matrix is rank deficient. ■

Below we show that matrices with orthonormal columns also have full column rank. Recall from Definition 3.37 that A has orthonormal columns if $A^*A = I$.

Fact 4.16. A matrix $A \in \mathbb{C}^{m\times n}$ with orthonormal columns has $\operatorname{rank}(A) = n$, and all singular values are equal to one.

Proof. Fact 4.14 implies $\operatorname{rank}(A) = \operatorname{rank}(A^*A) = \operatorname{rank}(I_n) = n$. Thus A has full column rank, and we can write its SVD as $A = U\begin{pmatrix}\Sigma\\ 0\end{pmatrix}V^*$. Then $I_n = A^*A = V\Sigma^2V^*$ implies $\Sigma = I_n$, so that all singular values of A are equal to one. □

Exercises

(i) Let $A \in \mathbb{C}^{m\times n}$. Show: If $Q \in \mathbb{C}^{m\times m}$ and $P \in \mathbb{C}^{n\times n}$ are unitary, then $\operatorname{rank}(A) = \operatorname{rank}(QAP)$.

(ii) What can you say about the rank of a nilpotent matrix, and the rank of an idempotent matrix?

(iii) Let $A \in \mathbb{C}^{m\times n}$. Show: If $\operatorname{rank}(A) = n$, then $\|(A^*A)^{-1}A^*\|_2 = 1/\sigma_n$, and if $\operatorname{rank}(A) = m$, then $\|(AA^*)^{-1}A\|_2 = 1/\sigma_m$.

(iv) Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$. Show that $A(A^*A)^{-1}A^*$ is idempotent and Hermitian, and $\|A(A^*A)^{-1}A^*\|_2 = 1$.

(v) Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = m$. Show that $A^*(AA^*)^{-1}A$ is idempotent and Hermitian, and $\|A^*(AA^*)^{-1}A\|_2 = 1$.

(vi) Nilpotent Matrices.
Let $A \in \mathbb{C}^{n\times n}$ be nilpotent so that $A^j = 0$ and $A^{j-1} \neq 0$ for some $j \geq 1$. Let $b \in \mathbb{C}^n$ with $A^{j-1}b \neq 0$. Show that $K = \begin{pmatrix} b & Ab & \ldots & A^{j-1}b\end{pmatrix}$ has full column rank.

(vii) In Fact 4.13 let B be a multiple of A_k, i.e., $B = \alpha A_k$. Determine $\|A - B\|_2$.

1. Let $A \in \mathbb{C}^{n\times n}$. Show that there exists a unitary matrix Q such that $A^* = QAQ$.
2. Polar Decomposition.
   Show: If $A \in \mathbb{C}^{m\times n}$ has $\operatorname{rank}(A) = n$, then there is a factorization $A = PH$, where $P \in \mathbb{C}^{m\times n}$ has orthonormal columns, and $H \in \mathbb{C}^{n\times n}$ is Hermitian positive definite.
3. The polar factor P is the closest matrix with orthonormal columns in the two norm.
   Let $A \in \mathbb{C}^{n\times n}$ have a polar decomposition $A = PH$. Show that $\|A - P\|_2 \leq \|A - Q\|_2$ for any unitary matrix Q.
4. The distance of a matrix A from its polar factor P is determined by how close the columns A are to being orthonormal.
   Let $A \in \mathbb{C}^{m\times n}$, with $\operatorname{rank}(A) = n$, have a polar decomposition $A = PH$. Show that
   $$\frac{\|A^*A - I_n\|_2}{1 + \|A\|_2} \leq \|A - P\|_2 \leq \frac{\|A^*A - I_n\|_2}{1 + \sigma_n}.$$

5. Let $A \in \mathbb{C}^{n\times n}$ and $\sigma > 0$. Show: σ is a singular value of A if and only if the matrix
$$\begin{pmatrix} A & -\sigma I \\ -\sigma I & A^* \end{pmatrix}$$
is singular.
6. Rank Revealing QR Factorization.
With an appropriate permutation of the columns, a QR factorization can almost reveal the smallest singular value of a full column rank matrix.
Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$ and smallest singular value σ_n. Let the corresponding singular vectors be $Av = \sigma_n u$, where $\|v\|_2 = \|u\|_2 = 1$. Choose a permutation P so that $w = P^* v$ and $|w_n| = \|w\|_\infty$, and let $AP = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$ be a QR decomposition of AP. Show: $|r_{nn}| \leq \sqrt{n}\sigma_n$.

4.3 Singular Vectors

The singular vectors of a matrix A give information about the column spaces and null spaces of A and A^*.

The column space of a matrix A is the set of all right-hand sides b for which the system $Ax = b$ has a solution, and the null space of A determines whether these solutions are unique.

Definition 4.17 (Column Space and Null Space). *If $A \in \mathbb{C}^{m\times n}$, then the set*
$$\mathcal{R}(A) = \{b \in \mathbb{C}^m : \ b = Ax \text{ for some } x \in \mathbb{C}^n\}$$
is the column space *or* range *of A, and the set*
$$\operatorname{Ker}(A) = \{x \in \mathbb{C}^n : \ Ax = 0\}$$
is the kernel *or* null space *of A.*

Example.

- The column space of an $m \times n$ zero matrix is the zero vector, and the null space is $\mathbb{C}^n$, i.e., $\mathcal{R}(0_{m\times n}) = \{0_{m\times 1}\}$ and $\operatorname{Ker}(0_{m\times n}) = \mathbb{C}^n$.
- The column space of an $n \times n$ nonsingular complex matrix is $\mathbb{C}^n$, and the null space consists of the single vector $0_{n\times 1}$.
- $\operatorname{Ker}(A) = \{0\}$ if and only if the columns of the matrix A are linearly independent.
- If $A \in \mathbb{C}^{m\times n}$, then for all $k \geq 1$
$$\mathcal{R}\begin{pmatrix} A & 0_{m\times k} \end{pmatrix} = \mathcal{R}(A), \qquad \operatorname{Ker}\begin{pmatrix} A \\ 0_{k\times n} \end{pmatrix} = \operatorname{Ker}(A).$$

- If $A \in \mathbb{C}^{n\times n}$ is nonsingular, then for any $B \in \mathbb{C}^{n\times p}$ and $C \in \mathbb{C}^{p\times n}$

$$\mathcal{R}\begin{pmatrix} A & B \end{pmatrix} = \mathcal{R}(A), \qquad \operatorname{Ker}\begin{pmatrix} A \\ C \end{pmatrix} = \{0_{n\times 1}\}. \qquad \blacksquare$$

The column and null spaces of A^* are also important, and we give them names that relate to the matrix A.

Definition 4.18 (Row Space and Left Null Space). *Let $A \in \mathbb{C}^{m\times n}$. The set*

$$\mathcal{R}(A^*) = \{d \in \mathbb{C}^n : \ d = A^* y \textit{ for some } y \in \mathbb{C}^m\}$$

is the row space *of A. The set*

$$\operatorname{Ker}(A^*) = \{y \in \mathbb{C}^m : \ A^* y = 0\}$$

is the left null space *of A.*

Note that all spaces of a matrix are defined by *column* vectors.

Example 4.19. If A is Hermitian, then $\mathcal{R}(A^*) = \mathcal{R}(A)$ and $\operatorname{Ker}(A^*) = \operatorname{Ker}(A)$. $\blacksquare$

The singular vectors reproduce the four spaces associated with a matrix. Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = r$ and SVD

$$A = U \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^*,$$

where Σ_r is nonsingular, and

$$\begin{array}{cc} & \begin{array}{cc} r & n-r \end{array} \\ \begin{array}{c} r \\ m-r \end{array} & \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} \end{array}, \qquad U = \begin{array}{c} \begin{array}{cc} r & m-r \end{array} \\ \begin{pmatrix} U_r & U_{m-r} \end{pmatrix} \end{array}, \qquad V = \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{pmatrix} V_r & V_{n-r} \end{pmatrix} \end{array}.$$

Fact 4.20 (Spaces of a Matrix and Singular Vectors). Let $A \in \mathbb{C}^{m\times n}$.

1. The leading r left singular vectors represent the column space of A: If $A \neq 0$, then $\mathcal{R}(U_r) = \mathcal{R}(A)$; otherwise $\mathcal{R}(A) = \{0_{m\times 1}\}$.
2. The trailing $n-r$ right singular vectors represent the null space of A: If $\operatorname{rank}(A) = r < n$, then $\mathcal{R}(V_{n-r}) = \operatorname{Ker}(A)$; otherwise $\operatorname{Ker}(A) = \{0_{n\times 1}\}$.
3. The leading r right singular vectors represent the row space of A: If $A \neq 0$, then $\mathcal{R}(A^*) = \mathcal{R}(V_r)$; otherwise $\mathcal{R}(A^*) = \{0_{n\times 1}\}$.
4. The trailing $m-r$ left singular vectors represent the left null space of A: If $r < m$, then $\mathcal{R}(U_{m-r}) = \operatorname{Ker}(A^*)$; otherwise $\operatorname{Ker}(A^*) = \{0_{m\times 1}\}$.

Proof. Although the statements may be intuitively obvious, they are proved rigorously in Section 6.1. □

The singular vectors help us to relate the spaces of A^*A and AA^* to those of the matrix A. Since A^*A and AA^* are Hermitian, we need to specify only two spaces; see Example 4.19.

Fact 4.21 (Spaces of A^*A and AA^*). Let $A \in \mathbb{C}^{m\times n}$.

1. $\operatorname{Ker}(A^*A) = \operatorname{Ker}(A)$ and $\mathcal{R}(A^*A) = \mathcal{R}(A^*)$.
2. $\mathcal{R}(AA^*) = \mathcal{R}(A)$ and $\operatorname{Ker}(AA^*) = \operatorname{Ker}(A^*)$.

Proof. Fact 4.14 implies that A^*A and AA^* have the same rank as A. Since A^*A has the same right singular vectors as A, Fact 4.20 implies $\operatorname{Ker}(A^*A) = \operatorname{Ker}(A)$ and $\mathcal{R}(A^*A) = \mathcal{R}(A^*)$. Since AA^* has the same left singular vectors as A, Fact 4.20 implies $\mathcal{R}(AA^*) = \mathcal{R}(A)$ and $\operatorname{Ker}(AA^*) = \operatorname{Ker}(A^*)$. □

In the special case when the rank of a matrix is equal to the number of rows, then the number of elements in the column space is as large as possible. When the rank of the matrix is equal to the number of columns, then the number of elements in the null space is as small as possible.

Fact 4.22 (Spaces of Full Rank Matrices). Let $A \in \mathbb{C}^{m\times n}$. Then

1. $\operatorname{rank}(A) = m$ if and only if $\mathcal{R}(A) = \mathbb{C}^m$;
2. $\operatorname{rank}(A) = n$ if and only if $\operatorname{Ker}(A) = \{0\}$.

Proof. Let $A = U\begin{pmatrix}\Sigma_r & 0\\ 0 & 0\end{pmatrix}V^*$ be an SVD of A, where Σ_r is nonsingular.

1. From Fact 4.20 follows $\mathcal{R}(A) = \mathcal{R}(U_r)$. Hence $r = m$ if and only if $U_r = U$, because U is nonsingular so that $\mathcal{R}(U) = \mathbb{C}^m$.
2. Fact 4.20 also implies $r = n$ if and only if V_{n-r} is empty, which means that $\operatorname{Ker}(A) = \{0\}$. □

If the matrix in a linear system has full rank, then existence or uniqueness of a solution is guaranteed.

Fact 4.23 (Solutions of Full Rank Linear Systems). Let $A \in \mathbb{C}^{m\times n}$.

1. If $\operatorname{rank}(A) = m$, then $Ax = b$ has a solution $x = A^*(AA^*)^{-1}b$ for every $b \in \mathbb{C}^m$.
2. If $\operatorname{rank}(A) = n$ and if $b \in \mathcal{R}(A)$, then $Ax = b$ has the unique solution $x = (A^*A)^{-1}A^*b$.

Proof.

1. Fact 4.22 implies that $Ax = b$ has a solution for every $b \in \mathbb{C}^m$, and Fact 4.14 implies that AA^* is nonsingular. Clearly, $x = A^*(AA^*)^{-1}b$ satisfies $Ax = b$.
2. Since $b \in \mathcal{R}(A)$, $Ax = b$ has a solution. Multiplying on the left by A^* gives $A^*Ax = A^*b$. According to Fact 4.14, A^*A is nonsingular, so that $x = (A^*A)^{-1}A^*b$.
   Suppose $Ax = b$ and $Ay = b$; then $A(x-y) = 0$. Fact 4.22 implies that $\operatorname{Ker}(A) = \{0\}$, so $x = y$, which proves uniqueness. □

Exercises

(i) Fredholm's Alternatives.

(a) The first alternative implies that $\mathcal{R}(A)$ and $\mathrm{Ker}(A^*)$ have only the zero vector in common. Assume $b \neq 0$ and show:

If $Ax = b$ has a solution, then $b^* A \neq 0$.

In other words, if $b \in \mathcal{R}(A)$, then $b \notin \mathrm{Ker}(A^*)$.

(b) The second alternative implies that $\mathrm{Ker}(A)$ and $\mathcal{R}(A^*)$ have only the zero vector in common. Assume $x \neq 0$ and show:

If $Ax = 0$, then there is no y such that $x = A^* y$.

In other words, if $x \in \mathrm{Ker}(A)$, then $x \notin \mathcal{R}(A^*)$,

(ii) Normal Matrices.
If $A \in \mathbb{C}^n$ is Hermitian, then $\mathcal{R}(A^*) = \mathcal{R}(A)$ and $\mathrm{Ker}(A^*) = \mathrm{Ker}(A)$. These equalities remain true for a larger class of matrices, the so-called *normal matrices*. A matrix $A \in \mathbb{C}^n$ is *normal* if $A^* A = AA^*$.
Show: If $A \in \mathbb{C}^{n \times n}$ is normal, then $\mathcal{R}(A^*) = \mathcal{R}(A)$ and $\mathrm{Ker}(A^*) = \mathrm{Ker}(A)$.

5. Least Squares Problems

Here we solve linear systems $Ax = b$ that do not have a solution. If b is not in the column space of A, there is no x such that $Ax = b$. The best we can do is to find a vector y that brings left- and right-hand sides of the linear system as close as possible; in other words y is chosen to make the distance between Ay and b as small as possible. That is, we want to minimize the distance $\|Ax - b\|_2$ over all x, and distance will again be measured in the two norm.

Definition 5.1 (Least Squares Problem). *Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. The* least squares problem *consists of finding a vector $y \in \mathbb{C}^n$ so that*

$$\min_x \|Ax - b\|_2 = \|Ay - b\|_2.$$

The vector $Ay - b$ is called the least squares residual.

The name comes about as follows:

$$\min_x \|Ax - b\|_2^2 = \underbrace{\min_x}_{\text{least}} \sum_i \underbrace{|(Ax - b)_i|^2}_{\text{squares}}.$$

5.1 Solutions of Least Squares Problems

We express the solutions of least squares problems in terms of the SVD.

Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = r$ and an SVD

$$A = U\begin{pmatrix}\Sigma_r & 0\\ 0 & 0\end{pmatrix}V^*, \qquad U = \begin{array}{cc} r & m-r \\ \big(U_r & U_{m-r}\big) \end{array}, \qquad V = \begin{array}{cc} r & n-r \\ \big(V_r & V_{n-r}\big) \end{array},$$

where $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary, and Σ_r is a diagonal matrix with diagonal elements $\sigma_1 \geq \cdots \geq \sigma_r > 0$, i.e., Σ_r is nonsingular.

Fact 5.2 (All Least Squares Solutions). Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. The solutions of $\min_x \|Ax-b\|_2$ are of the form $y = V_r\Sigma_r^{-1}U_r^*b + V_{n-r}z$ for any $z \in \mathbb{C}^{n-r}$.

Proof. Let y be a solution of the least squares problem, partition

$$V^*y = \begin{pmatrix} V_r^*y \\ V_{n-r}^*y \end{pmatrix} = \begin{pmatrix} w \\ z \end{pmatrix},$$

and substitute the SVD of A into the residual,

$$Ay-b = U\begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^*y - b = U\begin{pmatrix} \Sigma_r w - U_r^*b \\ -U_{m-r}^*b \end{pmatrix}.$$

Two norms are invariant under multiplication by unitary matrices, so that

$$\|Ay-b\|_2^2 = \|\Sigma_r w - U_r^*b\|_2^2 + \|U_{m-r}^*b\|_2^2.$$

Since the second summand is constant and independent of w and z, the residual is minimized if the first summand is zero, that is, if $w = \Sigma_r^{-1}U_r^*b$. Therefore, the solution of the least squares problem equals

$$y = V\begin{pmatrix} w \\ z \end{pmatrix} = V_r w + V_{n-r}z = V_r\Sigma_r^{-1}U_r^*b + V_{n-r}z.$$

Fact 4.20 implies that $V_{n-r}z \in \operatorname{Ker}(A)$ for any vector z. Hence $V_{n-r}z$ does not have any effect on the least squares residual, so that z can assume any value. □

Fact 5.1 shows that if A has rank $r < n$, then the least squares problem has infinitely many solutions. The first term in a least squares solution contains the matrix

$$V_r\Sigma_r^{-1}U_r^* = V\begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*$$

which is obtained by inverting only the nonsingular parts of an SVD. This matrix is almost an inverse, but not quite.

Definition 5.3 (Moore–Penrose Inverse). *If $A \in \mathbb{C}^{m\times n}$ and* $\operatorname{rank}(A) = r \geq 1$, *let $A = U\begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^*$ be an SVD where Σ_r is nonsingular. The $n\times m$ matrix*

$$A^\dagger = V\begin{pmatrix} \Sigma_r^{-1} & 0 \\ 0 & 0 \end{pmatrix} U^*$$

is called Moore–Penrose inverse *of A. If $A = 0_{m\times n}$, then $A^\dagger = 0_{n\times m}$.*

The Moore–Penrose inverse of a full rank matrix can be expressed in terms of the matrix itself.

Remark 5.4 (Moore–Penrose Inverses of Full Rank Matrices). *Let $A \in \mathbb{C}^{m\times n}$.*

- *If A is nonsingular, then $A^\dagger = A^{-1}$.*

- *If $A \in \mathbb{C}^{m \times n}$ and* $\operatorname{rank}(A) = n$*, then $A^\dagger = (A^*A)^{-1}A^*$.*
  This means $A^\dagger A = I_n$, so that $A^\dagger$ is a left inverse of A.
- *If $A \in \mathbb{C}^{m \times n}$ and* $\operatorname{rank}(A) = m$*, then $A^\dagger = A^*(AA^*)^{-1}$.*
  This means $AA^\dagger = I_m$, so that $A^\dagger$ is a right inverse of A.

Now we can express the least squares solutions in terms of the Moore–Penrose inverse, without reference to the SVD.

Corollary 5.5 (All Least Squares Solutions). *Let $A \in \mathbb{C}^{m \times n}$ and $b \in \mathbb{C}^{m \times n}$. The solutions of* $\min_x \|Ax - b\|_2$ *are of the form $y = A^\dagger b + q$, where $q \in$* $\operatorname{Ker}(A)$.

Proof. This follows from setting $q = V_{n-r}z \in \operatorname{Ker}(A)$ in Fact 4.20. □

Although a least squares problem can have infinitely many solutions, all solutions have the part $A^\dagger b$ in common, and they differ only in the part that belongs to $\operatorname{Ker}(A)$. As a result, all least squares solutions have not just residuals of the same norm, but they have the same residual.

Fact 5.6 (Uniqueness of the Least Squares Residual). Let $A \in \mathbb{C}^{m \times n}$ and $b \in \mathbb{C}^m$. All solutions y of $\min_x \|Ax - b\|_2$ have the same residual $b - Ay = (I - AA^\dagger)b$.

Proof. Let y_1 and y_2 be solutions to $\min_x \|Ax - b\|_2$. Corollary 5.5 implies $y_1 = A^\dagger b + q_1$ and $y_2 = A^\dagger b + q_2$, where $q_1, q_2 \in \operatorname{Ker}(A)$. Hence $Ay_1 = AA^\dagger b = Ay_2$, and both solutions have the same residual, $b - Ay_1 = b - Ay_2 = (I - AA^\dagger)b$. □

Besides being unique, the least squares residual has another important property: It is orthogonal to the column space of the matrix.

Fact 5.7 (Residual is Orthogonal to Column Space). Let $A \in \mathbb{C}^{m \times n}$, $b \in \mathbb{C}^m$, and y a solution of $\min_x \|Ax - b\|_2$ with residual $r = b - Ay$. Then $A^*r = 0$.

Proof. Fact 5.6 implies that the unique residual is $r = (I - AA^\dagger)b$. Let A have an SVD

$$A = U \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^*,$$

where U and V are unitary, and Σ_r is a diagonal matrix with positive diagonal elements. From Definition 5.3 of the Moore–Penrose inverse we obtain

$$AA^\dagger = U \begin{pmatrix} I_r & 0 \\ 0 & 0_{(m-r)\times(m-r)} \end{pmatrix} U^*, \qquad I - AA^\dagger = U \begin{pmatrix} 0_{r \times r} & 0 \\ 0 & I_{m-r} \end{pmatrix} U^*.$$

Hence $A^*(I - AA^\dagger) = 0_{n \times m}$ and $A^*r = 0$. □

The part of the least squares problem solution $y = A^\dagger b + q$ that is responsible for lack of uniqueness is the term $q \in \operatorname{Ker}(A)$. We can force the least squares problem to have a unique solution if we add the constraint $q = 0$. It turns out that the resulting solution $A^\dagger b$ has minimal norm among all least squares solutions.

Fact 5.8 (Minimal Norm Least Squares Solution). Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. Among all solutions of $\min_x \|Ax - b\|_2$ the one with minimal two norm is $y = A^\dagger b$.

Proof. From the proof of Fact 5.2 follows that any least squares solution has the form

$$y = V \begin{pmatrix} \Sigma_r^{-1} U_r^* b \\ z \end{pmatrix}.$$

Hence

$$\|y\|_2^2 = \|\Sigma_r^{-1} U_r^* b\|_2^2 + \|z\|_2^2 \geq \|\Sigma_r^{-1} U_r^* b\|_2^2 = \|V_r \Sigma_r^{-1} U_r^* b\|_2^2 = \|A^\dagger b\|_2^2.$$

Thus, any least squares solution y satisfies $\|y\|_2 \geq \|A^\dagger b\|_2$. This means $y = A^\dagger b$ is the least squares solution with minimal two norm. □

The most pleasant least squares problems are those where the matrix A has full column rank because then $\mathrm{Ker}(A) = \{0\}$ and the least squares solution is unique.

Fact 5.9 (Full Column Rank Least Squares). Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. If $\mathrm{rank}(A) = n$, then $\min_x \|Ax - b\|_2$ has the unique solution $y = (A^*A)^{-1}A^*b$.

Proof. From Fact 4.22 we know that $\mathrm{rank}(A) = n$ implies $\mathrm{Ker}(A) = \{0\}$. Hence $q = 0$ in Corollary 5.5. The expression for $A^\dagger$ follows from Remark 5.4. □

In particular, when A is nonsingular, then the Moore–Penrose inverse reduces to the ordinary inverse. This means, if we solve a least squares problem $\min_x \|Ax - b\|_2$ with a nonsingular matrix A, we obtain the solution $y = A^{-1}b$ of the linear system $Ax = b$.

Exercises

(i) What is the Moore–Penrose inverse of a nonzero column vector? of a nonzero row vector?

(ii) Let $u \in \mathbb{C}^{m\times n}$ and $v \in \mathbb{C}^n$ with $v \neq 0$. Show that $\|uv^\dagger\|_2 = \|u\|_2/\|v\|_2$.

(iii) Let $A \in \mathbb{C}^{m\times n}$. Show that the following matrices are idempotent:

$$AA^\dagger, \qquad A^\dagger A, \qquad I_m - AA^\dagger, \qquad I_n - A^\dagger A.$$

(iv) Let $A \in \mathbb{C}^{m\times n}$. Show: If $A \neq 0$, then $\|AA^\dagger\|_2 = \|A^\dagger A\|_2 = 1$.

(v) Let $A \in \mathbb{C}^{m\times n}$. Show:

$$(I_m - AA^\dagger)A = 0_{m\times n}, \qquad A(I_n - A^\dagger A) = 0_{m\times n}.$$

(vi) Let $A \in \mathbb{C}^{m\times n}$. Show: $\mathcal{R}(A^\dagger) = \mathcal{R}(A^*)$ and $\mathrm{Ker}(A^\dagger) = \mathrm{Ker}(A^*)$.

(vii) Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = r$. Show: $\|A^\dagger\|_2 = 1/\sigma_r$.

(viii) Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = n$. Show: $\|(A^*A)^{-1}\|_2 = \|A^\dagger\|_2^2$.

(ix) Let $A = BC$ where $B \in \mathbb{C}^{m\times n}$ has $\operatorname{rank}(B) = n$ and $C \in \mathbb{C}^{n\times n}$ is nonsingular. Show: $A^\dagger = C^{-1}B^\dagger$.

(x) Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$ and thin QR factorization $A = QR$, where $Q^*Q = I_n$ and R is upper triangular. Show: $A^\dagger = R^{-1}Q^*$.

(xi) Show: If A has orthonormal columns, then $A^\dagger = A^*$.

(xii) Partial Isometry.
A matrix $A \in \mathbb{C}^{m\times n}$ is called a partial isometry if $A^\dagger = A^*$. Show: A is a partial isometry if and only if all its singular values are 0 or 1.

(xiii) What is the minimal norm solution to $\min_x \|Ax - b\|_2$ when $A = 0$?

(xiv) If y is the minimal norm solution to $\min_x \|Ax - b\|_2$ and $A^*b = 0$, then what can you say about y?

(xv) Given an *approximate* solution z to a linear system $Ax = b$, this problem shows how to construct a linear system $(A + E)x = b$ for which z is the *exact* solution.
Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. Let $z \in \mathbb{C}^n$ with $z \neq 0$ and residual $r = b - Az$. Show: If $E = rz^\dagger$, then $(A + E)z = b$.

1. What is the minimal norm solution to $\min_x \|Ax - b\|_2$ when $A = uv^*$, where u and v are column vectors?
2. Let $A \in \mathbb{C}^{m\times n}$. Show: The singular values of $\begin{pmatrix} I_n \\ A \end{pmatrix}^\dagger$ are equal to $1/\sqrt{1+\sigma_j^2}$, $1 \le j \le n$.
3. Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = n$. Show: $\|I - AA^\dagger\|_2 = \min\{1, m-n\}$.
4. Let $A \in \mathbb{C}^{m\times n}$. Show: $A^\dagger$ is the Moore–Penrose inverse of A if and only if $A^\dagger$ satisfies

   MP1: $AA^\dagger A = A$, $A^\dagger AA^\dagger = A^\dagger$,

   MP2: $AA^\dagger$ and $A^\dagger A$ are Hermitian.
5. Partitioned Moore–Penrose Inverse.
   Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = n$ and be partitioned as $A = \begin{pmatrix} A_1 & A_2 \end{pmatrix}$. Show:

   (a)
   $$A^\dagger = \begin{pmatrix} B_1^\dagger \\ B_2^\dagger \end{pmatrix}, \qquad \text{where} \quad B_1 = (I - A_2A_2^\dagger)A_1, \quad B_2 = (I - A_1A_1^\dagger)A_2.$$

   (b) $\|B_1\|_2 = \min_Z \|A_1 - A_2Z\|_2$ and $\|B_2\|_2 = \min_Z \|A_2 - A_1Z\|_2$.

   (c) Let $1 \le k \le n$, and let V_{11} be the leading $k \times k$ principal submatrix of V. Show: If V_{11} is nonsingular, then $\|A_1^\dagger\|_2 \le \|V_{11}^{-1}\|_2/\sigma_k$.

5.2 Conditioning of Least Squares Problems

Least squares problems are much more sensitive to perturbations than linear systems. A least squares problem whose matrix is deficient in column rank is so sensitive that we cannot even define a condition number. The example below illustrates this.

Example 5.10 (Rank Deficient Least Squares Problems are Ill-Posed). Consider the least squares problem $\min_x \|Ax-b\|_2$ with

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = A^\dagger, \qquad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad y = A^\dagger b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The matrix A is rank deficient and y is the minimal norm solution. Let us perturb the matrix so that

$$A+E = \begin{pmatrix} 1 & 0 \\ 0 & \epsilon \end{pmatrix}, \qquad \text{where} \qquad 0 < \epsilon \ll 1.$$

The matrix $A+E$ has full column rank and $\min_x \|(A+E)x-b\|_2$ has the unique solution z where

$$z = (A+E)^\dagger b = (A+E)^{-1} b = \begin{pmatrix} 1 \\ 1/\epsilon \end{pmatrix}.$$

Comparing the two minimal norm solutions shows that the second element of z grows as the $(2,2)$ element of $A+E$ decreases, i.e., $z_2 = 1/\epsilon \rightarrow \infty$ as $\epsilon \rightarrow 0$. But at $\epsilon = 0$ we have $z_2 = 0$. Therefore, the least squares solution does not depend continuously on the $(2,2)$ element of the matrix. This is an *ill-posed* problem.

In an ill-posed problem the solution is not a continuous function of the inputs. The ill-posedness of a rank deficient least squares problem comes about because a small perturbation can increase the rank of the matrix. ■

To avoid ill-posedness we restrict ourselves to least squares problems where the exact and perturbed matrices have full column rank. Below we determine the sensitivity of the least squares solution to changes in the right-hand side.

Fact 5.11 (Right-Hand Side Perturbation). Let $A \in \mathbb{C}^{m \times n}$ have $\operatorname{rank}(A) = n$, let y be the solution to $\min_x \|Ax-b\|_2$, and let z be the solution to $\min_x \|Ax-(b+f)\|_2$. If $y \neq 0$, then

$$\frac{\|z-y\|_2}{\|y\|_2} \leq \kappa_2(A) \frac{\|f\|_2}{\|A\|_2 \|y\|_2},$$

and if $z \neq 0$, then

$$\frac{\|z-y\|_2}{\|z\|_2} \leq \kappa_2(A) \frac{\|f\|_2}{\|A\|_2 \|z\|_2},$$

where $\kappa_2(A) = \|A\|_2 \|A^\dagger\|_2$.

Proof. Fact 5.9 implies that $y = A^\dagger b$ and $z = A^\dagger(b+f)$ are the unique solutions to the respective least squares problems. From $y = A^\dagger b = (A^*A)^{-1}A^*b$, see

Remark 5.4, and the assumption $A^*b \neq 0$ follows $y \neq 0$. Applying the bound for matrix multiplication in Fact 2.22 yields

$$\frac{\|z-y\|_2}{\|y\|_2} \leq \frac{\|A^\dagger\|_2\|b\|_2}{\|A^\dagger b\|_2}\frac{\|f\|_2}{\|b\|_2} = \|A^\dagger\|_2\frac{\|f\|_2}{\|y\|_2}.$$

Now multiply and divide by $\|A\|_2$ on the right. □

In Fact 5.11 we have extended the two-norm condition number with respect to inversion from nonsingular matrices to matrices with full column rank.

Definition 5.12. *Let $A \in \mathbb{C}^{m\times n}$ with* $\text{rank}(A) = n$. *Then $\kappa_2(A) = \|A\|_2\|A^\dagger\|_2$ is the two-norm condition number of A with regard to left inversion.*

Fact 5.11 implies that $\kappa_2(A)$ is the normwise relative condition number of the least squares solution to changes in the right-hand side. If the columns of A are close to being linearly dependent, then A is close to being rank deficient and the least squares solution is sensitive to changes in the right-hand side.

With regard to changes in the matrix, though, the situation is much bleaker. It turns out that least squares problems are much more sensitive to changes in the matrix than linear systems.

Example 5.13 (Large Residual Norm). Let

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \\ 0 & 0 \end{pmatrix}, \qquad b = \begin{pmatrix} \beta_1 \\ 0 \\ \beta_3 \end{pmatrix}, \qquad \text{where} \quad 0 < \alpha \leq 1, \quad 0 < \beta_1, \beta_3.$$

The element β_3 represents the part of b outside $\mathcal{R}(A)$. The matrix A has full column rank, and the least squares problem $\min_x \|Ax-b\|_2$ has the unique solution y where

$$A^\dagger = (A^*A)^{-1}A^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1/\alpha & 0 \end{pmatrix}, \qquad y = A^\dagger b = \begin{pmatrix} \beta_1 \\ 0 \end{pmatrix}.$$

The residual norm is $\min_x \|Ax-b\|_2 = \|Ay-b\|_2 = \beta_3$.

Let us perturb the matrix and change its column space so that

$$A+E = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \\ 0 & \epsilon \end{pmatrix}, \qquad \text{where} \quad 0 < \epsilon \ll 1.$$

Note that $\mathcal{R}(A+E) \neq \mathcal{R}(A)$. The matrix $A+E$ has full column rank and Moore–Penrose inverse

$$(A+E)^\dagger = \left[(A+E)^*(A+E)\right]^{-1}(A+E)^* = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\alpha}{\alpha^2+\epsilon^2} & \frac{\epsilon}{\alpha^2+\epsilon^2} \end{pmatrix}.$$

The perturbed problem $\min_x \|(A+E)x-b\|_2$ has the unique solution z, where

$$z = (A+E)^\dagger b = \begin{pmatrix} \beta_1 \\ \epsilon\beta_3/(\alpha^2+\epsilon^2) \end{pmatrix}.$$

Since $\|y\|_2 = \beta_1$, the normwise relative error is

$$\frac{\|z-y\|_2}{\|y\|_2} = \frac{\beta_3\epsilon}{\beta_1(\alpha^2+\epsilon^2)} \leq \frac{\beta_3}{\alpha^2\beta_1}\epsilon.$$

If $\beta_3 \geq \beta_1$, then $\beta_3/(\alpha^2\beta_1) \geq 1/\alpha^2$. This means if more of b is outside $\mathcal{R}(A)$ than inside $\mathcal{R}(A)$, then the perturbation is amplified by at least $1/\alpha^2$.

In other words, since $\|E\| = \epsilon$, $\|A^\dagger\|_2 = 1/\alpha$, and $\beta_3/\beta_1 = \|Ay-b\|_2/\|y\|_2$, we can write

$$\frac{\|z-y\|_2}{\|y\|_2} \leq \|A^\dagger\|_2^2\,\frac{\|Ay-b\|_2}{\|y\|_2}\,\|E\|_2 = [\kappa_2(A)]^2\,\frac{\|r\|_2}{\|A\|_2\|y\|_2}\,\frac{\|E\|_2}{\|A\|_2},$$

where $r = Ay-b$ is the residual. This means, if the right-hand side is far away from the column space, then the condition number with respect to changes in the matrix is $[\kappa_2(A)]^2$, rather than just $\kappa_2(A)$.

We can give a geometric interpretation for the relative residual norm. If we bound

$$\frac{\|r\|_2}{\|A\|_2\|y\|_2} \leq \frac{\|r\|_2}{\|Ay\|_2},$$

then we can exploit the relation between $\|r\|_2$ and $\|Ay\|_2$ from Exercise (iii) below. There, it is shown that $\|b\|_2^2 = \|r\|_2^2 + \|Ay\|_2^2$, hence

$$1 = \left(\frac{\|r\|_2}{\|b\|_2}\right)^2 + \left(\frac{\|Ay\|_2}{\|b\|_2}\right)^2.$$

It follows that $\|r\|_2/\|b\|_2$ and $\|Ay\|_2/\|b\|_2$ behave like sine and cosine. Thus there is θ so that

$$1 = \sin\theta^2 + \cos\theta^2, \qquad \text{where} \quad \sin\theta = \frac{\|r\|_2}{\|b\|_2}, \quad \cos\theta = \frac{\|Ay\|_2}{\|b\|_2},$$

and θ can be interpreted as the angle between b and $\mathcal{R}(A)$. This allows us to bound the relative residual norm by

$$\frac{\|r\|_2}{\|A\|_2\|y\|_2} \leq \frac{\|r\|_2}{\|Ay\|_2} = \frac{\sin\theta}{\cos\theta} = \tan\theta.$$

This means if the angle between right-hand side and column space is large enough, then the least squares solution is sensitive to perturbations in the matrix, and this sensitivity is represented by $[\kappa_2(A)]^2$. ■

The matrix in Example 5.13 is representative of the situation in general. Least squares solutions are more sensitive to changes in the matrix when the right-hand side is too far from the column space. Below we present a bound for the relative error with regard to the perturbed solution z, because it is much easier to derive than a bound for the relative error with regard to the exact solution y.

Fact 5.14 (Matrix and Right-Hand Side Perturbation). Let $A, A+E \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = \operatorname{rank}(A+E) = n$, let y be the solution to $\min_x \|Ax-b\|_2$, and let $z \neq 0$ be the solution to $\min_x \|(A+E)x-(b+f)\|_2$. Then

$$\frac{\|z-y\|_2}{\|z\|_2} \leq \kappa_2(A)\left(\epsilon_A + \epsilon_f\right) + [\kappa_2(A)]^2\,\frac{\|s\|_2}{\|A\|_2\|z\|_2}\,\epsilon_A,$$

where

$$s = (A+E)z-(b+f), \qquad \epsilon_A = \frac{\|E\|_2}{\|A\|_2}, \qquad \epsilon_f = \frac{\|f\|_2}{\|A\|_2\|z\|_2}.$$

Proof. From Fact 5.9 follows that $y = A^\dagger b$ and $z = (A+E)^\dagger(b+f)$ are the unique solutions to the respective least squares problems. Applying Fact 5.7 to the perturbed least squares problem gives $(A+E)^* s = 0$, hence $A^* s = -E^* s$. Multiplying by $(A^*A)^{-1}$ and using $A^\dagger = (A^*A)^{-1}A^*$ from Remark 5.4 gives

$$-(A^*A)^{-1}E^*s = A^\dagger s = A^\dagger\left((A+E)z-(b+f)\right) = z-y+A^\dagger(Ez-f).$$

Solving for $z-y$ yields $z-y = -A^\dagger(Ez-f)-(A^*A)^{-1}E^*s$. Now take norms, and use the fact that $\|(A^*A)^{-1}\|_2 = \|A^\dagger\|_2^2$, see Exercise (viii) in Section 5.1, to obtain

$$\|z-y\|_2 \leq \|A^\dagger\|_2\,(\|E\|_2\|z\|_2 + \|f\|_2) + \|A^\dagger\|_2^2\,\|E\|_2\|s\|_2.$$

At last divide both sides of the inequality by $\|z\|_2$, and multiply and divide the right side by $\|A\|_2^2$. □

Remark 5.15.

- *If $E = 0$, then the bound in Fact* 5.14 *is identical to that in Fact* 5.11. *Therefore, the least squares solution is more sensitive to changes in the matrix than to changes in the right-hand side.*
- *The first term $\kappa_2(A)(\epsilon_A + \epsilon_f)$ in the above bound is the same as the perturbation bound for linear systems in Fact* 3.8. *It is because of the second term in Fact* 5.14 *that least squares problems are more sensitive than linear systems to perturbations in the matrix.*
- *We can interpret $\|s\|_2/(\|A\|_2\|z\|_2)$ as an approximation to the distance between perturbed right-hand side and perturbed matrix. From Exercise* (ii) *and Example* 5.13 *follows*

$$\frac{\|s\|_2}{\|A\|_2\|z\|_2} \leq \frac{\|s\|_2}{\|A+E\|_2\|z\|_2}(1+\epsilon_A) \leq \tan\tilde{\theta}\,(1+\epsilon_A),$$

  where $\tilde{\theta}$ is the angle between $b+f$ and $\mathcal{R}(A+E)$.
- *If most of the right-hand side lies in the column space, then the condition number of the least squares problem is $\kappa_2(A)$.*
  In particular, if $\frac{\|s\|_2}{\|A\|_2\|z\|_2} \approx \epsilon_A$, then the second term in the bound in Fact 5.14 *is about $[\kappa_2(A)]^2\epsilon_A^2$, and negligible for small enough ϵ_A.*

- *If the right-hand side is far away from the column space, then the condition number of the least squares problem is* $[\kappa_2(A)]^2$.
- *Therefore, the solution of the least squares is ill-conditioned in the normwise relative sense, if A is close to being rank deficient, i.e.,* $\kappa_2(A) \gg 1$, *or if the relative residual norm is large, i.e.,* $\|(A+E)z-(b+f)\|_2/(\|A\|_2\|z\|_2) \gg 0$.
- *If the perturbation does not change the column space so that* $\mathcal{R}(A+E) = \mathcal{R}(A)$, *then the least squares problem is no more sensitive than a linear system; see Exercise* 1 *below.*

Exercises

(i) Let $A \in \mathbb{C}^{m\times n}$ have orthonormal columns. Show that $\kappa_2(A) = 1$.

(ii) Under the assumptions of Fact 5.14 show that

$$\frac{\|s\|_2}{\|A+E\|_2\|z\|_2}(1-\epsilon_A) \le \frac{\|s\|_2}{\|A\|_2\|z\|_2} \le \frac{\|s\|_2}{\|A+E\|_2\|z\|_2}(1+\epsilon_A).$$

(iii) Let $A \in \mathbb{C}^{m\times n}$, and let y be a solution to the least squares problem $\min_x \|Ay-b\|_2$. Show:

$$\|b\|_2^2 = \|Ay-b\|_2^2 + \|Ay\|_2^2.$$

(iv) Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = n$. Show that the solution y of the least squares problem $\min_x \|Ax-b\|_2$ and the residual $r = b - Ay$ can be viewed as solutions to the linear system

$$\begin{pmatrix} I & A \\ A^* & 0 \end{pmatrix}\begin{pmatrix} r \\ y \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix},$$

and that

$$\begin{pmatrix} I & A \\ A^* & 0 \end{pmatrix}^{-1} = \begin{pmatrix} I - AA^\dagger & (A^\dagger)^* \\ A^\dagger & -(A^*A)^{-1} \end{pmatrix}.$$

(v) In addition to the assumptions of Exercise (ii), let $A + E \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A+E) = n$, and let z be the solution of the least squares problem $\min_x \|(A+E)x-(b+f)\|_2$ with residual $s = b + f - (A+E)z$. Show:

$$\begin{pmatrix} s-r \\ z-y \end{pmatrix} = \begin{pmatrix} I - AA^\dagger & (A^\dagger)^* \\ A^\dagger & -(A^*A)^{-1} \end{pmatrix}\begin{pmatrix} f - Ez \\ -E^*s \end{pmatrix}.$$

(vi) Let $A, A+E \in \mathbb{C}^{m\times n}$ and $\operatorname{rank}(A) = n$. Show: If $\|E\|_2\,\|A^\dagger\|_2 < 1$, then $\operatorname{rank}(A+E) = n$.

1. Matrices with the Same Column Space.
When the perturbed matrix has the same column space as the original matrix, then the least squares solution is less sensitive, and the error bound is the same as the one for linear systems in Fact 3.9.

Let $A, A+E \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = \operatorname{rank}(A+E) = n$. Let y be the solution to $\min_x \|Ax-b\|_2$, and let $z \neq 0$ be the solution to $\min_x \|(A+E)x-(b+f)\|_2$. Show: If $\mathcal{R}(A) = \mathcal{R}(A+E)$, then

$$\frac{\|z-y\|_2}{\|z\|_2} \le \kappa_2(A)\left(\epsilon_A + \epsilon_f\right), \qquad \text{where} \quad \epsilon_A = \frac{\|E\|_2}{\|A\|_2}, \quad \epsilon_f = \frac{\|f\|}{\|A\|_2\|z\|_2}.$$

2. Conditioning of the Least Squares Residual.
   This bound shows that the least squares residual is insensitive to changes in the right-hand side.
   Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = n$. Let y be the solution to $\min_x \|Ax-b\|_2$ with residual $r = Ay-b$, and let z be the solution to $\min_x \|Az-(b+f)\|_2$ with residual $s = Az-(b+f)$. Show:
   $$\|s-r\|_2 \le \|f\|_2.$$
3. Conditioning of the Least Squares Residual Norm.
   The following bound gives an indication of how sensitive the norm of the least squares residual may be to changes in the matrix and right-hand side.
   Let $A, A+E \in \mathbb{C}^{m\times n}$ so that $\operatorname{rank}(A) = \operatorname{rank}(A+E) = n$. Let y be the solution to $\min_x \|Ax-b\|_2$ with residual $r = Ay-b$, and let z be the solution to $\min_x \|(A+E)x-(b+f)\|_2$ with residual $s = (A+E)z-(b+f)$. Show: If $b \neq 0$, then
   $$\frac{\|s\|_2}{\|b\|_2} \le \frac{\|r\|_2}{\|b\|_2} + \kappa_2(A)\,\epsilon_A + \epsilon_b, \qquad \text{where} \quad \epsilon_A = \frac{\|E\|_2}{\|A\|_2}, \quad \epsilon_b = \frac{\|f\|_2}{\|b\|_2}.$$
4. This bound suggests that the error in the least squares solution depends on the error in the least squares residual.
   Under the conditions of Fact 5.14 show that
   $$\frac{\|z-y\|_2}{\|z\|_2} \le \kappa_2(A)\left[\frac{\|r-s\|_2}{\|A\|_2\|z\|_2} + \epsilon_A + \epsilon_f\right].$$
5. Given an approximate least squares solution z, this problem shows how to construct a least squares problem for which z is the exact solution.
   Let $z \neq 0$ be an approximate solution of the least squares problem $\min_x \|Ax-b\|_2$. Let $r_c = b - Az$ be the computable residual, h an arbitrary vector, and $F = -hh^\dagger A + (I - hh^\dagger)r_c z^\dagger$. Show that z is a least squares solution of $\min_x \|(A+F)x-b\|_2$.

5.3 Computation of Full Rank Least Squares Problems

We present two algorithms for computing the solution to a least squares problem with full column rank.

Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = n$ and an SVD

$$A = U\begin{pmatrix}\Sigma \\ 0\end{pmatrix}V^*, \qquad U = \begin{array}{cc} n & m-n \\ \big(U_n & U_{m-n}\big) \end{array},$$

where $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary, and $\Sigma \in \mathbb{C}^{n\times n}$ is a diagonal matrix with diagonal elements $\sigma_1 \geq \cdots \geq \sigma_n > 0$.

Fact 5.16 (Least Squares via SVD). Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$, let $b \in \mathbb{C}^m$, and let y be the solution to $\min_x \|Ax - b\|_2$. Then

$$y = V\Sigma^{-1}U_n^* b, \qquad \min_x \|Ax-b\|_2 = \|U_{m-n}^* b\|_2.$$

Proof. The expression for y follows from Fact 5.9. With regard to the residual,

$$\begin{aligned} Ay - b &= U\begin{pmatrix}\Sigma\\ 0\end{pmatrix} V^* V \Sigma^{-1} U_n^* b - b \\ &= U\left[\begin{pmatrix}U_n^* b\\ 0\end{pmatrix} - \begin{pmatrix}U_n^* b\\ U_{m-n}^* b\end{pmatrix}\right] \\ &= U\begin{pmatrix}0\\ -U_{m-n}^* b\end{pmatrix}. \end{aligned}$$

Therefore, $\min_x \|Ax-b\|_2 = \|Ay-b\|_2 = \|U_{m-n}^* b\|_2$. □

ALGORITHM 5.1. Least Squares Solution via SVD.

```
Input: Matrix $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$, vector $b \in \mathbb{C}^m$
Output: Solution $y$ of $\min_x \|Ax-b\|_2$, residual norm $\rho = \|Ay-b\|_2$
1. Compute an SVD $A = U\begin{pmatrix}\Sigma\\ 0\end{pmatrix}V^*$ where $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are
   unitary, and $\Sigma$ is diagonal.
2. Partition $U = \begin{pmatrix}U_n & U_{m-n}\end{pmatrix}$, where $U_n$ has $n$ columns.
3. Multiply $y \equiv V\Sigma^{-1}U_n^* b$.
4. Set $\rho \equiv \|U_{m-n}^* b\|_2$.
```

The least squares solution can also be computed from a QR factorization, which may be cheaper than an SVD. Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = n$ and a QR factorization

$$A = Q\begin{pmatrix}R\\ 0\end{pmatrix}, \qquad Q = \begin{array}{c} \begin{array}{cc} n & m-n \end{array} \\ \begin{pmatrix}Q_n & Q_{m-n}\end{pmatrix} \end{array},$$

where $Q \in \mathbb{C}^{m\times m}$ is unitary and $R \in \mathbb{C}^{n\times n}$ is upper triangular with positive diagonal elements.

Fact 5.17 (Least Squares Solution via QR). Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$, let $b \in \mathbb{C}^m$, and let y be the solution to $\min_x \|Ax-b\|_2$. Then

$$y = R^{-1}Q_n^* b, \qquad \min_x \|Ax-b\|_2 = \|Q_{m-n}^* b\|_2.$$

Proof. Fact 5.9 and Remark 5.4 imply for the solution

$$y = A^\dagger b = (A^*A)^{-1}A^* b = \begin{pmatrix}R^{-1} & 0\end{pmatrix} Q^* b = \begin{pmatrix}R^{-1} & 0\end{pmatrix}\begin{pmatrix}Q_n^* b\\ Q_{m-n}^* b\end{pmatrix} = R^{-1}Q_n^* b.$$

With regard to the residual,

$$Ay - b = Q\begin{pmatrix} R \\ 0 \end{pmatrix} R^{-1} Q_n^* b - b = Q\left[\begin{pmatrix} Q_n^* b \\ 0 \end{pmatrix} - \begin{pmatrix} Q_n^* b \\ Q_{m-n}^* b \end{pmatrix}\right] = Q\begin{pmatrix} 0 \\ -Q_{m-n}^* b \end{pmatrix}.$$

Therefore, $\min_x \|Ax - b\|_2 = \|Ay - b\|_2 = \|Q_{m-n}^* b\|_2$. □

ALGORITHM 5.2. Least Squares via QR.

Input: Matrix $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$, vector $b \in \mathbb{C}^m$
Output: Solution y of $\min_x \|Ax - b\|_2$, residual norm $\rho = \|Ay - b\|_2$

```
1. Factor $A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$ where $Q \in \mathbb{C}^{m\times m}$ is unitary and $R \in \mathbb{C}^{n\times n}$ is triangular.
2. Partition $Q = \begin{pmatrix} Q_n & Q_{m-n} \end{pmatrix}$, where $Q_n$ has $n$ columns.
3. Solve the triangular system $Ry = Q_n^* b$.
4. Set $\rho \equiv \|Q_{m-n}^* b\|_2$.
```

Exercises

1. Normal Equations.
   Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. Show: y is a solution of $\min_x \|Ax - b\|_2$ if and only if y is a solution of $A^*Ax = A^*b$.
2. Numerical Instability of Normal Equations.
   Show that the normal equations can be a numerically unstable method for solving the least squares problem.
   Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$, and let $A^*Ay = A^*b$ with $A^*b \neq 0$. Let z be a perturbed solution with $A^*Az = A^*b + f$. Show:

$$\frac{\|z - y\|_2}{\|y\|_2} \leq [\kappa_2(A)]^2 \frac{\|f\|_2}{\|A^*A\|_2 \|y\|_2}.$$

   That is, the numerical stability of the normal equations is always determined by $[\kappa_2(A)]^2$, even if the least squares residual is small.

6. Subspaces

We present properties of column, row, and null spaces; define operations on them; show how they are related to each other; and illustrate how they can be represented computationally.

Remark 6.1. *Column and null spaces of a matrix A are more than just ordinary sets.*

If $x, y \in \mathrm{Ker}(A)$, then $Ax = 0$ and $Ay = 0$. Hence $A(x+y) = 0$, and $A(\alpha x) = 0$ for $\alpha \in \mathbb{C}$. Therefore, $x + y \in \mathrm{Ker}(A)$ and $\alpha x \in \mathrm{Ker}(A)$.

Also, if $b, c \in \mathcal{R}(A)$, then $b = Ax$ and $c = Ay$ for some x and y. Hence $b + c = A(x+y)$ and $\alpha b = A(\alpha x)$ for $\alpha \in \mathbb{C}$. Therefore $b + c \in \mathcal{R}(A)$ and $\alpha b \in \mathcal{R}(A)$.

The above remark illustrates that we cannot "fall out of" the sets $\mathrm{Ker}(A)$ and $\mathcal{R}(A)$ by adding vectors from the set or by multiplying a vector from the set by a scalar. Sets with this property are called *subspaces*.

Definition 6.2 (Subspace). *A set $\mathcal{S} \subset \mathbb{C}^n$ is a* subspace of $\mathbb{C}^n$ *if $\mathcal{S}$ is closed under addition and scalar multiplication. That is, if $v, w \in \mathcal{S}$, then $v + w \in \mathcal{S}$ and $\alpha v \in \mathcal{S}$ for $\alpha \in \mathbb{C}$.*

A set $\mathcal{S} \subset \mathbb{R}^n$ is a subspace of $\mathbb{R}^n$ *if $v, w \in \mathcal{S}$ implies $v + w \in \mathcal{S}$ and $\alpha v \in \mathcal{S}$ for $\alpha \in \mathbb{R}$.*

A subspace is never empty. At the very least it contains the zero vector.

Example.

- Extreme cases: $\{0_{n\times 1}\}$ and $\mathbb{C}^n$ are subspaces of $\mathbb{C}^n$; and $\{0_{n\times 1}\}$ and $\mathbb{R}^n$ are subspaces of $\mathbb{R}^n$.
- If $A \in \mathbb{C}^{m\times n}$, then $\mathcal{R}(A)$ is a subspace of $\mathbb{C}^m$, and $\mathrm{Ker}(A)$ is a subspace of $\mathbb{C}^n$.

- If $A \in \mathbb{R}^{m\times n}$, then $\mathcal{R}(A)$ is a subspace of $\mathbb{R}^m$, and $\operatorname{Ker}(A)$ is a subspace of $\mathbb{R}^n$. ■

For simplicity, we will state subsequent results and definitions only for complex subspaces, but they hold also for real subspaces.

Exercises

(i) Let $\mathcal{S} \subset \mathbb{C}^3$ be the set of all vectors with first and third components equal to zero. Show that $\mathcal{S}$ is a subspace of $\mathbb{C}^3$.

(ii) Let $\mathcal{S} \subset \mathbb{C}^3$ be the set of all vectors with first component equal to 17. Show that $\mathcal{S}$ is not a subspace of $\mathbb{C}^3$.

(iii) Let $u \in \mathbb{C}^n$. Show that the set $\{x \in \mathbb{C}^n : x^*u = 0\}$ is a subspace of $\mathbb{C}^n$.

(iv) Let $u \in \mathbb{C}^n$ and $u \neq 0$. Show that the set $\{x \in \mathbb{C}^n : x^*u = 1\}$ is not a subspace of $\mathbb{C}^n$.

(v) Let $A \in \mathbb{C}^{m\times n}$. For which $b \in \mathbb{C}^m$ is the set of all solutions to $Ax = b$ a subspace of $\mathbb{C}^n$?

(vi) Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{m\times p}$. Show that the set $\left\{\begin{pmatrix} x \\ y \end{pmatrix} : Ax = By\right\}$ is a subspace of $\mathbb{C}^{n+p}$.

(vii) Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{m\times p}$. Show that the set

$$\{b : b = Ax + By \text{ for some } x \in \mathbb{C}^n, y \in \mathbb{C}^p\}$$

is a subspace of $\mathbb{C}^m$.

6.1 Spaces of Matrix Products

We give a rigorous proof of Fact 4.20, which shows that the four subspaces of a matrix are generated by singular vectors. In order to do so, we first relate column and null spaces of a product to those of the factors.

Fact 6.3 (Column Space and Null Space of a Product). Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times p}$. Then

1. $\mathcal{R}(AB) \subset \mathcal{R}(A)$. If B has linearly independent rows, then $\mathcal{R}(AB) = \mathcal{R}(A)$.
2. $\operatorname{Ker}(B) \subset \operatorname{Ker}(AB)$. If A has linearly independent columns, then $\operatorname{Ker}(B) = \operatorname{Ker}(AB)$.

Proof.

1. If $b \in \mathcal{R}(AB)$, then $b = ABx$ for some vector x. Setting $y = Bx$ implies $b = Ay$, which means that b is a linear combination of columns of A and $b \in \mathcal{R}(A)$. Thus $\mathcal{R}(AB) \subset \mathcal{R}(A)$.
   If B has linearly independent rows, then B has full row rank, and Fact 4.22 implies $\mathcal{R}(B) = \mathbb{C}^n$. Let $b \in \mathcal{R}(A)$ so that $b = Ax$ for some $x \in \mathbb{C}^n$. Since B has full row rank, there exists a $y \in \mathbb{C}^p$ so that $x = By$. Hence $b = ABy$ and $b \in \mathcal{R}(AB)$. Thus $\mathcal{R}(A) \subset \mathcal{R}(AB)$.

2. If $x \in \operatorname{Ker}(B)$, then $Bx = 0$. Hence $ABx = 0$ so that $x \in \operatorname{Ker}(AB)$. Thus $\operatorname{Ker}(B) \subset \operatorname{Ker}(AB)$.
   If A has linearly independent columns, then A has full column rank, and Fact 4.22 implies that $\operatorname{Ker}(A) = \{0_{n\times 1}\}$. Hence $ABx = 0$ implies that $Bx = 0$. Thus $\operatorname{Ker}(AB) \subset \operatorname{Ker}(B)$. □

Fact 6.3 implies in particular that $\operatorname{rank}(AB) = \operatorname{rank}(A)$ if B is nonsingular, and that $\operatorname{Ker}(AB) = \operatorname{Ker}(B)$ if A is nonsingular.

If we partition a nonsingular matrix and its inverse appropriately, then we can relate null spaces in the inverse to column spaces in the matrix proper.

Fact 6.4 (Partitioned Inverse). If $A \in \mathbb{C}^{n\times n}$ is nonsingular and

$$A = \begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{pmatrix} A_1 & A_2 \end{pmatrix} \end{array}, \qquad A^{-1} = \begin{array}{c} k \\ n-k \end{array} \begin{pmatrix} B_1^* \\ B_2^* \end{pmatrix},$$

then $\operatorname{Ker}(B_1^*) = \mathcal{R}(A_2)$ and $\operatorname{Ker}(B_2^*) = \mathcal{R}(A_1)$.

Proof. We will use the relations $B_1^* A_1 = I_k$ and $B_1^* A_2 = 0$, which follow from $A^{-1}A = I_n$.

If $b \in \mathcal{R}(A_2)$, then $b = A_2 x$ for some x and $B_1^* b = B_1^* A_2 x = 0$, so $b \in \operatorname{Ker}(B_1^*)$. Thus $\mathcal{R}(A_2) \subset \operatorname{Ker}(B_1^*)$.

If $b \in \operatorname{Ker}(B_1^*)$, then $B_1^* b = 0$. Write $b = AA^{-1}b = A_1 x_1 + A_2 x_2$, where $x_1 = B_1^* b$ and $x_2 = B_2^* b$. But $b \in \operatorname{Ker}(B_1^*)$ implies $x_1 = 0$, so $b = A_2 x_2$ and $b \in \mathcal{R}(A_2)$. Thus $\operatorname{Ker}(B_1^*) \subset \mathcal{R}(A_2)$.

The equality $\operatorname{Ker}(B_2^*) = \mathcal{R}(A_1)$ is shown in an analogous fashion. □

Example 6.5.

- Applying Fact 6.4 to the 3×3 identity matrix gives, for instance,

$$\operatorname{Ker}\begin{pmatrix} 1 & 0 & 0 \end{pmatrix} = \mathcal{R}\begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \operatorname{Ker}\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \mathcal{R}\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

- If $A \in \mathbb{C}^{n\times n}$ is unitary and $A = \begin{pmatrix} A_1 & A_2 \end{pmatrix}$, then

$$\operatorname{Ker}(A_1^*) = \mathcal{R}(A_2), \qquad \operatorname{Ker}(A_2^*) = \mathcal{R}(A_1). \qquad ■$$

Now we are ready to relate subspaces of a matrix to column spaces of singular vectors. Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = r$ and an SVD

$$A = U \begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix} V^*, \qquad U = \begin{array}{c} \begin{array}{cc} r & m-r \end{array} \\ \begin{pmatrix} U_r & U_{m-r} \end{pmatrix} \end{array}, \qquad V = \begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{pmatrix} V_r & V_{n-r} \end{pmatrix} \end{array},$$

where $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary, and Σ_r is a diagonal matrix with positive diagonal elements $\sigma_1 \geq \cdots \geq \sigma_r > 0$.

Fact 6.6 (Column Space). If $A \in \mathbb{C}^{m\times n}$ and $A \neq 0$, then $\mathcal{R}(A) = \mathcal{R}(U_r)$.

Proof. In the reduced SVD $A = U_r \Sigma_r V_r^*$, the matrices V_r^* and Σ_r have linearly independent rows. Fact 6.3 implies $\mathcal{R}(A) = \mathcal{R}(U_r)$. □

Fact 6.7 (Null Space). If $A \in \mathbb{C}^{m\times n}$ and $\operatorname{rank}(A) = r < n$, then $\mathcal{R}(V_{n-r}) = \operatorname{Ker}(A)$.

Proof. In the reduced SVD $A = U_r \Sigma_r V_r^*$, the matrices U_r and Σ_r have linearly independent columns. Fact 6.3 implies $\operatorname{Ker}(A) = \operatorname{Ker}(V_r^*)$. From Example 6.5 follows $\operatorname{Ker}(V_r^*) = \mathcal{R}(V_{n-r})$. □

The analogous statements for row space and left null space in Fact 4.20 can be proved by applying Facts 6.6 and 6.7 to A^*.

Exercises

(i) Let

$$A = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix},$$

where A_{11} and A_{22} are nonsingular. Show:

$$\mathcal{R}\begin{pmatrix} A_{11} \\ 0 \end{pmatrix} = \operatorname{Ker}\begin{pmatrix} 0 & A_{22}^{-1} \end{pmatrix}, \qquad \mathcal{R}\begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix} = \operatorname{Ker}\begin{pmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \end{pmatrix}.$$

(ii) Let $A \in \mathbb{C}^{n\times n}$ be idempotent. Show: $\mathcal{R}(I - A) = \operatorname{Ker}(A)$.

(iii) Let $A, B \in \mathbb{C}^{n\times n}$, and B idempotent. Show: $AB = A$ if and only if $\operatorname{Ker}(B) \subset \operatorname{Ker}(A)$.

(iv) Let $A \in \mathbb{C}^{n\times n}$ be idempotent. Show: $\mathcal{R}(A - AB)$ and $\mathcal{R}(AB - B)$ have only the zero vector in common.

1. QR Factorization.
Let $A \in \mathbb{C}^{m\times n}$ with $m \geq n$ have a QR decomposition

$$A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad Q = \begin{array}{cc} n & m-n \\ \big(Q_n & Q_{m-n} \big), \end{array}$$

where $Q \in \mathbb{C}^{m\times m}$ is unitary and $R \in \mathbb{C}^{n\times n}$ is upper triangular. Show:

$$\mathcal{R}(A) \subset \mathcal{R}(Q_n), \qquad \operatorname{Ker}(A) = \operatorname{Ker}(R), \qquad \mathcal{R}(Q_{m-n}) \subset \operatorname{Ker}(A^*).$$

If, in addition, $\operatorname{rank}(A) = n$, show: $\mathcal{R}(A) = \mathcal{R}(Q_n)$ and $\operatorname{Ker}(A^*) = \mathcal{R}(Q_{n-m})$.

6.2 Dimension

All subspaces of $\mathbb{C}^n$, except for $\{0\}$, have infinitely many elements. But some subspaces have more infinitely many elements than others. To quantify the "size" of a subspace we introduce the concept of dimension.

Definition 6.8 (Dimension). *Let $\mathcal{S}$ be a subspace of $\mathbb{C}^m$, and let $A \in \mathbb{C}^{m\times n}$ be a matrix so that $\mathcal{S} = \mathcal{R}(A)$. The* dimension *of $\mathcal{S}$ is* $\dim(\mathcal{S}) = \operatorname{rank}(A)$.

Example.

- $\dim(\mathbb{C}^n) = \dim(\mathbb{R}^n) = n$.
- $\dim(\{0_{n\times 1}\}) = 0$. ■

We show that the dimension of a subspace is unique and therefore well defined.

Fact 6.9 (Uniqueness of Dimension). Let $\mathcal{S}$ be a subspace of $\mathbb{C}^m$, and let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{m\times p}$ be matrices so that $\mathcal{S} = \mathcal{R}(A) = \mathcal{R}(B)$. Then $\operatorname{rank}(A) = \operatorname{rank}(B)$.

Proof. If $\mathcal{S} = \{0_{m\times 1}\}$, then $A = 0_{m\times n}$ and $B = 0_{m\times p}$ so that $\operatorname{rank}(A) = \operatorname{rank}(B) = 0$.

If $\mathcal{S} \neq \{0\}$, set $\alpha = \operatorname{rank}(A)$ and $\beta = \operatorname{rank}(B)$. Fact 6.6 implies that $\mathcal{R}(A) = \mathcal{R}(U_A)$, where U_A is an $m\times\alpha$ matrix of left singular vectors associated with the α nonzero singular values of A. Similarly, $\mathcal{R}(B) = \mathcal{R}(U_B)$ where U_B is an $m\times\beta$ matrix of left singular vectors associated with the β nonzero singular values of B.

Now suppose to the contrary that $\alpha > \beta$. Since $\mathcal{S} = \mathcal{R}(U_A) = \mathcal{R}(U_B)$, each of the α columns of U_A can be expressed as a linear combination of U_B. This means $U_A = U_B Y$, where Y is a $\beta\times\alpha$ matrix. Using the fact that U_A and U_B have orthonormal columns gives

$$I_\alpha = U_A^* U_A = Y^* U_B^* U_B Y = Y^* Y.$$

Fact 4.14 and Example 4.8 imply

$$\alpha = \operatorname{rank}(I_\alpha) = \operatorname{rank}(Y^*Y) = \operatorname{rank}(Y) \le \min\{\alpha,\beta\} = \beta.$$

Thus $\alpha \le \beta$, which contradicts the assumption $\alpha > \beta$. Therefore, we must have $\alpha = \beta$, so that the dimension of $\mathcal{S}$ is unique. □

The so-called dimension formula below is sometimes called the first part of the "fundamental theorem of linear algebra." The formula relates the dimensions of column and null spaces to the number of rows and columns.

Fact 6.10 (Dimension Formula). If $A \in \mathbb{C}^{m\times n}$, then

$$\operatorname{rank}(A) = \dim(\mathcal{R}(A)) = \dim(\mathcal{R}(A^*))$$

and

$$n = \operatorname{rank}(A) + \dim(\operatorname{Ker}(A)), \qquad m = \operatorname{rank}(A) + \dim(\operatorname{Ker}(A^*)).$$

Proof. The first set of equalities follows from $\operatorname{rank}(A) = \operatorname{rank}(A^*)$; see Fact 4.14. The remaining equalities follow from Fact 4.20, and from Fact 4.16 which implies that a matrix with k orthonormal columns has rank equal to k. □

Fact 6.10 implies that the column space and the row space of a matrix have the same dimension. Furthermore, for an $m \times n$ matrix, the null space has dimension $n - \operatorname{rank}(A)$, and the left null space has dimension $m - \operatorname{rank}(A)$.

Example 6.11.

- If $A \in \mathbb{C}^{n \times n}$ is nonsingular, then $\operatorname{rank}(A) = n$, and $\dim(\operatorname{Ker}(A)) = 0$.
- $\operatorname{rank}(0_{m \times n}) = 0$ and $\dim(\operatorname{Ker}(0_{m \times n})) = n$.
- If $u \in \mathbb{C}^m$, $v \in \mathbb{C}^n$, $u \neq 0$ and $v \neq 0$, then

$$\operatorname{rank}(uv^*) = 1, \quad \dim(\operatorname{Ker}(uv^*)) = n-1, \quad \dim(\operatorname{Ker}(vu^*)) = m-1. \quad \blacksquare$$

The following bound confirms that the dimension gives information about the "size" of a subspace: If a subspace $\mathcal{V}$ is contained in a subspace $\mathcal{W}$, then the dimension of $\mathcal{V}$ cannot exceed the dimension of $\mathcal{W}$ but it can be smaller.

Fact 6.12. If $\mathcal{V}$ and $\mathcal{W}$ are subspaces of $\mathbb{C}^n$, and $\mathcal{V} \subset \mathcal{W}$, then $\dim(\mathcal{V}) \leq \dim(\mathcal{W})$.

Proof. Let A and B be matrices so that $\mathcal{V} = \mathcal{R}(A)$ and $\mathcal{W} = \mathcal{R}(B)$. Since each element of $\mathcal{V}$ is also an element of $\mathcal{W}$, then in particular each column of A must be in $\mathcal{W}$. Thus there is a matrix X so that $A = BX$. Fact 6.13 implies that $\operatorname{rank}(A) \leq \operatorname{rank}(B)$. But from Fact 6.9 we know that $\operatorname{rank}(A) = \dim(\mathcal{V})$ and $\operatorname{rank}(B) = \dim(\mathcal{W})$. □

The rank of a product cannot exceed the rank of any factor.

Fact 6.13 (Rank of a Product). If $A \in \mathbb{C}^{m \times n}$ and $B \in \mathbb{C}^{m \times p}$, then

$$\operatorname{rank}(AB) \leq \min\{\operatorname{rank}(A), \operatorname{rank}(B)\}.$$

Proof. The inequality $\operatorname{rank}(AB) \leq \operatorname{rank}(A)$ follows from $\mathcal{R}(AB) \subset \mathcal{R}(A)$ in Fact 6.3, Fact 6.12, and $\operatorname{rank}(A) = \dim(\mathcal{R}(A))$. To derive $\operatorname{rank}(AB) \leq \operatorname{rank}(B)$, we use the fact that a matrix and its transpose have the same rank, see Fact 4.14, so that $\operatorname{rank}(AB) = \operatorname{rank}(B^*A^*)$. Now apply the first inequality. □

Exercises

(i) Let A be a 17×4 matrix with linearly independent columns. Determine the dimensions of the four spaces of A.

(ii) What can you say about the dimension of the left null space of a 25×7 matrix?

(iii) Let $A \in \mathbb{C}^{m \times n}$. Show: If $P \in \mathbb{C}^{m \times m}$ and $Q \in \mathbb{C}^{n \times n}$ are nonsingular, then $\operatorname{rank}(PAQ) = \operatorname{rank}(A)$.

(iv) Let $A \in \mathbb{C}^{n\times n}$. Show: $\operatorname{rank}(A^2 - I_n) \leq \min\{\operatorname{rank}(A+I_n), \operatorname{rank}(A-I_n)\}$.

(v) Let A and B be matrices with n columns, and let C and D be matrices with n rows. Show:

$$\operatorname{rank}\begin{pmatrix} AC & AD \\ BC & BD \end{pmatrix} \leq \min\left\{\operatorname{rank}\begin{pmatrix} A \\ B \end{pmatrix}, \begin{pmatrix} C & D \end{pmatrix}\right\}.$$

(vi) Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{m\times p}$. Show: $\operatorname{rank}(AA^* + BB^*) \leq \operatorname{rank}\begin{pmatrix} A & B \end{pmatrix}$.

(vii) Let $A, B \in \mathbb{C}^{m\times n}$. Show: $\operatorname{rank}(A+B) \leq \operatorname{rank}\begin{pmatrix} A & B \end{pmatrix}$.
Hint: Write $A+B$ as product.

(viii) Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times p}$. Show: If $AB = 0$, then $\operatorname{rank}(A) + \operatorname{rank}(B) \leq n$.

6.3 Intersection and Sum of Subspaces

We define operations on subspaces, so that we can relate column, row, and null spaces to those of submatrices.

Definition 6.14 (Intersection and Sum of Subspaces). *Let $\mathcal{V}$ and $\mathcal{W}$ be subspaces of $\mathbb{C}^n$. The* intersection *of two subspaces is defined as*

$$\mathcal{V} \cap \mathcal{W} = \{x : x \in \mathcal{V} \text{ and } x \in \mathcal{W}\},$$

and the sum *is defined as*

$$\mathcal{V} + \mathcal{W} = \{z :\ z = v + w,\ v \in \mathcal{V} \text{ and } w \in \mathcal{W}\}.$$

Example.

- Extreme cases: If $\mathcal{V}$ is a subspace of $\mathbb{C}^n$, then

$$\mathcal{V} \cap \{0_{n\times 1}\} = \{0_{n\times 1}\}, \qquad \mathcal{V} \cap \mathbb{C}^n = \mathcal{V}$$

and

$$\mathcal{V} + \{0_{n\times 1}\} = \mathcal{V}, \qquad \mathcal{V} + \mathbb{C}^n = \mathbb{C}^n.$$

- $$\mathcal{R}\begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \cap \mathcal{R}\begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathcal{R}\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

- $$\mathcal{R}\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} + \mathcal{R}\begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbb{C}^3.$$

- If $A \in \mathbb{C}^{n\times n}$ is nonsingular, and $A = \begin{pmatrix} A_1 & A_2 \end{pmatrix}$, then

$$\mathcal{R}(A_1) \cap \mathcal{R}(A_2) = \{0_{n\times 1}\}, \qquad \mathcal{R}(A_1) + \mathcal{R}(A_2) = \mathbb{C}^n. \qquad \blacksquare$$

Intersections and sums of subspaces produce again subspaces.

Fact 6.15 (Intersection and Sum of Subspaces). If $\mathcal{V}$ and $\mathcal{W}$ are subspaces of $\mathbb{C}^n$, then $\mathcal{V}\cap\mathcal{W}$ and $\mathcal{V}+\mathcal{W}$ are also subspaces of $\mathbb{C}^n$.

Proof. Let $x,y\in\mathcal{V}\cap\mathcal{W}$. Then $x,y\in\mathcal{V}$ and $x,y\in\mathcal{W}$. Since $\mathcal{V}$ and $\mathcal{W}$ are subspaces, this implies $x+y\in\mathcal{V}$ and $x+y\in\mathcal{W}$. Hence $x+y\in\mathcal{V}\cap\mathcal{W}$.

Let $x,y\in\mathcal{V}+\mathcal{W}$. Then $x=v_1+w_1$ and $y=v_2+w_2$ for some $v_1,v_2\in\mathcal{V}$ and $w_1,w_2\in\mathcal{W}$. Since $\mathcal{V}$ and $\mathcal{W}$ are subspaces, $v_1+v_2\in\mathcal{V}$ and $w_1+w_2\in\mathcal{W}$. Hence $x+y=(v_1+v_2)+(w_1+w_2)$ where $v_1+v_2\in\mathcal{V}$ and $w_1+w_2\in\mathcal{W}$.

The proofs for αx where $\alpha\in\mathbb{C}$ are analogous. □

With the sum of subspaces, we can express the column space of a matrix in terms of column spaces of subsets of columns.

Fact 6.16 (Sum of Column Spaces). If $A\in\mathbb{C}^{m\times n}$ and $B\in\mathbb{C}^{m\times p}$, then

$$\mathcal{R}\begin{pmatrix}A & B\end{pmatrix}=\mathcal{R}(A)+\mathcal{R}(B).$$

Proof. From Definition 4.17 of a column space, and second view of matrix times column vector in Section 1.5 we obtain the following equivalences:

$$b\in\mathcal{R}\begin{pmatrix}A & B\end{pmatrix}\Longleftrightarrow b=\begin{pmatrix}A & B\end{pmatrix}x=Ax_1+Bx_2 \text{ for some } x=\begin{pmatrix}x_1\\x_2\end{pmatrix}\in\mathbb{C}^{n+p}$$

$$\Longleftrightarrow b=v+w \text{ where } v=Ax_1\in\mathcal{R}(A) \text{ and } w=Bx_2\in\mathcal{R}(B).$$ □

Example.

- Let $A\in\mathbb{C}^{m\times n}$ and $B\in\mathbb{C}^{m\times p}$. If $\mathcal{R}(B)\subset\mathcal{R}(A)$, then
$$\mathcal{R}\begin{pmatrix}A & B\end{pmatrix}=\mathcal{R}(A)+\mathcal{R}(B)=\mathcal{R}(A).$$
- If $A\in\mathbb{C}^{m\times n}$, then
$$\mathcal{R}\begin{pmatrix}A & I_m\end{pmatrix}=\mathcal{R}(A)+\mathcal{R}(I_m)=\mathcal{R}(A)+\mathbb{C}^m=\mathbb{C}^m.$$ ■

With the help of sums of subspaces, we can now show that the row space and null space of an $m\times n$ matrix together make up all of $\mathbb{C}^n$, while the column space and left null space make up $\mathbb{C}^m$.

Fact 6.17 (Subspaces of a Matrix are Sums). If $A\in\mathbb{C}^{m\times n}$, then

$$\mathbb{C}^n=\mathcal{R}(A^*)+\mathrm{Ker}(A),\qquad \mathbb{C}^m=\mathcal{R}(A)+\mathrm{Ker}(A^*).$$

Proof. Facts 4.20 and 6.16 imply

$$\mathcal{R}(A^*)+\mathrm{Ker}(A)=\mathcal{R}(V_r)+\mathcal{R}(V_{n-r})=\mathcal{R}\begin{pmatrix}V_r & V_{n-r}\end{pmatrix}=\mathcal{R}(V)=\mathbb{C}^n$$

and

$$\mathcal{R}(A)+\mathrm{Ker}(A^*)=\mathcal{R}(U_r)+\mathcal{R}(U_{m-r})=\mathcal{R}\begin{pmatrix}U_r & U_{m-r}\end{pmatrix}=\mathcal{R}(U)=\mathbb{C}^m.$$ □

Fact 6.18 (Intersection of Null Spaces). If $A\in\mathbb{C}^{m\times n}$ and $B\in\mathbb{C}^{p\times n}$, then $\mathrm{Ker}\begin{pmatrix}A\\B\end{pmatrix}=\mathrm{Ker}(A)\cap\mathrm{Ker}(B)$.

Proof. From Definition 4.17 of a null space, and first view of matrix times column vector in Section 1.5 we obtain the following equivalences:

$$x \in \text{Ker}\begin{pmatrix} A \\ B \end{pmatrix} \iff 0 = \begin{pmatrix} A \\ B \end{pmatrix} x = \begin{pmatrix} Ax \\ Bx \end{pmatrix} \iff Ax = 0 \text{ and } Bx = 0$$
$$\iff x \in \text{Ker}(A) \text{ and } x \in \text{Ker}(B) \iff x \in \text{Ker}(A) \cap \text{Ker}(B). \qquad \square$$

Example.

- If $A \in \mathbb{C}^{m\times n}$, then
$$\text{Ker}\begin{pmatrix} A \\ 0_{k\times n} \end{pmatrix} = \text{Ker}(A) \cap \text{Ker}(0_{k\times n}) = \text{Ker}(A) \cap \mathbb{C}^n = \text{Ker}(A).$$
- If $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times n}$ is nonsingular, then
$$\text{Ker}\begin{pmatrix} A \\ B \end{pmatrix} = \text{Ker}(A) \cap \text{Ker}(B) = \text{Ker}(A) \cap \{0_{n\times 1}\} = \{0_{n\times 1}\}. \qquad \blacksquare$$

The rank of a submatrix cannot exceed the rank of a matrix. We already used this for proving the optimality of the SVD in Fact 4.13.

Fact 6.19 (Rank of Submatrix). If B is a submatrix of $A \in \mathbb{C}^{m\times n}$, then $\text{rank}(B) \le \text{rank}(A)$.

Proof. Let $P \in \mathbb{C}^{m\times m}$ and $Q \in \mathbb{C}^{n\times n}$ be permutation matrices that move the elements of B into the top left corner of the matrix,

$$PAQ = \begin{pmatrix} B & A_{12} \\ A_{21} & A_{22} \end{pmatrix}.$$

Since the permutation matrices P and Q can only affect singular vectors but not singular values, $\text{rank}(A) = \text{rank}(PAQ)$; see also Exercise (i) in Section 4.2.

We relate $\text{rank}(B)$ to $\text{rank}(PAQ)$ by gradually isolating B with the help of Fact 6.18. Partition

$$PAQ = \begin{pmatrix} C \\ D \end{pmatrix}, \qquad \text{where} \quad C = \begin{pmatrix} B & A_{12} \end{pmatrix}, \quad D = \begin{pmatrix} A_{21} & A_{22} \end{pmatrix}.$$

Fact 6.18 implies

$$\text{Ker}(PAQ) = \text{Ker}\begin{pmatrix} C \\ D \end{pmatrix} = \text{Ker}(C) \cap \text{Ker}(D) \subset \text{Ker}(C).$$

Hence $\text{Ker}(PAQ) \subset \text{Ker}(C)$. From Fact 6.12 follows $\dim(\text{Ker}(PAQ)) \le \dim(\text{Ker}(C))$. We use the dimension formula in Fact 6.10 to relate the dimension of $\text{Ker}(C)$ to $\text{rank}(C)$,

$$\text{rank}(A) = \text{rank}(PAQ) = n - \dim(\text{Ker}(PAQ)) \ge n - \dim(\text{Ker}(C)) = \text{rank}(C).$$

Thus $\text{rank}(C) \le \text{rank}(A)$.

In order to show that $\text{rank}(B) \le \text{rank}(C)$, we repeat the above argument for C^* and use the fact that a matrix has the same rank as its transpose; see Fact 4.14. $\square$

Exercises

(i) Solution of Linear Systems.
Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. Show: $Ax = b$ has a solution if and only if $\mathcal{R}\begin{pmatrix}A & b\end{pmatrix} = \mathcal{R}(A)$.

(ii) Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{m\times p}$. Show:

$$\mathbb{C}^m = \mathcal{R}(A) + \mathcal{R}(B) + \operatorname{Ker}(A^*) \cap \operatorname{Ker}(B^*).$$

(iii) Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{m\times p}$. Show that $\mathcal{R}(A) \cap \mathcal{R}(B)$ and $\operatorname{Ker}\begin{pmatrix}A & B\end{pmatrix}$ have the same number of elements.

(iv) Rank of Block Diagonal Matrix.
Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{p\times q}$. Show:

$$\operatorname{rank}\begin{pmatrix}A & 0\\ 0 & B\end{pmatrix} = \operatorname{rank}(A) + \operatorname{rank}(B).$$

(v) Rank of Block Triangular Matrix.
Let $A \in \mathbb{C}^{m\times n}$ and

$$A = \begin{pmatrix}A_{11} & A_{12}\\ 0 & A_{22}\end{pmatrix},$$

where A_{11} is nonsingular. Show: $\operatorname{rank}(A) \le \operatorname{rank}(A_{11}) + \operatorname{rank}(A_{22})$.
Give an example to illustrate that this inequality may not hold anymore when A_{11} is singular or not square.

(vi) Rank of Schur Complement.
Let $A \in \mathbb{C}^{m\times n}$ be partitioned so that

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

where A_{11} is nonsingular. For $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ show that

$$\operatorname{rank}(S) \le \operatorname{rank}(A) \le \operatorname{rank}(A_{11}) + \operatorname{rank}(S).$$

(vii) Let $A, B \in \mathbb{C}^{n\times n}$. Show:

$$\operatorname{rank}(AB) \ge \operatorname{rank}(A) + \operatorname{rank}(B) - n.$$

(viii) Properties of Intersections and Sums.
Intersections of subspaces can produce “smaller” subspaces, while sums can produce “larger” subspaces.
Let $\mathcal{V}$ and $\mathcal{W}$ be subspaces of $\mathbb{C}^n$. Show:

(a) $\mathcal{V} \cap \mathcal{W} \subset \mathcal{V}$, and $V \cap \mathcal{W} \subset \mathcal{W}$.

(b) $\mathcal{V} \cap \mathcal{W} = \mathcal{V}$ if and only if $\mathcal{V} \subset \mathcal{W}$.

(c) $\mathcal{V} \subset \mathcal{V} + \mathcal{W}$, and $\mathcal{W} \subset \mathcal{V} + \mathcal{W}$.

(d) $\mathcal{V} + \mathcal{W} = \mathcal{V}$ if and only if $\mathcal{W} \subset \mathcal{V}$.

1. Let $A, B \in \mathbb{C}^{n\times n}$ be idempotent and $AB = BA$. Show:

(a) $\mathcal{R}(AB) = \mathcal{R}(A) \cap \mathcal{R}(B)$.

(b) $\operatorname{Ker}(AB) = \operatorname{Ker}(A) + \operatorname{Ker}(B)$.

(c) If also $AB = 0$, then $A + B$ is idempotent.

6.4 Direct Sums and Orthogonal Subspaces

It turns out that the column and null space pairs in Fact 6.17 have only the minimal number of elements in common. Sums of subspaces that have minimal overlap are called direct sums.

Definition 6.20 (Direct Sum). *Let $\mathcal{V}$ and $\mathcal{W}$ be subspaces in $\mathbb{C}^n$ with $\mathcal{S} = \mathcal{V} + \mathcal{W}$. If $\mathcal{V} \cap \mathcal{W} = \{0\}$, then $\mathcal{S}$ is a* direct sum *of $\mathcal{V}$ and $\mathcal{W}$, and we write $\mathcal{S} = \mathcal{V} \oplus \mathcal{W}$. Subspaces $\mathcal{V}$ and $\mathcal{W}$ are also called* complementary subspaces.

Example.

- $$\mathcal{R}\begin{pmatrix}1\\0\\0\end{pmatrix} \oplus \begin{pmatrix}0\\1\\0\end{pmatrix} \oplus \begin{pmatrix}0\\0\\1\end{pmatrix} = \mathbb{C}^3.$$
- $$\mathcal{R}\begin{pmatrix}1 & 2\\-1 & -2\end{pmatrix} \oplus \mathcal{R}\begin{pmatrix}1 & 2 & 4\\-3 & -6 & -12\end{pmatrix} = \mathbb{C}^2.$$ ■

The example above illustrates that the columns of the identity matrix I_n form a direct sum of $\mathbb{C}^n$. In general, linearly independent columns form direct sums. That is, in a full column rank matrix, the columns form a direct sum of the column space.

Fact 6.21 (Full Column Rank Matrices). Let $A \in \mathbb{C}^{m\times n}$ with $A = \begin{pmatrix}A_1 & A_2\end{pmatrix}$. If $\operatorname{rank}(A) = n$, then $\mathcal{R}(A) = \mathcal{R}(A_1) \oplus \mathcal{R}(A_2)$.

Proof. Fact 6.16 implies $\mathcal{R}(A) = \mathcal{R}(A_1) + \mathcal{R}(A_2)$. To show that $\mathcal{R}(A_1) \cap \mathcal{R}(A_2) = \{0\}$, suppose that $b \in \mathcal{R}(A_1) \cap \mathcal{R}(A_2)$. Then $b = A_1x_1 = A_2x_2$, and

$$0 = A_1x_1 - A_2x_2 = \begin{pmatrix}A_1 & A_2\end{pmatrix}\begin{pmatrix}x_1\\-x_2\end{pmatrix}.$$

Since A has full column rank, Fact 4.22 implies $x_1 = 0$ and $x_2 = 0$, hence $b = 0$. □

We are ready to show that the row space and null space of a matrix have only minimal overlap, and so do the column space and left null space. In other words, for an $m \times n$ matrix, row space and null space form a direct sum of $\mathbb{C}^n$, while column space and left null space form a direct sum of $\mathbb{C}^m$.

Fact 6.22 (Subspaces of a Matrix are Direct Sums). If $A \in \mathbb{C}^{m\times n}$, then

$$\mathbb{C}^n = \mathcal{R}(A^*) \oplus \operatorname{Ker}(A), \qquad \mathbb{C}^m = \mathcal{R}(A) \oplus \operatorname{Ker}(A^*).$$

Proof. The proof of Fact 6.17 shows that

$$\mathcal{R}(A^*) + \operatorname{Ker}(A) = \mathcal{R}(V_r) + \mathcal{R}(V_{n-r}), \qquad \mathcal{R}(A) + \operatorname{Ker}(A^*) = \mathcal{R}(U_r) + \mathcal{R}(U_{m-r}).$$

Since the unitary matrices $V = \begin{pmatrix} V_r & V_{n-r} \end{pmatrix}$ and $U = \begin{pmatrix} U_r & U_{m-r} \end{pmatrix}$ have full column rank, Fact 6.21 implies $\mathcal{R}(A^*) \cap \operatorname{Ker}(A) = \{0_{n\times 1}\}$ and $\mathcal{R}(A) \cap \operatorname{Ker}(A^*) = \{0_{m\times 1}\}$. □

It is tempting to think that for a given subspace $\mathcal{V} \neq \{0_{n\times 1}\}$ of $\mathbb{C}^n$, there is only one way to complement $\mathcal{V}$ and fill up all of $\mathbb{C}^n$. However, that is not true—there are infinitely many complementary subspaces. Below is a very simple example.

Remark 6.23 (Complementary Subspaces Are Not Unique). *Let*

$$\mathcal{V} = \mathcal{R}\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad \mathcal{W} = \mathcal{R}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}.$$

Then for any $\beta \neq 0$, $\mathcal{V} \oplus \mathcal{W} = \mathbb{C}^2$.

This is because for $\beta \neq 0$ the matrix

$$A = \begin{pmatrix} 1 & \alpha \\ 0 & \beta \end{pmatrix}$$

is nonsingular, hence $\mathbb{C}^2 = \mathcal{R}(A) = \mathcal{V} + \mathcal{W}$. Since A has full column rank, Fact 6.21 *implies $\mathcal{V} \cap \mathcal{W} = \{0\}$.*

There is a particular type of direct sum, where the two subspaces are as "far apart" as possible.

Definition 6.24 (Orthogonal Subspaces). *Let $\mathcal{V}$ and $\mathcal{W}$ be subspaces in $\mathbb{C}^n$ with $\mathcal{V} + \mathcal{W} = \mathbb{C}^n$. If $v^*w = 0$ for all $v \in \mathcal{V}$ and $w \in \mathcal{W}$, then the spaces $\mathcal{V}$ and $\mathcal{W}$ are* orthogonal subspaces. *We write $\mathcal{V} = \mathcal{W}^{\perp}$, or equivalently, $\mathcal{W} = \mathcal{V}^{\perp}$.*

In particular, $(\mathbb{C}^n)^{\perp} = \{0_{n\times 1}\}$ and $\{0_{n\times 1}\}^{\perp} = \mathbb{C}^n$.

Below is an example of a matrix that produces orthogonal subspaces; it is a generalization of a unitary matrix.

Fact 6.25. Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and $A = \begin{pmatrix} A_1 & A_2 \end{pmatrix}$. If $A_1^* A_2 = 0$, then $\mathcal{R}(A_2)^{\perp} = \mathcal{R}(A_1)$.

Proof. Since A has full column rank, Fact 6.16 implies $\mathcal{R}(A_1) + \mathcal{R}(A_2) = \mathcal{R}(A) = \mathbb{C}^n$. From $A_1^* A_2 = 0$ follows $0 = x^* A_1^* A_2 y = (A_1 x)^*(A_2 y)$. With $v = A_1 x$ and $w = A_2 y$ we conclude that $v^* w = 0$ for all $v \in \mathcal{R}(A_1)$ and $w \in \mathcal{R}(A_2)$. □

Now we come to what is sometimes referred to as the second part of the "fundamental theorem of linear algebra." It says that any matrix has two pairs of orthogonal subspaces: column space and left null space are orthogonal subspaces, and row space and null space are orthogonal subspaces.

Fact 6.26 (Orthogonal Subspaces of a Matrix). If $A \in \mathbb{C}^{m\times n}$, then

$$\operatorname{Ker}(A) = \mathcal{R}(A^*)^{\perp}, \qquad \operatorname{Ker}(A^*) = \mathcal{R}(A)^{\perp}.$$

Proof. Facts 6.17 and 4.20 imply

$$\mathbb{C}^n = \mathcal{R}(A^*) + \operatorname{Ker}(A), \qquad \mathcal{R}(A^*) = \mathcal{R}(V_r), \quad \operatorname{Ker}(A) = \mathcal{R}(V_{n-r})$$

and

$$\mathbb{C}^m = \mathcal{R}(A) + \operatorname{Ker}(A^*), \qquad \mathcal{R}(A) = \mathcal{R}(U_r), \quad \operatorname{Ker}(A^*) = \mathcal{R}(U_{m-r}).$$

Now apply Fact 6.25 to the unitary matrices $V = \begin{pmatrix} V_r & V_{n-r} \end{pmatrix}$ and $U = \begin{pmatrix} U_r & U_{m-r} \end{pmatrix}$. □

Exercises

(i) Show: If $A \in \mathbb{C}^{n\times n}$ is Hermitian, then $\mathbb{C}^n = \mathcal{R}(A) \oplus \operatorname{Ker}(A)$.

(ii) Show: If $A \in \mathbb{C}^{n\times n}$ is idempotent, then

$$\mathcal{R}(A)^\perp = \mathcal{R}(I_n - A^*), \qquad \mathcal{R}(A^*)^\perp = \mathcal{R}(I_n - A).$$

(iii) Show: If $A \in \mathbb{C}^{n\times n}$ is idempotent, then $\mathbb{C}^n = \mathcal{R}(A) \oplus \mathcal{R}(I_n - A)$.

(iv) Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = n$ and a QR factorization $A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}$, where $Q = \begin{pmatrix} Q_n & Q_{m-n} \end{pmatrix}$ is unitary. Show: $\mathcal{R}(A)^\perp = \mathcal{R}(Q_{m-n})$.

(v) Let $A \in \mathbb{C}^{m\times n}$ have $\operatorname{rank}(A) = n$, and let y be the solution of the least squares problem $\min_x \|Ax - b\|_2$. Show:

$$\mathcal{R}\begin{pmatrix} A & b \end{pmatrix} = \mathcal{R}(A) \oplus \mathcal{R}(Ay - b).$$

(vi) Let $A \in \mathbb{C}^{n\times n}$ be a matrix all of whose rows sum to zero. Show: $\mathcal{R}(e) \subset \mathcal{R}(A^*)^\perp$, where e is the $n \times 1$ vector of all ones.

(vii) Orthogonal Subspaces Form Direct Sums.
Let $\mathcal{V}$ and $\mathcal{W}$ be subspaces of $\mathbb{C}^n$ so that $\mathcal{W} = \mathcal{V}^\perp$. Show: $\mathcal{V} \oplus \mathcal{W} = \mathbb{C}^n$.

(viii) Direct sums of subspaces produce unique representations in the following sense.
Let $\mathcal{S}$ be a subspace of $\mathbb{C}^m$ and $\mathcal{S} = \mathcal{V} + \mathcal{W}$. Show: $\mathcal{S} = \mathcal{V} \oplus \mathcal{W}$ if and only if for every $b \in \mathcal{S}$ there exist unique vectors $v \in \mathcal{V}$ and $w \in \mathcal{W}$ such that $b = v + w$.

(ix) Normal Matrices.
Show: If $A \in \mathbb{C}^n$ is normal, i.e., $A^*A = AA^*$, then $\operatorname{Ker}(A) = \mathcal{R}(A)^\perp$.

1. Let $A \in \mathbb{C}^{n\times n}$ with

$$X^{-1}AX = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix},$$

where A_1 and A_2 are square. Show: If A_1 is nonsingular and A_2 is nilpotent, then for k large enough we have $\mathbb{C}^n = \mathcal{R}(A^k) \oplus \operatorname{Ker}(A^k)$.

2. Properties of Orthogonal Subspaces.
Let $\mathcal{V}$ and $\mathcal{W}$ be subspaces of $\mathbb{C}^n$. Show:

(a) $(\mathcal{V}^\perp) = \mathcal{V}$.

(b) If $\mathcal{V} \subset \mathcal{W}$, then $\mathcal{W}^\perp \subset \mathcal{V}^\perp$.

(c) $(\mathcal{V}+\mathcal{W})^{\perp}=\mathcal{V}^{\perp}\cap\mathcal{W}^{\perp}$.

(d) $(\mathcal{V}\cap\mathcal{W})^{\perp}=\mathcal{V}^{\perp}+\mathcal{W}^{\perp}$.

6.5 Bases

A basis makes it possible to represent the infinitely many vectors of a subspace by just a finite number. The elements of a basis are much like members of parliament, with a few representatives standing for large constituency. A basis contains just enough vectors to capture the whole space, but sufficiently few to avoid redundancy.

Definition 6.27 (Basis). *The columns of a matrix $W\in\mathbb{C}^{m\times n}$ represent a* basis *for a subspace $\mathcal{S}$ of $\mathbb{C}^m$ if*

B1: $\operatorname{Ker}(W)=\{0\}$, *i.e.*, $\operatorname{rank}(W)=n$,

B2: $\mathcal{R}(W)=\mathcal{S}$.

If, in addition, W has orthonormal columns, then the columns of W represent an orthonormal basis *for $\mathcal{S}$.*

Example.

- The columns of a nonsingular matrix $A\in\mathbb{C}^{n\times n}$ represent a basis for $\mathbb{C}^n$. If A is unitary, then the columns of A represent an orthonormal basis for $\mathbb{C}^n$.
- Let $A\in\mathbb{C}^{n\times n}$ be nonsingular, and

$$A=\begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{pmatrix} A_1 & A_2 \end{pmatrix} \end{array}, \qquad A^{-1}=\begin{array}{c} k \\ n-k \end{array}\begin{pmatrix} B_1^* \\ B_2^* \end{pmatrix}.$$

  Then the columns of A_1 represent a basis for $\operatorname{Ker}(B_2^*)$, and the columns of A_2 represent a basis for $\operatorname{Ker}(B_1^*)$.
  This follows from Fact 6.4.
- Let $U\in\mathbb{C}^{n\times n}$ be unitary and $U=\begin{pmatrix} U_1 & U_2 \end{pmatrix}$. Then the columns of U_1 represent an orthonormal basis for $\operatorname{Ker}(U_2^*)$, and the columns of U_2 represent an orthonormal basis for $\operatorname{Ker}(U_1^*)$. $\blacksquare$

Remark 6.28. *Let $\mathcal{V}$ be a subspace of $\mathbb{C}^m$. If $\mathcal{V}\neq\{0_{m\times 1}\}$, then there are infinitely many different bases for $\mathcal{V}$. But all bases have the same number of vectors; this follows from Fact* 6.9.

The singular vectors furnish orthonormal bases for all four subspaces of a matrix. Let $A\in\mathbb{C}^{m\times n}$ have $\operatorname{rank}(A)=r$ and an SVD

$$A=U\begin{pmatrix} \Sigma_r & 0 \\ 0 & 0 \end{pmatrix}V^*, \qquad U=\begin{array}{c} \begin{array}{cc} r & m-r \end{array} \\ \begin{pmatrix} U_r & U_{m-r} \end{pmatrix} \end{array}, \qquad V=\begin{array}{c} \begin{array}{cc} r & n-r \end{array} \\ \begin{pmatrix} V_r & V_{n-r} \end{pmatrix} \end{array},$$

where $U\in\mathbb{C}^{m\times m}$ and $V\in\mathbb{C}^{n\times n}$ are unitary, and Σ_r is a diagonal matrix with positive diagonal elements $\sigma_1\geq\cdots\geq\sigma_r>0$.

Fact 6.29 (Orthonormal Bases for Spaces of a Matrix). Let $A \in \mathbb{C}^{m\times n}$.

- If $A \neq 0$, then the columns of U_r represent an orthonormal basis for $\mathcal{R}(A)$, and the columns of V_r represent an orthonormal basis for $\mathcal{R}(A^*)$.
- If $r < n$, then the columns of V_{n-r} represent an orthonormal basis for $\operatorname{Ker}(A)$.
- If $r < m$, then the columns of U_{m-r} represent an orthonormal basis for $\operatorname{Ker}(A^*)$.

Proof. This follows from applying Facts 6.6 and 6.7 to A and to A^*. □

Why Orthonormal Bases? Orthonormal bases are attractive because they are easy to work with, and they do not amplify errors. For instance, if x is the solution of the linear system $Ax = b$ where A has orthonormal columns, then $x = A^*b$ can be determined with only a matrix vector multiplication. The bound below justifies that orthonormal bases do not amplify errors.

Fact 6.30. Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$, and $b \in \mathbb{C}^m$ with $Ax = b$ and $b \neq 0$. Let z be an approximate solution with residual $r = Az - b$. Then

$$\frac{\|z-x\|_2}{\|x\|_2} \leq \kappa_2(A)\frac{\|r\|_2}{\|A\|_2\|x\|_2},$$

where $\kappa_2(A) = \|A^\dagger\|_2\|A\|_2$.

If A has orthonormal columns, then $\kappa_2(A) = 1$.

Proof. This follows from Fact 5.11. If A has orthonormal columns, then all singular values are equal to one, see Fact 4.16, so that $\kappa_2(A) = \sigma_1/\sigma_n = 1$. □

Exercises

(i) Let $u \in \mathbb{C}^m$ and $v \in \mathbb{C}^n$ with $u \neq 0$ and $v \neq 0$. Determine an orthonormal basis for $\mathcal{R}(uv^*)$.

(ii) Let $A \in \mathbb{C}^{m\times n}$ be nonsingular and $B \in \mathbb{C}^{m\times p}$. Prove: The columns of $\begin{pmatrix} -A^{-1}B \\ I_p \end{pmatrix}$ represent a basis for $\operatorname{Ker}\begin{pmatrix} A & B \end{pmatrix}$.

(iii) Let $A \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(A) = n$ have a QR decomposition

$$A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}, \qquad Q = \begin{array}{cc} n & m-n \\ \big(Q_n & Q_{m-n} \big) \end{array},$$

where $Q \in \mathbb{C}^{m\times m}$ is unitary and $R \in \mathbb{C}^{n\times n}$ is upper triangular. Show: The columns of Q_n represent an orthonormal basis for $\mathcal{R}(A)$, and the columns of Q_{m-n} represent an orthonormal basis for $\operatorname{Ker}(A^*)$.

Index

(Page numbers set in **bold** type indicate the definition of an entry.)