2

Solution of systems of linear equations

2.1 Introduction

In Chapter 1 we considered the solution of a single equation of the form $f(x) = 0$ where f is a real-valued function defined and continuous on a closed interval of the real line. The simplest example of this kind is the linear equation $ax = b$ where a and b are given real numbers, with $a \neq 0$, whose solution is

$$x = a^{-1}b\,, \tag{2.1}$$

trivially. Of course, we could have expressed the solution as $x = b/a$ as in Chapter 1, but, as you will see in a moment, writing $x = a^{-1}b$ is much more revealing in the present context. In this chapter we shall consider a different generalisation of this elementary problem:

Let A be an $n \times n$ matrix with a_{ij} as its entry in row i and column j
and $\boldsymbol{b}$ a given column vector of size n with jth entry b_j;
find a column vector $\boldsymbol{x}$ of size n such that $A\boldsymbol{x} = \boldsymbol{b}$.

Denoting by x_i the ith entry of the vector $\boldsymbol{x}$, we can also write $A\boldsymbol{x} = \boldsymbol{b}$ in the following expanded form:

$$\left.\begin{array}{l} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1\,, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2\,, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n\,. \end{array}\right\} \tag{2.2}$$

Recall that in order to ensure that for real numbers a and b the single linear equation $ax = b$ has a unique solution, we need to assume that $a \neq 0$. In the case of the simultaneous system (2.2) of n linear equations in n unknowns we shall have to make an analogous assumption on the matrix A.

To do so, we introduce the following definition.

Definition 2.1 *The set of all $m \times n$ matrices with real entries is denoted by $\mathbb{R}^{m\times n}$. A matrix of size $n \times n$ will be called a square matrix of order n, or simply a matrix of* **order** *n. The* **determinant** *of a square matrix $A \in \mathbb{R}^{n\times n}$ is the real number* $\det(A)$ *defined as follows:*

$$\det(A) = \sum_{\text{perm}} \text{sign}(\nu_1, \nu_2, \ldots, \nu_n) a_{1\nu_1} a_{2\nu_2} \ldots a_{n\nu_n} .$$

The summation is over all $n!$ permutations $(\nu_1, \nu_2, \ldots, \nu_n)$ of the integers $1, 2, \ldots, n$, and $\text{sign}(\nu_1, \nu_2, \ldots, \nu_n) = +1$ *or -1 depending on whether the n-tuple $(\nu_1, \nu_2, \ldots, \nu_n)$ is an even or odd permutation of $(1, 2, \ldots, n)$, respectively. An even (odd) permutation is obtained by an even (odd) number of exchanges of two adjacent elements in the array $(1, 2, \ldots, n)$. A matrix $A \in \mathbb{R}^{n\times n}$ is said to be* **nonsingular** *when its determinant* $\det(A)$ *is nonzero.*

The **inverse matrix** A^{-1} of a nonsingular matrix $A \in \mathbb{R}^{n\times n}$ is defined as the element of $\mathbb{R}^{n\times n}$ such that $A^{-1}A = AA^{-1} = I$, where I is the $n \times n$ identity matrix

$$I = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{pmatrix} . \tag{2.3}$$

In order to find an explicit expression for A^{-1} in terms of the elements of the matrix A, we recall from linear algebra that, for each $i = 1, 2, \ldots, n$,

$$a_{i1}A_{k1} + a_{i2}A_{k2} + \cdots + a_{in}A_{kn} = \begin{cases} \det(A) & \text{if } i = k\,, \\ 0 & \text{if } i \neq k\,, \end{cases} \tag{2.4}$$

where $A_{ij} = (-1)^{i+j}\text{Cof}(a_{ij})$ and $\text{Cof}(a_{ij})$, called the **cofactor** of a_{ij}, is the determinant of the $(n-1) \times (n-1)$ matrix obtained by erasing from $A \in \mathbb{R}^{n\times n}$ row i and column j. Then, it is a trivial matter to show using (2.4) that A^{-1} has the form

$$A^{-1} = \frac{1}{\det(\text{A})} \begin{pmatrix} A_{11} & A_{21} & \ldots & A_{n1} \\ A_{12} & A_{22} & \ldots & A_{n2} \\ \ldots & \ldots & \ldots & \ldots \\ A_{1n} & A_{2n} & \ldots & A_{nn} \end{pmatrix} . \tag{2.5}$$

Having found an explicit formula for the matrix A^{-1}, we now multiply both sides of the equation $A\boldsymbol{x} = \boldsymbol{b}$ on the left by A^{-1} to deduce that

$A^{-1}(A\boldsymbol{x}) = A^{-1}\boldsymbol{b}$; finally, since $A^{-1}(A\boldsymbol{x}) = (A^{-1}A)\boldsymbol{x} = I\boldsymbol{x} = \boldsymbol{x}$, it follows that

$$\boldsymbol{x} = A^{-1}\boldsymbol{b}\,, \tag{2.6}$$

where the inverse A^{-1} of the nonsingular matrix A is given in terms of the entries of A by (2.5).[1]

An alternative approach to the solution of the linear system $A\boldsymbol{x} = \boldsymbol{b}$, called Cramer's rule, proceeds by expressing the ith entry of $\boldsymbol{x}$ as

$$x_i = D_i/D\,, \qquad i = 1, 2, \ldots, n\,,$$

where $D = \det(A)$, and D_i is the $n \times n$ determinant obtained by replacing the ith column of D by the entries of $\boldsymbol{b}$. Evidently, we must require that A is nonsingular, *i.e.*, that $D = \det(A) \neq 0$. Thus, all we need to do to solve $A\boldsymbol{x} = \boldsymbol{b}$ is to evaluate the $n + 1$ determinants $D, D_1, \ldots, D_n$, each of them $n \times n$, and check that $D = \det(A)$ is nonzero; the final calculation of the elements $x_i, i = 1, 2, \ldots, n$, is then trivial.[2]

The purpose of our next example is to illustrate the application of Cramer's rule.

Example 2.1 *Suppose that we wish to solve the system of linear equations*

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

The solution of such a small system can easily be found in terms of determinants, by Cramer's rule. This gives

$$x_1 = D_1/D\,, \quad x_2 = D_2/D\,, \quad x_3 = D_3/D\,,$$

[1] By the way, on comparing (2.6) with (2.1) you will notice that (2.1) is a special case of (2.6) when $n = 1$.

[2] Gabriel Cramer (31 July 1704, Geneva, Switzerland – 4 January 1752, Bagnols-sur-Cèze, France). In the 1730s Colin Maclaurin (February 1698, Kilmodan, Cowal, Argyllshire, Scotland – 14 June 1746, Edinburgh, Scotland) wrote his *Treatise of Algebra* which was not published until 1748, two years after his death. It contained the first published results on determinants proving Cramer's rule for 2×2 and 3×3 systems and indicating how the 4×4 case would work. Cramer gave the general rule for $n \times n$ systems without proof in the Appendix to his paper 'Introduction to the analysis of algebraic curves' (1750), motivated by the desire to find the equation of a plane curve passing through a number of given points.

where

$$D = \begin{vmatrix} 1 & 1 & 1 \\ 2 & 4 & 2 \\ -1 & 5 & -4 \end{vmatrix}, \quad D_1 = \begin{vmatrix} 6 & 1 & 1 \\ 16 & 4 & 2 \\ -3 & 5 & -4 \end{vmatrix},$$

with similar expressions for D_2 and D_3. To obtain the solution we therefore need to evaluate four determinants. ◇

Now you may think that since, for $A \in \mathbb{R}^{n\times n}$ nonsingular, we have expressed the solution to $A\boldsymbol{x} = \boldsymbol{b}$ in the 'closed form'

$$\boldsymbol{x} = A^{-1}\boldsymbol{b}$$

and have even found a formula for A^{-1} in terms of the coefficients of A, or may simply compute the entries of $\boldsymbol{x}$ directly using Cramer's rule, the story about the simultaneous set of linear equations (2.2) has reached its happy ending. We are sorry to disappoint you: a disturbing tale is about to unfold.

Imagine the following example: let $n = 100$, say, and suppose that you have been given all 10000 entries of a 100×100 matrix A, together with the entries of a 100-component column vector $\boldsymbol{b}$. To avoid trivialities, let us suppose that none of the entries of A or $\boldsymbol{b}$ is equal to 0. Question: *Does the linear system $A\boldsymbol{x} = \boldsymbol{b}$ have a solution? If it does, how would you find, say, the 53rd entry of the solution vector $\boldsymbol{x}$?* Of course, you could calculate the determinant of A and check whether it is equal to zero; if not, you could then calculate the determinant D_{53} obtained by replacing the 53rd column of A by the vector $\boldsymbol{b}$, and the required result, by Cramer's rule, is then the ratio of these two determinants. How much time do you think you would need to accomplish this task? An hour? A day? A month?

I imagine that you do not have a large enough sheet of paper in front of you to write down this 100×100 matrix. Let us therefore start with a somewhat simpler setting. Assume that n is any integer, $n \geq 2$, and denote by d_n the number of arithmetic operations that are required to calculate $\det(A)$ for $A \in \mathbb{R}^{n\times n}$. For example, for a 2×2 matrix,

$$\det(A) = a_{11}a_{22} - a_{12}a_{21}\,;$$

this evaluation requires 3 arithmetic operations – 2 multiplications and 1 subtraction – giving $d_2 = 3$. In general, we can calculate $\det(A)$ by expanding it in the elements of its first row. This requires multiplying each of the n elements in the first row of A by a subdeterminant of size

$n-1$ (a total of $n(d_{n-1}+1)$ operations) and summing the n resulting numbers (another $n-1$ operations). Thus,

$$d_n = n(d_{n-1}+1)+n-1\,, \quad n \geq 3\,, \qquad d_2 = 3\,. \tag{2.7}$$

Let us write $d_n = c_n n!$ and substitute this into (2.7) to obtain

$$c_n = c_{n-1} + 2\frac{1}{(n-1)!} - \frac{1}{n!}\,, \quad n \geq 3\,, \qquad c_2 = \frac{3}{2}\,. \tag{2.8}$$

Now, summing (2.8) from $n=3$ to k for $k \geq 3$ yields, on letting $0!=1$,

$$c_k = \sum_{n=0}^{k-1} \frac{1}{n!} - \frac{1}{k!}.$$

As $\sum_{n=0}^{\infty}(1/n!) = \mathrm{e}$, it follows that

$$\lim_{k\to\infty} c_k = \mathrm{e}\,.$$

Thus,[1] $d_n \sim \mathrm{e}\,n!$ as $n \to \infty$. In order to compute the solution of a system of n simultaneous linear equations by Cramer's rule we need to evaluate $n+1$ determinants, each of size $n \times n$, so the total number of operations required is about $(n+1)d_n \sim \mathrm{e}\,(n+1)!$ as $n \to \infty$.

For $n = 100$, this means approximately $101!\,\mathrm{e} \approx 2.56\times 10^{160}$ arithmetic operations.[2] Today's fastest parallel computers are capable of teraflop speeds, *i.e.*, 10^{12} floating point operations per second; therefore, the computing time for our solution would be around $2.56 \times 10^{160}/10^{12} = 2.56 \times 10^{148}$ seconds, or a staggering 8.11×10^{140} years. According to the prevailing theoretical position, the Universe began in a violent explosion, the Big Bang, about $12.5(\pm 3) \times 10^9$ years ago. So please put that large sheet of paper away quickly! We need to discover a more efficient approach.

Incidentally, you might notice that in the expansion of all the determinants involved in Cramer's rule all the smaller subdeterminants occur many times over, so the number of operations involved can be reduced by avoiding such repetitions. However, a more careful analysis shows

[1] For two sequences (a_n) and (b_n), we shall write $a_n \sim b_n$ if $\lim_{n\to\infty}(a_n/b_n) = 1$.

[2] While on the subject of calculating factorials of large integers, let us mention **Stirling's formula** which states that $n! \sim \sqrt{2\pi}n^{n+1/2}\mathrm{e}^{-n}$ as $n \to \infty$ (J. Stirling, *Methodus differentialis*, 1730). Stirling's approximation can be made more precise as the double inequality

$$\sqrt{2\pi}n^{n+1/2}\mathrm{e}^{-n+1/(12n+1)} < n! < \sqrt{2\pi}n^{n+1/2}\mathrm{e}^{-n+1/(12n)}$$

(H. Robbins, A remark on Stirling's formula *Amer. Math. Monthly* **62**, 26–29, 1955).

that we cannot by this means reduce the total by more than a factor of about n, which hardly affects our conclusion.

Our other approach to solving $A\boldsymbol{x} = \boldsymbol{b}$, based on computing A^{-1} from (2.5) and writing $\boldsymbol{x} = A^{-1}\boldsymbol{b}$, is equally inefficient: in order to compute the inverse of an $n \times n$ matrix A using determinants, one has to calculate the determinant of A as well as n^2 determinants of size $n-1$ each of which then has to be divided by $\det(A)$, requiring a total of approximately

$$\mathrm{e}\, n! + n^2 \mathrm{e}\,(n-1)! + n^2 \sim \mathrm{e}\,(n+1)!$$

arithmetic operations, just the same as before.

The aim of this chapter is to develop alternative methods for the solution of the system of linear equations $A\boldsymbol{x} = \boldsymbol{b}$. We begin by considering a classical technique, Gaussian elimination.[1] We shall then explore its relationship to the factorisation $A = LU$ of the matrix A where L is lower triangular and U is upper triangular. It will be seen that by using the Gaussian elimination the number of arithmetic operations required to solve the linear system $A\boldsymbol{x} = \boldsymbol{b}$ with an $n \times n$ matrix A is approximately $\frac{2}{3}n^3$ – a dramatic reduction from the $\mathcal{O}(\mathrm{e}\,(n+1)!)$ operation count associated with matrix inversion using determinants.[2]

We conclude the chapter with a discussion of another classical idea attributed to Gauss:[3] the least squares method for the solution of the system of linear equations $A\boldsymbol{x} = \boldsymbol{b}$ where $A \in \mathbb{R}^{m \times n}$, $\boldsymbol{x}$ is the column vector of unknowns of size n and $\boldsymbol{b}$ a given column vector of size m.

2.2 Gaussian elimination

The technique for solving systems of linear algebraic equations that we shall describe in this section was developed by Carl Friedrich Gauss and was first published in his *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (1809), a major two-volume treatise on the motion of celestial bodies. Gauss was concerned with the study of

[1] Carl Friedrich Gauss (30 April 1777, Brunswick, Duchy of Brunswick, Holy Roman Empire (now Germany) – 23 February 1855, Göttingen, Hanover, Germany) made outstanding contributions to mathematics, physics and astronomy. He gave the first proof, in 1799, of the Fundamental Theorem of Algebra. Gauss worked in differential geometry, number theory, algebra and non-Euclidean geometry.

[2] Note, for example, that $\frac{2}{3}100^3 \approx 0.67 \times 10^6 \ll 101!\,\mathrm{e} \approx 2.56 \times 10^{160}$. On a computer that performs 10^{12} floating operations a second a calculation requiring 10^6 operations *via* Gaussian elimination would take 10^{-6} seconds, as opposed to the 8.11×10^{140} years using Cramer's rule or formula (2.5).

[3] See, however, the bibliographical notes at the end of the chapter about the priority dispute between Legendre and Gauss.

the asteroid Pallas, and derived a set of six linear equations with six unknowns, also giving a systematic method for its solution.

The method proceeds by successively eliminating the elements below the diagonal of the matrix of the linear system until the matrix becomes triangular, when the solution of the system is very easy. This technique is now known under the name **Gaussian elimination**.[1]

Before we embark on the general description of Gaussian elimination, let us illustrate its basic steps through a simple example; this is the same as Example 2.1 above, written out again for convenience.

Example 2.2 *Consider the system of linear equations*

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

It is convenient to rewrite this in the form $A\boldsymbol{x} = \boldsymbol{b}$ where $A \in \mathbb{R}^{3\times 3}$ and $\boldsymbol{x}$ and $\boldsymbol{b}$ are column vectors of size 3; thus,

$$\begin{pmatrix} 1 & 1 & 1 \\ 2 & 4 & 2 \\ -1 & 5 & -4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 16 \\ -3 \end{pmatrix}. \tag{2.9}$$

We begin by adding the first row, multiplied by -2, to the second row, and adding the first row to the third row, giving the new system

$$\begin{pmatrix} 1 & 1 & 1 \\ \mathit{0} & 2 & 0 \\ \mathit{0} & 6 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \\ 3 \end{pmatrix}. \tag{2.10}$$

The newly created 0 entries in the first column have been typeset in italics. Now adding the new second row, multiplied by -3, to the third row, we find

$$\begin{pmatrix} 1 & 1 & 1 \\ \mathit{0} & 2 & 0 \\ \mathit{0} & \mathit{0} & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 4 \\ -9 \end{pmatrix}, \tag{2.11}$$

[1] The idea of this elimination process was already known to the Chinese two thousand years ago. The book *Jiu zhang suan shu* (English translation, by K. Shen *et al.*: *The Nine Chapters on the Mathematical Art*, Oxford University Press, 1999) contained an example of the elimination for a system of five equations with five unknowns. This book was very influential in the history of Chinese mathematics, and is the earliest specialised mathematical work in China that survived to the present day. Although it is unclear when its mathematical content was produced, it is estimated that the book was assembled during the Han dynasty in the first century AD.

which can easily be solved for the unknowns in the reverse order, beginning with $x_3 = 3$. ◇

Each of these successive row operations can be expressed as a multiplication on the left of the matrix $A \in \mathbb{R}^{n\times n}$, $n \geq 2$ (in our example $n = 3$), of the system of linear equations by a transformation matrix. Writing $E^{(rs)}$ for the $n \times n$ matrix whose only nonzero element is $e_{rs} = 1$, we see that the product

$$(I + \mu_{rs} E^{(rs)})A \tag{2.12}$$

is the same as the original matrix A, except that the elements of row s, multiplied by a real number μ_{rs}, have been added to the corresponding elements of row r. Here I denotes the $n \times n$ identity matrix defined by (2.3). In the elimination process we always add a multiple of an earlier row to a later row in the matrix, so that $1 \leq s < r \leq n$ in (2.12); the transformation matrix $I + \mu_{rs} E^{(rs)}$ is therefore lower triangular in the following sense.

Definition 2.2 *Let n be an integer, $n \geq 2$. The matrix $L \in \mathbb{R}^{n\times n}$ is said to be* **lower triangular** *if $l_{ij} = 0$ for every i and j with $1 \leq i < j \leq n$. The matrix $L \in \mathbb{R}^{n\times n}$ is called* **unit lower triangular** *if it is lower triangular, and also the diagonal elements are all equal to unity, that is $l_{ii} = 1$ for $i = 1, 2, \ldots, n$.*

Thus the matrix $I+\mu_{rs}E^{(rs)} \in \mathbb{R}^{n\times n}$ appearing in (2.12) is unit lower triangular if $1 \leq s < r \leq n$, and the above elimination process can be expressed by multiplying A on the left successively by the unit lower triangular matrices $I+\mu_{rs}E^{(rs)}$ for $r = s+1, \ldots, n$ and $s = 1, \ldots, n-1$, with $\mu_{rs} \in \mathbb{R}$; there are $\frac{1}{2}n(n-1)$ of these matrices, one for each element of A below the diagonal (since there are n elements on the diagonal and, therefore, $1 + \cdots + (n-1) = \frac{1}{2}(n^2 - n)$ elements below the diagonal). The next theorem lists the technical tools which are required for proving that the resulting product is a lower triangular matrix.

Theorem 2.1 *The following statements hold for any integer $n \geq 2$:*

(i) *the product of two lower triangular matrices of order n is lower triangular of order n;*

(ii) *the product of two unit lower triangular matrices of order n is unit lower triangular of order n;*

(iii) *a lower triangular matrix is nonsingular if, and only if, all the*

diagonal elements are nonzero; in particular, a unit lower triangular matrix is nonsingular;

(iv) *the inverse of a nonsingular lower triangular matrix of order n is lower triangular of order n;*

(v) *the inverse of a unit lower triangular matrix of order n is unit lower triangular of order n.*

Proof The proofs of parts (i), (ii), (iii) and (v) are very straightforward, and are left as an exercise.

Part (iv) is proved by induction; it is easily verified for a nonsingular lower triangular matrix of order 2, using (2.5). Let $n > 2$, suppose that (iv) is true for all nonsingular lower triangular matrices of order k, with $2 \leq k < n$, and let L be a nonsingular lower triangular matrix of order $k+1$. Both L and its inverse L^{-1} can be partitioned by their last row and column:

$$L = \begin{pmatrix} L_1 & \mathbf{0} \\ \boldsymbol{r}^{\mathrm{T}} & \alpha \end{pmatrix}, \quad L^{-1} = \begin{pmatrix} X & \boldsymbol{y} \\ \boldsymbol{z}^{\mathrm{T}} & \beta \end{pmatrix},$$

where L_1 is a nonsingular lower triangular matrix of order k and $X \in \mathbb{R}^{k\times k}$; α and β are real numbers and $\boldsymbol{r}$, $\boldsymbol{z}$ and $\boldsymbol{y}$ are column vectors of size k. Since the product LL^{-1} is the identity matrix of order $k+1$, we have

$$L_1 X = I_k\,, \quad L_1 \boldsymbol{y} = \mathbf{0}\,, \quad \boldsymbol{r}^{\mathrm{T}} X + \alpha \boldsymbol{z}^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}\,, \quad \boldsymbol{r}^{\mathrm{T}} \boldsymbol{y} + \alpha\beta = 1\,;$$

here I_k signifies the identity matrix of order k. Thus $X = L_1^{-1}$, which is lower triangular of order k by the inductive hypothesis, and $\boldsymbol{y} = \mathbf{0}$ given that L_1 is nonsingular; the remaining two equations determine $\boldsymbol{z}$ and β on noting that $\alpha \neq 0$ (given that L is nonsingular). This shows that L^{-1} is lower triangular of order $k+1$, and the inductive step is complete; consequently, (iv) is true for any $n \geq 2$. □

We shall also require the concept of upper triangular matrix.

Definition 2.3 *Let n be an integer, $n \geq 2$. The matrix $U \in \mathbb{R}^{n\times n}$ is said to be* **upper triangular** *if $u_{ij} = 0$ for every i and j with $1 \leq j < i \leq n$.*

We note that results analogous to those in the preceding theorem concerning lower triangular matrices are also valid for upper triangular matrices (replacing the words 'lower triangular' by 'upper triangular' throughout).

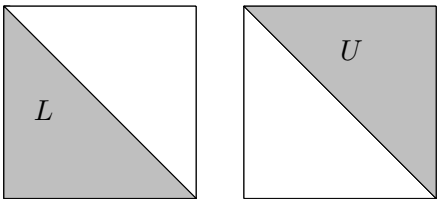

Fig. 2.1. LU factorisation of $A \in \mathbb{R}^{n\times n}$: $A = LU$. The matrix $L \in \mathbb{R}^{n\times n}$ is unit lower triangular and $U \in \mathbb{R}^{n\times n}$ is upper triangular.

The elimination process for $A \in \mathbb{R}^{n\times n}$ may now be written as follows:

$$L_{(N)}L_{(N-1)}\dots L_{(1)}A = U\,, \qquad N = \tfrac{1}{2}n(n-1)\,, \tag{2.13}$$

where $U \in \mathbb{R}^{n\times n}$ is an upper triangular matrix and each of the matrices $L_{(j)} \in \mathbb{R}^{n\times n}$, $j = 1,\dots,N$, is unit lower triangular of order n and has the form $I+\mu_{rs}E^{(rs)}$ with $1 \le s < r \le n$, where I is the identity matrix of order n. That is,

$$L_{(1)} = I+\mu_{21}E^{(21)},\ L_{(2)} = I+\mu_{31}E^{(31)},\ \dots,\ L_{(N)} = I+\mu_{n\,n-1}E^{(n\,n-1)}.$$

It is easy to see that $E^{(rs)}E^{(rs)} = \delta_{rs}E^{(rs)}$, where

$$\delta_{rs} = \begin{cases} 1 & \text{for } r = s\,, \\ 0 & \text{for } r \ne s \end{cases}$$

is known as the **Kronecker delta**.[1] Thus, for $1 \le s < r \le n$, the inverse of the matrix $I + \mu_{rs}E^{(rs)}$ is the lower triangular matrix $I - \mu_{rs}E^{(rs)}$, which corresponds to the subtraction of row s, multiplied by μ_{rs}, from row r. Hence

$$A = L_{(1)}^{-1}\dots L_{(N)}^{-1}U = LU\,, \tag{2.14}$$

where L, as the product of a finite number of unit lower triangular matrices of order n, is itself unit lower triangular of order n by Theorem 2.1(ii); see Figure 2.1.

2.3 LU factorisation

Having seen that the Gaussian elimination process gives rise to the factorisation $A = LU$ of the matrix $A \in \mathbb{R}^{n\times n}$, $n \ge 2$, where L is unit

[1] Leopold Kronecker (7 December 1823, Liegnitz, Prussia, Germany (now Legnica, Poland) – 29 December 1891, Berlin, Germany) made significant contributions to the theory of elliptic functions, the theory of ideals and the algebra of quadratic forms.

lower triangular and U is upper triangular, we shall now show how to calculate the elements of L and U directly. Equating the elements of A and LU we conclude that

$$a_{ij} = \sum_{k=1}^{n} l_{ik} u_{kj} \,, \qquad 1 \le i, j \le n \,. \tag{2.15}$$

Recalling that L and U are lower and upper triangular respectively, we see that, in fact, the range of k in this sum extends only up to $\min\{i, j\}$, the smaller of the numbers i and j. Taking the two cases separately gives

$$a_{ij} = \sum_{k=1}^{j} l_{ik} u_{kj} \,, \qquad 1 \le j < i \le n \,, \tag{2.16}$$

$$a_{ij} = \sum_{k=1}^{i} l_{ik} u_{kj} \,, \qquad 1 \le i \le j \le n \,. \tag{2.17}$$

Rearranging these equations, and using the fact that $l_{ii} = 1$ for all $i = 1, 2, \ldots, n$, we find that

$$l_{ij} = \frac{1}{u_{jj}} \left\{ a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \right\}, \quad \begin{array}{l} i = 2, \ldots, n \,, \\ j = 1, \ldots, i-1 \,, \end{array} \tag{2.18}$$

$$u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj} \,, \quad \begin{array}{l} i = 1, \ldots, n \,, \\ j = i, \ldots, n \,, \end{array} \tag{2.19}$$

with the convention that sums over empty index sets are equal to zero. Thus, the elements of U in the first row are $u_{1j} = a_{1j}$, $j = 1, 2, \ldots, n$, and the elements of L in the first column are $l_{11} = 1$ and $l_{i1} = a_{i1}/u_{11}$, $i = 2, \ldots, n$.

The equations (2.18) and (2.19) can now be used for the calculation of the elements l_{ij} and u_{ij}. For each value of i, starting with $i = 2$, we calculate first l_{ij}, for $j = 1, \ldots, i-1$ in order, and then the values of u_{ij}, for $j = i, \ldots, n$, again in increasing order. We then move on to the same calculation for $i + 1$, and so on until $i = n$. In the calculation of l_{ij} we need the values of u_{kj}, $1 \le k \le j < i - 1$, from previous rows, and we also need the values of l_{ik}, $1 \le k \le j - 1$, in the same row but in previous columns; a similar argument applies to the calculation of u_{ij}. When carried out in this order, all the values required at each step have already been calculated.

Of course, we must ensure that the calculation does not fail because of division by zero; this requires that none of the u_{jj}, $j = 1, \ldots, n - 1$,

in the formula (2.18) is zero. To investigate this possibility we use the properties of certain submatrices of A.

Definition 2.4 *Suppose that $A \in \mathbb{R}^{n\times n}$ with $n \geq 2$, and let $1 \leq k \leq n$. The* **leading principal submatrix** *of order k of A is defined as the matrix $A^{(k)} \in \mathbb{R}^{k\times k}$ whose element in row i and column j is equal to the element of the matrix A in row i and column j for $1 \leq i, j \leq k$.*

Armed with this definition, we can now formulate the main result of this section. It provides a sufficient condition for ensuring that the algorithm (2.18), (2.19) for calculating the entries of the matrices L and U in the LU factorisation $A = LU$ of a matrix $A \in \mathbb{R}^{n\times n}$ does not break down due to division by zero in (2.18).

Theorem 2.2 *Let $n \geq 2$, and suppose that $A \in \mathbb{R}^{n\times n}$ is such that every leading principal submatrix $A^{(k)} \in \mathbb{R}^{k\times k}$ of A of order k, with $1 \leq k < n$, is nonsingular. (Note that A itself is not required to be nonsingular.) Then, A can be factorised in the form $A = LU$, where $L \in \mathbb{R}^{n\times n}$ is unit lower triangular and $U \in \mathbb{R}^{n\times n}$ is upper triangular.*

Proof The proof is by induction on the order n. Let us begin by verifying the statement of the theorem for $n = 2$. We intend to show that any 2×2 matrix

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

with $a \neq 0$, is equal to the product of a unit lower triangular matrix L of order 2 and an upper triangular matrix U of order 2; that is, we wish to establish the existence of

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

such that $LU = A$, where m, u, v and η are four real numbers, to be determined. Equating the product LU with A, we deduce that

$$u = a\,, \quad v = b\,, \quad mu = c\,, \quad mv + \eta = d\,.$$

Since $a \neq 0$ by hypothesis, the first of these equalities implies that $u \neq 0$ also; hence $m = c/u$, $v = b$, and $\eta = d - mv$. Thus we have shown the existence of the required matrices L and U in $\mathbb{R}^{2\times 2}$ and completed the proof for $n = 2$.

Now, suppose that the statement of the theorem has already been verified for matrices of order k, $2 \le k < n$; suppose that $A \in \mathbb{R}^{(k+1)\times(k+1)}$ and all leading principal submatrices of A of order k and less are nonsingular. We mimic the proof in the case of $n = 2$ by partitioning A into blocks by the last row and column:

$$A = \begin{pmatrix} A^{(k)} & \boldsymbol{b} \\ \boldsymbol{c}^{\mathrm{T}} & d \end{pmatrix},$$

where $A^{(k)} \in \mathbb{R}^{k\times k}$ is a nonsingular matrix (all of whose leading principal submatrices are themselves nonsingular), $\boldsymbol{b}$, $\boldsymbol{c}$ are column vectors of size k, and d is a real number. According to our inductive hypothesis, there exist a unit lower triangular matrix $L^{(k)}$ of order k and an upper triangular matrix $U^{(k)}$ of order k such that $A^{(k)} = L^{(k)}U^{(k)}$. Thus we shall seek the desired unit lower triangular matrix L of order $k+1$ and the upper triangular matrix U of order $k+1$ in the form

$$L = \begin{pmatrix} L^{(k)} & \mathbf{0} \\ \boldsymbol{m}^{\mathrm{T}} & 1 \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} U^{(k)} & \boldsymbol{v} \\ \mathbf{0}^{\mathrm{T}} & \eta \end{pmatrix}$$

where $\boldsymbol{m}$ and $\boldsymbol{v}$ are column vectors of size k and η is a real number, to be determined from the requirement that the product LU be equal to the matrix A. On equating LU with A, we obtain

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

The first of these four equalities provides no new information. However, we can use the remaining three to determine the column vectors $\boldsymbol{v}$ and $\boldsymbol{m}$ and the real number η. Since $L^{(k)}$ is unit lower triangular, its determinant is equal to 1; therefore $L^{(k)}$ is nonsingular. This means that the second equation uniquely determines the unknown column vector $\boldsymbol{v}$. Further, since by hypothesis $A^{(k)}$ is nonsingular and $A^{(k)} = L^{(k)}U^{(k)}$, we conclude that

$$\det(A^{(k)}) = \det(L^{(k)}U^{(k)}) = \det(L^{(k)})\det(U^{(k)}) = \det(U^{(k)})\,;$$

given that $\det(A^{(k)}) \neq 0$ by the inductive hypothesis, this implies that $\det(U^{(k)}) \neq 0$ also, and therefore the third equation uniquely determines $\boldsymbol{m}$. Having found $\boldsymbol{v}$ and $\boldsymbol{m}$, the fourth equation yields $\eta = d - \boldsymbol{m}^{\mathrm{T}}\boldsymbol{v}$. Thus we have shown the existence of the desired matrices L and U of order $k+1$, and the inductive step is complete.[1] □

[1] In the last paragraph we made use of the **Binet–Cauchy Theorem** which states that for three matrices A, B, C in $\mathbb{R}^{k\times k}$ with $A = BC$, we have $\det(A) = \det(B)\det(C)$. This result was proved in 1812 independently by Augustin-Louis Cauchy (1789–1857) and Jacques Philippe Marie Binet (1786–1856).

2.4 Pivoting

The aim of this section is to show that even if the matrix A does not satisfy the conditions of Theorem 2.2, by permuting rows and columns it can be transformed into a new matrix $\tilde{A}$ of the same size so that $\tilde{A}$ admits an LU factorisation.

Example 2.3 *Consider, for example, the system obtained from (2.9) by replacing the coefficient of x_1 in the first equation by zero. Then, the leading element in the matrix A is zero, the computation fails at the first step, and the LU factorisation of A does not exist. However if we interchange the first two equations we obtain a new matrix $\tilde{A}$ which is the same as A but with the first two rows interchanged,*

$$\tilde{A} = \begin{pmatrix} 2 & 4 & 2 \\ 0 & 1 & 1 \\ -1 & 5 & -4 \end{pmatrix}. \tag{2.20}$$

Since the leading principal submatrices of order 1 and 2 of $\tilde{A}$ are non-singular, by Theorem 2.2 the matrix $\tilde{A}$ now has the required LU factorisation, which is easily computed.

A computation which fails when an element is exactly zero is also likely to run into difficulties when that element is nonzero but of very small absolute value; the problem stems from the presence of rounding errors. The basic operation in the elimination process consists of multiplying the elements of one row of the matrix by a scalar μ_{rs}, and adding to the elements of another row. The multiplication operation will always introduce a rounding error, so the elements which are multiplied by μ_{rs} will already contain a rounding error from operations with earlier rows of the matrix; these errors will therefore themselves be multiplied by μ_{rs} before adding to the new row. The errors will be magnified if $|\mu_{rs}| > 1$, and will be greatly magnified if $|\mu_{rs}| \gg 1$.

The accumulation of rounding errors alluded to in the previous paragraph can be alleviated by permuting the rows of the matrix. Thus, at each stage of the elimination process we interchange two rows, if necessary, so that the largest element in the current column lies on the diagonal. This process is known as **pivoting**. Clearly, when pivoting is performed none of the multipliers μ_{rs} have absolute value greater than unity. The process is easily formalised by introducing permutation matrices. This leads us to our next definition.

Definition 2.5 *Suppose that $n \geq 2$. A matrix $P \in \mathbb{R}^{n\times n}$ in which every element is either 0 or 1, and whose every row and every column contain exactly one nonzero element, is called a* **permutation matrix**.

Example 2.4 *Here are three of the possible* 3! *permutation matrices in* $\mathbb{R}^{3\times 3}$:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

The proof of our next result is elementary and is left to the reader.

Lemma 2.1 *Let $n \geq 2$ and suppose that $P \in \mathbb{R}^{n\times n}$ is a permutation matrix. Then, the following statements hold:*

(i) *given that I is the identity matrix of order n, the matrix P can be obtained from I by permuting rows;*

(ii) *if $Q \in \mathbb{R}^{n\times n}$ is another permutation matrix, then the products PQ and QP are also permutation matrices;*

(iii) *let $P^{(rs)} \in \mathbb{R}^{n\times n}$ denote the* **interchange matrix**, *obtained from the identity matrix $I \in \mathbb{R}^{n\times n}$ by interchanging rows r and s; any interchange matrix is a permutation matrix; moreover, any permutation matrix of order n can be written as a product of interchange matrices of order n;*

(iv) *the determinant of a permutation matrix $P \in \mathbb{R}^{n\times n}$ is equal to 1 or -1, depending on whether P is obtained from the identity matrix of order n by an even or odd number of permutations of rows, respectively; in particular, a permutation matrix is nonsingular.*

Now we are ready to prove the next theorem.

Theorem 2.3 *Let $n \geq 2$ and $A \in \mathbb{R}^{n\times n}$. There exist a permutation matrix P, a unit lower triangular matrix L, and an upper triangular matrix U, all three in $\mathbb{R}^{n\times n}$, such that*

$$PA = LU\,. \tag{2.21}$$

Proof The proof is by induction on the order n. Let $n = 2$ and consider the matrix

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

If $a \neq 0$, the proof follows from Theorem 2.2 with P taken as the 2×2 identity matrix. If $a = 0$ but $c \neq 0$, we take

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

and write

$$PA = \begin{pmatrix} c & d \\ 0 & b \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} c & d \\ 0 & b \end{pmatrix} \equiv LU\,.$$

If $a = 0$ and $c = 0$, the result trivially follows by writing

$$\begin{pmatrix} 0 & b \\ 0 & d \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & b \\ 0 & d \end{pmatrix} \equiv LU$$

and taking P as the 2×2 identity matrix. That completes the proof for $n = 2$.

Now, suppose that $A \in \mathbb{R}^{(k+1)\times(k+1)}$ and assume that the theorem holds for every matrix of order k with $2 \leq k < n$. We begin by locating the element in the first column of A which has the largest absolute value, or any one of them if there is more than one such element, and interchange rows if required; if the largest element is in row r we interchange rows 1 and r. We then partition the new matrix according to the first row and column, writing

$$P^{(1r)}A = \begin{pmatrix} \alpha & \boldsymbol{w}^{\mathrm{T}} \\ \boldsymbol{p} & B \end{pmatrix} = \begin{pmatrix} 1 & \boldsymbol{0}^{\mathrm{T}} \\ \boldsymbol{m} & I \end{pmatrix} \begin{pmatrix} \alpha & \boldsymbol{v}^{\mathrm{T}} \\ \boldsymbol{0} & C \end{pmatrix} \tag{2.22}$$

where α is the element of largest absolute value in the first column, $B, C \in \mathbb{R}^{k\times k}$, and $\boldsymbol{p}$, $\boldsymbol{w}$, $\boldsymbol{m}$ and $\boldsymbol{v}$ are column vectors of size k, with $\boldsymbol{m}$, $\boldsymbol{v}$ and C to be determined. Writing out the product we find that

$$\left.\begin{aligned} \boldsymbol{v}^{\mathrm{T}} &= \boldsymbol{w}^{\mathrm{T}}\,, \\ \alpha\boldsymbol{m} &= \boldsymbol{p}\,, \\ C &= B - \boldsymbol{m}\boldsymbol{v}^{\mathrm{T}}\,. \end{aligned}\right\} \tag{2.23}$$

If $\alpha = 0$, then the first column of A consists entirely of zeros ($\boldsymbol{p} = \boldsymbol{0}$); in this case we can evidently choose $\boldsymbol{m} = \boldsymbol{0}$, $\boldsymbol{v} = \boldsymbol{w}$ and $C = B$. Suppose now that $\alpha \neq 0$; then $\boldsymbol{m} = (1/\alpha)\boldsymbol{p}$, so that all the elements of $\boldsymbol{m}$ have absolute value less than or equal to unity, since α is the largest in absolute value element in the first column. By the inductive hypothesis we can now write

$$P^*C = L^*U^*\,, \tag{2.24}$$

where P^*, L^*, $U^* \in \mathbb{R}^{k\times k}$, P^* is a permutation matrix, L^* is unit lower triangular, and U^* is upper triangular. Hence, by (2.23),

$$P^{(1r)}A = \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P^* \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ P^*\boldsymbol{m} & L^* \end{pmatrix} \begin{pmatrix} \alpha & \boldsymbol{v}^{\mathrm{T}} \\ \mathbf{0} & U^* \end{pmatrix} \tag{2.25}$$

since $P^*P^* = I$. Now, defining the permutation matrix P by

$$P = \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ \mathbf{0} & P^* \end{pmatrix} P^{(1r)}, \tag{2.26}$$

we obtain

$$PA = \begin{pmatrix} 1 & \mathbf{0}^{\mathrm{T}} \\ P^*\boldsymbol{m} & L^* \end{pmatrix} \begin{pmatrix} \alpha & \boldsymbol{v}^{\mathrm{T}} \\ \mathbf{0} & U^* \end{pmatrix}, \tag{2.27}$$

which is the required factorisation of $A \in \mathbb{R}^{(k+1)\times(k+1)}$. This completes the inductive step. The theorem therefore holds for every matrix of order $n \geq 2$. □

The proof of this theorem also contains an algorithm for constructing the permutation matrix P, and the matrices L and U. The permutation matrix is conveniently described by specifying the sequence of interchanges: given the $n-1$ integers $p_1, p_2, \ldots, p_{n-1}$, the matrix P is the product of the permutation matrices which interchange rows 1 and p_1, 2 and p_2, and so on.

2.5 Solution of systems of equations

Consider the linear system $A\boldsymbol{x} = \boldsymbol{b}$ where $A \in \mathbb{R}^{n\times n}$ and $\boldsymbol{x}$ and $\boldsymbol{b}$ are column vectors of size n. According to Theorem 2.3 there exist a permutation matrix $P \in \mathbb{R}^{n\times n}$, a unit lower triangular matrix $L \in \mathbb{R}^{n\times n}$ and an upper triangular matrix $U \in \mathbb{R}^{n\times n}$ such that $PA = LU$. Having obtained the LU factorisation of the matrix PA, the solution of the system of linear equations $A\boldsymbol{x} = \boldsymbol{b}$ is straightforward: multiplying both sides of $A\boldsymbol{x} = \boldsymbol{b}$ on the left by the permutation matrix P, we obtain that

$$PA\boldsymbol{x} = P\boldsymbol{b}; \tag{2.28}$$

equivalently, $LU\boldsymbol{x} = P\boldsymbol{b}$. On defining $\boldsymbol{y} = U\boldsymbol{x}$ we can rewrite (2.28) as the following coupled set of linear equations:

$$L\boldsymbol{y} = P\boldsymbol{b}, \qquad U\boldsymbol{x} = \boldsymbol{y}. \tag{2.29}$$

Assuming that the matrix P and the LU factorisation of PA are already known, there are three stages to the calculation of $\boldsymbol{x}$:

`Step 1.` First we apply the sequence of permutations to the vector $\boldsymbol{b}$, to produce $P\boldsymbol{b}$;

`Step 2.[Forward substitution]` We then solve the lower triangular system $L\boldsymbol{y} = P\boldsymbol{b}$, calculating the elements in the order $y_1,\ y_2, \ldots, y_n$;

`Step 3. [Backsubstitution]` Finally the required solution $\boldsymbol{x}$ is obtained from the upper triangular system $U\boldsymbol{x} = \boldsymbol{y}$, calculating the elements of $\boldsymbol{x}$ in the reverse order, $x_n,\ x_{n-1}, \ldots, x_1$.

`Step 3` will break down if any of the diagonal elements of U are zero, but if this happens the matrix A is singular.

The next section is devoted to assessing the amount of computational work for this algorithm.

2.6 Computational work

In this section we shall show that the work involved in factorising an $n \times n$ matrix in the form $A = LU$ is proportional to n^3. An estimate of the amount of computational work of this kind is important in deciding in advance how long a calculation would take for a very large matrix, and is also useful in comparing different methods for the solution of a given problem. For example, in the next chapter we shall derive a method for solving a system of equations with a symmetric positive definite matrix; that method requires only half the amount of work involved in the standard LU factorisation algorithm which takes no account of symmetry.

Accurate estimates of the time taken by a computation are very complicated and require some detailed knowledge of the computer being used. The estimates which we shall give are simple but crude; they are normally good enough for the types of comparisons we have just mentioned.

We see from (2.18) that the calculation of l_{ij} requires $j-1$ multiplications, $j-2$ additions, 1 subtraction and 1 division, a total of $2j-1$ operations. In the same way, (2.19) shows that the calculation of u_{ij} requires $2i-2$ operations.[1] Recalling that, for any integer $k \geq 2$,

$$1 + \cdots + k = \tfrac{1}{2}k(k+1) \quad \text{and} \quad 1^2 + \cdots + k^2 = \tfrac{1}{6}k(k+1)(2k+1)\,,$$

we then deduce that the total number of operations involved in the LU

[1] We do not count the row interchanges in the number of 'operations'.

factorisation is

$$\sum_{i=2}^{n}\sum_{j=1}^{i-1}(2j-1)+\sum_{i=1}^{n}\sum_{j=i}^{n}2(i-1)=\tfrac{1}{6}n(n-1)(4n+1)\,.$$

It is enough to say that the number of multiplications required is about $\frac{2}{3}n^3-\frac{1}{2}n^2$, for moderately large values of n.

Having constructed the factorisation we can now count the number of operations required to compute the vectors $\boldsymbol{y}$ and $\boldsymbol{x}$ in (2.29). Given the vector $P\boldsymbol{b}$, the elements of $\boldsymbol{y}$ are obtained from

$$y_1=(P\boldsymbol{b})_1\,,\qquad y_i=(P\boldsymbol{b})_i-\sum_{j=1}^{i-1}l_{ij}y_j\,,\quad i=2,3,\ldots,n\,,\qquad(2.30)$$

which requires $2i-2$ operations. Summing over i this gives a total of $n(n-1)$. The calculation of the elements of $\boldsymbol{x}$ is similar:

$$x_i=\frac{1}{u_{ii}}\left(y_i-\sum_{j=i+1}^{n}u_{ij}x_j\right)\,,\qquad i=1,2,\ldots,n\,.\qquad(2.31)$$

This requires $2(n-i)+1$ operations, giving a total of n^2.

The total number of operations involved in the solution of the system of equations is therefore approximately $\frac{2}{3}n^3-\frac{1}{2}n^2$ for the factorisation, followed by $n(n-1)+n^2=2n^2-n$ for the solution of the two triangular systems, that is, approximately $\frac{2}{3}n^3+\frac{3}{2}n^2$, ignoring terms of size $\mathcal{O}(n)$.

We often need to solve a number of systems of this kind, all with different right-hand sides, but with the same matrix. We then need only factorise the matrix once, and the total number of multiplications required for k right-hand sides becomes approximately $\frac{2}{3}n^3+\left(2k-\frac{1}{2}\right)n^2$. When k is fairly large it might appear that it would be more efficient to form the inverse matrix A^{-1}, and then multiply each right-hand side by the inverse; but we shall show that it is not so.

To form the inverse matrix we first factorise the matrix A, and then solve n systems, with the right-hand sides being the vectors which constitute the columns of the identity matrix. Because these right-hand sides have a special form, there is the possibility of saving some work; some careful counting shows that the total can be reduced from $\frac{2}{3}n^3+2n^3=\frac{8}{3}n^3$ to an approximate total of $2n^3$ operations. It is easy to see that the operation of multiplying a vector by the inverse matrix requires $n(2n-1)$ operations; hence the whole computation of first constructing the inverse matrix, and then multiplying each right-hand side by the inverse, requires a total of $2n^3+2kn^2$ multiplications (ignoring terms of size

$\mathcal{O}(n)$). This is always greater than the previous value $\frac{2}{3}n^3 + \left(2k - \frac{1}{2}\right) n^2$, whether k is small or large. The most efficient way of solving this problem is to construct and save the L and U factors of A, rather than to form the inverse of A.

2.7 Norms and condition numbers

The analysis of the effects of rounding error on solutions of systems of linear equations requires an appropriate measure. This is provided by the concept of **norm** defined below. In order to motivate the axioms of norm stated in Definition 2.6, we note that the set $\mathbb{R}$ of real numbers is a linear space, and that the **absolute value** function

$$v \in \mathbb{R} \mapsto |v| = \begin{cases} v & \text{if } v \geq 0, \\ -v & \text{if } v < 0 \end{cases}$$

has the following properties:

- $|v| \geq 0$ for any $v \in \mathbb{R}$, and $|v| = 0$ if, and only if, $v = 0$;
- $|\lambda v| = |\lambda|\,|v|$ for all $\lambda \in \mathbb{R}$ and all $v \in \mathbb{R}$;
- $|u + v| \leq |u| + |v|$ for all u and v in $\mathbb{R}$.

The absolute value $|v|$ of a real number v measures the distance between v and 0 (the zero element of the linear space $\mathbb{R}$). Our next definition aims to generalise this idea to an arbitrary linear space $\mathcal{V}$ over the field $\mathbb{R}$ of real numbers: even though the discussion in the present chapter is confined to finite-dimensional linear spaces of vectors ($\mathcal{V} = \mathbb{R}^n$) and square matrices ($\mathcal{V} = \mathbb{R}^{n \times n}$), norms over other linear spaces, including infinite-dimensional function spaces, will appear elsewhere in the text (see Chapters 8, 9, 11 and 14).

Definition 2.6 *Suppose that $\mathcal{V}$ is a linear space over the field $\mathbb{R}$ of real numbers. The nonnegative real-valued function $\|\cdot\|$ is said to be a* **norm** *on the space $\mathcal{V}$ provided that it satisfies the following axioms:*

❶ $\|v\| = 0$ *if, and only if,* $v = 0$ *in* $\mathcal{V}$*;*
❷ $\|\lambda v\| = |\lambda|\,\|v\|$ *for all* $\lambda \in \mathbb{R}$ *and all* v *in* $\mathcal{V}$*;*
❸ $\|u + v\| \leq \|u\| + \|v\|$ *for all* u *and* v *in* $\mathcal{V}$ *(the triangle inequality).*

A linear space $\mathcal{V}$, equipped with a norm, is called a **normed linear space**.

Remark 2.1 *If $\mathcal{V}$ is a linear space over the field $\mathbb{C}$ of complex numbers, then $\mathbb{R}$ in the second axiom of Definition 2.6 should be replaced by $\mathbb{C}$, with $|\lambda|$ signifying the modulus of $\lambda \in \mathbb{C}$.*

Any norm on the linear space $\mathcal{V} = \mathbb{R}^n$ will be called a **vector norm**. Three vector norms are in common use in numerical linear algebra: the 1-norm $\|\cdot\|_1$, the 2-norm (or Euclidean norm) $\|\cdot\|_2$, and the ∞-norm $\|\cdot\|_\infty$; these are defined below.

Definition 2.7 *The* 1**-norm** *of the vector* $\boldsymbol{v} = (v_1, \ldots, v_n)^{\mathrm{T}} \in \mathbb{R}^n$ *is defined by*

$$\|\boldsymbol{v}\|_1 = \sum_{i=1}^{n} |v_i| \,. \tag{2.32}$$

Definition 2.8 *The* 2**-norm** *of the vector* $\boldsymbol{v} = (v_1, \ldots, v_n)^{\mathrm{T}} \in \mathbb{R}^n$ *is defined by* $\|\boldsymbol{v}\|_2 = \left(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}\right)^{1/2}$. *In other words,*

$$\|\boldsymbol{v}\|_2 = \left\{ \sum_{i=1}^{n} |v_i|^2 \right\}^{1/2} . \tag{2.33}$$

Definition 2.9 *The* ∞**-norm** *of the vector* $\boldsymbol{v} = (v_1, \ldots, v_n)^{\mathrm{T}} \in \mathbb{R}^n$ *is defined by*

$$\|\boldsymbol{v}\|_\infty = \max_{i=1}^{n} |v_i| \,. \tag{2.34}$$

When $n = 1$, each of these norms collapses to the absolute value, $|\cdot|$, the simplest example of a norm on $\mathcal{V} = \mathbb{R}$.

It is easy to show that $\|\cdot\|_1$ and $\|\cdot\|_\infty$ obey all axioms of a norm. For the 2-norm the first two axioms are still trivial to verify; to show that the triangle inequality is satisfied by the 2-norm requires use of the Cauchy[1]–Schwarz[2] inequality.

Lemma 2.2 (**Cauchy–Schwarz inequality**)

$$\left| \sum_{i=1}^{n} u_i v_i \right| \leq \|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2 \qquad \forall\, \boldsymbol{u},\, \boldsymbol{v} \in \mathbb{R}^n \,. \tag{2.35}$$

[1] Augustin-Louis Cauchy (21 August 1789, Paris, France – 23 May 1857, Sceaux (near Paris), France) made very significant contributions to algebra and number theory. He was one of the founders of modern mathematical analysis, the theory of complex functions, and the mathematics of elasticity theory.

[2] Karl Herman Amandus Schwarz (25 January 1843, Hermsdorf, Silesia, Germany (now in Poland) – 30 November 1921, Berlin, Germany) succeeded Karl Weierstrass as Professor of Mathematics at Berlin in 1892. Outside mathematics he acted as captain of the local Voluntary Fire Brigade, and helped the station-master at the local railway station by closing the doors of the trains.

Proof The proof of this inequality is rather simple: for any $\boldsymbol{u}$ and $\boldsymbol{v}$ in $\mathbb{R}^n$, and all $\lambda \in \mathbb{R}$,

$$\begin{aligned} 0 &\leq \|\lambda \boldsymbol{u} + \boldsymbol{v}\|_2^2 = \sum_{i=1}^{n} (\lambda u_i + v_i)^2 \\ &= \lambda^2 \sum_{i=1}^{n} |u_i|^2 + 2\lambda \sum_{i=1}^{n} u_i v_i + \sum_{i=1}^{n} |v_i|^2 . \end{aligned} \tag{2.36}$$

Hence, the expression on the right is a nonnegative quadratic polynomial in $\lambda \in \mathbb{R}$, of the form $A\lambda^2 + B\lambda + C$; therefore, the associated discriminant,

$$B^2 - 4AC = \left(2 \sum_{i=1}^{n} u_i v_i \right)^2 - 4 \left(\sum_{i=1}^{n} |u_i|^2 \right) \left(\sum_{i=1}^{n} |v_i|^2 \right),$$

is nonpositive. This implies (2.35) on recalling Definition 2.8. □

The triangle inequality for the 2-norm is now deduced as follows: letting $\lambda = 1$ in (2.36) and using (2.35), it follows that

$$\begin{aligned} \|\boldsymbol{u} + \boldsymbol{v}\|_2^2 &= \|\boldsymbol{u}\|_2^2 + 2 \sum_{i=1}^{n} u_i v_i + \|\boldsymbol{v}\|_2^2 \\ &\leq \|\boldsymbol{u}\|_2^2 + 2\|\boldsymbol{u}\|_2 \|\boldsymbol{v}\|_2 + \|\boldsymbol{v}\|_2^2 \\ &= \left(\|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2 \right)^2 , \end{aligned}$$

which yields the triangle inequality in the 2-norm on taking square roots. Hence $\| \cdot \|_2$ satisfies all three axioms of norm.

The 1-norm and the 2-norm on $\mathbb{R}^n$ are special cases of the p-norm, defined on $\mathbb{R}^n$, for $p \geq 1$, by

$$\|\boldsymbol{v}\|_p = \left\{ \sum_{i=1}^{n} |v_i|^p \right\}^{1/p} . \tag{2.37}$$

The first two axioms of norm are trivial to verify for $\| \cdot \|_p$; however, showing the triangle inequality is less straightforward (except for $p = 1$, and for $p = 2$, as we have already seen before); we shall now sketch the proof of this for $p > 1$. The starting point is the following result, known as Young's inequality.[1]

[1] William Henry Young (20 October 1863, London, England – 7 July 1942, Lausanne, Switzerland) studied mathematics at Peterhouse, Cambridge. His most important contributions were to the calculus of functions of several variables. Young was elected Fellow of the Royal Society in 1907; he was president of the London Mathematical Society (1922–1924) and president of the International Union of Mathematicians (1929–1936).

Theorem 2.4 (**Young's inequality**) *Let $p, q > 1$, $(1/p) + (1/q) = 1$. Then, for any two nonnegative real numbers a and b,*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}\,.$$

Proof If either $a = 0$ or $b = 0$ the inequality holds trivially. Let us therefore suppose that $a > 0$ and $b > 0$. We recall that a function $x \in \mathbb{R} \mapsto f(x) \in \mathbb{R}$ is said to be **convex** if

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

for all $\theta \in [0,1]$, and all x and y in $\mathbb{R}$; *i.e.*, for any x and y in $\mathbb{R}$ the graph of the function f between the points $(x, f(x))$ and $(y, f(y))$ lies below the chord that connects these two points. Note that the function $x \mapsto \mathrm{e}^x$ is convex. Therefore, with $\theta = 1/p$ and $1 - \theta = 1/q$, we get that

$$ab = \mathrm{e}^{\ln a + \ln b} = \mathrm{e}^{(1/p)\ln a^p + (1/q)\ln b^q} \leq \frac{1}{p}\,\mathrm{e}^{\ln a^p} + \frac{1}{q}\,\mathrm{e}^{\ln a^q} = \frac{a^p}{p} + \frac{b^q}{q}\,,$$

and the proof is complete. (When $p = q = 2$ the proof is trivial: as $(a-b)^2 \geq 0$ also $2ab \leq a^2 + b^2$, and hence the required result.) □

The next step is to establish Hölder's inequality;[1] it is a generalisation of the Cauchy–Schwarz inequality.

Theorem 2.5 (**Hölder's inequality**) *Let $p, q > 1$, $(1/p) + (1/q) = 1$. Then, for any $\boldsymbol{u} \in \mathbb{R}^n$ and $\boldsymbol{v} \in \mathbb{R}^n$, we have*

$$\left| \sum_{i=1}^{n} u_i v_i \right| \leq \|\boldsymbol{u}\|_p \|\boldsymbol{v}\|_q\,.$$

Proof If either $\boldsymbol{u} = \boldsymbol{0}$ or $\boldsymbol{v} = \boldsymbol{0}$ the inequality holds trivially. Let us therefore suppose that $\boldsymbol{u} \neq \boldsymbol{0}$ and $\boldsymbol{v} \neq \boldsymbol{0}$, and consider the vectors $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{v}}$ in $\mathbb{R}^n$ with components $\tilde{u}_i = u_i / \|\boldsymbol{u}\|_p$ and $\tilde{v}_i = v_i / \|\boldsymbol{v}\|_q$, respectively, $i = 1, 2, \ldots, n$. By Young's inequality,

$$\left| \sum_{i=1}^{n} \tilde{u}_i \tilde{v}_i \right| \leq \sum_{i=1}^{n} |\tilde{u}_i \tilde{v}_i| \leq \frac{1}{p} \sum_{i=1}^{n} |\tilde{u}_i|^p + \frac{1}{q} \sum_{i=1}^{n} |\tilde{v}_i|^q = \frac{1}{p} + \frac{1}{q} = 1\,.$$

Inserting the defining expressions for $\tilde{u}_i$ and $\tilde{v}_i$ into the left-most expression in this chain, the result follows. □

[1] Otto Ludwig Hölder (22 December 1859, Stuttgart, Germany – 29 August 1937, Leipzig, Germany) contributed to group theory; we owe him the concepts of factor group, and inner and outer automorphisms. Hölder discovered the inequality now named after him in 1884 while working on the convergence of Fourier series.

The triangle inequality in the p-norm is referred to as Minkowski's inequality.[1]

Theorem 2.6 (**Minkowski's inequality**) *Let* $1 \le p \le \infty$ *and* $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$. *Then,*

$$\|\boldsymbol{u} + \boldsymbol{v}\|_p \le \|\boldsymbol{u}\|_p + \|\boldsymbol{v}\|_p \,.$$

Proof As we noted earlier, the proof of this inequality for $p = 1$ and $p = \infty$ is easy. Let us therefore focus on the case $1 < p < \infty$. In the nontrivial case of $\boldsymbol{u} \neq \boldsymbol{0}$ and $\boldsymbol{v} \neq \boldsymbol{0}$, Hölder's inequality yields

$$\begin{aligned} \|\boldsymbol{u} + \boldsymbol{v}\|_p^p &= \sum_{i=1}^n |u_i + v_i|^p \le \sum_{i=1}^n |u_i + v_i|^{p-1} \left(|u_i| + |v_i|\right) \\ &\le \left(\sum_{i=1}^n |u_i + v_i|^p\right)^{\frac{p-1}{p}} \left(\left(\sum_{i=1}^n |u_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^n |v_i|^p\right)^{\frac{1}{p}}\right) \\ &= \|\boldsymbol{u} + \boldsymbol{v}\|_p^{p-1} \left(\|\boldsymbol{u}\|_p + \|\boldsymbol{v}\|_p\right) , \end{aligned}$$

and hence the desired result on dividing through by $\|\boldsymbol{u} + \boldsymbol{v}\|_p^{p-1}$. □

Remark 2.2 *For a nonzero element* $\boldsymbol{u}$ *in* $\mathbb{R}^n$, *let* $\tilde{\boldsymbol{u}} = (\|\boldsymbol{u}\|_\infty)^{-1}\boldsymbol{u}$. *Clearly,* $1 \le \|\tilde{\boldsymbol{u}}\|_p \le n^{1/p}$, *and hence* $\lim_{p\to\infty} \|\tilde{\boldsymbol{u}}\|_p = 1$. *Therefore,*

$$\|\boldsymbol{u}\|_\infty = \lim_{p\to\infty} \|\boldsymbol{u}\|_p \,, \qquad \boldsymbol{u} \in \mathbb{R}^n \,.$$

This identity justifies our use of the notation $\|\cdot\|_\infty$ *for the maximum norm, defined by* $\|\boldsymbol{u}\|_\infty = \max_{i=1}^n |u_i|$, *and our terminology:* ∞**-norm**.

Remark 2.3 *We note here that* $\|\cdot\|_p$, $1 \le p \le \infty$, *is also a norm on the linear space* $\mathbb{C}^n$ *of* n*-component vectors with complex entries, over the field* $\mathbb{C}$ *of complex numbers, provided that* $|v_i|$ *in the definition (2.37) of* $\|\cdot\|_p$ *is interpreted as the modulus of the complex number* v_i.

In order to highlight the difference between $\|\cdot\|_1$, $\|\cdot\|_2$ and $\|\cdot\|_\infty$, in Figure 2.2 we plot the 'unit spheres' (or 'unit circles', in the case of $n = 2$) corresponding to these three norms on $\mathcal{V} = \mathbb{R}^2$. We recall that

[1] Hermann Minkowski (22 June 1864, Alexotas, Russia (now Kaunas, Lithuania) – 12 January 1909, Göttingen, Germany) held a chair at the University of Göttingen, where he was exposed to Hilbert's work on mathematical physics. Minkowski realised that the ideas of Lorentz and Einstein can be best understood in terms of non-Euclidean geometry, with space and time coupled into a four-dimensional continuum. He died at the age of 44 from a ruptured appendix.

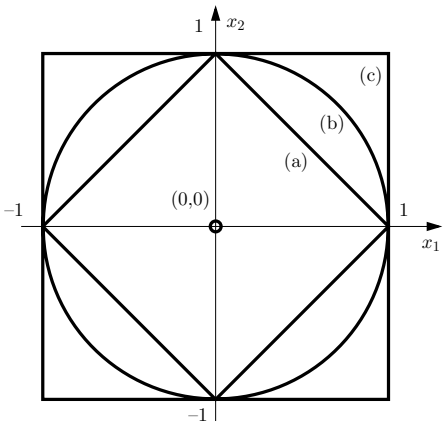

Fig. 2.2. 'Unit circles' in the linear space $\mathcal{V} = \mathbb{R}^2$ with respect to three vector norms: (a) the 1-norm; (b) the 2-norm; (c) the ∞-norm.

the unit sphere in a normed linear space $\mathcal{V}$, with norm $\|\cdot\|$, is defined as the set $\{\boldsymbol{v} \in \mathcal{V}\colon \|\boldsymbol{v}\| = 1\}$. It can be seen from Figure 2.2 that

$$\{\boldsymbol{v} \in \mathbb{R}^2\colon \|\boldsymbol{v}\|_1 \le 1\} \subset \{\boldsymbol{v} \in \mathbb{R}^2\colon \|\boldsymbol{v}\|_2 \le 1\} \subset \{\boldsymbol{v} \in \mathbb{R}^2\colon \|\boldsymbol{v}\|_\infty \le 1\}\,.$$

We leave it to the reader as an exercise to show that analogous inclusions hold in $\mathbb{R}^n$ for any $n \ge 1$. (See Exercise 8.)

The unit sphere in a normed linear space $\mathcal{V}$ with norm $\|\cdot\|$ is the boundary of the closed unit ball $\bar{B}_1(\mathbf{0})$ centred at $\mathbf{0}$ defined by

$$\bar{B}_1(\mathbf{0}) = \{\boldsymbol{v} \in \mathcal{V}\colon \|\boldsymbol{v}\| \le 1\}\,.$$

Analogously, the open unit ball centred at $\mathbf{0}$ is defined by

$$B_1(\mathbf{0}) = \{\boldsymbol{v} \in \mathcal{V}\colon \|\boldsymbol{v}\| < 1\}\,.$$

More generally, for $\varepsilon > 0$ and $\boldsymbol{\xi} \in \mathcal{V}$,

$$\bar{B}_\varepsilon(\boldsymbol{\xi}) = \{\boldsymbol{v} \in \mathcal{V}\colon \|\boldsymbol{v} - \boldsymbol{\xi}\| \le \varepsilon\}$$

is the **closed ball** of radius ε centred at $\boldsymbol{\xi}$; analogously,

$$B_\varepsilon(\boldsymbol{\xi}) = \{\boldsymbol{v} \in \mathcal{V}\colon \|\boldsymbol{v} - \boldsymbol{\xi}\| < \varepsilon\}$$

is the **open ball** of radius ε centred at $\boldsymbol{\xi}$.

Any norm on the linear space $\mathbb{R}^{n\times n}$ of $n \times n$ matrices with real entries will be referred to as a **matrix norm**. In particular, we shall now

consider matrix norms which are induced by vector norms in a sense that will be made precise in the next definition.

Definition 2.10 *Given any norm $\|\cdot\|$ on the space $\mathbb{R}^n$ of n-dimensional vectors with real entries, the* **subordinate matrix norm** *on the space $\mathbb{R}^{n\times n}$ of $n\times n$ matrices with real entries is defined by*

$$\|A\| = \max_{\boldsymbol{v}\in\mathbb{R}^n_*} \frac{\|A\boldsymbol{v}\|}{\|\boldsymbol{v}\|}\,. \tag{2.38}$$

In (2.38) we used $\mathbb{R}^n_*$ to denote $\mathbb{R}^n \setminus \{\mathbf{0}\}$, where, for sets A and B, $A \setminus B = \{x \in A\colon x \notin B\}$.

Remark 2.4 *Let $\mathbb{C}^{n\times n}$ denote the linear space of $n\times n$ matrices with complex entries over the field $\mathbb{C}$ of complex numbers. Given any norm $\|\cdot\|$ on the linear space $\mathbb{C}^n$, the* **subordinate matrix norm** *on $\mathbb{C}^{n\times n}$ is defined by*

$$\|A\| = \max_{\boldsymbol{v}\in\mathbb{C}^n_*} \frac{\|A\boldsymbol{v}\,\|}{\|\boldsymbol{v}\,\|}\,,$$

where $\mathbb{C}^n_* = \mathbb{C}^n \setminus \{\mathbf{0}\}$.

It is easy to show that a subordinate matrix norm satisfies the axioms of norm listed in Definition 2.6; the details are left as an exercise. Definition 2.10 implies that, for $A \in \mathbb{R}^{n\times n}$,

$$\|A\boldsymbol{v}\,\| \le \|A\|\,\|\boldsymbol{v}\,\|\,, \qquad \text{for all } \boldsymbol{v} \in \mathbb{R}^n\,.$$

In a relation like this any vector norm may be used, but of course it is necessary to use the same norm throughout. It follows from Definition 2.10 that, in any subordinate matrix norm $\|\cdot\|$ on $\mathbb{R}^{n\times n}$,

$$\|I\| = 1$$

where I is the $n\times n$ identity matrix.

Given any vector $\boldsymbol{v}$ in $\mathbb{R}^n$, it is a trivial matter to evaluate each of the three norms $\|\boldsymbol{v}\,\|_1$, $\|\boldsymbol{v}\,\|_2$, $\|\boldsymbol{v}\,\|_\infty$; however, it is not yet obvious how one can calculate the corresponding subordinate matrix norm of a given matrix A in $\mathbb{R}^{n\times n}$. Definition 2.10 is unhelpful in this respect: calculating $\|A\|$ *via* (2.38) would involve the unpleasant task of maximising the function $\boldsymbol{v} \mapsto \|A\boldsymbol{v}\,\|/\|\boldsymbol{v}\,\|$ over $\mathbb{R}^n_*$ (or, equivalently, maximising $\boldsymbol{w} \mapsto \|A\boldsymbol{w}\|$ over the unit sphere $\{\boldsymbol{w} \in \mathbb{R}^n\colon \|\boldsymbol{w}\| = 1\}$). This difficulty is resolved by the following three theorems.

Theorem 2.7 *The matrix norm subordinate to the vector norm* $\|\cdot\|_\infty$ *can be expressed, for an* $n \times n$ *matrix* $A = (a_{ij})_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$, *as*

$$\|A\|_\infty = \max_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}| \,. \tag{2.39}$$

This result is often loosely expressed by saying that the ∞*-norm of a matrix is its largest row-sum.*

Proof Given an arbitrary vector $\boldsymbol{v}$ in $\mathbb{R}^n_*$, write $K = \|\boldsymbol{v}\|_\infty$, so that $|v_j| \le K$ for $j = 1, 2, \ldots, n$. Then,

$$|(A\boldsymbol{v})_i| = \left| \sum_{j=1}^{n} a_{ij} v_j \right| \le \sum_{j=1}^{n} |a_{ij}|\,|v_j| \le K \sum_{j=1}^{n} |a_{ij}| \,, \qquad i = 1, 2, \ldots, n \,.$$

Now we define

$$C = \max_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}| \tag{2.40}$$

and note that

$$\frac{\|A\boldsymbol{v}\|_\infty}{\|\boldsymbol{v}\|_\infty} = \frac{\max_{i=1}^n |(A\boldsymbol{v})_i|}{\|\boldsymbol{v}\|_\infty} = \frac{\max_{i=1}^n |(A\boldsymbol{v})_i|}{K} \le C \qquad \forall\, \boldsymbol{v} \in \mathbb{R}^n_* \,.$$

Hence, $\|A\|_\infty \le C$.

Next we show that $\|A\|_\infty \ge C$. To do so, we take $\boldsymbol{v}$ to be a vector in $\mathbb{R}^n_*$ each of whose entries is ± 1, with the choice of sign to be made clear below. In the definition of C, equation (2.40), let m be the value of i for which the maximum is attained, or any one of the values if there is more than one. Then, in the vector $\boldsymbol{v}$ we give the element v_j the same sign as that of a_{mj}; if a_{mj} happens to be zero, the choice of the sign of v_j is irrelevant. With this definition of $\boldsymbol{v}$ we see at once that

$$\|A\boldsymbol{v}\|_\infty = \max_{i=1}^{n} \left| \sum_{j=1}^{n} a_{ij} v_j \right| \ge \left| \sum_{j=1}^{n} a_{mj} v_j \right| = \sum_{j=1}^{n} |a_{mj}|\,|v_j| = \sum_{j=1}^{n} |a_{mj}| = C \,.$$

As $\|\boldsymbol{v}\|_\infty = 1$, it follows that

$$\|A\boldsymbol{v}\|_\infty \ge C \|\boldsymbol{v}\|_\infty \,,$$

which means that $\|A\|_\infty \ge C$. Hence $\|A\|_\infty = C$, as required. □

Theorem 2.8 *The matrix norm subordinate to the vector norm* $\|\cdot\|_1$ *can be expressed, for an* $n \times n$ *matrix* $A = (a_{ij})_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$, *as*

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

This is often loosely expressed by saying that the 1-norm of a matrix is its largest column-sum. The proof of this theorem is very similar to that of the previous one, and is left as an exercise (see Exercise 7). Note that Theorems 2.7 and 2.8 mean that the 1-norm of a matrix $A = (a_{ij})_{1\le i,j\le n}$ is the ∞-norm of the transpose $A^{\mathrm{T}} = (a_{ji})_{1\le i,j\le n}$ of the matrix.

Before we state a characterisation of the subordinate matrix 2-norm, we recall the following definition from linear algebra.

Definition 2.11 *Suppose that* $A \in \mathbb{R}^{n\times n}$. *A complex number* λ, *for which the set of linear equations*

$$A\boldsymbol{x} = \lambda\boldsymbol{x}$$

has a nontrivial solution $\boldsymbol{x} \in \mathbb{C}^n_* = \mathbb{C}^n \setminus \{\boldsymbol{0}\}$, *is called an* **eigenvalue** *of* A; *the associated solution* $\boldsymbol{x} \in \mathbb{C}^n_*$ *is called an* **eigenvector** *of* A *(corresponding to* λ*).*

Now we are ready to state our result.

Theorem 2.9 *Let* $A \in \mathbb{R}^{n\times n}$ *and denote the eigenvalues of the matrix* $B = A^{\mathrm{T}}A$ *by* λ_i, $i = 1, 2, \ldots, n$. *Then,*

$$\|A\|_2 = \max_{i=1}^{n} \lambda_i^{1/2}.$$

Proof Note first that the matrix B is symmetric, *i.e.*, $B = B^{\mathrm{T}}$; therefore all of its eigenvalues are real and the associated eigenvectors belong to $\mathbb{R}^n_*$. (You may wish to prove this: consult the proof of Theorem 3.1, part (ii), for a hint.) Moreover, all eigenvalues of B are nonnegative, since if $\boldsymbol{v} \in \mathbb{R}^n_*$ is an eigenvector of B and λ is the associated eigenvalue λ, then

$$A^{\mathrm{T}}A\boldsymbol{v} = B\boldsymbol{v} = \lambda\boldsymbol{v}$$

and therefore

$$\lambda = \frac{\boldsymbol{v}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}} = \frac{\|A\boldsymbol{v}\|_2^2}{\|\boldsymbol{v}\|_2^2} \ge 0 .$$

Suppose that the vectors $\boldsymbol{w}_i \in \mathbb{R}^n_*$, $i = 1, 2, \ldots, n$, are eigenvectors of B corresponding to the eigenvalues λ_i, $i = 1, 2, \ldots, n$. Since B is symmetric

we may assume that the vectors $\boldsymbol{w}_i$ are orthogonal, *i.e.*, $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_j = 0$ for $i \neq j$, and we can normalise them so that $\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{w}_i = 1$ for $i = 1, 2, \ldots, n$. Now choose an arbitrary vector $\boldsymbol{u}$ in $\mathbb{R}^n_*$ and express it as a linear combination of the vectors $\boldsymbol{w}_i$, $i = 1, 2, \ldots, n$:

$$\boldsymbol{u} = c_1\boldsymbol{w}_1 + \cdots + c_n\boldsymbol{w}_n\,.$$

Then,

$$B\boldsymbol{u} = c_1\lambda_1\boldsymbol{w}_1 + \cdots + c_n\lambda_n\boldsymbol{w}_n\,.$$

We may assume, without loss of generality, that

$$(0 \leq)\ \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n\,.$$

Using the orthonormality of the vectors $\boldsymbol{w}_i$, $i = 1, 2, \ldots, n$, we get that

$$\begin{aligned} \|A\boldsymbol{u}\|_2^2 &= \boldsymbol{u}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{u} = \boldsymbol{u}^{\mathrm{T}}B\boldsymbol{u} \\ &= c_1^2\lambda_1 + \cdots + c_n^2\lambda_n \\ &\leq (c_1^2 + \cdots + c_n^2)\lambda_n \\ &= \lambda_n\|\boldsymbol{u}\|_2^2\,, \end{aligned} \tag{2.41}$$

for any vector $\boldsymbol{u} \in \mathbb{R}^n_*$. Hence $\|A\|_2^2 \leq \lambda_n$. To prove equality we simply choose $\boldsymbol{u} = \boldsymbol{w}_n$ in (2.41), so that $c_1 = \cdots = c_{n-1} = 0$ and $c_n = 1$. □

The square roots of the (nonnegative) eigenvalues of $A^{\mathrm{T}}A$ are referred to as the **singular values** of A. Thus we have shown that the 2-norm of a matrix A is equal to the largest singular value of A.

If the matrix A is symmetric, then $B = A^{\mathrm{T}}A = A^2$, and the eigenvalues of B are just the squares of the eigenvalues of A. In this special case the 2-norm of A is the largest of the absolute values of its eigenvalues.

Theorem 2.10 *Given that $\|\cdot\|$ is a subordinate matrix norm on $\mathbb{R}^{n\times n}$,*

$$\|AB\| \leq \|A\|\,\|B\|$$

for any two matrices A and B in $\mathbb{R}^{n\times n}$.

Proof From the definition of subordinate matrix norm,

$$\|AB\| = \max_{\boldsymbol{v}\in\mathbb{R}^n_*} \frac{\|AB\boldsymbol{v}\,\|}{\|\boldsymbol{v}\,\|}\,.$$

As

$$\|AB\boldsymbol{v}\,\| \leq \|A\|\,\|B\boldsymbol{v}\,\|$$

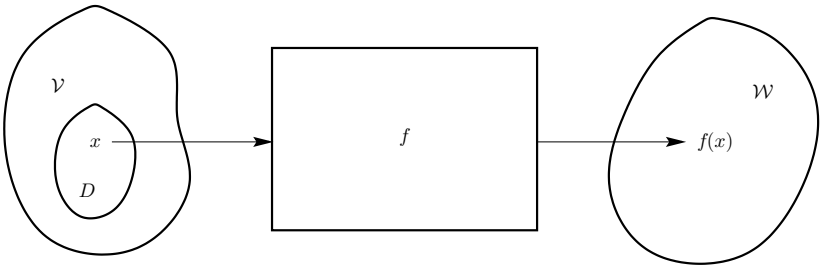

Fig. 2.3. 'Input' $x \in D \subset \mathcal{V}$ and 'output' $f(x) \in \mathcal{W}$ for a mapping $f\colon \mathcal{V} \to \mathcal{W}$.

for all $\boldsymbol{v} \in \mathbb{R}^n_*$, we have

$$\begin{aligned} \|AB\| &\leq \max_{\boldsymbol{v}\in\mathbb{R}^n_*} \frac{\|A\|\,\|B\boldsymbol{v}\,\|}{\|\boldsymbol{v}\,\|} \\ &= \|A\| \max_{\boldsymbol{v}\in\mathbb{R}^n_*} \frac{\|B\boldsymbol{v}\,\|}{\|\boldsymbol{v}\,\|} \\ &= \|A\|\,\|B\|\,, \end{aligned}$$

and hence the desired result. □

Now we are ready to embark on the study of sensitivity to perturbations in the problem of matrix inversion. In order to motivate the concept of *condition number* of a matrix which will play a key role in the analysis, we begin with a discussion of 'conditioning' in a slightly more general context.

Consider a mapping f from a subset D of a normed linear space $\mathcal{V}$ with norm $\|\cdot\|_{\mathcal{V}}$ into another normed linear space $\mathcal{W}$ with norm $\|\cdot\|_{\mathcal{W}}$, depicted in Figure 2.3, where $x \in D \subset \mathcal{V}$ is regarded as the 'input' for f and $f(x) \in \mathcal{W}$ is the 'output'. We shall be concerned with the sensitivity of the output to perturbations in the input; therefore, as a measure of sensitivity, we define the **absolute condition number** of f by

$$\mathrm{Cond}(f) = \sup_{\substack{x,y\in D\subset\mathcal{V}\\ x\neq y}} \frac{\|f(y)-f(x)\|_{\mathcal{W}}}{\|y-x\|_{\mathcal{V}}}\,. \tag{2.42}$$

If $\mathrm{Cond}(f) = +\infty$ or if $1 \ll \mathrm{Cond}(f) < +\infty$, we say that the mapping f is **ill-conditioned**.

Example 2.5 *Consider the function* $f\colon x \in D \mapsto \sqrt{x}$*, where* D *is a closed subinterval of* $[0,\infty)$*. Clearly, if* $D = [1,2]$*, then* $\mathrm{Cond}(f) = 1/2$*,*

while if $D = [0,1]$, *then* $\mathrm{Cond}(f) = +\infty$. *Indeed, in the latter case, perturbing* $x = 0$ *to* $x = \varepsilon^2$, $0 < \varepsilon \ll 1$, *leads to a perturbation of the function value* $f(0) = 0$ *to* $f(\varepsilon^2) = \varepsilon = \frac{1}{\varepsilon}\varepsilon^2$*: a magnification by a factor* $\frac{1}{\varepsilon} \gg 1$ *in comparison with the size of the perturbation in* x.

When $\|f(y)-f(x)\|_{\mathcal{W}}/\|y-x\|_{\mathcal{V}}$ exhibits large variation as (x,y) ranges through $D \times D$, it is more helpful to consider a finer, local measure of conditioning, the **absolute local condition number**, at $x \in D \subset \mathcal{V}$, of the function f, defined by

$$\mathrm{Cond}_x(f) = \sup_{\substack{\delta x \in \mathcal{V}\setminus\{0\} \\ x+\delta x \in D}} \frac{\|f(x+\delta x) - f(x)\|_{\mathcal{W}}}{\|\delta x\|_{\mathcal{V}}}\,. \tag{2.43}$$

Example 2.6 *Let us consider the function* $f\colon x \in D \mapsto \sqrt{x}$, *defined on the interval* $D = (0,\infty)$. *The absolute local condition number of* f *at* $x \in D$ *is* $\mathrm{Cond}_x(f) = 1/(2\sqrt{x})$. *Clearly,* $\lim_{x\to 0+} \mathrm{Cond}_x(f) = +\infty$, $\lim_{x\to+\infty} \mathrm{Cond}_x(f) = 0$.

Although the definitions (2.42) and (2.43) seem intuitive, they are not always satisfactory from the practical point of view since they depend on the magnitudes of $f(x)$ and x. A more convenient definition of conditioning is arrived at by rescaling (2.43) by the norms of $f(x)$ and x. This leads us to the notion of **relative local condition number**

$$\mathrm{cond}_x(f) = \sup_{\substack{\delta x \in \mathcal{V}\setminus\{0\} \\ x+\delta x \in D}} \frac{\|f(x+\delta x) - f(x)\|_{\mathcal{W}}/\|f(x)\|_{\mathcal{W}}}{\|\delta x\|_{\mathcal{V}}/\|x\|_{\mathcal{V}}}\,,$$

where it is implicitly assumed that $x \in \mathcal{V}\setminus\{0\}$ and $f(x) \in \mathcal{W}\setminus\{0\}$. The next example highlights the difference between the absolute local condition number and the relative local condition number of f.

Example 2.7 *Let us consider the function* $f\colon x \in D \mapsto \sqrt{x}$, *defined on the interval* $D = (0,\infty)$. *Recall from the preceding example that the absolute local condition number of* f *at* $x \in D$ *approaches* $+\infty$ *as* x *tends to zero. In contrast with this, the relative local condition number of* f *is* $\mathrm{cond}_x(f) = 1/2$ *for all* $x \in D$.

You may also wish to ponder the following, seemingly paradoxical, observation: $\lim_{\varepsilon\to 0} \mathrm{cond}_\varepsilon(\sin) = 1$ *and* $\lim_{\varepsilon\to 0} \mathrm{cond}_{\pi-\varepsilon}(\sin) = \infty$, *even though* $\sin 0 = \sin\pi = 0$ *and* $\mathrm{Cond}_0(\sin) = \mathrm{Cond}_\pi(\sin) = 1$.

Since the present section is concerned with the solution of the linear system $A\boldsymbol{x} = \boldsymbol{b}$, where $A \in \mathbb{R}^{n\times n}$ is nonsingular and $\boldsymbol{b} \in \mathbb{R}^n$, let us

consider the relative local condition number of the mapping

$$A^{-1}\cdot : \boldsymbol{b} \in \mathbb{R}^n \mapsto A^{-1}\boldsymbol{b} \in \mathbb{R}^n$$

at $\boldsymbol{b} \in \mathbb{R}^n_* = \mathbb{R}^n \setminus \{\mathbf{0}\}$. We suppose that $\mathbb{R}^n$ has been equipped with a vector norm $\|\cdot\|$ and, since there is no danger of confusion, we denote the associated subordinate matrix norm by $\|\cdot\|$ also. Noting that $A^{-1}\cdot$ is defined on the whole of $\mathbb{R}^n$, it follows that $D = \mathcal{V} = \mathbb{R}^n$, $\mathcal{W} = \mathbb{R}^n$ and we deduce that

$$\begin{aligned}\mathrm{cond}_{\boldsymbol{b}}(A^{-1}\cdot) &= \sup_{\boldsymbol{\delta b}\in\mathbb{R}^n_*} \frac{\|A^{-1}(\boldsymbol{b}+\boldsymbol{\delta b}) - A^{-1}\boldsymbol{b}\| \,/\, \|A^{-1}\boldsymbol{b}\|}{\|\boldsymbol{\delta b}\| \,/\, \|\boldsymbol{b}\|} \\ &= \|A^{-1}\| \frac{\|\boldsymbol{b}\|}{\|A^{-1}\boldsymbol{b}\|}\,.\end{aligned}$$

Since $\|\boldsymbol{b}\| = \|A(A^{-1}\boldsymbol{b})\| \le \|A\|\,\|A^{-1}\boldsymbol{b}\|$, we conclude that

$$\mathrm{cond}_{\boldsymbol{b}}(A^{-1}\cdot) \le \|A^{-1}\|\,\|A\|\,. \tag{2.44}$$

If now, instead, we consider the mapping

$$A\cdot : \boldsymbol{x} \in \mathbb{R}^n \mapsto A\boldsymbol{x} \in \mathbb{R}^n,$$

an identical argument shows that, for $\boldsymbol{x} \in \mathbb{R}^n_*$,

$$\mathrm{cond}_{\boldsymbol{x}}(A\cdot) \le \|A\|\,\|A^{-1}\|\,. \tag{2.45}$$

The inequalities (2.44) and (2.45) indicate that the number $\|A^{-1}\|\,\|A\| = \|A\|\,\|A^{-1}\|$ plays a relevant role in the analysis of sensitivity to perturbations in numerical linear algebra; therefore we adopt the following definition.

Definition 2.12 *The* **condition number** *of a nonsingular matrix A is defined by*

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

Clearly, $\kappa(A^{-1}) = \kappa(A)$. Further, since $AA^{-1} = I$, it follows from Theorem 2.10 that $\kappa(A) \ge 1$ for every matrix A. If $\kappa(A) \gg 1$, the matrix is said to be **ill-conditioned**. Evidently the condition number of a matrix is unaffected by scaling all its elements by multiplying by a nonzero constant.[1]

[1] We note in passing that, more generally, the condition number of a matrix $A \in \mathbb{R}^{m\times n}$ is defined by $\kappa(A) = \|A\|\,\|A^+\|$ where A^+ is the Moore–Penrose generalised inverse of A. In the special case when $m = n$ and A is nonsingular, $A^+ = A^{-1}$. For further details in this direction, we refer to the Notes at the end of the chapter. Here, the norm $\|\cdot\|$ on $\mathbb{R}^{m\times n}$ is defined as in (2.38). Theorems 2.7 and 2.8 are

There is a condition number for each norm; for example, if we use the 2-norm, then $\kappa_2(A) = \|A\|_2 \, \|A^{-1}\|_2$, and so on. Indeed, the size of the condition number of a matrix $A \in \mathbb{R}^{n\times n}$ is strongly dependent on the choice of the norm in $\mathbb{R}^n$. In order to illustrate the last point, let us consider the unit lower triangular matrix $A \in \mathbb{R}^{n\times n}$ defined by

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}, \tag{2.46}$$

and note that its inverse is

$$A^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 \\ -1 & 1 & 0 & 0 & \dots & 0 \\ -1 & 0 & 1 & 0 & \dots & 0 \\ -1 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ -1 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}.$$

Since

$$\|A\|_1 = n \qquad \text{and} \qquad \|A^{-1}\|_1 = n\,,$$

it follows that $\kappa_1(A) = n^2$. On the other hand,

$$\|A\|_\infty = 2 \qquad \text{and} \qquad \|A^{-1}\|_\infty = 2\,.$$

so that $\kappa_\infty(A) = 4 \ll n^2 = \kappa_1(A)$ when $n \gg 1$. (A question for the curious: how does the condition number $\kappa_2(A)$ of the matrix A in (2.46) depend on the size n of A? See Exercise 11.)

It is left as an exercise to show that for a symmetric matrix A (*i.e.*, when $A^{\mathrm{T}} = A$), the 2-norm condition number $\kappa_2(A)$ is the ratio of the largest of the absolute values of the eigenvalues of A to the smallest of the absolute values of the eigenvalues (see Exercise 9).

easily extended to show that, for $A \in \mathbb{R}^{m\times n}$,

$$\|A\|_\infty = \max_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}| \quad \text{and} \quad \|A\|_1 = \max_{j=1}^{n} \sum_{i=1}^{m} |a_{ij}|\,.$$

The 2-norm of A, $\|A\|_2$, is equal to the largest singular value of A, *i.e.*, the square root of the largest eigenvalue of the matrix $A^{\mathrm{T}}A \in \mathbb{R}^{n\times n}$, just as in Theorem 2.9.

We can now assess the sensitivity of the solution of the system $A\boldsymbol{x} = \boldsymbol{b}$ to changes in the right-hand side vector $\boldsymbol{b}$.

Theorem 2.11 *Suppose that $A \in \mathbb{R}^{n\times n}$ is a nonsingular matrix, $\boldsymbol{b} \in \mathbb{R}^n_*$, $A\boldsymbol{x} = \boldsymbol{b}$ and $A(\boldsymbol{x} + \boldsymbol{\delta x}) = \boldsymbol{b} + \boldsymbol{\delta b}$, with $\boldsymbol{\delta x}$, $\boldsymbol{\delta b} \in \mathbb{R}^n$. Then, $\boldsymbol{x} \in \mathbb{R}^n_*$ and*

$$\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} \le \kappa(A)\frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}\,.$$

Proof Evidently,

$$\boldsymbol{b} = A\boldsymbol{x} \quad \text{and} \quad \boldsymbol{\delta x} = A^{-1}(\boldsymbol{b} + \boldsymbol{\delta b}) - \boldsymbol{x} = A^{-1}\boldsymbol{\delta b}\,.$$

As $\boldsymbol{b} \neq \boldsymbol{0}$ and A is nonsingular, the first of these implies that $\boldsymbol{x} \neq \boldsymbol{0}$. Further,

$$\|\boldsymbol{b}\| \le \|A\|\,\|\boldsymbol{x}\| \quad \text{and} \quad \|\boldsymbol{\delta x}\| \le \|A^{-1}\|\,\|\boldsymbol{\delta b}\|\,.$$

The result follows immediately by multiplying these inequalities. □

Owing to the effect of rounding errors during the calculation, the numerical solution of $A\boldsymbol{x} = \boldsymbol{b}$ will not be exact. The numerical solution may be written $\boldsymbol{x}+\boldsymbol{\delta x}$, and we shall usually find that this vector satisfies the equation $A(\boldsymbol{x}+\boldsymbol{\delta x}) = \boldsymbol{b}+\boldsymbol{\delta b}$, where the elements of $\boldsymbol{\delta b}$ are very small. If the matrix A has a large condition number, however, the elements of $\boldsymbol{\delta x}$ may not be so small. An example of this will be presented in the next section.

2.8 Hilbert matrix

We consider the Hilbert matrix[1] H_n of order n, whose elements are

$$h_{ij} = \frac{1}{i+j-1}\,, \qquad i,j = 1,2,\ldots,n\,.$$

This matrix is symmetric and positive definite (*i.e.*, $H_n^{\mathrm{T}} = H_n$, and $\boldsymbol{x}^{\mathrm{T}}H_n\boldsymbol{x} > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n_*$), and therefore all of its eigenvalues are real and positive (cf. Theorem 3.1, part (ii)). However, H_n becomes very nearly singular as n increases. Table 2.1 shows the largest and smallest eigenvalues, and the 2-norm condition number $\kappa_2(H_n)$ of H_n, for various values of n.

[1] David Hilbert (23 January 1862, Königsberg, Prussia (now Kaliningrad, Russia) – 14 February 1943, Göttingen, Germany) was the most prominent member of the Göttingen school of mathematics. He made significant contributions to many areas of the subject, including algebra, geometry, number theory, calculus of variations, functional analysis, integral equations, and the foundations of mathematics.

Table 2.1. *Eigenvalues and condition number of the Hilbert matrix* H_n.

n	$\lambda_{\max}$	$\lambda_{\min}$	$\kappa_2(H_n)$
5	1.6	3.3×10^{-6}	4.8×10^{5}
10	1.8	1.1×10^{-13}	1.6×10^{13}
15	1.8	3.0×10^{-21}	6.1×10^{20}
20	1.9	7.8×10^{-29}	2.5×10^{28}
25	2.0	1.9×10^{-36}	1.0×10^{36}

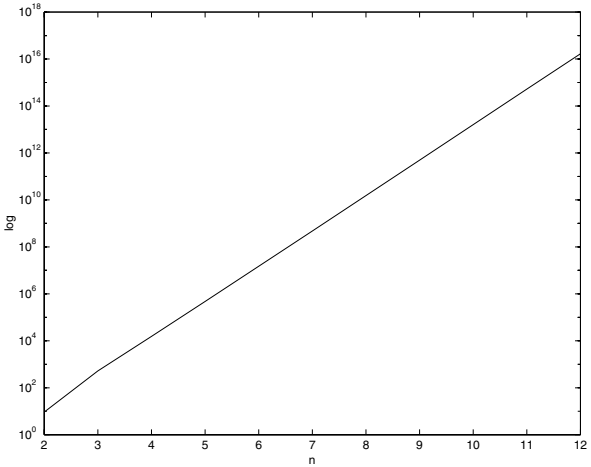

Fig. 2.4. Condition number $\kappa_2(H_n)$ of the Hilbert matrix H_n of size $n = 2, 3, \ldots, 12$ in the 2-norm, against n, in a semilogarithmic-scale plot.

Figure 2.4 depicts the logarithm of the condition number $\kappa_2(H_n)$ in the 2-norm of the Hilbert matrix H_n against its order, n; the straight line in our semilogarithmic-scale plot indicates that $\kappa_2(H_n)$, as a function of n, exhibits exponential growth. Indeed, it can be shown that

$$\kappa_2(H_n) \sim \frac{\left(\sqrt{2}+1\right)^{4n+4}}{2^{15/4}\sqrt{\pi n}} \quad \text{as } n \to \infty\,.$$

We now define the vector $\boldsymbol{b}$ with elements $b_i = \sum_{j=1}^{n} (j/(i+j-1))$, $i = 1, 2, \ldots, n$, chosen so that the solution of $A\boldsymbol{x} = \boldsymbol{b}$, with $A = H_n$, is the vector $\boldsymbol{x}$ with elements $x_i = i$, $i = 1, 2, \ldots, n$. We obtain a numerical solution of the system, using the method described in Section

2.5 to give the calculated vector $\boldsymbol{x}+\boldsymbol{\delta x}$, and then compute the residual $\boldsymbol{\delta b}$ from $A(\boldsymbol{x}+\boldsymbol{\delta x})=\boldsymbol{b}+\boldsymbol{\delta b}$. The calculation uses arithmetic operations correct to 15 decimal digits, which is roughly the accuracy used by many computer systems. The results are listed in Table 2.2.

Table 2.2. *Rounding errors in the solution of $H_n\boldsymbol{x}=\boldsymbol{b}$, where H_n is the Hilbert matrix of order n and $\boldsymbol{b}=(1,2,\ldots,n)^{\mathrm{T}}$.*

n	$\|\boldsymbol{\delta b}\|_2/\|\boldsymbol{b}\|_2$	$\|\boldsymbol{\delta x}\|_2/\|\boldsymbol{x}\|_2$
5	1.2×10^{-15}	8.5×10^{-11}
10	1.7×10^{-15}	1.3×10^{-3}
15	2.8×10^{-15}	4.1
20	6.3×10^{-15}	8.7
25	1.9×10^{-13}	5.5×10^{2}

The relative size of the residual is, in nearly every case, about the size of the basic rounding error, 10^{-15}. The resulting errors in $\boldsymbol{x}$ are smaller than the bound given by Theorem 2.11, as might be expected, since that bound corresponds to the worst possible case. In any case, for the Hilbert matrix of order greater than 14 the error is larger than the calculated solution itself, which renders the calculated solution meaningless. For matrices of this kind the condition number and the bound given by Theorem 2.11 are so large that they have little practical relevance, though they do indicate that, due to sensitivity to rounding errors, the numerical calculations are of unreliable accuracy.

The Hilbert matrix is, of course, a rather extreme example of an ill-conditioned matrix. However, we shall meet it in an important problem in Section 9.3 concerning the least squares approximation of a function by polynomials, where we shall see how a reformulation of the problem using an orthonormal basis avoids the disastrous loss of accuracy that would otherwise occur. In the next section, we introduce the idea of least squares approximation in the context of linear algebra and consider the solution of the resulting system of linear equations using the QR algorithm; this, too, relies on the notion of (ortho)normalisation.

2.9 Least squares method

Up to now, we have been dealing with systems of linear equations of the form $A\boldsymbol{x}=\boldsymbol{b}$ where $A\in\mathbb{R}^{n\times n}$. However, it is frequently the case

in practical problems (typically, in problems of data-fitting) that the matrix A is not square but rectangular, and we have to solve a linear system of equations $A\boldsymbol{x} = \boldsymbol{b}$ with $A \in \mathbb{R}^{m\times n}$, $\boldsymbol{b} \in \mathbb{R}^m$, with $m > n$; since there are more equations than unknowns, in general such a system will have no solution. Consider, for example, the linear system (with $m = 3$, $n = 2$)

$$\begin{pmatrix} 3 & 1 \\ 1 & 1 \\ 4 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix};$$

by adding the first two of the three equations and comparing the result with the third, it is easily seen that there is no solution. If, on the other hand, $m < n$, then the situation is reversed and there may be an infinite number of solutions. Consider, for example, the linear system (with $m = 1$, $n = 2$)

$$(3\ \ 1) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 1;$$

any vector $\boldsymbol{x} = (\mu, 1 - 3\mu)^{\mathrm{T}}$, with $\mu \in \mathbb{R}$, is a solution to this system.

Suppose that $m \geq n$; we may then need to find a vector $\boldsymbol{x} \in \mathbb{R}^n$ which satisfies $A\boldsymbol{x} - \boldsymbol{b} \approx \boldsymbol{0}$ in $\mathbb{R}^m$ as nearly as possible in some sense. This suggests that we define the residual vector $\boldsymbol{r} = A\boldsymbol{x} - \boldsymbol{b}$ and require to minimise a certain norm of $\boldsymbol{r}$ in $\mathbb{R}^m$. From the practical point of view, it is particularly convenient to minimise the residual vector $\boldsymbol{r}$ in the 2-norm on $\mathbb{R}^m$; this leads to the **least squares** problem:

$$\underset{\boldsymbol{x}\,\in\,\mathbb{R}^n}{\text{Minimise}} \quad \|A\boldsymbol{x} - \boldsymbol{b}\|_2 .$$

This is clearly equivalent to minimising the square of the norm; so, on noting that

$$\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 = (A\boldsymbol{x} - \boldsymbol{b})^{\mathrm{T}}(A\boldsymbol{x} - \boldsymbol{b}),$$

the problem may be restated as

$$\underset{\boldsymbol{x}\,\in\,\mathbb{R}^n}{\text{Minimise}} \quad (A\boldsymbol{x} - \boldsymbol{b})^{\mathrm{T}}(A\boldsymbol{x} - \boldsymbol{b}) .$$

Since

$$(A\boldsymbol{x} - \boldsymbol{b})^{\mathrm{T}}(A\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{x}^{\mathrm{T}}A^{\mathrm{T}}A\boldsymbol{x} - 2\boldsymbol{x}^{\mathrm{T}}A^{\mathrm{T}}\boldsymbol{b} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{b},$$

the quantity to be minimised is a nonnegative quadratic function of the n components of the vector $\boldsymbol{x}$; the minimum therefore exists, and may

be found by equating to zero the partial derivatives with respect to the components. This leads to the system of equations

$$B\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}\,, \qquad \text{where } B = A^{\mathrm{T}}A\,.$$

The matrix B is symmetric, and if A has full rank, n, then B is nonsingular; it is called the **normal** matrix, and the system $B\boldsymbol{x} = A^{\mathrm{T}}\boldsymbol{b}$ is called the system of **normal equations**.

The normal equations have important theoretical properties, but do not lead to a satisfactory numerical algorithm, except for fairly small problems. The difficulty is that in a practical least squares problem the matrix A is likely to be quite ill-conditioned, and $B = A^{\mathrm{T}}A$ will then be extremely ill-conditioned. For example, if

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

where $\varepsilon \in (0,1)$, then $\kappa_2(A) = \varepsilon^{-1} > 1$, while

$$\kappa_2(B) = \kappa_2(A^{\mathrm{T}}A) = \varepsilon^{-2} = \varepsilon^{-1}\kappa_2(A) \gg \kappa_2(A)$$

when $0 < \varepsilon \ll 1$. If possible, one should avoid using a method which leads to such a dramatic deterioration of the condition number.

There are various alternative techniques which avoid the direct construction of the normal matrix $A^{\mathrm{T}}A$, and so do not lead to this extreme ill-conditioning. Here we shall describe just one algorithm, which begins by factorising the matrix A, but using an orthogonal matrix rather than the lower triangular factor as in Section 2.3.

Theorem 2.12 *Suppose that $A \in \mathbb{R}^{m\times n}$ where $m \geq n$. Then, A can be written in the form*

$$A = \hat{Q}\hat{R}\,,$$

where $\hat{R}$ is an upper triangular $n \times n$ matrix, and $\hat{Q}$ is an $m \times n$ matrix which satisfies

$$\hat{Q}^{\mathrm{T}}\hat{Q} = I_n\,, \tag{2.47}$$

where I_n is the $n \times n$ identity matrix; see Figure 2.5. If $\mathrm{rank}(A) = n$, then $\hat{R}$ is nonsingular.

Proof We use induction on n, the number of columns in A. The theorem clearly holds when $n = 1$ so that A has only one column. Indeed, writing $\boldsymbol{c}$ for this column vector and assuming that $\boldsymbol{c} \neq \boldsymbol{0}$, the matrix $\hat{Q}$ has just

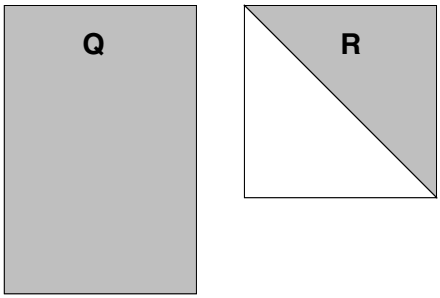

Fig. 2.5. QR factorisation of $A \in \mathbb{R}^{m\times n}$, $m \geq n$: $A = \hat{Q}\hat{R}$, $\hat{Q} \in \mathbb{R}^{m\times n}$, $\hat{Q}^{\mathrm{T}}\hat{Q} = I_n$, and the matrix $\hat{R} \in \mathbb{R}^{n\times n}$ is upper triangular.

one column, the vector $\boldsymbol{c}/\|\boldsymbol{c}\|_2$, and $\hat{R}$ has a single element, $\|\boldsymbol{c}\|_2$. In the special case where $\boldsymbol{c}$ is the zero vector we can choose $\hat{R}$ to have the single element 0, and $\hat{Q}$ to have a single column which can be an arbitrary vector in $\mathbb{R}^m$ whose 2-norm is equal to 1.

Suppose that the theorem is true when $n = k$, where $1 \leq k < m$. Consider a matrix A which has m rows and $k+1$ columns, partitioned as

$$A = (A_k \;\; \boldsymbol{a}) \,,$$

where $\boldsymbol{a} \in \mathbb{R}^m$ is a column vector and A_k has k columns. To obtain the desired factorisation $\hat{Q}\hat{R}$ of A we seek $\hat{Q} = (\hat{Q}_k \;\; \boldsymbol{q})$ and

$$\hat{R} = \begin{pmatrix} \hat{R}_k & \boldsymbol{r} \\ \mathbf{0} & \alpha \end{pmatrix}$$

such that

$$A = (A_k \;\; \boldsymbol{a}) = (\hat{Q}_k \;\; \boldsymbol{q}) \begin{pmatrix} \hat{R}_k & \boldsymbol{r} \\ \mathbf{0} & \alpha \end{pmatrix}.$$

Multiplying this out and requiring that $\hat{Q}^{\mathrm{T}}\hat{Q} = I_{k+1}$, the identity matrix of order $k+1$, we conclude that

$$\begin{aligned} A_k &= \hat{Q}_k\hat{R}_k \,, && (2.48)\\ \boldsymbol{a} &= \hat{Q}_k\boldsymbol{r} + \boldsymbol{q}\alpha \,, && (2.49)\\ \hat{Q}_k^{\mathrm{T}}\hat{Q}_k &= I_k \,, && (2.50)\\ \boldsymbol{q}^{\mathrm{T}}\hat{Q}_k &= \mathbf{0}^{\mathrm{T}} \,, && (2.51)\\ \boldsymbol{q}^{\mathrm{T}}\boldsymbol{q} &= 1 \,. && (2.52) \end{aligned}$$

These equations show that $\hat{Q}_k\hat{R}_k$ is the factorisation of A_k, which exists by the inductive hypothesis, and then lead to

$$\begin{aligned} \boldsymbol{r} &= \hat{Q}_k^{\mathrm{T}}\boldsymbol{a}\,, \\ \boldsymbol{q} &= (1/\alpha)(\boldsymbol{a} - \hat{Q}_k\hat{Q}_k^{\mathrm{T}}\boldsymbol{a})\,, \end{aligned}$$

where $\alpha = \|\boldsymbol{a} - \hat{Q}_k\hat{Q}_k^{\mathrm{T}}\boldsymbol{a}\|_2$. The number α is the constant required to ensure that the vector $\boldsymbol{q}$ is normalised.

The construction fails when $\boldsymbol{a} - \hat{Q}_k\hat{Q}_k^{\mathrm{T}}\boldsymbol{a} = \boldsymbol{0}$, for then the vector $\boldsymbol{q}$ cannot be normalised. In this case we choose $\boldsymbol{q}$ to be any normalised vector in $\mathbb{R}^m$ which is orthogonal in $\mathbb{R}^m$ to all the columns of $\hat{Q}_k$, for then $\boldsymbol{q}^{\mathrm{T}}\hat{Q}_k = \boldsymbol{0}^{\mathrm{T}}$ as required. The condition at the beginning of the proof, that $k < m$, is required by the fact that when $k = m$ the matrix $\hat{Q}_m$ is a square orthogonal matrix, and there is no vector $\boldsymbol{q}$ in $\mathbb{R}^m \setminus \{\boldsymbol{0}\}$ such that $\boldsymbol{q}^{\mathrm{T}}\hat{Q}_m = \boldsymbol{0}^{\mathrm{T}}$.

With these definitions of $\boldsymbol{q}, \boldsymbol{r}, \alpha, \hat{Q}_k$ and $\hat{R}_k$ we have constructed the required factors of A, showing that the theorem is true when $n = k+1$. Since it holds when $n = 1$ the induction is complete.

Now, for the final part, suppose that $\mathrm{rank}(A) = n$. If $\hat{R}$ were singular, there would exist a nonzero vector $\boldsymbol{p} \in \mathbb{R}^n$ such that $\hat{R}\boldsymbol{p} = \boldsymbol{0}$; then, $A\boldsymbol{p} = \hat{Q}\hat{R}\boldsymbol{p} = \boldsymbol{0}$, and hence $\mathrm{rank}(A) < n$, contradicting our hypothesis that $\mathrm{rank}(A) = n$. Therefore, if $\mathrm{rank}(A) = n$, then $\hat{R}$ is nonsingular. □

The matrix factorisation whose existence is asserted in Theorem 2.12 is called the **QR factorisation**. Here, we shall present its use in the solution of least squares problems. In Chapter 5 we shall revisit the idea in a different context which concerns the numerical solution of eigenvalue problems.

Theorem 2.13 *Suppose that $A \in \mathbb{R}^{m\times n}$, with $m \geq n$ and $\mathrm{rank}(A) = n$, and let $\boldsymbol{b} \in \mathbb{R}^m$. Then, there exists a unique least squares solution of the system of equations $A\boldsymbol{x} = \boldsymbol{b}$: a vector $\boldsymbol{x}$ in $\mathbb{R}^n$ which minimises the function $\boldsymbol{y} \mapsto \|A\boldsymbol{y} - \boldsymbol{b}\|_2$ over all $\boldsymbol{y}$ in $\mathbb{R}^n$. The vector $\boldsymbol{x}$ can be obtained by finding the factors $\hat{Q}$ and $\hat{R}$ of A defined in Theorem 2.12, and then solving the nonsingular upper triangular system $\hat{R}\boldsymbol{x} = \hat{Q}^{\mathrm{T}}\boldsymbol{b}$.*

Proof The matrix $\hat{Q}$ has m rows and n columns, with $m \geq n$, and it satisfies

$$\hat{Q}^{\mathrm{T}}\hat{Q} = I_n\,.$$

We shall suppose that $m > n$, the case $m = n$ being a trivial special case with

$$\boldsymbol{x} = A^{-1}\boldsymbol{b} = (\hat{Q}\hat{R})^{-1}\boldsymbol{b} = \hat{R}^{-1}\hat{Q}^{-1}\boldsymbol{b} = \hat{R}^{-1}\hat{Q}^{\mathrm{T}}\boldsymbol{b}\,,$$

and hence $\hat{R}\boldsymbol{x} = \hat{Q}^{\mathrm{T}}\boldsymbol{b}$, as required.

For $m > n$ now, the vector $\boldsymbol{b} \in \mathbb{R}^m$ can be written as the sum of two vectors:

$$\boldsymbol{b} = \boldsymbol{b}_q + \boldsymbol{b}_r\,,$$

where $\boldsymbol{b}_q$ is in the linear space spanned by the n columns of the matrix $\hat{Q}$, and $\boldsymbol{b}_r$ is in the orthogonal complement of this space in $\mathbb{R}^m$. The vector $\boldsymbol{b}_q$ is a linear combination of the columns of $\hat{Q}$, and $\boldsymbol{b}_r$ is orthogonal to every column of $\hat{Q}$; *i.e.*, there exists $\boldsymbol{c} \in \mathbb{R}^n$ such that

$$\boldsymbol{b} = \boldsymbol{b}_q + \boldsymbol{b}_r\,, \quad \boldsymbol{b}_q = \hat{Q}\boldsymbol{c}\,, \quad \hat{Q}^{\mathrm{T}}\boldsymbol{b}_r = \boldsymbol{0}\,. \tag{2.53}$$

Now, suppose that $\boldsymbol{x}$ is the solution of $\hat{R}\boldsymbol{x} = \hat{Q}^{\mathrm{T}}\boldsymbol{b}$, and that $\boldsymbol{y}$ is any vector in $\mathbb{R}^n$. Then,

$$\begin{aligned}
A\boldsymbol{y} - \boldsymbol{b} &= \hat{Q}\hat{R}\boldsymbol{y} - \boldsymbol{b} \\
&= \hat{Q}\hat{R}(\boldsymbol{y} - \boldsymbol{x}) + \hat{Q}\hat{R}\boldsymbol{x} - \boldsymbol{b} \\
&= \hat{Q}\hat{R}(\boldsymbol{y} - \boldsymbol{x}) + \hat{Q}\hat{Q}^{\mathrm{T}}\boldsymbol{b} - \boldsymbol{b} \\
&= \hat{Q}\hat{R}(\boldsymbol{y} - \boldsymbol{x}) + \hat{Q}\hat{Q}^{\mathrm{T}}\boldsymbol{b}_q - \boldsymbol{b}_q + \hat{Q}\hat{Q}^{\mathrm{T}}\boldsymbol{b}_r - \boldsymbol{b}_r \\
&= \hat{Q}\hat{R}(\boldsymbol{y} - \boldsymbol{x}) + \hat{Q}\hat{Q}^{\mathrm{T}}\hat{Q}\boldsymbol{c} - \boldsymbol{b}_q - \boldsymbol{b}_r \\
&= \hat{Q}\hat{R}(\boldsymbol{y} - \boldsymbol{x}) - \boldsymbol{b}_r\,,
\end{aligned}$$

where we have used (2.53) repeatedly; in particular, the last equality follows by noting that $\hat{Q}^{\mathrm{T}}\hat{Q} = I_n$. Hence

$$\begin{aligned}
\|A\boldsymbol{y} - \boldsymbol{b}\|_2^2 &= (\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}}\hat{R}^{\mathrm{T}}\hat{Q}^{\mathrm{T}}\hat{Q}\hat{R}(\boldsymbol{y} - \boldsymbol{x}) + \boldsymbol{b}_r^{\mathrm{T}}\boldsymbol{b}_r - 2(\boldsymbol{y} - \boldsymbol{x})^{\mathrm{T}}\hat{R}^{\mathrm{T}}\hat{Q}^{\mathrm{T}}\boldsymbol{b}_r \\
&= \|\hat{R}(\boldsymbol{y} - \boldsymbol{x})\|_2^2 + \|\boldsymbol{b}_r\|^2 \\
&\geq \|\boldsymbol{b}_r\|^2
\end{aligned}$$

since $\hat{Q}^{\mathrm{T}}\boldsymbol{b}_r = \boldsymbol{0}$. Thus $\|A\boldsymbol{y} - \boldsymbol{b}\|_2$ is smallest when $\hat{R}(\boldsymbol{y} - \boldsymbol{x}) = \boldsymbol{0}$, which implies that $\boldsymbol{y} = \boldsymbol{x}$, since the matrix $\hat{R}$ is nonsingular. Hence $\boldsymbol{x}$, defined as the solution of $\hat{R}\boldsymbol{x} = \hat{Q}^{\mathrm{T}}\boldsymbol{b}$, is the required least squares solution. □

2.10 Notes

There are many good books on the subject of numerical linear algebra which cover the topics discussed in this chapter in much greater detail,

and address questions which we have not touched on here. Without any attempt to be exhaustive, we single out four texts from the vast literature. The first two books on the list below are well-known monographs on the subject, while the last two are excellent textbooks.

- G.H. Golub and C.F. Van Loan, *Matrix Computations,* Third Edition, Johns Hopkins University Press, Baltimore, 1996.
- N.J. Higham, *Accuracy and Stability of Numerical Algorithms,* SIAM, Philadelphia, 1996.
- L.N. Trefethen and D. Bau, III, *Numerical Linear Algebra,* SIAM, Philadelphia, 1997.
- P.G. Ciarlet, *Introduction to Numerical Linear Algebra and Optimisation*, Cambridge University Press, Cambridge, 1989.

As we have already noted in Section 2.2, the invention of the elimination technique is attributed to Gauss who published the method in his *Theoria motus* (1809), although the idea was already known to the Chinese two thousand years ago. Gauss himself was concerned with positive definite systems. The method was extended to linear systems with general matrices by Jacobi.[1] The interpretation of Gaussian elimination as matrix factorisation is due to P.S. Dwyer: A matrix presentation of least squares and correlation theory with matrix justification of improved methods of solutions, *Ann. Math. Stat.* **15**, 82–89, 1944.

The sensitivity of Gaussian elimination to rounding errors was studied by Wilkinson[2] in Error analysis of direct methods of matrix inversion, *J. Assoc. Comput. Math.* **8**, 281–330, 1961. The idea of pivoting was used as early as 1947 by von Neumann[3] and Goldstein.[4] The concept of the condition number of a matrix was introduced by Turing[5] in Rounding-off errors in matrix processes, *Quart. J. Mech. Appl. Math.* **1**, 287–308, 1948. Our treatment of condition numbers follows the textbook of Trefethen and Bau, cited above.

[1] Carl Gustav Jacob Jacobi (10 December 1804, Potsdam, Prussia, Holy Roman Empire (now Germany) – 18 February 1851, Berlin, Germany) had made important contributions to the theory of elliptic functions and differential equations. The English translation, by G.W. Stuart, of Jacobi's German original article is available from the Internet on `ftp://thales.cs.umd.edu/pub/biographical/xhist.html`

[2] James Hardy Wilkinson (27 September 1919, Strood, Kent, England – 5 October 1986, London, England).

[3] John von Neumann (28 December 1903, Budapest, Austria–Hungary (now in Hungary) – 8 February 1957, Washington DC, USA).

[4] Sydney Goldstein (3 December 1903, Hull, England – 22 January 1989, Belmont, Massachusetts, USA).

[5] Alan Mathison Turing (23 June 1912, London, England – 7 June 1954, Wilmslow, Cheshire, England).

Normed linear spaces play a key role in functional analysis (see, for example, K. Yosida, *Functional Analysis*, Third Edition, Springer, Berlin, 1971, page 30). Here, we have concentrated on finite-dimensional normed linear spaces over the field of real numbers.

The relevance of norms in numerical linear algebra was highlighted by Householder[1] in his book *The Theory of Matrices in Numerical Analysis,* Blaisdell, New York, 1964.

The idea of least squares fitting is due to Gauss, who invented the method in the 1790s. However, it was the French mathematician Legendre[2] who first published the method in 1806 in a book on determining the orbits of comets. Legendre's method involved a number of observations taken at equal intervals and he assumed that the comet followed a parabolic path, so he ended up with more equations than there were unknowns. Legendre then applied his methods to the data known for two comets. In an Appendix to the book Legendre described the least squares method of fitting a curve to the data available. Gauss published his version of the least squares method in 1809 and, although acknowledging that it had already appeared in Legendre's book, Gauss nevertheless claimed priority for himself. This greatly hurt Legendre, leading to one of the infamous priority disputes in the history of mathematics. A recent exhaustive monograph on numerical algorithms for least squares problems is due to Å. Björk: *Numerical Methods for Least Squares Problems,* SIAM, Philadelphia, 1996.

The version of the QR factorisation considered here is the *reduced version*, following the terminology in Chapter 7 of Trefethen and Bau. In the *full version* of the QR factorisation for a matrix $A \in \mathbb{R}^{m\times n}$, we have $A = QR$, where $Q \in \mathbb{R}^{m\times m}$, $R \in \mathbb{R}^{m\times n}$ (cf. Chapter 5).

In a footnote to Definition 2.12 we mentioned the Moore–Penrose generalised inverse A^+ of a matrix $A \in \mathbb{R}^{m\times n}$. A^+ can be defined through the singular value decomposition of A (cf. L.N. Trefethen and D. Bau, III: *Numerical Linear Algebra*, SIAM, Philadelphia, 1997). Recall that the singular values of A are the square roots of the (nonnegative) eigenvalues of the matrix $A^{\mathrm{T}}A$.

[1] Alton Scott Householder (5 May 1904, Rockford, Illinois, USA – 4 July 1993, Malibu, California, USA) was one of the pioneers of numerical linear algebra. Householder's obituary by G.W. Stuart, published in *SIAM News*, is available from `http://www.inf.ethz.ch/research/wr/conferences/householder/stewart.html`

[2] Adrien-Marie Legendre (18 September 1752, Paris, France – 10 January 1833, Paris, France).

Theorem 2.14 (Singular value decomposition) *Let $A \in \mathbb{R}^{m\times n}$; then, there exist $U \in \mathbb{R}^{m\times n}$, $\Sigma \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{n\times n}$ such that*

$$A = U\Sigma V^{\mathrm{T}},$$

where Σ is a diagonal matrix whose diagonal entries, σ_{ii}, $i = 1, 2, \ldots, n$, are the singular values of A, $U^{\mathrm{T}}U = I_n$ and $V^{\mathrm{T}}V = I_n$, with I_n denoting the $n \times n$ identity matrix.

The Moore–Penrose generalised inverse of the diagonal matrix $\Sigma \in \mathbb{R}^{n\times n}$ is defined as the diagonal matrix $\Sigma^{+} \in \mathbb{R}^{n\times n}$ whose diagonal entries are

$$\sigma_{ii}^{+} = \begin{cases} \sigma_{ii}^{-1} & \text{if } \sigma_{ii} \neq 0\,, \\ 0 & \text{if } \sigma_{ii} = 0\,. \end{cases}$$

The generalised inverse $A^{+} \in \mathbb{R}^{n\times m}$ of a matrix $A \in \mathbb{R}^{m\times n}$ with singular value decomposition $A = U\Sigma V^{\mathrm{T}}$ is defined by

$$A^{+} = V\Sigma^{+}U^{\mathrm{T}}\,.$$

In the special case when $m = n$ and $A \in \mathbb{R}^{n\times n}$ is nonsingular, the n singular values of A are all nonzero and therefore $\Sigma^{+} = \Sigma^{-1}$. Hence, also, $A^{+} = A^{-1}$, which then justifies the use of the terminology 'generalised inverse' for the matrix A^{+} defined above.

Exercises

2.1 Let $n \geq 2$. Given the matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$, the permutation matrix $Q \in \mathbb{R}^{n\times n}$ reverses the order of the rows of A, so that $(QA)_{i,j} = a_{n+1-i,j}$. If $L \in \mathbb{R}^{n\times n}$ is a lower triangular matrix, what is the structure of the matrix QLQ?

Show how to factorise $A \in \mathbb{R}^{n\times n}$ in the form $A = UL$, where $U \in \mathbb{R}^{n\times n}$ is unit upper triangular and $L \in \mathbb{R}^{n\times n}$ is lower triangular. What conditions on A will ensure that the factorisation exists? Give an example of a square matrix A which cannot be factorised in this way.

2.2 Let $n \geq 2$. Consider a matrix $A \in \mathbb{R}^{n\times n}$ whose every leading principal submatrix of order less than n is nonsingular. Show that A can be factored in the form $A = LDU$, where $L \in \mathbb{R}^{n\times n}$ is unit lower triangular, $D \in \mathbb{R}^{n\times n}$ is diagonal and $U \in \mathbb{R}^{n\times n}$ is unit upper triangular.

If the factorisation $A = LU$ is known, where L is unit lower

triangular and U is upper triangular, show how to find the factors of the transpose A^{T}.

2.3 Let $n \geq 2$ and suppose that the matrix $A \in \mathbb{R}^{n\times n}$ is nonsingular. Show by induction, as in Theorem 2.3, that there are a permutation matrix $P \in \mathbb{R}^{n\times n}$, a lower triangular matrix $L \in \mathbb{R}^{n\times n}$, and a unit upper triangular matrix $U \in \mathbb{R}^{n\times n}$ such that $PA = LU$.

By finding a suitable 2×2 matrix A, or otherwise, show that this may not be true if A is singular.

2.4 The lower triangular matrix $L \in \mathbb{R}^{n\times n}$, $n \geq 2$, is nonsingular, and the vector $\boldsymbol{b} \in \mathbb{R}^n$ is such that $b_i = 0,\ i = 1, 2, \ldots, k$, with $1 \leq k \leq n$. The vector $\boldsymbol{y} \in \mathbb{R}^n$ is the solution of $L\boldsymbol{y} = \boldsymbol{b}$. Show, by partitioning L, that $y_j = 0,\ j = 1, 2, \ldots, k$. Hence give an alternative proof of Theorem 2.1(iv), that the inverse of a nonsingular lower triangular matrix is itself lower triangular.

2.5 Given a matrix $A \in \mathbb{R}^{n\times n}$, define the matrix $B \in \mathbb{R}^{n\times 2n}$ in which the first n columns are the columns of A, and the last n columns are the columns of the identity matrix I_n. Consider the following computational scheme. Treat the rows of the matrix B in order, so that $j = 1, 2, \ldots, n$. Multiply every element in row j by the reciprocal of the diagonal element, $1/b_{jj}$; then, replace every element b_{ik} which is not in row j, so that $i \neq j$, by $b_{ik} - b_{ij}b_{jk}$.

Show that the result is equivalent to multiplying B on the left by a sequence of matrices. Explain why, at the end of the computation, the first n columns of B are the columns of the identity matrix I_n, and the last n columns are the columns of the inverse matrix A^{-1}. Give a condition on the matrix A which will ensure that the computation does not break down.

Show that the process as described requires approximately $2n^3$ multiplications, but that, if the multiplications in which one of the factors is zero are not counted, the total is approximately n^3.

2.6 Use the method of Exercise 5 to find the inverse of the matrix

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

2.7 Suppose that for a matrix $A \in \mathbb{R}^{n\times n}$,

$$\sum_{i=1}^{n} |a_{ij}| \le C\,, \quad j = 1, 2, \ldots, n\,.$$

Show that, for any vector $\boldsymbol{x} \in \mathbb{R}^n$,

$$\sum_{i=1}^{n} |(A\boldsymbol{x})_i| \le C\|\boldsymbol{x}\|_1\,.$$

Find a nonzero vector $\boldsymbol{x}$ for which equality can be achieved, and deduce that

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

2.8 (i) Show that, for any vector $\boldsymbol{v} = (v_1, \ldots, v_n)^{\mathrm{T}} \in \mathbb{R}^n$,

$$\|\boldsymbol{v}\|_\infty \le \|\boldsymbol{v}\|_2 \ \text{ and } \ \|\boldsymbol{v}\|_2^2 \le \|\boldsymbol{v}\|_1\,\|\boldsymbol{v}\|_\infty\,.$$

In each case give an example of a nonzero vector $\boldsymbol{v}$ for which equality is attained. Deduce that $\|\boldsymbol{v}\|_\infty \le \|\boldsymbol{v}\|_2 \le \|\boldsymbol{v}\|_1$. Show also that $\|\boldsymbol{v}\|_2 \le \sqrt{n}\,\|\boldsymbol{v}\|_\infty$.
(ii) Show that, for any matrix $A \in \mathbb{R}^{m\times n}$,

$$\|A\|_\infty \le \sqrt{n}\,\|A\|_2 \ \text{ and } \ \|A\|_2 \le \sqrt{m}\,\|A\|_\infty\,.$$

In each case give an example of a matrix A for which equality is attained. (See the footnote following Definition 2.12 for the meaning of $\|A\|_1$, $\|A\|_2$ and $\|A\|_\infty$ when $A \in \mathbb{R}^{m\times n}$.)

2.9 Prove that, for any nonsingular matrix $A \in \mathbb{R}^{n\times n}$,

$$\kappa_2(A) = \left(\frac{\lambda_n}{\lambda_1}\right)^{1/2},$$

where λ_1 is the smallest and λ_n is the largest eigenvalue of the matrix $A^{\mathrm{T}}A$.

Show that the condition number $\kappa_2(Q)$ of an orthogonal matrix Q is equal to 1. Conversely, if $\kappa_2(A) = 1$ for the matrix A, show that all the eigenvalues of $A^{\mathrm{T}}A$ are equal; deduce that A is a scalar multiple of an orthogonal matrix.

2.10 Let $A \in \mathbb{R}^{n\times n}$. Show that if λ is an eigenvalue of $A^{\mathrm{T}}A$, then

$$0 \le \lambda \le \|A^{\mathrm{T}}\|\,\|A\|\,,$$

provided that the same subordinate matrix norm is used for

both A and A^{T}. Hence show that, for any nonsingular $n \times n$ matrix A,

$$\kappa_2(A) \leq \{\kappa_1(A)\,\kappa_\infty(A)\}^{1/2}\,.$$

2.11 For the matrix defined by (2.46) write down the matrix $A^{\mathrm{T}}A$. Show that any vector $\boldsymbol{x} \neq \boldsymbol{0}$ is an eigenvector of $A^{\mathrm{T}}A$ with eigenvalue $\lambda = 1$, provided that $x_1 = 0$ and $x_2 + \cdots + x_n = 0$. Show also that there are two eigenvectors with $x_2 = \cdots = x_n$ and find the corresponding eigenvalues. Deduce that

$$\kappa_2(A) = \tfrac{1}{2}(n+1)\left(1 + \sqrt{1 - \tfrac{4}{(n+1)^2}}\right)\,.$$

2.12 Let $B \in \mathbb{R}^{n\times n}$ and denote by I the identity matrix of order n. Show that if the matrix $I - B$ is singular, then there exists a nonzero vector $\boldsymbol{x} \in \mathbb{R}^n$ such that $(I - B)\boldsymbol{x} = \boldsymbol{0}$; deduce that $\|B\| \geq 1$, and hence that, if $\|A\| < 1$, then the matrix $I - A$ is nonsingular.

Now suppose that $A \in \mathbb{R}^{n\times n}$ with $\|A\| < 1$. Show that

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

and hence that

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

Deduce that

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

2.13 Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix and $\boldsymbol{b} \in \mathbb{R}^n_*$. Suppose that $A\boldsymbol{x} = \boldsymbol{b}$ and $(A+\delta A)(\boldsymbol{x}+\boldsymbol{\delta x}) = \boldsymbol{b}$, and that $\|A^{-1}\,\delta A\| < 1$. Use the result of Exercise 12 to show that

$$\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} \leq \frac{\|A^{-1}\,\delta A\|}{1 - \|A^{-1}\,\delta A\|}\,.$$

2.14 Suppose that $A \in \mathbb{R}^{n\times n}$ is a nonsingular matrix, and $\boldsymbol{b} \in \mathbb{R}^n_*$. Given that $A\boldsymbol{x} = \boldsymbol{b}$ and $A(\boldsymbol{x} + \boldsymbol{\delta x}) = \boldsymbol{b} + \boldsymbol{\delta b}$, Theorem 2.11 states that

$$\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} \leq \kappa(A)\frac{\|\boldsymbol{\delta b}\|}{\|\boldsymbol{b}\|}\,.$$

By considering the eigenvectors of $A^{\mathrm{T}}A$, show how to find vectors $\boldsymbol{b}$ and $\boldsymbol{\delta b}$ for which equality is attained, when using the 2-norm.

2.15 Find the QR factorisation of the matrix

$$A = \begin{pmatrix} 9 & -6 \\ 12 & -8 \\ 0 & 20 \end{pmatrix},$$

and hence find the least squares solution of the system of linear equations

$$\begin{aligned} 9x - 6y &= 300\,, \\ 12x - 8y &= 600\,, \\ 20y &= 900\,. \end{aligned}$$