3 Systems of Linear Equations

In this chapter, we will be concerned with the following problem: Given the matrix $\mathrm{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ and the vector $\boldsymbol{f} = [f_i] \in \mathbb{C}^n$, find $\boldsymbol{x} = [x_i] \in \mathbb{C}^n$ such that

$$\mathrm{A}\boldsymbol{x} = \boldsymbol{f}. \tag{3.1}$$

Of course, this is shorthand for the following system of linear equations:

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = f_1, \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n = f_2, \\ \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,n}x_n = f_n. \end{cases}$$

We call A the *coefficient matrix*. First of all, we need to make sure that a solution exists and is unique. The following result is nothing but a recapitulation of statements that the reader will have encountered before.

Theorem. *The system of linear equations (3.1) has a unique solution if and only if* $\det(\mathrm{A}) \neq 0$ *if and only if* $\mathrm{A}\boldsymbol{x} = \boldsymbol{0}$ *has only the trivial solution if and only if* A^{-1} *exists.*

This theorem gives necessary and sufficient conditions for the inverse of the coefficient matrix, A^{-1}, to exist. If it does, then the solution is $\boldsymbol{x} = \mathrm{A}^{-1}\boldsymbol{f}$, but it is not usually computationally tractable to calculate the inverse, as we discuss later. Thus, we will look for ways of computing $\boldsymbol{x}$ without first explicitly finding A^{-1}. Most of us, in a course on linear algebra, learned of a method called Gaussian elimination. This simple, powerful, and sometimes mysterious technique will be the basis of most of what we do in this chapter. In particular, Gaussian elimination is the foundation for some well-known factorization techniques, such as the LU decomposition method and the Cholesky factorization method for positive definite matrices.

Why is Gaussian elimination mysterious? The reason for this is that one of the big open questions of numerical linear algebra is

Why is Gaussian elimination (with partial pivoting) usually so numerically stable in practice?

Except for certain classes of truly pathological matrices, our best generic estimates for the growth of roundoff error in the algorithm tend to be overly pessimistic. In other words, this simple algorithm usually performs much better than the worst-case scenario for the average matrix. Gaussian elimination is much more reliable

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

than numerical analysts would expect. A discussion of this topic is beyond the scope of our text, but see [34, 96].

3.1 Solution of Simple Systems

Before we describe the general case, in this section, we develop some algorithms to find the solution to (3.1) for some simple cases, all of which avoid the direct construction of A^{-1}.

3.1.1 Diagonal Matrices

If the coefficient matrix A is diagonal, i.e., $\mathsf{A} = \operatorname{diag}(a_1, \ldots, a_n)$ with $a_k \neq 0$, $k = 1, \ldots, n$, then the solution can be easily found by $x_k = f_k / a_k$.

3.1.2 Triangular Matrices

Let us, to be definite, consider the case when A is upper triangular. The system of equations reads

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = f_1, \\ \qquad\quad a_{2,2}x_2 + \cdots + a_{2,n}x_n = f_2, \\ \qquad\qquad\qquad\qquad \vdots \\ \qquad\qquad\qquad\quad a_{n,n}x_n = f_n. \end{cases}$$

A unique solution exists if and only if $a_{i,i} \neq 0$ for all $i = 1, \ldots, n$. In this case, the solution can be easily found by first computing the value of the last variable,

$$x_n = f_n / a_{n,n},$$

and, after that, recursively computing

$$x_k = \frac{1}{a_{k,k}} \left(f_k - \sum_{j=k+1}^{n} a_{k,j} x_j \right), \quad k = n-1, n-2, \ldots, 2, 1.$$

The order of execution of this algorithm is vital: one must start with $k = n - 1$ and proceed in reverse order, finishing with $k = 1$. This algorithm is known as *back substitution*.

Remark 3.1 (forward substitution)**.** A similar procedure, known as *forward substitution*, can be applied to lower triangular matrices.

3.1.3 Tridiagonal Matrices

We begin with a definition.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Definition 3.2 (tridiagonal matrix)**.** Let $\mathsf{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$. We say that A is **tridiagonal** if and only if when $i, j \in \{1, \ldots, n\}$ and $|i - j| > 1$ then $a_{i,j} = 0$.

A generic system of equations with a tridiagonal coefficient matrix can be conveniently expressed as

$$a_k x_{k-1} + b_k x_k + c_k x_{k+1} = f_k, \quad k = 1, \ldots, n, \tag{3.2}$$

with $a_1 = c_n = 0$. This can be visualized as

$$\begin{bmatrix} b_1 & c_1 & 0 & \cdots & 0 & 0 \\ a_2 & b_2 & c_2 & 0 & & 0 \\ 0 & a_3 & b_3 & \ddots & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots & b_{n-1} & c_{n-1} \\ 0 & 0 & \cdots & 0 & a_n & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-2} \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{n-2} \\ f_{n-1} \\ f_n \end{bmatrix}.$$

To find the solution — assuming that a unique solution exists — we begin by assuming that it has the following form:

$$x_k = \alpha_k x_{k+1} + \beta_k.$$

This seems reasonable since, for $k = 1$, we have

$$x_1 = -\frac{c_1}{b_1} x_2 + \frac{f_1}{b_1},$$

which conforms to our solution ansatz with

$$\alpha_1 = -\frac{c_1}{b_1}, \quad \beta_1 = \frac{f_1}{b_1}.$$

Substituting our solution expression into the general form of the equations gives

$$a_k(\alpha_{k-1} x_k + \beta_{k-1}) + b_k x_k + c_k x_{k+1} = f_k,$$

from which we get

$$x_k = -\frac{c_k}{a_k \alpha_{k-1} + b_k} x_{k+1} + \frac{f_k - a_k \beta_{k-1}}{a_k \alpha_{k-1} + b_k} = \alpha_k x_{k+1} + \beta_k. \tag{3.3}$$

Then, since $c_n = 0$,

$$x_n = \frac{f_n - a_n \beta_{n-1}}{b_n + a_n \alpha_{n-1}}.$$

Then, for $k = n - 1, \ldots, 1$, we can use (3.3) to find the remaining components of the solution.

An implementation of the just described algorithm is presented in Listing 3.1. The reader can easily verify that the obtained $\boldsymbol{x}$ is indeed a solution to system (3.2). In the literature, this algorithm is sometimes called the *Thomas algorithm*. [1]

[1] Named in honor of the British physicist and applied mathematician Llewellyn Hilleth Thomas (1903–1992).

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Remark 3.3 (structure)**.** The reader may wonder how useful the algorithm that we just devised may be, as it requires a very special structure on the system matrix, namely that it is tridiagonal. Later, in Chapters 24 and 28, we will see that many schemes for the solution of one-dimensional boundary and initial boundary problems entail solving (a sequence of) systems of linear equations with tridiagonal matrices.

3.1.4 Cyclically Tridiagonal Matrices

Consider a system of the form

$$\begin{cases} b_1x_1 + c_1x_2 + a_1x_n = f_1, \\ a_kx_{k-1} + b_kx_k + c_kx_{k+1} = f_k, \quad k = 2, \dots, n-1, \\ c_nx_1 + a_nx_{n-1} + b_nx_n = f_n. \end{cases} \tag{3.4}$$

The coefficient matrix for this system is said to be *cyclically tridiagonal*. Equation (3.4) can also be visualized as

$$\begin{bmatrix} b_1 & c_1 & 0 & \cdots & 0 & a_1 \\ a_2 & b_2 & c_2 & 0 & & 0 \\ 0 & a_3 & b_3 & \ddots & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots & b_{n-1} & c_{n-1} \\ c_n & 0 & \cdots & 0 & a_n & b_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-2} \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ \vdots \\ f_{n-2} \\ f_{n-1} \\ f_n \end{bmatrix}.$$

Notice that the coefficient matrix differs from a tridiagonal one, only in the $(1, n)$ and $(n, 1)$ entries. This hints at the fact that, to solve (3.4), we will make use of the solution of systems with tridiagonal matrices.

Indeed, to find the solution of (3.4), we will first solve the tridiagonal systems

$$\begin{cases} b_2u_2 + c_2u_3 = f_2, \\ a_3u_2 + b_3u_3 + c_3u_4 = f_3, \\ \qquad\qquad \vdots \\ a_nu_{n-1} + b_nu_n = f_n \end{cases} \tag{3.5}$$

and

$$\begin{cases} b_2v_2 + c_2v_3 = -a_2, \\ a_3v_2 + b_3v_3 + c_3v_4 = 0, \\ \qquad\qquad \vdots \\ a_nv_{n-1} + b_nv_n = -c_n. \end{cases} \tag{3.6}$$

Then we set

$$x_k = u_k + x_1v_k, \quad k = 1, \dots, n, \tag{3.7}$$

with $u_1 = 0$ and $v_1 = 1$. Substituting this representation in the first equation yields

$$b_1x_1 + c_1(u_2 + x_1v_2) + a_1(u_n + x_1v_n) = f_1,$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

which implies that

$$x_1 = \frac{f_1 - c_1 u_2 - a_1 u_n}{b_1 + c_1 v_2 + a_1 v_n}. \tag{3.8}$$

Let us verify that (3.7) and (3.8) are indeed the solution to (3.4). To do so, multiply the first equation of system (3.6) by x_1 and add it to the first equation of system (3.5) to obtain

$$a_2 x_1 + b_2(u_2 + x_1 v_2) + c_2(u_3 + x_1 v_3) = f_2,$$

which implies that

$$a_2 x_1 + b_2 x_2 + c_2 x_3 = f_2.$$

A similar calculation can be made for the remaining equations, up until the last one, where we get

$$c_n x_1 + a_n(u_{n-1} + x_1 v_{n-1}) + b_n(u_n + x_1 v_n) = f_n,$$

which implies that

$$c_n x_1 + a_n x_{n-1} + b_n x_n = f_n.$$

An implementation of this procedure is presented in Listing 3.2. Once again, the reader may wonder how often one encounters cyclically tridiagonal matrices. Many discretization schemes for one-dimensional boundary and initial boundary value problems with periodic boundary conditions entail the solution of (a collection of) systems of linear equations with cyclically tridiagonal matrices; see Chapters 24–28 for more details.

3.2 LU Factorization

In this section, we give a practical and theoretical description of the method of LU factorization for solving a square system of linear equations. The idea is based upon the very familiar concept of Gaussian elimination. We need the following preliminary results.

Theorem 3.4 (properties of triangular matrices)**.** *Let the matrices* $\mathsf{T}, \mathsf{T}_k \in \mathbb{C}^{n\times n}$*, for* $k = 1, 2$*, be lower (upper) triangular. Then the following are true.*

1. *The product* $\mathsf{T}_1\mathsf{T}_2$ *is lower (upper) triangular.*
2. *If, in addition* $[\mathsf{T}_k]_{i,i} = 1$*, for* $k = 1, 2$ *and* $i = 1, \ldots, n$ *— i.e.,* $\mathsf{T}_1, \mathsf{T}_2 \in \mathbb{C}^{n\times n}$ *are unit lower (unit upper) triangular — then the product* $\mathsf{T}_1\mathsf{T}_2$ *is unit lower (unit upper) triangular.*
3. *The matrix* T *is nonsingular if and only if* $[\mathsf{T}]_{i,i} \neq 0$ *for all* $i = 1, \ldots, n$*.*
4. *If* T *is nonsingular,* $\mathsf{T}^{-1} \in \mathbb{C}^{n\times n}$ *is lower (upper) triangular.*
5. *If* T *is unit lower (unit upper) triangular, then it is invertible and* T^{-1} *is unit lower (unit upper) triangular.*
6. *If* $[\mathsf{T}]_{i,i} > 0$*, then* $[\mathsf{T}_k^{-1}]_{i,i} = \frac{1}{[\mathsf{T}]_{i,i}} > 0$*.*

Proof. See Problem 1.32. □

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Definition 3.5 (sub-matrix)**.** Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ and $S \subseteq \{1, 2, \ldots, n\}$ is nonempty with cardinality $k = \#(S) > 0$. The **sub-matrix** $\mathsf{A}(S) \in \mathbb{C}^{k\times k}$ is that matrix obtained by deleting the columns and rows of A whose indices are not in S. In symbols,

$$[\mathsf{A}(S)]_{i,j} = [\mathsf{A}]_{m_i,m_j}, \quad i, j = 1, \ldots, k,$$

where

$$S = \{m_1, \ldots, m_k\} \quad \text{and} \quad 1 \leq m_1 < m_2 < \cdots < m_k \leq n.$$

Example 3.1 Suppose that

$$\mathsf{A} = \begin{bmatrix} 1 & -7 & 12 & 4 \\ 6 & 9 & -3 & -4 \\ 1 & -6 & 8 & 9 \\ 4 & 4 & -11 & 17 \end{bmatrix}, \quad S = \{2, 4\}.$$

Then $m_1 = 2$ and $m_2 = 4$ and

$$\mathsf{A}(S) = \begin{bmatrix} 9 & -4 \\ 4 & 17 \end{bmatrix}.$$

Definition 3.6 (leading principal sub-matrix)**.** Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ and $S = \{1, 2, \ldots, k\}$ with $k \leq n$. Then we define

$$\mathsf{A}^{(k)} = \mathsf{A}(S) \in \mathbb{C}^{k\times k}$$

and we call $\mathsf{A}^{(k)}$ the **leading principal sub-matrix** of A of order k.

Typically, the LU factorization is produced via Gaussian elimination. But the proof of the following theorem obscures this fact and guarantees, independently of Gaussian elimination, the existence and uniqueness of the LU factorization.

Theorem 3.7 (LU factorization)**.** *Let $n \geq 2$ and $\mathsf{A} \in \mathbb{C}^{n\times n}$. Suppose that all the leading principal sub-matrices of A are nonsingular, i.e., $\det(\mathsf{A}^{(k)}) \neq 0$ for all $k = 1, \ldots, n-1$. Then there exists a unit lower triangular matrix $\mathsf{L} \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $\mathsf{U} \in \mathbb{C}^{n\times n}$ such that*

$$\mathsf{A} = \mathsf{LU}.$$

Proof. The proof is by induction on n, the size of the matrix.
($n = 2$) Consider

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

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

with $a \neq 0$, by assumption. Define

$$\mathsf{L} = \begin{bmatrix} 1 & 0 \\ m & 1 \end{bmatrix}, \qquad \mathsf{U} = \begin{bmatrix} u & v \\ 0 & \eta \end{bmatrix},$$

where

$$u = a, \quad v = b, \quad m = \frac{c}{a}, \quad \eta = d - b\frac{c}{a}.$$

Then

$$mu = c, \qquad mv + \eta = d,$$

and consequently $\mathsf{A} = \mathsf{LU}$, as is easily confirmed.
($n = m$) The induction hypothesis is as follows: suppose that the result is valid for any $\mathsf{A} \in \mathbb{C}^{m\times m}$, provided that $\mathsf{A}^{(k)}$ is nonsingular for all $k = 1, \ldots, m-1$.
($n = m+1$) Suppose that $\mathsf{A}^{(k)}$ is nonsingular for $k = 1, \ldots, m$. Set

$$\mathsf{A} = \begin{bmatrix} \mathsf{A}^{(m)} & \boldsymbol{b} \\ \boldsymbol{c}^{\mathsf{T}} & d \end{bmatrix} \in \mathbb{C}^{(m+1)\times(m+1)}.$$

From the induction hypothesis, there is a unit lower triangular matrix $\mathsf{L}^{(m)}$ and an upper triangular matrix $\mathsf{U}^{(m)}$ such that $\mathsf{A}^{(m)} = \mathsf{L}^{(m)}\mathsf{U}^{(m)}$, where $\mathsf{A}^{(m)}$ is the leading principal sub-matrix of A of order m. Define

$$\mathsf{L} = \begin{bmatrix} \mathsf{L}^{(m)} & \mathbf{0} \\ \boldsymbol{m}^{\mathsf{T}} & 1 \end{bmatrix}, \quad \mathsf{U} = \begin{bmatrix} \mathsf{U}^{(m)} & \boldsymbol{v} \\ \mathbf{0}^{\mathsf{T}} & \eta \end{bmatrix},$$

where $\boldsymbol{b}, \boldsymbol{c}, \boldsymbol{m}, \boldsymbol{v}, \mathbf{0} \in \mathbb{C}^m$. Then

$$\mathsf{LU} = \begin{bmatrix} \mathsf{L}^{(m)}\mathsf{U}^{(m)} & \mathsf{L}^{(m)}\boldsymbol{v} \\ \boldsymbol{m}^{\mathsf{T}}\mathsf{U}^{(m)} & \boldsymbol{m}^{\mathsf{T}}\boldsymbol{v} + \eta \end{bmatrix}.$$

Let us set this equal to A and determine whether or not the resulting equations are solvable. It is easy to see that $\mathsf{A} = \mathsf{LU}$ if and only if

$$\mathsf{L}^{(m)}\mathsf{U}^{(m)} = \mathsf{A}^{(m)}, \qquad \mathsf{L}^{(m)}\boldsymbol{v} = \boldsymbol{b},$$
$$\boldsymbol{m}^{\mathsf{T}}\mathsf{U}^{(m)} = \boldsymbol{c}^{\mathsf{T}}, \qquad \boldsymbol{m}^{\mathsf{T}}\boldsymbol{v} + \eta = d.$$

The last three equations are uniquely solvable, as we now show: since $\mathsf{L}^{(m)}$ is invertible,

$$\boldsymbol{v} = \left(\mathsf{L}^{(m)}\right)^{-1}\boldsymbol{b}.$$

The matrix $\mathsf{U}^{(m)}$ is invertible since

$$0 \neq \det(\mathsf{A}^{(m)}) = \det(\mathsf{L}^{(m)}\mathsf{U}^{(m)}) = \det(\mathsf{U}^{(m)}).$$

Hence,

$$\boldsymbol{m}^{\mathsf{T}} = \boldsymbol{c}^{\mathsf{T}}\left(\mathsf{U}^{(m)}\right)^{-1} \quad \text{or} \quad \boldsymbol{m} = \left(\mathsf{U}^{(m)}\right)^{-\mathsf{T}}\boldsymbol{c}.$$

Finally,

$$\eta = d - \boldsymbol{m}^{\mathsf{T}}\boldsymbol{v}.$$

The proof by induction is complete. □

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Before we go any further, we ought to say why it is that an LU factorization of a matrix is useful. Suppose that we want to solve the indexed family of problems

$$\mathsf{A}\boldsymbol{x}^{(k)} = \boldsymbol{f}^{(k)}, \quad k = 1, \ldots, K,$$

and that there exists a unit lower triangular matrix $\mathsf{L} \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $\mathsf{U} \in \mathbb{C}^{n\times n}$ such that $\mathsf{A} = \mathsf{LU}$. To find the solutions $\boldsymbol{x}^{(k)}$, we solve the following equivalent family:

$$\mathsf{L}\boldsymbol{y}^{(k)} = \boldsymbol{f}^{(k)}, \qquad \mathsf{U}\boldsymbol{x}^{(k)} = \boldsymbol{y}^{(k)}, \quad k = 1, \ldots, K.$$

The vector $\boldsymbol{y}^{(k)}$ can be obtained easily and cheaply via forward substitution. Subsequently, the vector $\boldsymbol{x}^{(k)}$ can be obtained by back substitution; see Section 3.1.2.

Now we show a practical connection between the LU factorization and what is commonly called Gaussian elimination. We use an example to motivate our discussion.

Example 3.2 Consider the following system of linear equations:

$$\begin{cases} x_1 + & x_2 + & x_3 = & 6, \\ 2x_1 + & 4x_2 + & 2x_3 = & 16, \\ -x_1 + & 5x_2 - & 4x_3 = & -3. \end{cases}$$

Of course, we can represent this as a matrix–vector equation $\mathsf{A}\boldsymbol{x} = \boldsymbol{f}$. We write this as an augmented matrix and perform Gaussian elimination to put the system into so-callled row echelon form,

$$[\mathsf{A}|\boldsymbol{f}] = \left[\begin{array}{ccc|c} \boxed{1} & 1 & 1 & 6 \\ 2 & 4 & 2 & 16 \\ -1 & 5 & -4 & -3 \end{array}\right] \xrightarrow[1R_1+R_3\to R_3]{-2R_1+R_2\to R_2} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & \boxed{2} & 0 & 4 \\ 0 & 6 & -3 & 3 \end{array}\right] \xrightarrow{-3R_2+R_3\to R_3} \left[\begin{array}{ccc|c} 1 & 1 & 1 & 6 \\ 0 & 2 & 0 & 4 \\ 0 & 0 & -3 & -9 \end{array}\right].$$

The boxed entries indicate the so-called pivot elements. The values of the pivot elements help to determine the row multipliers in the algorithm. As long as these are nonzero, the algorithm can run to completion. Gaussian elimination uses elementary row operations to produce an equivalent upper triangular system of linear equations, $\mathsf{U}\boldsymbol{x} = \boldsymbol{b}$. By *equivalent*, we mean that the solution sets are exactly the same, even though the coefficient matrix and right-hand-side vector are changed.

Let us focus on the left-hand side of the augmented system, as this will be the important part with respect to the LU factorization. We have

$$\mathsf{L}^{(3,2)}\mathsf{L}^{(3,1)}\mathsf{L}^{(2,1)}\mathsf{A} = \mathsf{U} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{bmatrix},$$

where $\mathsf{L}^{(2,1)}, \mathsf{L}^{(3,1)}, \mathsf{L}^{(3,2)}$ are elementary matrices encoding the elementary row operations performed in our Gaussian elimination process. To produce the matrix

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

representations of these operations, recall that we need only to apply the corresponding elementary row operations on the identity matrix:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow{-2R_1+R_2\to R_2} \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathsf{L}^{(2,1)}.$$

Likewise,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow{1R_1+R_3\to R_3} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \mathsf{L}^{(3,1)}$$

and

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \xrightarrow{-3R_2+R_3\to R_3} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3 & 1 \end{bmatrix} = \mathsf{L}^{(3,2)}.$$

Then it is easy to see that

$$\mathsf{L}^{(3,1)}\mathsf{L}^{(2,1)}\mathsf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 6 & -3 \end{bmatrix} \quad \text{and} \quad \mathsf{L}^{(3,2)}\mathsf{L}^{(3,1)}\mathsf{L}^{(2,1)}\mathsf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{bmatrix}.$$

Observe that the order of application of $\mathsf{L}^{(2,1)}$ and $\mathsf{L}^{(3,1)}$ does not matter:

$$\mathsf{L}^{(3,1)}\mathsf{L}^{(2,1)} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} = \mathsf{L}^{(2,1)}\mathsf{L}^{(3,1)}.$$

This will be shown to be true in general.

Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$, where $n \geq 2$. If Gaussian elimination for A proceeds to completion without encountering any zero pivots, one gets

$$\mathsf{L}^{(n,n-1)} \cdots \mathsf{L}^{(n,2)} \cdots \mathsf{L}^{(3,2)}\mathsf{L}^{(n,1)} \cdots \mathsf{L}^{(2,1)}\mathsf{A} = \mathsf{U},$$

where U is square and upper triangular. Moreover, since we assume that no zero pivot entries are encountered, $[\mathsf{U}]_{i,i} \neq 0$, for $i = 1, \cdots, n-1$. However, it is possible that $[\mathsf{U}]_{n,n} = 0$. We can group the elementary operations into column operations as follows:

$$\begin{aligned}
&\text{(column 1):} && \mathsf{L}_1 = \mathsf{L}^{(n,1)} \cdots \mathsf{L}^{(2,1)}, \\
&\text{(column 2):} && \mathsf{L}_2 = \mathsf{L}^{(n,2)} \cdots \mathsf{L}^{(3,2)}, \\
& && \vdots \\
&\text{(column } n-2\text{):} && \mathsf{L}_{n-2} = \mathsf{L}^{(n,n-2)}\mathsf{L}^{(n-1,n-2)}, \\
&\text{(column } n-1\text{):} && \mathsf{L}_{n-1} = \mathsf{L}^{(n,n-1)},
\end{aligned}$$

so that

$$\mathsf{L}_{n-1}\mathsf{L}_{n-2} \cdots \mathsf{L}_2\ \mathsf{L}_1\mathsf{A} = \mathsf{U}.$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Furthermore, we can prove that the matrices defining the L_i matrices commute and can be multiplied in any order, as we will show. The matrices L_i are examples of what we will call column-i complete elementary matrices.

In any case,

$$\mathsf{A} = \mathsf{L}_1^{-1} \cdots \mathsf{L}_{n-1}^{-1} \mathsf{U} = \mathsf{L}\mathsf{U}.$$

It only remains to show that L is unit lower triangular. If so, we have connected Gaussian elimination to the LU factorization.

Now consider the implications of the following example.

Example 3.3 What is the inverse of an elementary matrix? Suppose that

$$\mathsf{L} = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathsf{L}' = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then it is easy to see that $\mathsf{L}'\mathsf{L} = \mathsf{I}_3$. Or, in other words, $\mathsf{L}' = \mathsf{L}^{-1}$.

Definition 3.8 (elementary matrix)**.** A matrix $\mathsf{E} \in \mathbb{C}^{n\times n}$ is called **elementary** if and only if $\mathsf{E} = \mathsf{I} + \mu_{r,s}\mathsf{M}^{(r,s)}$ for some $\mu_{r,s} \in \mathbb{C}$ and for some $1 \leq s < r \leq n$, where

$$\mathsf{M}^{(r,s)} = \boldsymbol{e}_r \boldsymbol{e}_s^{\mathsf{T}},$$

i.e.,

$$\left[\mathsf{M}^{(r,s)}\right]_{i,j} = \delta_{i,r}\delta_{j,s}.$$

Remark 3.9 (generality)**.** We remark that our last definition is more restrictive than might be found in some other references, specifically in our requirement that r is strictly greater than s. But this is all that is needed for our purposes.

Proposition 3.10 (properties of elementary matrices)**.** *Suppose that*

$$\mathsf{E}_k = \mathsf{I} + \mu_{r_k,s}\mathsf{M}^{(r_k,s)}, \quad k = 1, 2$$

are two elementary matrices with $r_1 \neq r_2$. Then both matrices are invertible, the inverses are elementary, and the matrices commute. Furthermore,

$$\mathsf{E}_k^{-1} = \mathsf{I} - \mu_{r_k,s}\mathsf{M}^{(r_k,s)},$$

$$(\mathsf{E}_1\mathsf{E}_2)^{-1} = \mathsf{E}_2^{-1}\mathsf{E}_1^{-1} = \mathsf{E}_1^{-1}\mathsf{E}_2^{-1} = \mathsf{I} - \mu_{r_1,s}\mathsf{M}^{(r_1,s)} - \mu_{r_2,s}\mathsf{M}^{(r_2,s)},$$

and

$$\mathsf{E}_1\mathsf{E}_2 = \mathsf{E}_2\mathsf{E}_1 = \mathsf{I} + \mu_{r_1,s}\mathsf{M}^{(r_1,s)} + \mu_{r_2,s}\mathsf{M}^{(r_2,s)},$$

Proof. See Problem 3.5. □

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Definition 3.11 (complete elementary matrix)**.** Suppose that $n \geq 2$. Let the index $s \in \{1, 2, \ldots, n-1\}$ be given. The matrix $\mathsf{F} \in \mathbb{C}^{n \times n}$ is called a **column-s complete elementary matrix** if and only if

$$\mathsf{F} = \mathsf{I} + \sum_{r=s+1}^{n} \mu_{r,s} \mathsf{M}^{(r,s)}$$

for some scalars $\mu_{r,s} \in \mathbb{C}$, $r = s+1, \ldots, n$. In other words, F is a unit lower triangular matrix of the form

$$\mathsf{F} = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & \mu_{s+1,s} & 1 & \\ & & \vdots & & \ddots \\ & & \mu_{n,s} & & 1 \end{bmatrix}.$$

Definition 3.12 (Gaussian elimination[2])**.** Let $\mathsf{A} \in \mathbb{C}^{n \times n}$ be given with $n \geq 2$. We define the **Gaussian elimination** algorithm recursively as follows. Suppose that k stages of Gaussian elimination have been completed, where $k \in \{0, \ldots, n-1\}$, such that no zero pivots have been encountered, producing the matrix factorization

$$\mathsf{L}_k \cdots \mathsf{L}_1 \mathsf{A} = \mathsf{A}^{(k)}, \quad k = 1, \ldots, n-1,$$

where $\mathsf{A}^{(0)} = \mathsf{A}$ and, for $k = 1, \ldots, n-1$,

$$\mathsf{A}^{(k)} = \begin{bmatrix} a_{1,1}^{(0)} & a_{1,2}^{(0)} & a_{1,3}^{(0)} & a_{1,4}^{(0)} & \cdots & a_{1,k+1}^{(0)} & \cdots & a_{1,n}^{(0)} \\ 0 & a_{2,2}^{(1)} & a_{2,3}^{(1)} & a_{2,4}^{(1)} & \cdots & a_{2,k+1}^{(1)} & \cdots & a_{2,n}^{(1)} \\ 0 & 0 & a_{3,3}^{(2)} & a_{3,4}^{(2)} & \cdots & a_{3,k+1}^{(2)} & \cdots & a_{3,n}^{(2)} \\ \vdots & & \ddots & \ddots & \ddots & \vdots & & \vdots \\ 0 & & & 0 & a_{k,k}^{(k-1)} & a_{k,k+1}^{(k-1)} & \cdots & a_{k,n}^{(k-1)} \\ 0 & & & 0 & 0 & a_{k+1,k+1}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ \vdots & & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & a_{n,k+1}^{(k)} & \cdots & a_{n,n}^{(k)} \end{bmatrix}.$$

If $k = n-1$, we are done and we set $\mathsf{U} = \mathsf{A}^{(n-1)}$. Otherwise, if the $(k+1)$st pivot entry, $a_{k+1,k+1}^{(k)}$, is not equal to zero, the algorithm may proceed. Construct the column-$(k+1)$ complete elementary matrix

$$\mathsf{L}_{k+1} = \mathsf{I} + \sum_{r=k+2}^{n} \mu_{r,k+1} \mathsf{M}^{(r,k+1)},$$

where

$$\mu_{r,k+1} = -\frac{a_{r,k+1}^{(k)}}{a_{k+1,k+1}^{(k)}}, \quad r = k+2, \ldots, n.$$

[2] Named in honor of the German mathematician and physicist Johann Carl Friedrich Gauss (1777–1855).

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Then set

$$\mathsf{L}_{k+1}\mathsf{A}^{(k)} = \mathsf{A}^{(k+1)},$$

obtaining

$$\mathsf{A}^{(k+1)} = \begin{bmatrix} a_{1,1}^{(0)} & a_{1,2}^{(0)} & a_{1,3}^{(0)} & a_{1,4}^{(0)} & \cdots & a_{1,k+1}^{(0)} & a_{1,k+2}^{(0)} & \cdots & a_{1,n}^{(0)} \\ 0 & a_{2,2}^{(1)} & a_{2,3}^{(1)} & a_{2,4}^{(1)} & \cdots & a_{2,k+1}^{(1)} & a_{2,k+2}^{(1)} & \cdots & a_{2,n}^{(1)} \\ 0 & 0 & a_{3,3}^{(2)} & a_{3,4}^{(2)} & \cdots & a_{3,k+1}^{(2)} & a_{3,k+2}^{(2)} & \cdots & a_{3,n}^{(2)} \\ \vdots & & \ddots & \ddots & \ddots & \vdots & \vdots & & \vdots \\ 0 & & & 0 & a_{k,k}^{(k-1)} & a_{k,k+1}^{(k-1)} & a_{k,k+2}^{(k-1)} & \cdots & a_{k,n}^{(k-1)} \\ 0 & & & 0 & 0 & a_{k+1,k+1}^{(k)} & a_{k+1,k+2}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ 0 & & & 0 & 0 & 0 & a_{k+2,k+2}^{(k+1)} & \cdots & a_{k+2,n}^{(k+1)} \\ \vdots & & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & a_{n,k+2}^{(k+1)} & \cdots & a_{n,n}^{(k+1)} \end{bmatrix}.$$

Based on our previous computations, the following result should be clear.

Theorem 3.13 (Gaussian elimination)**.** *Let* $\mathsf{A} \in \mathbb{C}^{n\times n}$*. If Gaussian elimination proceeds to completion without encountering any zero pivots, then there are column-k complete elementary matrices* $\mathsf{L}_k \in \mathbb{C}^{n\times n}$*, for* $k = 1, \ldots, n-1$*, such that*

$$\mathsf{L}_{n-1}\cdots\mathsf{L}_2\mathsf{L}_1\mathsf{A} = \mathsf{U},$$

where $\mathsf{U} \in \mathbb{C}^{n\times n}$ *is upper triangular and*

$$[\mathsf{U}]_{i,i} \neq 0, \quad i = 1, \ldots, n-1,$$

since no zero pivots are encountered. Furthermore,

$$\mathsf{A} = \mathsf{L}_1^{-1}\cdots\mathsf{L}_{n-1}^{-1}\mathsf{U} = \mathsf{L}\mathsf{U},$$

where L *is unit lower triangular. Writing*

$$\mathsf{L}_k = \mathsf{I} + \sum_{r=k+1}^{n} \mu_{r,s}\mathsf{M}^{(r,k)},$$

it follows that

$$\mathsf{L} = \mathsf{I} - \sum_{k=1}^{n-1}\sum_{r=k+1}^{n} \mu_{r,k}\mathsf{M}^{(r,k)}.$$

In other words,

$$\mathsf{L} = \begin{bmatrix} 1 & & & & \\ -\mu_{2,1} & 1 & & & \\ -\mu_{3,1} & -\mu_{3,2} & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ -\mu_{n,1} & -\mu_{n,2} & \cdots & -\mu_{n,n-1} & 1 \end{bmatrix}.$$

Proof. See Problem 3.7. □

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Theorem 3.14 (uniqueness)**.** *Suppose that $n \geq 2$ and $A \in \mathbb{C}^{n\times n}$ is invertible. Suppose that there is a unit lower triangular matrix $L \in \mathbb{C}^{n\times n}$ and an upper triangular matrix $U \in \mathbb{C}^{n\times n}$ such that $A = LU$. Then this LU factorization is unique.*

Proof. Suppose that there are two factorizations with the desired properties:

$$L_1U_1 = A = L_2U_2.$$

Since A is invertible, U_1 and U_2 must be invertible, i.e., there are no zeros on their diagonals. Furthermore,

$$L_2^{-1}L_1 = U_2U_1^{-1} = D,$$

where D is by necessity diagonal. Therefore,

$$L_1 = L_2D.$$

But it must be that $D = I_n$, since the diagonal elements of L_1 and L_2 are all ones. □

Listing 3.3 provides a more streamlined, computable version of the LU factorization algorithm. The algorithm proceeds to completion provided that no zero pivots are encountered. This listing can also be used to estimate the complexity of the LU factorization algorithm.

Proposition 3.15 (complexity of LU)**.** *Let $A \in \mathbb{C}^{n\times n}$. Then the LU factorization algorithm requires, to leading order, $\frac{2}{3}n^3$ operations.*

Proof. We only care about the leading order of operations, which, from Listing 3.3, can easily be seen to be roughly

$$\begin{aligned}\sum_{k=1}^{n-1}\sum_{j=k}^{n}\sum_{t=k}^{n} 2 &= 2\sum_{k=1}^{n-1}(n-k)\sum_{j=k+1}^{n} 1 \\ &\approx 2\sum_{k=1}^{n-1}(n-k)^2 \\ &= \frac{1}{3}(n-1)n(2n-1) \\ &\approx \frac{2}{3}n^3.\end{aligned}$$

□

3.3 Gaussian Elimination with Column Pivoting

In the last section, we did not consider what one should do if a zero pivot is encountered. Let us examine a simple situation where zero pivots appear.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Example 3.4 Suppose that $A \in \mathbb{C}^{3\times 3}$ is given by

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

We want to use Gaussian elimination to obtain an LU factorization of A. However, we notice from the beginning that there is a zero in the first pivot location. But a simple row interchange operation will fix this situation. In the following algorithm, let us agree to interchange rows, so that the element with largest modulus in the column at or below the pivot position moves into the pivot position. This is called Gaussian elimination with maximal column pivoting:

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

Our procedure may be expressed as

$$L_2P_2L_1P_1A = U,$$

where P_1 and P_2 are simple permutation matrices and L_1 and L_2 are column-1 and column-2 complete elementary matrices, respectively.

Definition 3.16 (permutation)**.** A matrix $P \in \mathbb{C}^{n\times n}$ is called a **simple permutation** matrix if and only if it is obtained from the $n \times n$ identity matrix I by interchanging exactly two rows of I. P is called a **regular permutation** (or just a **permutation**) matrix if and only if P is the product of simple permutation matrices.

Proposition 3.17 (action of permutations)**.** *Let $n \geq 2$. Suppose that $A \in \mathbb{C}^{n\times n}$ is any matrix and $P \in \mathbb{C}^{n\times n}$ is a simple permutation matrix obtained by interchanging rows r and s of the identity matrix with $1 \leq r < s \leq n$. Then* PA *is identical to* A*, except with rows r and s interchanged. Furthermore,* AP *is identical to* A*, except with columns r and s interchanged.*

Proof. See Problem 3.8. □

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Lemma 3.18 (properties of permutations)**.** *Suppose that* $\mathsf{P}, \mathsf{Q} \in \mathbb{C}^{n\times n}$*, with* $n \geq 2$*, are permutation matrices. Then*

1. *The product* PQ *is a permutation matrix.*
2. $\det(\mathsf{P}) = \pm 1$ *according to whether* P *is the product of an even* ($\det(\mathsf{P}) = 1$) *or an odd* ($\det(\mathsf{P}) = -1$) *number of simple permutation matrices.*
3. *The inverse of a simple permutation matrix is itself. Any regular permutation matrix* P *is invertible, and, if*

$$\mathsf{P} = \mathsf{P}_1\mathsf{P}_2\cdots\mathsf{P}_k,$$

where P_i *is a simple permutation matrix, for* $1 \leq i \leq k$*, then*

$$\mathsf{P}^{-1} = \mathsf{P}_k\cdots\mathsf{P}_2\mathsf{P}_1 = \mathsf{P}^{\mathsf{T}}.$$

Proof. See Problem 3.9. □

Example 3.5 Let us continue with our 3×3 example, but in general terms. We have

$$\mathsf{L}_2\mathsf{P}_2\mathsf{L}_1\mathsf{P}_1\mathsf{A} = \mathsf{U},$$

where P_j, $j = 1, 2$ are simple permutation matrices or the identity matrix (in the case that no row interchange took place) and L_j are column-j complete elementary matrices, $j = 1, 2$. Now observe that

$$\mathsf{L}_2\mathsf{P}_2\mathsf{L}_1\mathsf{P}_2\mathsf{P}_2\mathsf{P}_1\mathsf{A} = \mathsf{U}.$$

Therefore,

$$\hat{\mathsf{L}}_2\hat{\mathsf{L}}_1\mathsf{PA} = \mathsf{U},$$

where

$$\hat{\mathsf{L}}_2 = \mathsf{L}_2, \quad \hat{\mathsf{L}}_1 = \mathsf{P}_2\mathsf{L}_1\mathsf{P}_2, \quad \mathsf{P} = \mathsf{P}_2\mathsf{P}_1.$$

Example 3.6 Suppose that Gaussian elimination with maximal column pivoting is applied to $\mathsf{A} \in \mathbb{C}^{4\times 4}$. Then it should be clear that one obtains

$$\mathsf{L}_3\mathsf{P}_3\mathsf{L}_2\mathsf{P}_2\mathsf{L}_1\mathsf{P}_1\mathsf{A} = \mathsf{U},$$

which can be rewritten as

$$\hat{\mathsf{L}}_3\hat{\mathsf{L}}_2\hat{\mathsf{L}}_1\mathsf{PA} = \mathsf{U},$$

where

$$\hat{\mathsf{L}}_3 = \mathsf{L}_3, \quad \hat{\mathsf{L}}_2 = \mathsf{P}_3\mathsf{L}_2\mathsf{P}_3, \quad \hat{\mathsf{L}}_1 = \mathsf{P}_3\mathsf{P}_2\mathsf{L}_1\mathsf{P}_2\mathsf{P}_3, \quad \mathsf{P} = \mathsf{P}_3\mathsf{P}_2\mathsf{P}_1.$$

It turns out — and it is probably not so hard to see — that the $\hat{\mathsf{L}}_k$ matrices constructed above are still column-k complete elementary matrices.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Proposition 3.19 (permutations and elementary matrices)**.** *Suppose that* $\mathsf{L}_k \in \mathbb{C}^{n\times n}$ *is a column-*k *complete elementary matrix,*

$$\mathsf{L}_k = \mathsf{I}_n + \sum_{r=k+1}^{n} \mu_{r,k}\mathsf{M}^{(r,k)},$$

for some constants $\mu_{r,k} \in \mathbb{C}$*, for* $k = k+1,\ldots,n$*. Assume that* $\mathsf{Q} \in \mathbb{C}^{n\times n}$ *is a simple permutation matrix encoding the interchange of rows* r' *and* s'*, where* $k < r' < s' \leq n$*. Then the matrix* $\mathsf{Q}\mathsf{L}_k\mathsf{Q}$ *is a column-*k *complete matrix. In particular,* $\mathsf{Q}\mathsf{L}_k\mathsf{Q}$ *is identical to* L_k*, except that entries* $\mu_{r',k}$ *and* $\mu_{s',k}$ *are interchanged.*

Proof. It follows that

$$\mathsf{Q}\mathsf{L}_k\mathsf{Q} = \mathsf{Q}\mathsf{I}_n\mathsf{Q} + \sum_{r=k+1}^{n} \mu_{r,k}\mathsf{Q}\boldsymbol{e}_r\boldsymbol{e}_k^{\mathsf{T}}\mathsf{Q}.$$

But observe that $\boldsymbol{e}_k^{\mathsf{T}}\mathsf{Q} = \boldsymbol{e}_k^{\mathsf{T}}$ and

$$\mathsf{Q}\boldsymbol{e}_r = \begin{cases} \boldsymbol{e}_{s'}, & r = r', \\ \boldsymbol{e}_{r'}, & r = s', \\ \boldsymbol{e}_r, & r \in \{1,\ldots,n\}\backslash\{r',s'\}. \end{cases}$$

Therefore,

$$\mathsf{Q}\mathsf{L}_k\mathsf{Q} = \mathsf{I}_n + \sum_{\substack{r=k+1 \\ r\neq r',s'}}^{n} \mu_{r,k}\boldsymbol{e}_r\boldsymbol{e}_k^{\mathsf{T}} + \mu_{s',k}\boldsymbol{e}_{r'}\boldsymbol{e}_k^{\mathsf{T}} + \mu_{r',k}\boldsymbol{e}_{s'}\boldsymbol{e}_k^{\mathsf{T}}.$$

In other words, $\mathsf{Q}\mathsf{L}_k\mathsf{Q}$ is a column-k complete elementary matrix that is identical to L_k, except that the positions of $\mu_{r',k}$ and $\mu_{s',k}$ are swapped. □

Definition 3.20 (Gaussian elimination with maximal column pivoting)**.** Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be given with $n \geq 2$. We define the **Gaussian elimination with maximal column pivoting** algorithm recursively as follows. Suppose that k stages of Gaussian elimination with maximal column pivoting have been completed, where $k = 0,\ldots,n-1$, producing the matrix decomposition

$$\mathsf{L}_k\mathsf{P}_k\cdots\mathsf{L}_1\mathsf{P}_1\mathsf{A} = \mathsf{A}^{(k)}, \quad k = 1,\ldots,n-1,$$

where $\mathsf{A}^{(0)} = \mathsf{A}$ and, for $k = 1,\ldots,n-1$,

$$\mathsf{A}^{(k)} = \begin{bmatrix} a_{1,1}^{(0,1)} & a_{1,2}^{(0,1)} & a_{1,3}^{(0,1)} & a_{1,4}^{(0,1)} & \cdots & a_{1,k+1}^{(0,1)} & \cdots & a_{1,n}^{(0,1)} \\ 0 & a_{2,2}^{(1,1)} & a_{2,3}^{(1,1)} & a_{2,4}^{(1,1)} & \cdots & a_{2,k+1}^{(1,1)} & \cdots & a_{2,n}^{(1,1)} \\ 0 & 0 & a_{3,3}^{(2,1)} & a_{3,4}^{(2,1)} & \cdots & a_{3,k+1}^{(2,1)} & \cdots & a_{3,n}^{(2,1)} \\ \vdots & & \ddots & \ddots & \ddots & \vdots & & \vdots \\ 0 & & & 0 & a_{k,k}^{(k-1,1)} & a_{k,k+1}^{(k-1,1)} & \cdots & a_{k,n}^{(k-1,1)} \\ 0 & & & 0 & 0 & a_{k+1,k+1}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ \vdots & & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & a_{n,k+1}^{(k)} & \cdots & a_{n,n}^{(k)} \end{bmatrix}.$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

If $k = n-1$, we are done, and we set $\mathsf{U} = \mathsf{A}^{(n-1)}$. Otherwise, use a simple permutation matrix to interchange rows $k+1$ and r, with $r \geq k+1$ and

$$\left|a_{r,k+1}^{(k)}\right| \geq \left|a_{j,k+1}^{(k)}\right|, \quad j = k+1, \ldots, n,$$

with the understanding that the simple permutation is the identity matrix if $r = k+1$, obtaining

$$\mathsf{P}_{k+1}\mathsf{A}^{(k)} = \mathsf{A}^{(k,1)},$$

where

$$\mathsf{A}^{(k,1)} = \begin{bmatrix} a_{1,1}^{(0,1)} & a_{1,2}^{(0,1)} & a_{1,3}^{(0,1)} & a_{1,4}^{(0,1)} & \cdots & a_{1,k+1}^{(0,1)} & \cdots & a_{1,n}^{(0,1)} \\ 0 & a_{2,2}^{(1,1)} & a_{2,3}^{(1,1)} & a_{2,4}^{(1,1)} & \cdots & a_{2,k+1}^{(1,1)} & \cdots & a_{2,n}^{(1,1)} \\ 0 & 0 & a_{3,3}^{(2,1)} & a_{3,4}^{(2,1)} & \cdots & a_{3,k+1}^{(2,1)} & \cdots & a_{3,n}^{(2,1)} \\ \vdots & & \ddots & \ddots & \ddots & \vdots & & \vdots \\ 0 & & & 0 & a_{k,k}^{(k-1,1)} & a_{k,k+1}^{(k-1,1)} & \cdots & a_{k,n}^{(k-1,1)} \\ 0 & & & 0 & 0 & a_{k+1,k+1}^{(k,1)} & \cdots & a_{k+1,n}^{(k,1)} \\ \vdots & & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & a_{n,k+1}^{(k,1)} & \cdots & a_{n,n}^{(k,1)} \end{bmatrix}.$$

If the updated $(k+1)$st pivot entry, $a_{k+1,k+1}^{(k,1)}$, is equal to zero, we set $\mathsf{L}_{k+1} = \mathsf{I}_n$. Otherwise, construct the column-$(k+1)$ complete elementary matrix

$$\mathsf{L}_{k+1} = \mathsf{I}_n + \sum_{r=k+2}^{n} \mu_{r,k+1}\mathsf{M}^{(r,k+1)},$$

where

$$\mu_{r,k+1} = -\frac{a_{r,k+1}^{(k,1)}}{a_{k+1,k+1}^{(k,1)}}.$$

Then set

$$\mathsf{L}_{k+1}\mathsf{A}^{(k,1)} = \mathsf{A}^{(k+1)},$$

obtaining

$$\mathsf{A}^{(k+1)} = \begin{bmatrix} a_{1,1}^{(0,1)} & a_{1,2}^{(0,1)} & a_{1,3}^{(0,1)} & a_{1,4}^{(0,1)} & \cdots & a_{1,k+1}^{(0,1)} & a_{1,k+2}^{(0,1)} & \cdots & a_{1,n}^{(0,1)} \\ 0 & a_{2,2}^{(1,1)} & a_{2,3}^{(1,1)} & a_{2,4}^{(1,1)} & \cdots & a_{2,k+1}^{(1,1)} & a_{2,k+2}^{(1,1)} & \cdots & a_{2,n}^{(1,1)} \\ 0 & 0 & a_{3,3}^{(2,1)} & a_{3,4}^{(2,1)} & \cdots & a_{3,k+1}^{(2,1)} & a_{3,k+2}^{(2,1)} & \cdots & a_{3,n}^{(2,1)} \\ \vdots & & \ddots & \ddots & \ddots & \vdots & \vdots & & \vdots \\ 0 & & & 0 & a_{k,k}^{(k-1,1)} & a_{k,k+1}^{(k-1,1)} & a_{k,k+2}^{(k-1,1)} & \cdots & a_{k,n}^{(k-1,1)} \\ 0 & & & 0 & 0 & a_{k+1,k+1}^{(k,1)} & a_{k+1,k+2}^{(k,1)} & \cdots & a_{k+1,n}^{(k,1)} \\ 0 & & & 0 & 0 & 0 & a_{k+2,k+2}^{(k+1)} & \cdots & a_{k+2,n}^{(k+1)} \\ \vdots & & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & a_{n,k+2}^{(k+1)} & \cdots & a_{n,n}^{(k+1)} \end{bmatrix}.$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Theorem 3.21 (LU factorization with pivoting)**.** *Suppose that $n \geq 2$ and $\mathsf{A} \in \mathbb{C}^{n\times n}$. The Gaussian elimination with maximal column pivoting algorithm always proceeds to completion to yield an upper triangular matrix U. In particular, there are matrices $\mathsf{L}_j, \mathsf{P}_j \in \mathbb{C}^{n\times n}$, $j = 1, \ldots, n-1$ such that*

$$\mathsf{L}_{n-1}\mathsf{P}_{n-1}\cdots\mathsf{L}_2\mathsf{P}_2\mathsf{L}_1\mathsf{P}_1\mathsf{A} = \mathsf{U},$$

where L_j is a column-j complete elementary matrix and P_j is either the $n \times n$ identity or a simple permutation matrix. Furthermore, there are column-j complete elementary matrices $\hat{\mathsf{L}}_j$, for $j = 1, \ldots, n-1$, and a permutation matrix P such that

$$\hat{\mathsf{L}}_{n-1}\cdots\hat{\mathsf{L}}_1\mathsf{P}\mathsf{A} = \mathsf{U},$$

where

$$\mathsf{P} = \mathsf{P}_{n-1}\cdots\mathsf{P}_1,$$

$$\hat{\mathsf{L}}_j = \mathsf{P}_{n-1}\cdots\mathsf{P}_{j+1}\mathsf{L}_j\mathsf{P}_{j+1}\cdots\mathsf{P}_{n-1}, \quad j = 1, \ldots, n-2,$$

and

$$\hat{\mathsf{L}}_{n-1} = \mathsf{L}_{n-1}.$$

Finally, there is a unit lower triangular matrix L such that

$$\mathsf{P}\mathsf{A} = \mathsf{L}\mathsf{U}.$$

Proof. This is nothing but an exercise in applying our previous results and definitions; see Problem 3.10. □

Listing 3.4 computes the LU factorization with pivoting. As before, and for a different perspective, it is also possible to prove the existence of the factorization $\mathsf{A} = \mathsf{P}^{\mathsf{T}}\mathsf{L}\mathsf{U}$ independent of the consideration of the Gaussian elimination algorithm. Here, we follow the presentation in [89].

Theorem 3.22 (LU factorization with pivoting)**.** *Suppose that $n \geq 2$. Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be given. There exists a permutation matrix $\mathsf{P} \in \mathbb{R}^{n\times n}$, a unit lower triangular matrix $\mathsf{L} \in \mathbb{C}^{n\times n}$, and an upper triangular matrix $\mathsf{U} \in \mathbb{C}^{n\times n}$ such that*

$$\mathsf{P}\mathsf{A} = \mathsf{L}\mathsf{U}.$$

Proof. We proceed by induction.

($n = 2$) Consider

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

If $a \neq 0$, then set $\mathsf{P} = \mathsf{I}_2$,

$$\mathsf{L} = \begin{bmatrix} 1 & 0 \\ \frac{c}{a} & 1 \end{bmatrix}, \qquad \mathsf{U} = \begin{bmatrix} a & b \\ 0 & d - \frac{c}{a}b \end{bmatrix}.$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Clearly, $\mathsf{PA} = \mathsf{LU}$. On the other hand, if $a = 0$, but $c \neq 0$, set

$$\mathsf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \mathsf{L} = \mathsf{I}_2, \qquad \mathsf{U} = \begin{bmatrix} c & d \\ 0 & b \end{bmatrix},$$

and again observe that $\mathsf{PA} = \mathsf{LU}$. Finally, if $a = 0 = c$, there is not much to do: set

$$\mathsf{P} = \mathsf{I}_2 = \mathsf{L}, \qquad \mathsf{U} = \begin{bmatrix} 0 & b \\ 0 & d \end{bmatrix},$$

and conclude the result.

($n = m$) The induction hypothesis is to suppose that the result is true for every matrix $\mathsf{A} \in \mathbb{C}^{k \times k}$ for all $k = 2, \ldots, m$.

($n = m+1$) Suppose that $\mathsf{A} \in \mathbb{C}^{(m+1)\times(m+1)}$ is arbitrary. Suppose that, in the first column of A, the largest element by modulus is contained in row r. Set P_1 as the simple permutation matrix interchanging rows 1 and r. Then the result is, for some $\boldsymbol{p}, \boldsymbol{w} \in \mathbb{C}^m$ and $\mathsf{B} \in \mathbb{C}^{m \times m}$,

$$\mathsf{P}_1\mathsf{A} = \begin{bmatrix} \alpha & \boldsymbol{w}^\mathsf{T} \\ \boldsymbol{p} & \mathsf{B} \end{bmatrix},$$

where

$$|\alpha| \geq |[\boldsymbol{p}]_i|, \quad i = 1, \ldots, m.$$

Observe that it is possible that $\alpha = 0$, in which case $\boldsymbol{p} = \boldsymbol{0}$. Next, we seek a solution, if possible, to the following intermediate problem:

$$\mathsf{P}_1\mathsf{A} = \begin{bmatrix} \alpha & \boldsymbol{w}^\mathsf{T} \\ \boldsymbol{p} & \mathsf{B} \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{0}^\mathsf{T} \\ \boldsymbol{m} & \mathsf{I}_m \end{bmatrix} \begin{bmatrix} \alpha & \boldsymbol{v}^\mathsf{T} \\ \boldsymbol{0} & \mathsf{C} \end{bmatrix},$$

where

$$\mathsf{C} \in \mathbb{C}^{m \times m}, \quad \boldsymbol{m}, \boldsymbol{v}, \boldsymbol{0} \in \mathbb{C}^m.$$

The block matrix equation is satisfied if and only if

$$\boldsymbol{v} = \boldsymbol{w}, \quad \alpha\boldsymbol{m} = \boldsymbol{p}, \quad \mathsf{C} = \mathsf{B} - \boldsymbol{m}\boldsymbol{v}^\mathsf{T}.$$

Recall that, if $\alpha = 0$, $\boldsymbol{p} = \boldsymbol{0}$. In this case,

$$\boldsymbol{m} = \boldsymbol{0}, \quad \boldsymbol{v} = \boldsymbol{w}, \quad \mathsf{C} = \mathsf{B}$$

is one possible solution. If $\alpha \neq 0$, then

$$\boldsymbol{m} = \frac{1}{\alpha}\boldsymbol{p}, \quad \boldsymbol{v} = \boldsymbol{w}, \quad \mathsf{C} = \mathsf{B} - \frac{1}{\alpha}\boldsymbol{p}\boldsymbol{w}^\mathsf{T}.$$

Observe that

$$|[\boldsymbol{m}]_i| \leq 1, \quad i = 1, \ldots, m.$$

Now, from the induction hypothesis, there is a permutation matrix $\tilde{\mathsf{P}} \in \mathbb{C}^{m \times m}$, a unit upper triangular matrix $\tilde{\mathsf{L}} \in \mathbb{C}^{m \times m}$, and an upper triangular matrix $\tilde{\mathsf{U}} \in \mathbb{C}^{m \times m}$ such that

$$\tilde{\mathsf{P}}\mathsf{C} = \tilde{\mathsf{L}}\tilde{\mathsf{U}}.$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Finally, the reader will observe that

$$\begin{bmatrix} 1 & \mathbf{0}^\intercal \\ \mathbf{0} & \tilde{\mathsf{P}}^\intercal \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^\intercal \\ \tilde{\mathsf{P}}\boldsymbol{m} & \tilde{\mathsf{L}} \end{bmatrix} \begin{bmatrix} \alpha & \boldsymbol{v}^\intercal \\ \mathbf{0} & \tilde{\mathsf{U}} \end{bmatrix} = \mathsf{P}_1\mathsf{A},$$

using the fact that $\tilde{\mathsf{P}}^\intercal = \tilde{\mathsf{P}}^{-1}$. Therefore,

$$\mathsf{LU} = \mathsf{PA},$$

where

$$\mathsf{P} = \begin{bmatrix} 1 & \mathbf{0}^\intercal \\ \mathbf{0} & \tilde{\mathsf{P}} \end{bmatrix} \mathsf{P}_1, \qquad \mathsf{L} = \begin{bmatrix} 1 & \mathbf{0}^\intercal \\ \tilde{\mathsf{P}}\boldsymbol{m} & \tilde{\mathsf{L}} \end{bmatrix}, \qquad \mathsf{U} = \begin{bmatrix} \alpha & \boldsymbol{v}^\intercal \\ \mathbf{0} & \tilde{\mathsf{U}} \end{bmatrix},$$

and the matrices are of the required types. □

The LU factorization can be used to efficiently compute the solution to (3.1). Let us suppose that A is nonsingular. The algorithm is as follows: if $\mathsf{PA} = \mathsf{LU}$, then we have the equivalent system

$$\mathsf{L}\boldsymbol{y} = \mathsf{P}\boldsymbol{f} = \boldsymbol{q}, \qquad \mathsf{U}\boldsymbol{x} = \boldsymbol{y}. \tag{3.9}$$

We use the forward substitution algorithm to solve $\mathsf{L}\boldsymbol{y} = \boldsymbol{q}$ for $\boldsymbol{y}$ and, then, we use the backward substitution algorithm to solve $\mathsf{U}\boldsymbol{x} = \boldsymbol{y}$ for $\boldsymbol{x}$. These algorithms were covered in Section 3.1.2.

3.4 Implementation of the LU Factorization

We conclude the discussion of Gaussian elimination with some practical considerations. First of all, one does not need to store the permutation matrix P. Instead one only needs to remember which rows were swapped. This means that we only need to store a vector $\boldsymbol{s} \in \mathbb{N}^n$ of indices which is used to indirectly reference the entries of the matrix. Second, since the matrices L and U are lower and upper triangular, respectively, and the diagonal entries of L are always equal to one, both of these matrices can be conveniently stored in the already allocated array for A. This is a convenient and efficient way of storing the LU factorization of the system matrix A. Listing 3.5 provides an implementation of this idea.

Once this factorization has taken place, we can use it, together with the swap vector $\boldsymbol{s}$, to perform a back substitution and find the solution to system (3.1), as described by (3.9). Listing 3.6 provides the implementation details for this. It is important to note that the LU factorization, the most expensive part of this procedure, needs to only be called once. After that, solving several systems of the form (3.1) where the system matrix A does not change, but the right-hand-side $\boldsymbol{f}$ does, is rather efficient, as it only requires n^2 operations. See Problem 3.2. This is a situation that is very common in practice. See, for instance, Chapters 24–28.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

3.5 Special Matrices

While Theorems 3.21 and 3.22 provide, in general, the existence and uniqueness of an LU factorization with pivoting, it is of interest to study this process for some special kinds of matrix that often appear in applications.

3.5.1 Diagonally Dominant Matrices

Definition 3.23 (diagonal dominance)**.** A matrix $\mathsf{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ is called **diagonally dominant** if and only if

$$|a_{i,i}| \geq \sum_{\substack{k=1\\k\neq i}}^{n} |a_{i,k}| \,, \quad \forall i = 1, \ldots, n.$$

A is **strictly diagonally dominant** (SDD) if and only if

$$|a_{i,i}| > \sum_{\substack{k=1\\k\neq i}}^{n} |a_{i,k}| \,, \quad \forall i = 1, \ldots, n.$$

A is called **strictly diagonally dominant of dominance** δ if and only if there is a $\delta > 0$ such that

$$|a_{i,i}| \geq \delta + \sum_{\substack{k=1\\k\neq i}}^{n} |a_{i,k}| \,, \quad \forall i = 1, \ldots, n.$$

The reader should verify that, essentially, the last two definitions are equivalent. It turns out that SDD matrices are always invertible, and we have an easy bound on the norm of its inverse.

Theorem 3.24 (properties of an SDD matrix)**.** *If* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is SDD, then* A *is invertible. If* A *is SDD of dominance* $\delta > 0$*, then*

$$\left\|\mathsf{A}^{-1}\right\|_\infty < \frac{1}{\delta}.$$

Proof. Suppose that A is singular. If that is the case there is an $\boldsymbol{x} = [x_i] \in \mathbb{C}^n_\star$ such that $\mathsf{A}\boldsymbol{x} = \mathbf{0}$. Suppose that $k \in \{1, \ldots, n\}$ is an index for which $|x_k| = \|\boldsymbol{x}\|_\infty$. Since $\mathsf{A}\boldsymbol{x} = \mathbf{0}$, we must have that, for each $i = 1, \ldots, n$,

$$\sum_{j=1}^{n} a_{i,j}x_j = 0.$$

In particular, $\sum_{j=1}^{n} a_{k,j}x_j = 0$. Then, from the triangle inequality,

$$|a_{k,k}| \cdot \|\boldsymbol{x}\|_\infty = |a_{k,k}x_k| = \left| -\sum_{\substack{j=1\\j\neq k}}^{n} a_{k,j}x_j \right| \leq \sum_{\substack{j=1\\j\neq k}}^{n} |a_{k,j}| \cdot |x_j| \leq \|\boldsymbol{x}\|_\infty \sum_{\substack{j=1\\j\neq k}}^{n} |a_{k,j}|.$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Since $\|\boldsymbol{x}\|_\infty > 0$, we have

$$|a_{k,k}| \leq \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{k,j}|.$$

This proves that A is not SDD, a contradiction.

Next, suppose that A has dominance $\delta > 0$. Let $\boldsymbol{x}$ be arbitrary. Set $\mathsf{A}\boldsymbol{x} = \boldsymbol{f}$. Assume that $\|\boldsymbol{x}\|_\infty = |x_k|$, for some $k = 1, \ldots, n$. Then

$$a_{k,1}x_1 + \cdots + a_{k,k}x_k + \cdots + a_{k,n}x_n = f_k$$

and, using the reverse triangle inequality,

$$|f_k| \geq |a_{k,k}||x_k| - \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{j,k}||x_j| \geq \left(|a_{k,k}| - \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{j,k}| \right) |x_k| \geq \delta \|\boldsymbol{x}\|_\infty.$$

This shows that

$$\frac{\|\mathsf{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} \geq \delta, \quad \forall \boldsymbol{x} \in \mathbb{C}^n,$$

which is equivalent to

$$\frac{1}{\delta} \geq \frac{\left\|\mathsf{A}^{-1}\boldsymbol{w}\right\|_\infty}{\|\boldsymbol{w}\|_\infty}, \quad \forall \boldsymbol{w} \in \mathbb{C}^n_\star.$$

This, in turn, implies that

$$\frac{1}{\delta} \geq \sup_{\boldsymbol{w}\in\mathbb{C}^n_\star} \frac{\left\|\mathsf{A}^{-1}\boldsymbol{w}\right\|_\infty}{\|\boldsymbol{w}\|_\infty} = \left\|\mathsf{A}^{-1}\right\|_\infty,$$

as we intended to show. □

Theorem 3.25 (Gaussian elimination and SDD)**.** *Let* $\mathsf{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ *be SDD and assume that it is represented as*

$$\mathsf{A} = \begin{bmatrix} \alpha & \boldsymbol{v}^\mathsf{T} \\ \boldsymbol{p} & \hat{\mathsf{A}} \end{bmatrix},$$

where $\alpha \in \mathbb{C}$, $\boldsymbol{p}, \boldsymbol{v} \in \mathbb{C}^{n-1}$, *and* $\hat{\mathsf{A}} = [\hat{a}_{i,j}] \in \mathbb{C}^{(n-1)\times(n-1)}$. *After one step of Gaussian elimination (without pivoting),* A *will be reduced to the matrix*

$$\begin{bmatrix} \alpha & \boldsymbol{v}^\mathsf{T} \\ \boldsymbol{0} & \mathsf{B} \end{bmatrix},$$

where $\mathsf{B} = [b_{i,j}] \in \mathbb{C}^{(n-1)\times(n-1)}$ *is SDD.*

Proof. Let us construct a matrix $\mathsf{L} \in \mathbb{C}^{n\times n}$ such that

$$\mathsf{LA} = \begin{bmatrix} \alpha & \boldsymbol{v}^\mathsf{T} \\ \boldsymbol{0} & \mathsf{B} \end{bmatrix},$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

if possible. Consider

$$\mathsf{L} = \begin{bmatrix} 1 & \mathbf{0}^\mathsf{T} \\ \boldsymbol{m} & \mathsf{I}_{n-1} \end{bmatrix}.$$

Then

$$\mathsf{LA} = \begin{bmatrix} 1 & \mathbf{0}^\mathsf{T} \\ \boldsymbol{m} & \mathsf{I}_{n-1} \end{bmatrix} \begin{bmatrix} \alpha & \boldsymbol{v}^\mathsf{T} \\ \boldsymbol{p} & \hat{\mathsf{A}} \end{bmatrix} = \begin{bmatrix} \alpha & \boldsymbol{v}^\mathsf{T} \\ \alpha\boldsymbol{m} + \boldsymbol{p} & \boldsymbol{m}\boldsymbol{v}^\mathsf{T} + \hat{\mathsf{A}} \end{bmatrix}.$$

Note that, since A is SDD, $\alpha \neq 0$ and $\hat{\mathsf{A}}$ is SDD. Choosing $\boldsymbol{m} = -\alpha^{-1}\boldsymbol{p}$, we have

$$\mathsf{LA} = \begin{bmatrix} \alpha & \boldsymbol{v}^\mathsf{T} \\ \mathbf{0} & \hat{\mathsf{A}} - \alpha^{-1}\boldsymbol{p}\boldsymbol{v}^\mathsf{T} \end{bmatrix}.$$

Having successfully constructed L, we find $\mathsf{B} = \hat{\mathsf{A}} - \alpha^{-1}\boldsymbol{p}\boldsymbol{v}^\mathsf{T}$.

All that remains is to show that B is SDD. To see that this is the case, consider for row i of B,

$$\begin{aligned}
\sum_{\substack{j=1 \\ j\neq i}}^{n-1} |b_{i,j}| &= \sum_{\substack{j=1 \\ j\neq i}}^{n-1} \left|\hat{a}_{i,j} - \alpha^{-1} p_i v_j\right| \\
&\leq \sum_{\substack{j=1 \\ j\neq i}}^{n-1} |\hat{a}_{i,j}| + \sum_{\substack{j=1 \\ j\neq i}}^{n-1} \left|\alpha^{-1} p_i v_j\right| \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{n-1} |\hat{a}_{i,j}| + \frac{|p_i|}{|\alpha|} \sum_{\substack{j=1 \\ j\neq i}}^{n-1} |v_j| \\
&= \sum_{\substack{j=1 \\ j\neq i}}^{n-1} |a_{i+1,j+1}| + \frac{|a_{i+1,1}|}{|a_{1,1}|} \sum_{\substack{j=1 \\ j\neq i}}^{n-1} |a_{1,j+1}| \\
&= \sum_{\substack{j=2 \\ j\neq i+1}}^{n} |a_{i+1,j}| + \frac{|a_{i+1,1}|}{|a_{1,1}|} \sum_{\substack{j=2 \\ j\neq i+1}}^{n} |a_{1,j}| \\
&= \sum_{\substack{j=1 \\ j\neq i+1}}^{n} |a_{i+1,j}| - |a_{i+1,1}| + \frac{|a_{i+1,1}|}{|a_{1,1}|} \sum_{j=2}^{n} |a_{1,j}| - \frac{|a_{i+1,1}| \cdot |a_{1,i+1}|}{|a_{1,1}|}.
\end{aligned}$$

Now, since A is SDD, we can continue this string of inequalities to obtain

$$\begin{aligned}
\sum_{\substack{j=1 \\ j\neq i}}^{n-1} |b_{i,j}| &< |a_{i+1,i+1}| - |a_{i+1,1}| + |a_{i+1,1}| - \frac{|a_{i+1,1}| \cdot |a_{1,i+1}|}{|a_{1,1}|} \\
&= |a_{i+1,i+1}| - \frac{|a_{i+1,1}| \cdot |a_{1,i+1}|}{|a_{1,1}|} \\
&\leq \left| a_{i+1,i+1} - \frac{a_{i+1,1} a_{1,i+1}}{a_{1,1}} \right| \\
&= |b_{i,i}|,
\end{aligned}$$

where we used the reverse triangle inequality. Thus, as claimed, B is SDD. □

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Corollary 3.26 (SDD of magnitude δ)**.** *If* $A \in \mathbb{C}^{n\times n}$ *is SDD of magnitude* $\delta > 0$*, then* $B \in \mathbb{C}^{(n-1)\times(n-1)}$*, introduced above, is SDD of magnitude* δ*.*

Proof. The details of our last proof still hold up to the point where we apply the fact that A is SDD. Thus,

$$\begin{aligned}
\sum_{\substack{j=1\\ j\neq i}}^{n-1} |b_{i,j}| &\leq \sum_{\substack{j=1\\ j\neq i+1}}^{n} |a_{i+1,j}| - |a_{i+1,1}| + \frac{|a_{i+1,1}|}{|a_{1,1}|}\sum_{j=2}^{n} |a_{1,j}| - \frac{|a_{i+1,1}|\cdot|a_{1,i+1}|}{|a_{1,1}|} \\
&\leq |a_{i+1,i+1}| - \delta - |a_{i+1,1}| + \frac{|a_{i+1,1}|}{|a_{1,1}|}(|a_{1,1}| - \delta) - \frac{|a_{i+1,1}|\cdot|a_{1,i+1}|}{|a_{1,1}|} \\
&\leq |a_{i+1,i+1}| - \delta - |a_{i+1,1}| + |a_{i+1,1}| - \frac{|a_{i+1,1}|\cdot|a_{1,i+1}|}{|a_{1,1}|} \\
&= |a_{i+1,i+1}| - \frac{|a_{i+1,1}|\cdot|a_{1,i+1}|}{|a_{1,1}|} - \delta \\
&\leq \left| a_{i+1,i+1} - \frac{a_{i+1,1}a_{1,i+1}}{a_{1,1}} \right| - \delta \\
&= |b_{i,i}| - \delta.
\end{aligned}$$

Thus, B is SDD of magnitude δ. □

Corollary 3.27 (Gaussian elimination and SDD)**.** *If* $A \in \mathbb{C}^{n\times n}$ *is SDD, then Gaussian elimination without pivoting applied to* A *proceeds to completion without encountering any zero pivot elements.*

Proof. We only need to proceed recursively using the previous two results. □

Remark 3.28 (modified Gaussian elimination)**.** In a variant of Gaussian elimination, the pivot entry is normalized to one. Namely, given the system of linear equations

$$\begin{cases}
a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = f_1, \\
a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n = f_2, \\
\qquad\qquad\qquad\qquad\qquad\quad \vdots \\
a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,n}x_n = f_n,
\end{cases}$$

we could proceed as follows.

1. Using the first equation, express x_1 in terms of all the other variables to obtain
$$x_1 = a_{1,2}^{(1)}x_2 + \cdots + a_{1,n}^{(1)}x_n + f_1^{(1)}.$$
Eliminate x_1 from equations indexed 2 through n.
2. Using the second equation, express x_2 only in terms of $x_3, \ldots, x_n$:
$$x_2 = a_{2,3}^{(2)}x_3 + \cdots + a_{2,n}^{(2)}x_n + f_2^{(2)}.$$
Eliminate x_2 from equations indexed 3 through n.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

3. Using the third equation, express x_3 only in terms of $x_4, \ldots, x_n$:

$$x_3 = a_{3,4}^{(3)} x_4 + \cdots + a_{3,n}^{(3)} x_n + f_3^{(3)}.$$

Eliminate x_3 from equations indexed 4 through n.

i. For $i = 4, \ldots, n-1$, using the ith equation, express x_i only in terms of the variables $x_{i+1}, \ldots, x_n$:

$$x_i = a_{i,i+1}^{(i)} x_{i+1} + \cdots + a_{i,n}^{(i)} x_n + f_i^{(i)}. \tag{3.10}$$

Eliminate x_i from equations indexed $i+1$ through n.

n. Using the last remaining equation, express x_n as

$$x_n = f_n^{(n)}.$$

$n+1$. Use back substitution to solve for $\boldsymbol{x}$.

We will call this the *modified Gaussian elimination* process. It works as long as no zero pivots are encountered. For systems whose coefficient matrix is SDD, this process yields an interesting and desirable property. The following two results address this case and give a new perspective to our methodology.

Lemma 3.29 (modified Gaussian elimination)**.** *Let $n \geq 2$ and $\mathsf{A}\boldsymbol{x} = \boldsymbol{f}$ be a system of n equations with n unknowns, where the coefficient matrix $\mathsf{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ is SDD of magnitude $\delta > 0$. Then one may reduce the first equation to the form*

$$x_1 = a_{1,2}^{(1)} x_2 + \cdots + a_{1,n}^{(1)} x_n + f_1^{(1)} \tag{3.11}$$

for some coefficients $a_{1,j}^{(1)}$, $j = 2, \ldots, n$, and $f_1^{(1)}$, with no division by zero. Moreover,

$$\sum_{j=2}^{n} |a_{1,j}^{(1)}| < 1.$$

Finally, x_1 can be eliminated from equations indexed 2 through n to obtain a subsystem, $\mathsf{A}^{(1)}\boldsymbol{x}^{(1)} = \boldsymbol{f}^{(1)}$, of $n-1$ equations with $n-1$ unknowns. The coefficient matrix $\mathsf{A}^{(1)}$ has diagonal dominance of magnitude $\delta > 0$.

Proof. Since the matrix A has diagonal dominance of magnitude $\delta > 0$, we observe that

$$|a_{1,1}| \geq \delta + |a_{1,2}| + \cdots + |a_{1,n}| > 0,$$

so that $a_{1,1} \neq 0$, and we can define $a_{1,j}^{(1)} = -a_{1,j}/a_{1,1}$, for $j = 2, \ldots, n$, and $f_1^{(1)} = f_1/a_{1,1}$. Notice also that

$$\sum_{j=2}^{n} |a_{1,j}^{(1)}| = \frac{|a_{1,2}| + \cdots + |a_{1,n}|}{|a_{1,1}|} < 1.$$

Now, to eliminate x_1 from the system, we substitute (3.11) into all the remaining equations. The ith equation, for $i = 2, \ldots, n$, reads

$$a_{i,1}x_1 + a_{i,2}x_2 + \cdots + a_{i,n}x_n = f_i,$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

and, when we substitute for x_1, we find, for $i = 2, \ldots, n$,

$$(a_{i,2} + a_{i,1}a_{1,2}^{(1)})x_2 + (a_{i,3} + a_{i,1}a_{1,3}^{(1)})x_3 + \cdots + (a_{i,n} + a_{i,1}a_{1,n}^{(1)})x_n = f_i^{(1)},$$

where

$$f_i^{(1)} = f_i - a_{i,1}f_1^{(1)}, \quad i = 2, \ldots, n.$$

It is convenient to index the matrix $\mathsf{A}^{(1)}$ starting at row and column 2 and ending at row and column n. In this case, the entries of $\mathsf{A}^{(1)}$ are

$$\left[\mathsf{A}^{(1)}\right]_{i,j} = a_{i,j} + a_{i,1}a_{1,j}^{(1)}, \quad i, j = 2, \ldots, n.$$

The proof that $\mathsf{A}^{(1)}$ is SDD of magnitude $\delta > 0$ follows as before. □

We can now provide a sufficient condition for our modified Gaussian elimination process to proceed to completion and include some further results.

Theorem 3.30 (modified Gaussian elimination, SDD case)**.** *Let $n \geq 2$. Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ has diagonal dominance of magnitude $\delta > 0$ and $\boldsymbol{f} \in \mathbb{C}^n$ is given. Then the modified Gaussian elimination process (without pivoting) used to solve $\mathsf{A}\boldsymbol{x} = \boldsymbol{f}$ does not fail. Moreover, we have that, at every step,*

$$\sum_{j=i+1}^{n} |a_{i,j}^{(i)}| < 1, \quad i = 1, \ldots, n-1,$$

and

$$|f_i^{(i)}| \leq \frac{2}{\delta}\|\boldsymbol{f}\|_\infty, \quad i = 1, \ldots, n.$$

Proof. To prove the first part, proceed by induction, using the previous lemma to show that the process does not fail. Invoking the fact that $\mathsf{A}^{(i-1)}$ is SDD of magnitude $\delta > 0$, we can always conclude that

$$\sum_{j=i+1}^{n} |a_{i,j}^{(i)}| < 1, \quad i = 1, \ldots, n-1.$$

For the second part, notice that, since A has diagonal dominance of magnitude $\delta > 0$, we can recall Theorem 3.24, which indicates

$$\|\boldsymbol{x}\|_\infty = \|\mathsf{A}^{-1}\boldsymbol{f}\|_\infty \leq \|\mathsf{A}^{-1}\|_\infty\|\boldsymbol{f}\|_\infty \leq \frac{1}{\delta}\|\boldsymbol{f}\|_\infty.$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Now, from (3.10), we see that

$$\begin{aligned}
|f_i^{(i)}| &= \left| x_i - \sum_{j=i+1}^{n} a_{i,j}^{(i)} x_j \right| \\
&\leq |x_i| + \sum_{j=i+1}^{n} |a_{i,j}^{(i)}||x_j| \\
&\leq \left(1 + \sum_{j=i+1}^{n} |a_{i,j}^{(i)}| \right) \|\boldsymbol{x}\|_\infty \\
&\leq 2\|\boldsymbol{x}\|_\infty \\
&\leq \frac{2}{\delta}\|\boldsymbol{f}\|_\infty,
\end{aligned}$$

as we intended to show. □

3.5.2 Positive Definite Matrices

Definition 3.31 (HPD matrices)**.** A matrix $\mathsf{A} \in \mathbb{C}^{n\times n}$ is called **Hermitian positive semi-definite** (HPSD) if and only if $\mathsf{A} = \mathsf{A}^{\mathsf{H}}$ and

$$\boldsymbol{x}^{\mathsf{H}}\mathsf{A}\boldsymbol{x} \geq 0, \quad \forall \boldsymbol{x} \in \mathbb{C}^n.$$

A is called **Hermitian positive definite** (HPD) if and only if $\mathsf{A} = \mathsf{A}^{\mathsf{H}}$ and

$$\boldsymbol{x}^{\mathsf{H}}\mathsf{A}\boldsymbol{x} > 0, \quad \forall \boldsymbol{x} \in \mathbb{C}_\star^n.$$

Remark 3.32 (notation)**.** We often use the symbol $\mathbb{C}_{\text{Her}}^{n\times n}$ to denote the vector space of Hermitian matrices. For real, symmetric matrices we use the notation $\mathbb{R}_{\text{sym}}^{n\times n}$.

Theorem 3.33 (properties of HPD matrices)**.** *Suppose that* $\mathsf{A} = [a_{i,j}] \in \mathbb{C}^{n\times n}$ *is HPD. Then*

1. $a_{i,i} > 0$ *for all* $i = 1, \ldots, n$.
2. $\sigma(\mathsf{A}) \subset (0, \infty)$.
3. $\det(\mathsf{A}) > 0$ *and* $\operatorname{tr}(\mathsf{A}) = \sum_{i=1}^{n} a_{i,i} > 0$.
4. *For all* $\emptyset \neq S \subseteq \{1, \ldots, n\}$, *we have that* $\mathsf{A}(S)$ *is HPD.*
5. $|a_{i,j}|^2 \leq a_{i,i}a_{j,j}$ *for all* $i \neq j$.
6. $\max_{1\leq i,j\leq n} |a_{i,j}| \leq \max_{1\leq i\leq n} |a_{i,i}|$.

Proof. We will prove statements 4–6 and leave the others for exercises; see Problem 3.13.
4: Given $S \subseteq \{1, \ldots, n\}$ nonempty. Suppose that $\boldsymbol{y} \in \mathbb{C}_\star^{\#(S)}$ is arbitrary. Define $\boldsymbol{x} \in \mathbb{C}_\star^n$ such that $x_i = 0$ for all $i \notin S$; otherwise,

$$x_{i_k} = y_k, \quad k = 1, \ldots, \#(S),$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

where

$$S = \{i_1, \ldots, i_{\#(S)}\}$$

and

$$i_k < i_{k+1}, \quad k = 1, \ldots, \#(S) - 1.$$

Then $\mathsf{A}(S) = \mathsf{A}(S)^{\mathsf{H}}$ and

$$\boldsymbol{y}^{\mathsf{H}}\mathsf{A}(S)\boldsymbol{y} = \boldsymbol{x}^{\mathsf{H}}\mathsf{A}\boldsymbol{x} > 0, \quad \boldsymbol{y} \in \mathbb{C}_{\star}^{\#(S)}.$$

5: Let $S = \{r, s\}$, $r < s$. Then

$$\mathsf{A}(S) = \begin{bmatrix} a_{r,r} & a_{r,s} \\ a_{s,r} & a_{s,s} \end{bmatrix};$$

consequently,

$$0 < \det(\mathsf{A}(S)) = a_{r,r}a_{s,s} - |a_{r,s}|^2.$$

6: We argue by contradiction. Suppose that the element with largest modulus is off-diagonal in, say, row r and column s, $r \neq s$. Then

$$|a_{r,s}| > a_{r,r}$$

and

$$|a_{r,s}| > a_{s,s}.$$

Thus,

$$|a_{r,s}|^2 > a_{r,r}a_{s,s},$$

which is a contradiction. □

Theorem 3.34 (factorization of HPD matrices)**.** *Let $n \geq 2$. Suppose that $\mathsf{A} \in \mathbb{C}^{n\times n}$ is HPD. There exists a unit lower triangular matrix $\mathsf{L} \in \mathbb{C}^{n\times n}$ and a diagonal matrix $\mathsf{D} \in \mathbb{R}^{n\times n}$ with positive diagonal entries such that*

$$\mathsf{A} = \mathsf{L}\mathsf{D}\mathsf{L}^{\mathsf{H}}.$$

Proof. Since A is HPD, for $k = 1, \ldots, n-1$ the principal sub-matrices $\mathsf{A}^{(k)} \in \mathbb{C}^{k\times k}$ are invertible. By Theorem 3.7 there exists a unit lower triangular matrix L and an upper triangular matrix U such that $\mathsf{A} = \mathsf{L}\mathsf{U}$.

Next, we claim that all of the diagonal elements of U are real and positive. This follows by an induction argument, as in the proof of Theorem 3.7, and the fact that $\det(\mathsf{A}) = \det(\mathsf{U})$. The details are left to the reader see Problem 3.14.

Now set $\mathsf{D} = \operatorname{diag}(u_{1,1}, \ldots, u_{n,n})$ and $\tilde{\mathsf{U}} = \mathsf{D}^{-1}\mathsf{U}$. Then

$$\mathsf{A} = \mathsf{L}\mathsf{D}\mathsf{D}^{-1}\mathsf{U} = \mathsf{L}\mathsf{D}\tilde{\mathsf{U}}.$$

It follows that, for $i = 1, \ldots, n$, $\tilde{u}_{i,i} = 1$. Taking the conjugate transpose, we have

$$\tilde{\mathsf{U}}^{\mathsf{H}}\mathsf{D}\mathsf{L}^{\mathsf{H}} = \mathsf{A}^{\mathsf{H}} = \mathsf{A} = \mathsf{L}\mathsf{D}\tilde{\mathsf{U}}.$$

It follows that

$$\mathsf{L}^{-1}\tilde{\mathsf{U}}^{\mathsf{H}}\mathsf{D} = \mathsf{D}\tilde{\mathsf{U}}\mathsf{L}^{-\mathsf{H}}. \tag{3.12}$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Recall that, by Theorem 3.4, $L^{-1}\tilde{U}^H$ must be unit lower triangular and $\tilde{U}L^{-H}$ must be unit upper triangular. The only way for (3.12) to hold is for $L^{-1}\tilde{U}^H$ and $\tilde{U}L^{-H}$ to be diagonal. But, as these products must be unit triangular, they are both equal to the identity. In other words,

$$L = \tilde{U}^H,$$

and we have proven that

$$A = LDL^H. \qquad \square$$

Corollary 3.35 (Cholesky factorization[3])**.** *Let $n \geq 2$. Suppose that $A \in \mathbb{C}^{n\times n}$ is HPD. Then there is a lower triangular matrix $L \in \mathbb{C}^{n\times n}$ such that*

$$A = LL^H.$$

This is known as the Cholesky factorization.

Proof. From the last theorem, there is a unit lower triangular matrix $\tilde{L}$ and a diagonal matrix D with positive real diagonal entries such that

$$A = \tilde{L}D\tilde{L}^H.$$

Suppose that $D = \operatorname{diag}(d_1, \ldots, d_n)$. Define $\tilde{D} = \operatorname{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$. Then, setting $L = \tilde{L}\tilde{D}$, we see that

$$A = LL^H.$$

The proof is complete. $\square$

Theorem 3.36 (uniqueness of Cholesky factorization)**.** *Let $n \geq 2$. Suppose that $A \in \mathbb{C}^{n\times n}$ is HPD. Then there is a unique lower triangular matrix L such that the diagonal entries of L are positive real numbers and*

$$A = LL^H.$$

In other words, the Cholesky factorization is unique.

Proof. Suppose that there are two lower triangular matrices L_1 and L_2 with positive real diagonal entries and

$$L_1L_1^H = A = L_2L_2^H.$$

Then

$$L_2^{-1}L_1 = L_2^HL_1^{-H};$$

by Theorem 3.4, $L_2^{-1}L_1$ is lower triangular and $L_2^HL_1^{-H}$ is upper triangular. Thus, there is a diagonal matrix D such that

$$L_2^{-1}L_1 = D = L_2^HL_1^{-H}.$$

Therefore,

$$L_1 = L_2D \tag{3.13}$$

[3] Named in honor of the French military officer and mathematician André-Louis Cholesky (1875–1918).

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

and

$$DL_1^H = L_2^H,$$

or, equivalently,

$$L_2 = L_1 D^H. \tag{3.14}$$

Combining (3.13) and (3.14), we have

$$L_1 = L_2 D = L_1 D^H D.$$

Since L_1 is invertible, the cancellation property holds and $D^H D = I_n$. But since L_1 and L_2 have positive diagonal entries, so must D have positive diagonal entries. It follows that $D = I_n$, which implies that $L_1 = L_2$. □

Theorem 3.37 (HPD and spectrum)**.** *Suppose that* $A \in \mathbb{C}^{n\times n}$ *is Hermitian. Then* A *is HPD if and only if* $\sigma(A) \subset (0, \infty)$.

Proof. We only prove one direction here, as the other has already been proven. Since $A \in \mathbb{C}^{n\times n}_{\text{Her}}$, there exists a unitary matrix U and a diagonal matrix $D = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_i \in \sigma(A)$, for $i = 1, \ldots, n$, such that $A = U^H D U$. Let $\boldsymbol{x} \in \mathbb{C}^n_\star$. Set $\boldsymbol{y} = U\boldsymbol{x}$ and note that $\boldsymbol{y} \neq \boldsymbol{0}$. Then

$$\boldsymbol{x}^H A \boldsymbol{x} = (U\boldsymbol{x})^H D U \boldsymbol{x} = \boldsymbol{y}^H D \boldsymbol{y} = \sum_{i=1}^{n} \lambda_i |y_i|^2 \geq \min_{1\leq i\leq n} \lambda_i \boldsymbol{y}^H \boldsymbol{y} = \min_{1\leq i\leq n} \lambda_i \|\boldsymbol{y}\|_2^2 > 0.$$

This proves that A is HPD. □

The matrix encountered in the next result comes up repeatedly in the text. It is an example of an HPD matrix, but with all real entries. Such matrices are called symmetric positive definite (SPD). This matrix is also an example of a *Toeplitz symmetric tridiagonal*[4] (TST) matrix.

Theorem 3.38 (TST matrix)**.** *Define* $A \in \mathbb{R}^{(n-1)\times(n-1)}$ *via*

$$A = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & -1 & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}.$$

Then A *is SPD. Let* $h = 1/n$. *The set*

$$S = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{n-1}\},$$

where the i*th component of* $\boldsymbol{w}_k$ *is defined via*

$$[\boldsymbol{w}_k]_i = \sin(k\pi i h),$$

is an orthogonal set of eigenvectors of A.

[4] Named in honor of the German mathematician Otto Toeplitz (1881–1940).

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Proof. Note that for $k = 1, \ldots, n-1$,

$$\begin{aligned}[\mathsf{A}\boldsymbol{w}_k]_i &= -\sin(k\pi(i-1)h) + 2\sin(k\pi ih) - \sin(k\pi(i+1)h) \\ &= 2\sin(k\pi ih) - 2\cos(k\pi h)\sin(k\pi ih) \\ &= (2 - 2\cos(k\pi h))\sin(k\pi ih) \\ &= 2(1-\cos(k\pi h))[\boldsymbol{w}_k]_i.\end{aligned}$$

Hence, the distinct eigenvalues are $\lambda_k = 2 - 2\cos(k\pi h)$. To see that these are strictly positive for $k = 1, \ldots, n-1$, note that

$$1 > \cos(k\pi h) > -1,$$

which implies that

$$-2 < -2\cos(k\pi h) < 2,$$

which implies that

$$0 = 2 - 2 < 2 - 2\cos(k\pi h) < 2 + 2 = 4.$$

Since A is symmetric, the eigenvectors associated with distinct eigenvalues are orthogonal. A is SPD since its eigenvalues are strictly positive. □

Proposition 3.39 (HPD and similarity transformations)**.** *Let $\mathsf{A} \in \mathbb{C}^{m\times m}$ be HPD and $\mathsf{X} \in \mathbb{C}^{m\times n}$ with $m \geq n$ have full rank. Then $\mathsf{X}^{\mathsf{H}}\mathsf{A}\mathsf{X} \in \mathbb{C}^{n\times n}$ is HPD.*

Proof. Notice that

$$(\mathsf{X}^{\mathsf{H}}\mathsf{A}\mathsf{X})^{\mathsf{H}} = \mathsf{X}^{\mathsf{H}}\mathsf{A}^{\mathsf{H}}\mathsf{X} = \mathsf{X}^{\mathsf{H}}\mathsf{A}\mathsf{X}.$$

Suppose that $\boldsymbol{x} \in \mathbb{C}^n_\star$ is arbitrary. Since X is full rank, it follows that $\boldsymbol{y} = \mathsf{X}\boldsymbol{x} \neq \boldsymbol{0}$, i.e., $\boldsymbol{y} \in \mathbb{C}^m_\star$. Then

$$\boldsymbol{x}^{\mathsf{H}}\mathsf{X}^{\mathsf{H}}\mathsf{A}\mathsf{X}\boldsymbol{x} = \boldsymbol{y}^{\mathsf{H}}\mathsf{A}\boldsymbol{y} > 0,$$

since A is HPD. □

Theorem 3.40 (HPD and Gaussian elimination)**.** *Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be HPD and represented as*

$$\mathsf{A} = \begin{bmatrix} \alpha & \boldsymbol{p}^{\mathsf{H}} \\ \boldsymbol{p} & \hat{\mathsf{A}} \end{bmatrix},$$

where $\alpha \in \mathbb{C}$, $\boldsymbol{p} \in \mathbb{C}^{n-1}$, and $\hat{\mathsf{A}} \in \mathbb{C}^{(n-1)\times(n-1)}$. After one step of Gaussian elimination (without pivoting), A *will be reduced to the matrix*

$$\begin{bmatrix} \alpha & \boldsymbol{p}^{\mathsf{H}} \\ \boldsymbol{0} & \mathsf{B} \end{bmatrix},$$

where $\mathsf{B} \in \mathbb{C}^{(n-1)\times(n-1)}$. Then B *is HPD and the corresponding diagonal elements of* B *are smaller than those of* $\hat{\mathsf{A}}$.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Proof. Let us construct a matrix $\mathsf{L} \in \mathbb{C}^{n\times n}$ such that

$$\mathsf{LA} = \begin{bmatrix} \alpha & \boldsymbol{p}^{\mathsf{H}} \\ \mathbf{0} & \mathsf{B} \end{bmatrix},$$

if possible. Consider

$$\mathsf{L} = \begin{bmatrix} 1 & \mathbf{0}^{\mathsf{T}} \\ \boldsymbol{m} & \mathsf{I}_{n-1} \end{bmatrix}.$$

Then

$$\mathsf{LA} = \begin{bmatrix} 1 & \mathbf{0}^{\mathsf{T}} \\ \boldsymbol{m} & \mathsf{I}_{n-1} \end{bmatrix} \begin{bmatrix} \alpha & \boldsymbol{p}^{\mathsf{H}} \\ \boldsymbol{p} & \hat{\mathsf{A}} \end{bmatrix} = \begin{bmatrix} \alpha & \boldsymbol{p}^{\mathsf{H}} \\ \alpha\boldsymbol{m}+\boldsymbol{p} & \boldsymbol{m}\boldsymbol{p}^{\mathsf{H}}+\hat{\mathsf{A}} \end{bmatrix}.$$

Note that, since A is HPD, $\alpha > 0$ and $\hat{\mathsf{A}}$ is HPD. Choosing $\boldsymbol{m} = -\alpha^{-1}\boldsymbol{p}$, we have

$$\mathsf{LA} = \begin{bmatrix} \alpha & \boldsymbol{p}^{\mathsf{H}} \\ \mathbf{0} & \hat{\mathsf{A}} - \alpha^{-1}\boldsymbol{p}\boldsymbol{p}^{\mathsf{H}} \end{bmatrix}.$$

Having successfully constructed L, we find $\mathsf{B} = \hat{\mathsf{A}} - \alpha^{-1}\boldsymbol{p}\boldsymbol{p}^{\mathsf{H}}$. Notice that this is not the only way to find the matrix B.

Now let $\boldsymbol{x} \in \mathbb{C}^{n-1}_{\star}$ be arbitrary. Define $\boldsymbol{y} \in \mathbb{C}^{n}_{\star}$ via

$$\boldsymbol{y} = \begin{bmatrix} \gamma \\ \boldsymbol{x} \end{bmatrix},$$

where $\gamma \in \mathbb{C}$ is arbitrary. Then, since A is HPD,

$$\begin{aligned} 0 &< \boldsymbol{y}^{\mathsf{H}}\mathsf{A}\boldsymbol{y} \\ &= \begin{bmatrix} \overline{\gamma} & \boldsymbol{x}^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} \alpha & \boldsymbol{p}^{\mathsf{H}} \\ \boldsymbol{p} & \hat{\mathsf{A}} \end{bmatrix} \begin{bmatrix} \gamma \\ \boldsymbol{x} \end{bmatrix} \\ &= \begin{bmatrix} \overline{\gamma} & \boldsymbol{x}^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} \alpha\gamma + \boldsymbol{p}^{\mathsf{H}}\boldsymbol{x} \\ \gamma\boldsymbol{p} + \hat{\mathsf{A}}\boldsymbol{x} \end{bmatrix} \\ &= \alpha|\gamma|^2 + \overline{\gamma}\boldsymbol{p}^{\mathsf{H}}\boldsymbol{x} + \gamma\boldsymbol{x}^{\mathsf{H}}\boldsymbol{p} + \boldsymbol{x}^{\mathsf{H}}\hat{\mathsf{A}}\boldsymbol{x}. \end{aligned}$$

Now we set $\gamma = -\alpha^{-1}\boldsymbol{p}^{\mathsf{H}}\boldsymbol{x}$. From the last calculation

$$\begin{aligned} 0 &< \boldsymbol{y}^{\mathsf{H}}\mathsf{A}\boldsymbol{y} \\ &= \alpha^{-1}|\boldsymbol{p}^{\mathsf{H}}\boldsymbol{x}|^2 - \alpha^{-1}|\boldsymbol{p}^{\mathsf{H}}\boldsymbol{x}|^2 - \alpha^{-1}|\boldsymbol{p}^{\mathsf{H}}\boldsymbol{x}|^2 + \boldsymbol{x}^{\mathsf{H}}\hat{\mathsf{A}}\boldsymbol{x} \\ &= \boldsymbol{x}^{\mathsf{H}}\hat{\mathsf{A}}\boldsymbol{x} - \alpha^{-1}|\boldsymbol{p}^{\mathsf{H}}\boldsymbol{x}|^2 \\ &= \boldsymbol{x}^{\mathsf{H}}\mathsf{B}\boldsymbol{x}. \end{aligned}$$

This proves that B is HPD.

Now the diagonal elements of B, which must be positive since B is HPD, are precisely $[\mathsf{B}]_{i,i} = [\hat{\mathsf{A}}]_{i,i} - \alpha^{-1}|[\boldsymbol{p}]_i|^2$. Hence, $0 < [\mathsf{B}]_{i,i} \leq [\hat{\mathsf{A}}]_{ii}$, since $\alpha^{-1}|[\boldsymbol{p}]_i|^2 \geq 0$. □

Corollary 3.41 (HPD and Gaussian elimination)**.** *Suppose that* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD. Then Gaussian elimination without pivoting proceeds to completion to produce a unit lower triangular matrix* $\mathsf{L} \in \mathbb{C}^{n\times n}$ *and an upper triangular matrix* $\mathsf{U} \in \mathbb{C}^{n\times n}$ *with positive diagonal elements such that* $\mathsf{A} = \mathsf{LU}$.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Proof. Apply recursively the previous result. □

Theorem 3.42 (HPD criterion)**.** $\mathsf{A} \in \mathbb{C}^{n\times n}$ *is HPD if and only if* $\mathsf{A} = \mathsf{L}\mathsf{L}^{\mathsf{H}}$*, where* $\mathsf{L} \in \mathbb{C}^{n\times n}$ *is invertible.*

Proof. Suppose that $\mathsf{A} = \mathsf{L}\mathsf{L}^{\mathsf{H}}$, where L is invertible. Let $\boldsymbol{x} \in \mathbb{C}^n$ be arbitrary. Set $\boldsymbol{y} = \mathsf{L}^{\mathsf{H}}\boldsymbol{x}$. Since L is invertible, L^{H} is invertible; and $\boldsymbol{y} = \boldsymbol{0}$ if and only if $\boldsymbol{x} = \boldsymbol{0}$. Then

$$\boldsymbol{x}^{\mathsf{H}}\mathsf{A}\boldsymbol{x} = \boldsymbol{x}^{\mathsf{H}}\mathsf{L}\mathsf{L}^{\mathsf{H}}\boldsymbol{x} = \left(\mathsf{L}^{\mathsf{H}}\boldsymbol{x}\right)^{\mathsf{H}}\mathsf{L}^{\mathsf{H}}\boldsymbol{x} = \boldsymbol{y}^{\mathsf{H}}\boldsymbol{y} = \|\boldsymbol{y}\|_2^2 \geq 0,$$

with equality if and only if $\boldsymbol{x} = \boldsymbol{0}$. This proves that A is HPD.

The converse direction follows from the Cholesky factorization provided in Corollary 3.35. □

Theorem 3.43 (block matrices)**.** *Let* $k, m \in \mathbb{N}$*. Set* $n = k + m$*. Suppose that* $\mathsf{A} \in \mathbb{C}^{n\times n}$ *has the decomposition*

$$\mathsf{A} = \begin{bmatrix} \mathsf{B} & \mathsf{C}^{\mathsf{H}} \\ \mathsf{C} & \mathsf{D} \end{bmatrix},$$

where $\mathsf{B} \in \mathbb{C}^{k\times k}$*,* $\mathsf{C} \in \mathbb{C}^{m\times k}$*, and* $\mathsf{D} \in \mathbb{C}^{m\times m}$*.*

1. *If* A *is HPD, then* B*,* D*, and* $\mathsf{S} = \mathsf{D} - \mathsf{C}\mathsf{B}^{-1}\mathsf{C}^{\mathsf{H}}$ *are HPD.* S *is called the* Schur complement *of* B *in* A*.*
2. *If* A *is HPD, the Cholesky factorization of* A *may be expressed in terms of the matrix* C *and the Cholesky factorizations of* B *and* S*.*

Proof. We prove each statement separately.

1. Since A is HPD, $\boldsymbol{x}^{\mathsf{H}}\mathsf{A}\boldsymbol{x} > 0$ for any $\boldsymbol{x} \in \mathbb{C}_\star^n$. Let $\boldsymbol{y} \in \mathbb{C}_\star^k$ and $\boldsymbol{w} \in \mathbb{C}_\star^m$ be arbitrary. Setting $\boldsymbol{x} = \left[\boldsymbol{y}^{\mathsf{H}}, \boldsymbol{0}^{\mathsf{H}}\right]^{\mathsf{H}} \in \mathbb{C}_\star^n$, we have

$$0 < \boldsymbol{x}^{\mathsf{H}}\mathsf{A}\boldsymbol{x} = \boldsymbol{y}^{\mathsf{H}}\mathsf{B}\boldsymbol{y};$$

on the other hand, setting $\boldsymbol{x} = \left[\boldsymbol{0}^{\mathsf{H}}, \boldsymbol{w}^{\mathsf{H}}\right]^{\mathsf{H}} \in \mathbb{C}_\star^n$, we have

$$0 < \boldsymbol{x}^{\mathsf{H}}\mathsf{A}\boldsymbol{x} = \boldsymbol{w}^{\mathsf{H}}\mathsf{D}\boldsymbol{w}.$$

Thus, B and D are HPD. More generally, suppose that $\boldsymbol{x} = \left[\boldsymbol{x}_1^{\mathsf{H}}, \boldsymbol{x}_2^{\mathsf{H}}\right]^{\mathsf{H}} \in \mathbb{C}_\star^n$. Then

$$0 < \boldsymbol{x}^{\mathsf{H}}\mathsf{A}\boldsymbol{x} = \boldsymbol{x}_1^{\mathsf{H}}\mathsf{B}\boldsymbol{x}_1 + \boldsymbol{x}_1^{\mathsf{H}}\mathsf{C}^{\mathsf{H}}\boldsymbol{x}_2 + \boldsymbol{x}_2^{\mathsf{H}}\mathsf{C}\boldsymbol{x}_1 + \boldsymbol{x}_2^{\mathsf{H}}\mathsf{D}\boldsymbol{x}_2.$$

Now pick $\boldsymbol{x}_1 = -\mathsf{B}^{-1}\mathsf{C}^{\mathsf{H}}\boldsymbol{x}_2$ and suppose that $\boldsymbol{x}_2 \neq \boldsymbol{0}$. Then

$$\begin{aligned} 0 &< \boldsymbol{x}^{\mathsf{H}}\mathsf{A}\boldsymbol{x} \\ &= \boldsymbol{x}_2^{\mathsf{H}}\mathsf{C}\mathsf{B}^{-\mathsf{H}}\mathsf{B}\mathsf{B}^{-1}\mathsf{C}^{\mathsf{H}}\boldsymbol{x}_2 - \boldsymbol{x}_2^{\mathsf{H}}\mathsf{C}\mathsf{B}^{-\mathsf{H}}\mathsf{C}^{\mathsf{H}}\boldsymbol{x}_2 - \boldsymbol{x}_2^{\mathsf{H}}\mathsf{C}\mathsf{B}^{-1}\mathsf{C}^{\mathsf{H}}\boldsymbol{x}_2 + \boldsymbol{x}_2^{\mathsf{H}}\mathsf{D}\boldsymbol{x}_2 \\ &= \boldsymbol{x}_2^{\mathsf{H}}\left(\mathsf{D} - \mathsf{C}\mathsf{B}^{-1}\mathsf{C}^{\mathsf{H}}\right)\boldsymbol{x}_2, \end{aligned}$$

where we used the fact that $\mathsf{B}^{-1} = \mathsf{B}^{-\mathsf{H}}$ — since B is HPD — on the last step. It follows that S is HPD.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

2. Since A is HPD, there is a unique lower triangular matrix $L_A \in \mathbb{C}^{n\times n}$ with positive diagonal entries, such that $A = L_A L_A^H$. Likewise there are unique lower triangular matrices $L_B \in \mathbb{C}^{k\times k}$ and $L_S \in \mathbb{C}^{m\times m}$, both with positive diagonal entries, such that $B = L_B L_B^H$ and $S = L_S L_S^H$. Suppose that

$$L_A = \begin{bmatrix} L_1 & O \\ M & L_2 \end{bmatrix},$$

where $L_1 \in \mathbb{C}^{k\times k}$ is lower triangular with positive diagonal entries; $M \in \mathbb{C}^{m\times k}$; and $L_2 \in \mathbb{C}^{m\times m}$ is lower triangular with positive diagonal entries. Then

$$L_A L_A^H = \begin{bmatrix} L_1 & O \\ M & L_2 \end{bmatrix} \begin{bmatrix} L_1^H & M^H \\ O^H & L_2^H \end{bmatrix} = \begin{bmatrix} L_1 L_1^H & L_1 M^H \\ M L_1^H & M M^H + L_2 L_2^H \end{bmatrix} = \begin{bmatrix} B & C^H \\ C & D \end{bmatrix}.$$

Comparing entries we conclude that

$$\begin{aligned} L_1 L_1^H = B \quad &\Longrightarrow \quad L_1 = L_B, \\ M L_1^H = C \quad &\Longrightarrow \quad M = C L_B^{-H}, \\ M M^H + L_2 L_2^H = D \quad &\Longrightarrow \quad L_2 L_2^H = D - C L_B^{-H} L_B^{-1} C^H \\ &\Longrightarrow \quad L_2 L_2^H = S \\ &\Longrightarrow \quad L_2 = L_S, \end{aligned}$$

and the result is proven. □

3.5.3 Cholesky Factorization, Revisited

In this section, our aim is to produce a practical, computable algorithm for the Cholesky factorization. Let us first recall how Gaussian elimination acts on an HPD matrix. Let us consider a simplified HPD matrix $A \in \mathbb{C}^{n\times n}$:

$$A = \begin{bmatrix} 1 & \boldsymbol{w}^H \\ \boldsymbol{w} & K \end{bmatrix} \to L_1 A = \begin{bmatrix} 1 & \boldsymbol{w}^H \\ \boldsymbol{0} & \hat{K} \end{bmatrix}$$

with

$$L_1 = \begin{bmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{m} & I_{n-1} \end{bmatrix},$$

so that

$$L_1 A = \begin{bmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{m} & I_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \boldsymbol{w}^H \\ \boldsymbol{w} & K \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{w}^H \\ \boldsymbol{m} + \boldsymbol{w} & K + \boldsymbol{m}\boldsymbol{w}^H \end{bmatrix}.$$

In other words, we must have $\boldsymbol{m} = -\boldsymbol{w}$ and $\hat{K} = K - \boldsymbol{w}\boldsymbol{w}^H$. Therefore,

$$A = L_1^{-1} \begin{bmatrix} 1 & \boldsymbol{w}^H \\ \boldsymbol{0} & \hat{K} \end{bmatrix} = \begin{bmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{w} & I_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \boldsymbol{w}^H \\ \boldsymbol{0} & K - \boldsymbol{w}\boldsymbol{w}^H \end{bmatrix},$$

so that, in the process, we lost that the matrix is Hermitian.

But being Hermitian is an important feature, and it would be desirable to keep it throughout the Gaussian elimination process. To do so, we will introduce zeros

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

on the first row to match the ones we introduced on the first column. To achieve this, notice that

$$\begin{bmatrix} 1 & \boldsymbol{w}^{\mathsf{H}} \\ \mathbf{0} & \mathsf{K} - \boldsymbol{w}\boldsymbol{w}^{\mathsf{H}} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{0}^{\mathsf{T}} \\ \mathbf{0} & \mathsf{K} - \boldsymbol{w}\boldsymbol{w}^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} 1 & \boldsymbol{w}^{\mathsf{H}} \\ \mathbf{0} & \mathsf{I}_{n-1} \end{bmatrix}.$$

Notice that the factor on the right is nothing but $\mathsf{L}_1^{-\mathsf{H}}$. In combining these two operations, we get

$$\mathsf{A} = \begin{bmatrix} 1 & \mathbf{0}^{\mathsf{T}} \\ \boldsymbol{w} & \mathsf{I}_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathsf{T}} \\ \mathbf{0} & \mathsf{K} - \boldsymbol{w}\boldsymbol{w}^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} 1 & \boldsymbol{w}^{\mathsf{H}} \\ \mathbf{0} & \mathsf{I}_{n-1} \end{bmatrix}.$$

This is the main idea behind the algorithmic version of the Cholesky factorization.

In general, since $\alpha^2 = a_{1,1} > 0$, we can proceed as follows:

$$\mathsf{A} = \begin{bmatrix} \alpha^2 & \boldsymbol{w}^{\mathsf{H}} \\ \boldsymbol{w} & \mathsf{K} \end{bmatrix} = \begin{bmatrix} \alpha & \mathbf{0}^{\mathsf{T}} \\ \frac{1}{\alpha}\boldsymbol{w} & \mathsf{I}_{n-1} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^{\mathsf{T}} \\ \mathbf{0} & \mathsf{K} - \frac{1}{a_{1,1}}\boldsymbol{w}\boldsymbol{w}^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} \alpha & \frac{1}{\alpha}\boldsymbol{w}^{\mathsf{H}} \\ \mathbf{0} & \mathsf{I}_{n-1} \end{bmatrix} = \mathsf{R}_1^{\mathsf{H}} \mathsf{A}_1 \mathsf{R}_1,$$

where R_1 is an upper triangular matrix. Applying this repeatedly we obtain a computable version of the Cholesky factorization, yielding an upper triangular matrix R with positive diagonal entries such that

$$\mathsf{A} = \mathsf{R}^{\mathsf{H}} \mathsf{I}_n \mathsf{R} = \mathsf{R}^{\mathsf{H}} \mathsf{R}.$$

Remark 3.44 (terminology)**.** Since $\alpha_k = \sqrt{a_{k,k}}$ uniquely determines the algorithm, this Cholesky factorization is sometimes also referred to as the *square root method*.

Listing 3.7 describes how to compute this factorization. Once this factorization is computed, system (3.1), in the case that A is HPD, can be solved rather efficiently.

Proposition 3.45 (complexity of Cholesky factorization)**.** *Let $\mathsf{A} \in \mathbb{C}^{n\times n}$ be HPD. The Cholesky factorization algorithm requires of the order of $\frac{1}{3}n^3$ operations.*

Proof. We just need to notice that the innermost loop of Listing 3.7 requires two (2) operations. Thus, to leading order, the total complexity is

$$\sum_{i=2}^{n} \sum_{j=1}^{i-1} \sum_{k=1}^{j-1} 2 \approx \frac{1}{3} n^3. \qquad \square$$

Problems

3.1 Write the computational formulas used to solve a system of equations with a lower triangular coefficient matrix. In other words, describe the forward substitution algorithm.

3.2 Show that when the generic back (forward) substitution algorithm is applied to solve a system of $n \in \mathbb{N}$ unknowns, there are n divisions, $\frac{n^2-n}{2}$ multiplications, and $\frac{n^2-n}{2}$ additions/subtractions. Use this to show that *complexity*, i.e., the number of operations of this algorithm, is

$$n + \frac{n^2 - n}{2} + \frac{n^2 - n}{2} = n^2.$$

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

3.3 Find the complexity of the algorithm presented in Listing 3.1 to solve a system with a tridiagonal coefficient matrix.

3.4 Given a nonsingular matrix $A \in \mathbb{C}^{n\times n}$ propose an algorithm, based on Gaussian elimination, to find A^{-1}.

3.5 Prove Proposition 3.10.

3.6 Show that a column-s complete elementary matrix is invertible and find its inverse.

3.7 Prove Theorem 3.13.

3.8 Prove Proposition 3.17.

3.9 Prove Lemma 3.18.

3.10 Complete the proof of Theorem 3.21.

3.11 Suppose that $A \in \mathbb{R}^{n\times n}$ is a nonsingular matrix whose leading principal sub-matrices are all nonsingular. Partition A as

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

where $A_{11} \in \mathbb{R}^{k\times k}$.

a) Show that there is a matrix M such that

$$\begin{bmatrix} I & O \\ -M & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ O & \widetilde{A}_{22} \end{bmatrix}$$

and write out the explicit expresssions for M and $\widetilde{A}_{22}$.

b) Show that

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} I & O \\ M & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ O & \widetilde{A}_{22} \end{bmatrix}.$$

c) The leading principal sub-matrices of A_{11} are, of course, all nonsingular. Prove that $\widetilde{A}_{22}$ is also nonsingular.

d) By the previous statement, both A_{11} and $\widetilde{A}_{22}$ have LU decompositions, say $A_{11} = L_1U_1$ and $\widetilde{A}_{22} = L_2U_2$. Show that

$$A = \begin{bmatrix} L_1 & O \\ ML_1 & L_2 \end{bmatrix} \begin{bmatrix} U_1 & L_1^{-1}A_{12} \\ O & U_2 \end{bmatrix},$$

which is the LU decomposition of A.

3.12 Suppose that $A \in \mathbb{R}^{n\times n}$ is SDD with positive diagonal entries and nonpositive off-diagonal entries. Show that, in this case, A^{-1} exists and contains only nonnegative elements.

Hint: Use the procedure for inverting a matrix using Gaussian elimination.

3.13 Complete the proof of Theorem 3.33.

3.14 Complete the proof of Theorem 3.34.

3.15 Suppose that $A \in \mathbb{C}^{n\times n}$ is Hermitian, has nonnegative (real) diagonal elements, and is SDD. Prove that it is HPD.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

Listings

```
function [x, err] = TriDiagonal( a, b, c, f )
% Solution of a linear system of equations with a tridiagonal
% matrix.
%
%   a(k) x(k-1) + b(k) x(k) + c(k) x(k+1) = f(k)
%
% with a(1) = c(n) = 0.
%
% Input
%   a(1:n), b(1:n), c(1:n) : the coefficients of the system
%                            matrix
%   f(1:n) : the right hand side vector
%
% Output
%   x(1:n) : the solution to the linear system of equations, if
%            no division by zero occurs
%   err : = 0, if no division by zero occurs
%         = 1, if division by zero is encountered
  n = length(f);
  alpha = zeros(n,1);
  beta = zeros(n,1);
  err = 0;
  if abs(b(1)) > eps( b(1) )
    alpha(1) = -c(1)/b(1);
    beta(1) = f(1)/b(1);
  else
    err = 1;
    return;
  end
  for k=2:n
    denominator = a(k)*alpha(k-1) + b(k);
    if abs(denominator) > eps( denominator )
      alpha(k) = - c(k)/denominator;
      beta(k) = ( f(k) - a(k)*beta(k-1) )/denominator;
    else
      err = 1;
      return;
    end
  end
  if abs(a(n)*alpha(n-1) + b(n)) > eps( b(n) )
     x(n) = ( f(n) - a(n)*beta(n-1) )/( a(n)*alpha(n-1) + b(n) );
  else
    err = 1;
    return;
  end
  for k=n-1:-1:1
    x(k) = alpha(k)*x(k+1) + beta(k);
  end
end
```

Listing 3.1 Solution of a system of equations with a tridiagonal coefficient matrix.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

```
function [x, err] = CyclicallyTriDiagonal( a, b, c, f )
% Solution of a cyclically tridiagonal linear system of equations:
%
%   b(1) x(  1) + c(1) x(  2) + a(1) x(  n) = f(1),
%   a(k) x(k-1) + b(k) x(  k) + c(k) x(k+1) = f(k),
%   c(n) x(  1) + a(n) x(n-1) + b(n) x(  n) = f(n).
%
% Input
%   a(1:n), b(1:n), c(1:n) : the coefficients of the system
%                            matrix
%   f(1:n) : the right hand side vector
%
% Output
%   x(1:n) : the solution to the linear system of equations, if
%            no division by zero is encountered
%   err : = 0, if division by zero does not occur
%         = 1, if division by zero occurs
  n = length(f);
  err = 0;
  newa = a(2:n);
  newa(1) = 0;
  newb = b(2:n);
  newc = c(2:n);
  newc(n-1) = 0;
  newf = f(2:n);
  [u, err] = TriDiagonal( newa, newb, newc, newf );
  if err == 1
    return;
  end
  newf = zeros(n-1, 1);
  newf(1) = - a(2);
  newf(n-1) = -c(n);
  [v, err] = TriDiagonal( newa, newb, newc, newf );
  if err == 1
    return;
  end
  x = zeros(n,1);
  denmoniator = b(1) + c(1)*v(1) + a(1)*v(n-1);
  if abs(denominator) > eps( denorminator )
    x(1) =  ( f(1) - c(1)*u(1) - a(1)*u(n-1) )/denominator;
  else
    err = 1;
    return;
  end
  for k=2:n
    x(k) = u(k-1) + x(1)*v(k-1);
  end
end
```

Listing 3.2 Solution of a system of equations with a cyclically tridiagonal coefficient matrix.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

```
function [L, U, err] = LUFactSimple( A )
% LU factorization of a square matrix.
%
% Input
%   A(1:n,1:n) : the matrix to be factorized
%
% Output
%   L(1:n,1:n), U(1:n,1:n) : the factors in A = LU, if Gaussian
%                            elimination proceeds to completion
%   err : = 0 if Gaussian elimination proceeds to completion
%         = 1 if a zero pivot is encountered
  n = size(A,1);
  U = A;
  L = eye(n);
  err = 0;
  for k=1:n-1
    for j=k+1:n
      if abs( U(k,k) ) > eps( U(k,k) )
        L(j,k) = U(j,k)/U(k,k);
      else
        err = 1;
        return;
      end
      for t = k:n
        U(j,t) = U(j,t)-L(j,k)*U(k,t);
      end
    end
  end
end
```

Listing 3.3 LU factorization.

```
function [L, U, P, err] = LUFactPivot( A )
% LU factorization with pivoting of a square matrix.
%
% Input
%   A(1:n,1:n) : the matrix to be factorized
%
% Output
%   L(1:n,1:n), U(1:n,1:n), P(1:n,1:n) : the factors in A = P'LU
%   err : = 0 if no error was encountered
%         = 1 if a division by zero occurred
  n = size(A,1);
  U = A;
  L = eye(n);
  P = eye(n);
  err = 0;
  for k=1:n-1
    i=k;
    for t=k:n
      if abs( U(t,k) ) > abs( U(i,k) )
        i=t;
      end
    end
```

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

```
    for t=k:n
      temp = U(k,t);
      U(k,t) = U(i,t);
      U(i,t) = temp;
    end
    for t=1:k-1
      temp = L(k,t);
      L(k,t) = L(i,t);
      L(i,t) = temp;
    end
    for t=1:n
      temp = P(k,t);
      P(k,t) = P(i,t);
      P(i,t) = temp;
    end
    if abs( U(k,k) ) > eps( U(k,k) )
      for j=k+1:n
        L(j,k) = U(j,k)/U(k,k);
        for t=k:n
          U(j,t) = U(j,t) - L(j,k)*U(k,t);
        end
      end
    else
      err = 1;
      return;
    end
  end
end
```

Listing 3.4 LU factorization with pivoting.

```
function [Afact, swaps, err] = LUFactEfficient( A )
% Efficient implementation of LU factorization with partial
% pivoting.
%
% Input
%   A(1:n,1:n) : the matrix to be factorized into A = P'LU
%
% Output
%   Afact(1:n,1:n) : a matrix containing the matrices L and U
%                    below and above the diagonal, respectively
%   swaps : the indices that indicate the permutations
%           (row swaps)
%   err : = 0 if no the factorization finished successfully
%         = 1 if a division by zero was encountered
  n=size(A,1);
  swaps = 1:n;
  err = 0;
  for k=1:n-1
    i = k;
    for t=k:n
      if abs( A(t,k) ) > abs( A(i,k) )
        i=t;
      end
    end
```

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

```
    tt = swaps(k);
    swaps(k) = swaps(i);
    swaps(i) = tt;
    if abs( A( swaps(k), k ) <= eps( A( swaps(k), k) )
      err = 1;
      return;
    end
    for i=k+1:n
      xmult = A(swaps(i),k)/A(swaps(k),k);
      A( swaps(i), k ) = xmult;
      for j=k+1:n
        A(swaps(i),j) = A(swaps(i),j) - xmult*A(swaps(k),j);
      end
    end
  end
  Afact = A;
end
```

Listing 3.5 Efficient implementation of LU factorization with pivoting.

```
function [x, err] = Solve( A, f )
% Solves the system Ax = f.
%
% Input
%   A(1:n,1:n) : the coefficient matrix
%   f(1:n) : the RHS vector
%
% Output:
%   x(1:n) : the solution vector
%   err : = 0 if the solution was found successfully
%         = 1 if an error occurred during the process
  n = length(f);
  err = 0;
  [AA, swaps, err] = LUFactEfficient( A );
  if err == 1
    return
  end
  for k=1:n-1
    for i=k+1:n
      f( swaps(i) ) = f( swaps(i) ) - AA( swaps(i), k ) ...
        *f( swaps(k) );
    end
  end
  if abs( AA( swaps(n), n ) ) > eps( AA(swaps(n),n) )
    x(n) = f( swaps(n) )/AA( swaps(n), n );
  else
    err = 1;
    return;
  end
  for i=n-1:-1:1
    xsum = f( swaps(i) );
    for j=i+1:n
      xsum = xsum - AA( swaps(i), j )*x(j);
    end
    if abs( AA( swaps(i), i ) ) > eps( AA( swaps(i), i ) )
```

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press

```
      x(i) = xsum/AA( swaps(i), i );
    else
      err = 1;
      return;
    end
  end
end
```

Listing 3.6 Solution of (3.1) after LU factorization.

```
function [R, err] = CholeskyDecomposition( A )
% Choleksy factorization of the HPD matrix A:
%   A = R'*R
%
% Input:
%   A(1:n,1:n) : an HPD matrix
%
% Output:
%   R(1:n,1:n) : The upper triangular matrix satisfying A = R'*R
%   err : = 0 if the decomposition finished successfully
%         = 1 if an error occurred
  err = 0;
  n = size(A,1);
  R = zeros(n,n);
  R(1,1) = sqrt( A(1,1) );
  for i=2:n
    for j=1:i-1
      sum = A(i,j);
      for k=1:j-1
        sum = sum - R(k,i)* conj( R(k,j) );
      end
      if abs( R(j,j) ) > eps( R(j,j) )
        R(j,i) = sum/R(j,j);
      else
        err = 1;
        return;
      end
    end
    sum = A(i,i);
    for k=1:i-1
      sum = sum - R(k,i)*conj( R(k,i) );
    end
    R(i,i) = sqrt( sum );
  end
end
```

Listing 3.7 Computation of the Cholesky factorization for an HPD matrix.

https://doi.org/10.1017/9781108942607.006 Published online by Cambridge University Press