Chapter 7

Gaussian Elimination, LU-Factorization, Cholesky Factorization, Reduced Row Echelon Form

In this chapter we assume that all vector spaces are over the field $\mathbb{R}$. All results that do not rely on the ordering on $\mathbb{R}$ or on taking square roots hold for arbitrary fields.

7.1 Motivating Example: Curve Interpolation

Curve interpolation is a problem that arises frequently in computer graphics and in robotics (path planning). There are many ways of tackling this problem and in this section we will describe a solution using *cubic splines*. Such splines consist of cubic Bézier curves. They are often used because they are cheap to implement and give more flexibility than quadratic Bézier curves.

A *cubic Bézier curve* $C(t)$ (in $\mathbb{R}^2$ or $\mathbb{R}^3$) is specified by a list of four *control points* (b_0, b_2, b_2, b_3) and is given parametrically by the equation

$$C(t) = (1-t)^3\, b_0 + 3(1-t)^2 t\, b_1 + 3(1-t)t^2\, b_2 + t^3\, b_3.$$

Clearly, $C(0) = b_0$, $C(1) = b_3$, and for $t \in [0, 1]$, the point $C(t)$ belongs to the convex hull of the control points b_0, b_1, b_2, b_3. The polynomials

$$(1-t)^3, \quad 3(1-t)^2 t, \quad 3(1-t)t^2, \quad t^3$$

are the *Bernstein polynomials* of degree 3.

Typically, we are only interested in the curve segment corresponding to the values of t in the interval $[0, 1]$. Still, the placement of the control points drastically affects the shape of the curve segment, which can even have a self-intersection; See Figures 7.1, 7.2, 7.3 illustrating various configurations.

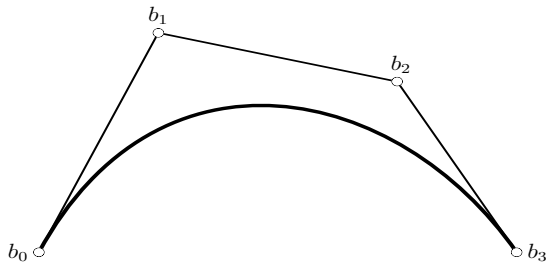

Figure 7.1: A "standard" Bézier curve.

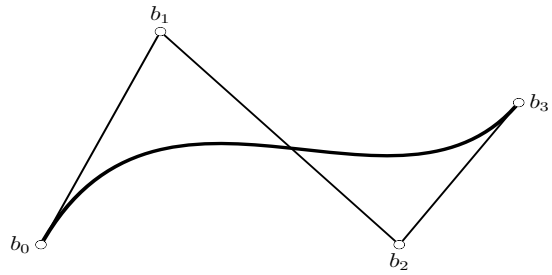

Figure 7.2: A Bézier curve with an inflection point.

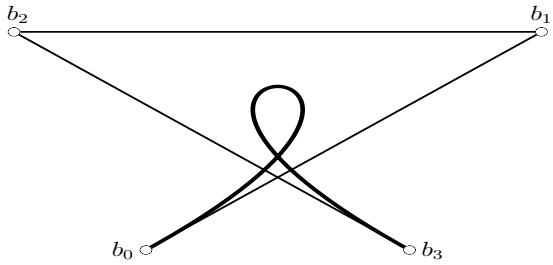

Figure 7.3: A self-intersecting Bézier curve.

Interpolation problems require finding curves passing through some given data points and possibly satisfying some extra constraints.

A *Bézier spline curve* F is a curve which is made up of curve segments which are Bézier curves, say $C_1, \ldots, C_m$ $(m \geq 2)$. We will assume that F defined on $[0, m]$, so that for $i = 1, \ldots, m$,

$$F(t) = C_i(t - i + 1), \quad i - 1 \leq t \leq i.$$

Typically, some smoothness is required between any two junction points, that is, between any two points $C_i(1)$ and $C_{i+1}(0)$, for $i = 1, \ldots, m-1$. We require that $C_i(1) = C_{i+1}(0)$ (C^0-*continuity*), and typically that the derivatives of C_i at 1 and of C_{i+1} at 0 agree up to second order derivatives. This is called C^2-*continuity*, and it ensures that the tangents agree as well as the curvatures.

There are a number of interpolation problems, and we consider one of the most common problems which can be stated as follows:

Problem: Given $N+1$ data points $x_0, \ldots, x_N$, find a C^2 cubic spline curve F such that $F(i) = x_i$ for all i, $0 \leq i \leq N$ ($N \geq 2$).

A way to solve this problem is to find $N+3$ auxiliary points $d_{-1}, \ldots, d_{N+1}$, called *de Boor control points*, from which N Bézier curves can be found. Actually,

$$d_{-1} = x_0 \quad \text{and} \quad d_{N+1} = x_N$$

so we only need to find $N+1$ points $d_0, \ldots, d_N$.

It turns out that the C^2-continuity constraints on the N Bézier curves yield only $N-1$ equations, so d_0 and d_N can be chosen arbitrarily. In practice, d_0 and d_N are chosen according to various *end conditions*, such as prescribed velocities at x_0 and x_N. For the time being, we will assume that d_0 and d_N are given.

Figure 7.4 illustrates an interpolation problem involving $N+1 = 7+1 = 8$ data points. The control points d_0 and d_7 were chosen arbitrarily.

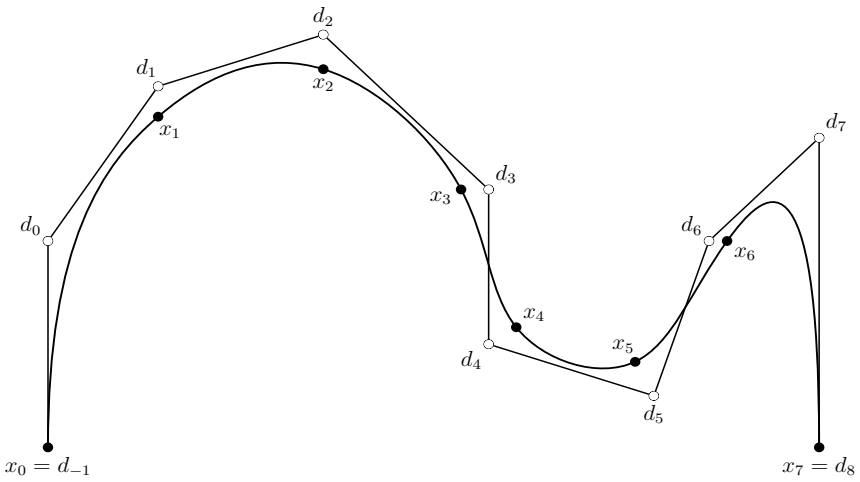

Figure 7.4: A C^2 cubic interpolation spline curve passing through the points x_0, x_1, x_2, x_3, x_4, x_5, x_6, x_7.

It can be shown that $d_1, \ldots, d_{N-1}$ are given by the linear system

$$\begin{pmatrix} \frac{7}{2} & 1 & & & \\ 1 & 4 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & 4 & 1 \\ & & & 1 & \frac{7}{2} \end{pmatrix} \begin{pmatrix} d_1 \\ d_2 \\ \vdots \\ d_{N-2} \\ d_{N-1} \end{pmatrix} = \begin{pmatrix} 6x_1 - \frac{3}{2}d_0 \\ 6x_2 \\ \vdots \\ 6x_{N-2} \\ 6x_{N-1} - \frac{3}{2}d_N \end{pmatrix}.$$

We will show later that the above matrix is invertible because it is strictly diagonally dominant.

Once the above system is solved, the Bézier cubics $C_1, \ldots, C_N$ are determined as follows (we assume $N \geq 2$): For $2 \leq i \leq N-1$, the control points $(b_0^i, b_1^i, b_2^i, b_3^i)$ of C_i are given by

$$\begin{aligned} b_0^i &= x_{i-1} \\ b_1^i &= \frac{2}{3}d_{i-1} + \frac{1}{3}d_i \\ b_2^i &= \frac{1}{3}d_{i-1} + \frac{2}{3}d_i \\ b_3^i &= x_i. \end{aligned}$$

The control points $(b_0^1, b_1^1, b_2^1, b_3^1)$ of C_1 are given by

$$\begin{aligned} b_0^1 &= x_0 \\ b_1^1 &= d_0 \\ b_2^1 &= \frac{1}{2}d_0 + \frac{1}{2}d_1 \\ b_3^1 &= x_1, \end{aligned}$$

and the control points $(b_0^N, b_1^N, b_2^N, b_3^N)$ of C_N are given by

$$\begin{aligned} b_0^N &= x_{N-1} \\ b_1^N &= \frac{1}{2}d_{N-1} + \frac{1}{2}d_N \\ b_2^N &= d_N \\ b_3^N &= x_N. \end{aligned}$$

Figure 7.5 illustrates this process spline interpolation for $N = 7$.

We will now describe various methods for solving linear systems. Since the matrix of the above system is tridiagonal, there are specialized methods which are more efficient than the general methods. We will discuss a few of these methods.

7.2 Gaussian Elimination

Let A be an $n \times n$ matrix, let $b \in \mathbb{R}^n$ be an n-dimensional vector and assume that A is invertible. Our goal is to solve the system $Ax = b$. Since A is assumed to be invertible,

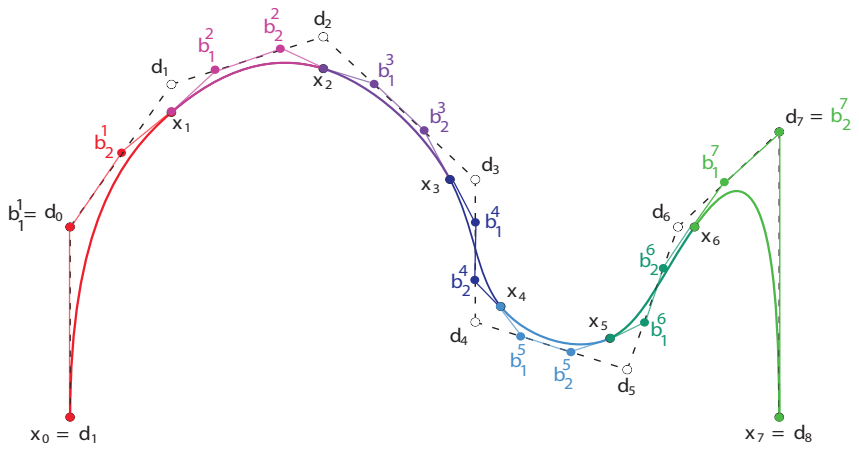

Figure 7.5: A C^2 cubic interpolation of $x_0, x_1,\ x_2, x_3,\ x_4, x_5,\ x_6, x_7$ with associated color coded Bézier cubics.

we know that this system has a unique solution $x = A^{-1}b$. Experience shows that two counter-intuitive facts are revealed:

(1) One should avoid computing the inverse A^{-1} of A explicitly. This is inefficient since it would amount to solving the n linear systems $Au^{(j)} = e_j$ for $j = 1, \ldots, n$, where $e_j = (0, \ldots, 1, \ldots, 0)$ is the jth canonical basis vector of $\mathbb{R}^n$ (with a 1 is the jth slot). By doing so, we would replace the resolution of a single system by the resolution of n systems, and we would still have to multiply A^{-1} by b.

(2) One does not solve (large) linear systems by computing determinants (using Cramer's formulae) since this method requires a number of additions (resp. multiplications) proportional to $(n+1)!$ (resp. $(n+2)!$).

The key idea on which most direct methods (as opposed to iterative methods, that look for an approximation of the solution) are based is that if A is an upper-triangular matrix, which means that $a_{ij} = 0$ for $1 \leq j < i \leq n$ (resp. lower-triangular, which means that $a_{ij} = 0$ for $1 \leq i < j \leq n$), then computing the solution x is trivial. Indeed, say A is an upper-triangular matrix

$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \cdots & a_{1\,n-2} & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{2\,2} & \cdots & a_{2\,n-2} & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & \ddots & \vdots & \vdots & \vdots \\ & & & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & \cdots & 0 & 0 & a_{n\,n} \end{pmatrix}.$$

Then $\det(A) = a_{11}a_{22}\cdots a_{nn} \neq 0$, which implies that $a_{ii} \neq 0$ for $i = 1, \ldots, n$, and we can solve the system $Ax = b$ from bottom-up by *back-substitution*. That is, first we compute x_n from the last equation, next plug this value of x_n into the next to the last equation and compute x_{n-1} from it, *etc.* This yields

$$\begin{aligned} x_n &= a_{n\,n}^{-1} b_n \\ x_{n-1} &= a_{n-1\,n-1}^{-1}(b_{n-1} - a_{n-1\,n}x_n) \\ &\vdots \\ x_1 &= a_{1\,1}^{-1}(b_1 - a_{1\,2}x_2 - \cdots - a_{1\,n}x_n). \end{aligned}$$

Note that the use of determinants can be avoided to prove that if A is invertible then $a_{ii} \neq 0$ for $i = 1, \ldots, n$. Indeed, it can be shown directly (by induction) that an upper (or lower) triangular matrix is invertible iff all its diagonal entries are nonzero.

If A is lower-triangular, we solve the system from top-down by *forward-substitution*.

Thus, what we need is a method for transforming a matrix to an equivalent one in upper-triangular form. This can be done by *elimination*. Let us illustrate this method on the following example:

$$\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 5 \\ 4x & - & 6y & & & = & -2 \\ -2x & + & 7y & + & 2z & = & 9. \end{array}$$

We can eliminate the variable x from the second and the third equation as follows: Subtract twice the first equation from the second and add the first equation to the third. We get the new system

$$\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 5 \\ & - & 8y & - & 2z & = & -12 \\ & & 8y & + & 3z & = & 14. \end{array}$$

This time we can eliminate the variable y from the third equation by adding the second equation to the third:

$$\begin{array}{rcrcrcr} 2x & + & y & + & z & = & 5 \\ & - & 8y & - & 2z & = & -12 \\ & & & & z & = & 2. \end{array}$$

This last system is upper-triangular. Using back-substitution, we find the solution: $z = 2$, $y = 1$, $x = 1$.

Observe that we have performed only *row operations*. The general method is to iteratively eliminate variables using simple row operations (namely, adding or subtracting a multiple of a row to another row of the matrix) while simultaneously applying these operations to the vector b, to obtain a system, $MAx = Mb$, where MA is upper-triangular. Such a method is called *Gaussian elimination*. However, one extra twist is needed for the method to work in all cases: It may be necessary to permute rows, as illustrated by the following example:

$$\begin{array}{rcrcrl} x & + & y & + & z & = 1 \\ x & + & y & + & 3z & = 1 \\ 2x & + & 5y & + & 8z & = 1. \end{array}$$

In order to eliminate x from the second and third row, we subtract the first row from the second and we subtract twice the first row from the third:

$$\begin{aligned} x + y + z &= 1 \\ 2z &= 0 \\ 3y + 6z &= -1. \end{aligned}$$

Now the trouble is that y does not occur in the second row; so, we can't eliminate y from the third row by adding or subtracting a multiple of the second row to it. The remedy is simple: Permute the second and the third row! We get the system:

$$\begin{aligned} x + y + z &= 1 \\ 3y + 6z &= -1 \\ 2z &= 0, \end{aligned}$$

which is already in triangular form. Another example where some permutations are needed is:

$$\begin{aligned} z &= 1 \\ -2x + 7y + 2z &= 1 \\ 4x - 6y &= -1. \end{aligned}$$

First we permute the first and the second row, obtaining

$$\begin{aligned} -2x + 7y + 2z &= 1 \\ z &= 1 \\ 4x - 6y &= -1, \end{aligned}$$

and then we add twice the first row to the third, obtaining:

$$\begin{aligned} -2x + 7y + 2z &= 1 \\ z &= 1 \\ 8y + 4z &= 1. \end{aligned}$$

Again we permute the second and the third row, getting

$$\begin{aligned} -2x + 7y + 2z &= 1 \\ 8y + 4z &= 1 \\ z &= 1, \end{aligned}$$

an upper-triangular system. Of course, in this example, z is already solved and we could have eliminated it first, but for the general method, we need to proceed in a systematic fashion.

We now describe the method of *Gaussian elimination* applied to a linear system $Ax = b$, where A is assumed to be invertible. We use the variable k to keep track of the stages of elimination. Initially, $k = 1$.

(1) The first step is to pick some nonzero entry $a_{i\,1}$ in the first column of A. Such an entry must exist, since A is invertible (otherwise, the first column of A would be the zero vector, and the columns of A would not be linearly independent. Equivalently, we would have $\det(A) = 0$). The actual choice of such an element has some impact on the numerical stability of the method, but this will be examined later. For the time being, we assume that some arbitrary choice is made. This chosen element is called the *pivot* of the elimination step and is denoted π_1 (so, in this first step, $\pi_1 = a_{i\,1}$).

(2) Next we permute the row (i) corresponding to the pivot with the first row. Such a step is called *pivoting*. So after this permutation, the first element of the first row is nonzero.

(3) We now eliminate the variable x_1 from all rows except the first by adding suitable multiples of the first row to these rows. More precisely we add $-a_{i\,1}/\pi_1$ times the first row to the ith row for $i = 2, \ldots, n$. At the end of this step, all entries in the first column are zero except the first.

(4) Increment k by 1. If $k = n$, stop. Otherwise, $k < n$, and then iteratively repeat Steps (1), (2), (3) on the $(n - k + 1) \times (n - k + 1)$ subsystem obtained by deleting the first $k - 1$ rows and $k - 1$ columns from the current system.

If we let $A_1 = A$ and $A_k = (a_{i\,j}^{(k)})$ be the matrix obtained after $k - 1$ elimination steps $(2 \leq k \leq n)$, then the kth elimination step is applied to the matrix A_k of the form

$$A_k = \begin{pmatrix} a_{1\,1}^{(k)} & a_{1\,2}^{(k)} & \cdots & \cdots & \cdots & a_{1\,n}^{(k)} \\ 0 & a_{2\,2}^{(k)} & \cdots & \cdots & \cdots & a_{2\,n}^{(k)} \\ \vdots & \ddots & \ddots & \vdots & & \vdots \\ 0 & 0 & 0 & a_{k\,k}^{(k)} & \cdots & a_{k\,n}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & a_{n\,k}^{(k)} & \cdots & a_{n\,n}^{(k)} \end{pmatrix}.$$

Actually, note that

$$a_{i\,j}^{(k)} = a_{i\,j}^{(i)}$$

for all i, j with $1 \leq i \leq k - 2$ and $i \leq j \leq n$, since the first $k - 1$ rows remain unchanged after the $(k - 1)$th step.

We will prove later that $\det(A_k) = \pm \det(A)$. Consequently, A_k is invertible. The fact that A_k is invertible iff A is invertible can also be shown without determinants from the fact that there is some invertible matrix M_k such that $A_k = M_k A$, as we will see shortly.

Since A_k is invertible, some entry $a_{i\,k}^{(k)}$ with $k \leq i \leq n$ is nonzero. Otherwise, the last $n - k + 1$ entries in the first k columns of A_k would be zero, and the first k columns of A_k would yield k vectors in $\mathbb{R}^{k-1}$. But then the first k columns of A_k would be linearly

dependent and A_k would not be invertible, a contradiction. This situation is illustrated by the following matrix for $n = 5$ and $k = 3$:

$$\begin{pmatrix} a_{11}^{(3)} & a_{12}^{(3)} & a_{13}^{(3)} & a_{13}^{(3)} & a_{15}^{(3)} \\ 0 & a_{22}^{(3)} & a_{23}^{(3)} & a_{24}^{(3)} & a_{25}^{(3)} \\ 0 & 0 & 0 & a_{34}^{(3)} & a_{35}^{(3)} \\ 0 & 0 & 0 & a_{44}^{(3)} & a_{4n}^{(3)} \\ 0 & 0 & 0 & a_{54}^{(3)} & a_{55}^{(3)} \end{pmatrix}.$$

The first three columns of the above matrix are linearly dependent.

So one of the entries $a_{ik}^{(k)}$ with $k \leq i \leq n$ can be chosen as pivot, and we permute the kth row with the ith row, obtaining the matrix $\alpha^{(k)} = (\alpha_{jl}^{(k)})$. The new pivot is $\pi_k = \alpha_{kk}^{(k)}$, and we zero the entries $i = k+1, \ldots, n$ in column k by adding $-\alpha_{ik}^{(k)}/\pi_k$ times row k to row i. At the end of this step, we have A_{k+1}. Observe that the first $k-1$ rows of A_k are identical to the first $k-1$ rows of A_{k+1}.

The process of Gaussian elimination is illustrated in schematic form below:

$$\begin{pmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix} \Longrightarrow \begin{pmatrix} \times & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \\ \mathbf{0} & \times & \times & \times \end{pmatrix} \Longrightarrow \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times \\ 0 & \mathbf{0} & \times & \times \end{pmatrix} \Longrightarrow \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \mathbf{0} & \times \end{pmatrix}.$$

7.3 Elementary Matrices and Row Operations

It is easy to figure out what kind of matrices perform the elementary row operations used during Gaussian elimination. The key point is that if $A = PB$, where A, B are $m \times n$ matrices and P is a square matrix of dimension m, if (as usual) we denote the rows of A and B by $A_1, \ldots, A_m$ and $B_1, \ldots, B_m$, then the formula

$$a_{ij} = \sum_{k=1}^{m} p_{ik} b_{kj}$$

giving the (i, j)th entry in A shows that the ith row of A is a *linear combination* of the rows of B:

$$A_i = p_{i1} B_1 + \cdots + p_{im} B_m.$$

Therefore, *multiplication of a matrix on the left by a square matrix performs row operations*. Similarly, multiplication of a matrix on the right by a square matrix performs column operations

The permutation of the kth row with the ith row is achieved by multiplying A on the left by the *transposition matrix* $P(i, k)$, which is the matrix obtained from the identity matrix

by permuting rows i and k, *i.e.*,

$$P(i,k) = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 0 & & & & 1 & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & 1 & & & & 0 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix}.$$

For example, if $m = 3$,

$$P(1,3) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix},$$

then

$$P(1,3)B = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & \cdots & \cdots b_{1n} \\ b_{21} & b_{22} & \cdots & \cdots & \cdots b_{2n} \\ b_{31} & b_{32} & \cdots & \cdots & \cdots b_{3n} \end{pmatrix} = \begin{pmatrix} b_{31} & b_{32} & \cdots & \cdots & \cdots b_{3n} \\ b_{21} & b_{22} & \cdots & \cdots & \cdots b_{2n} \\ b_{11} & b_{12} & \cdots & \cdots & \cdots b_{1n} \end{pmatrix}.$$

Observe that $\det(P(i,k)) = -1$. Furthermore, $P(i,k)$ is *symmetric* ($P(i,k)^\top = P(i,k)$), and

$$P(i,k)^{-1} = P(i,k).$$

During the permutation Step (2), if row k and row i need to be permuted, the matrix A is multiplied on the left by the matrix P_k such that $P_k = P(i,k)$, else we set $P_k = I$.

Adding β times row j to row i (with $i \neq j$) is achieved by multiplying A on the left by the *elementary matrix*,

$$E_{i,j;\beta} = I + \beta e_{ij},$$

where

$$(e_{ij})_{kl} = \begin{cases} 1 & \text{if } k = i \text{ and } l = j \\ 0 & \text{if } k \neq i \text{ or } l \neq j, \end{cases}$$

i.e.,

$$E_{i,j;\beta} = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 1 & & & & & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & \beta & & & & 1 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix} \quad \text{or} \quad E_{i,j;\beta} = \begin{pmatrix} 1 & & & & & & & & \\ & 1 & & & & & & & \\ & & 1 & & & & \beta & & \\ & & & 1 & & & & & \\ & & & & \ddots & & & & \\ & & & & & 1 & & & \\ & & & & & & 1 & & \\ & & & & & & & 1 & \\ & & & & & & & & 1 \end{pmatrix},$$

on the left, $i > j$, and on the right, $i < j$. The index i is the index of the row that is *changed* by the multiplication. For example, if $m = 3$ and we want to add twice row 1 to row 3, since $\beta = 2$, $j = 1$ and $i = 3$, we form

$$E_{3,1;2} = I + 2e_{31} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 2 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix},$$

and calculate

$$\begin{aligned} E_{3,1;2}B &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & \cdots & \cdots b_{1n} \\ b_{21} & b_{22} & \cdots & \cdots & \cdots b_{2n} \\ b_{31} & b_{32} & \cdots & \cdots & \cdots b_{3n} \end{pmatrix} \\ &= \begin{pmatrix} b_{11} & b_{12} & \cdots & \cdots & \cdots b_{1n} \\ b_{21} & b_{22} & \cdots & \cdots & \cdots b_{2n} \\ 2b_{11} + b_{31} & 2b_{12} + b_{32} & \cdots & \cdots & \cdots 2b_{1n} + b_{3n} \end{pmatrix}. \end{aligned}$$

Observe that the inverse of $E_{i,j;\beta} = I + \beta e_{i\,j}$ is $E_{i,j;-\beta} = I - \beta e_{i\,j}$ and that $\det(E_{i,j;\beta}) = 1$. Therefore, during Step 3 (the elimination step), the matrix A is multiplied on the left by a product E_k of matrices of the form $E_{i,k;\beta_{i,k}}$, with $i > k$.

Consequently, we see that

$$A_{k+1} = E_k P_k A_k,$$

and then

$$A_k = E_{k-1} P_{k-1} \cdots E_1 P_1 A.$$

This justifies the claim made earlier that $A_k = M_k A$ for some invertible matrix M_k; we can pick

$$M_k = E_{k-1} P_{k-1} \cdots E_1 P_1,$$

a product of invertible matrices.

The fact that $\det(P(i,k)) = -1$ and that $\det(E_{i,j;\beta}) = 1$ implies immediately the fact claimed above: We always have

$$\det(A_k) = \pm \det(A).$$

Furthermore, since

$$A_k = E_{k-1} P_{k-1} \cdots E_1 P_1 A$$

and since Gaussian elimination stops for $k = n$, the matrix

$$A_n = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1 A$$

is upper-triangular. Also note that if we let $M = E_{n-1} P_{n-1} \cdots E_2 P_2 E_1 P_1$, then $\det(M) = \pm 1$, and

$$\det(A) = \pm \det(A_n).$$

The matrices $P(i,k)$ and $E_{i,j;\beta}$ are called *elementary matrices*. We can summarize the above discussion in the following theorem:

Theorem 7.1. *(Gaussian elimination) Let A be an $n \times n$ matrix (invertible or not). Then there is some invertible matrix M so that $U = MA$ is upper-triangular. The pivots are all nonzero iff A is invertible.*

Proof. We already proved the theorem when A is invertible, as well as the last assertion. Now A is singular iff some pivot is zero, say at Stage k of the elimination. If so, we must have $a_{i\,k}^{(k)} = 0$ for $i = k, \ldots, n$; but in this case, $A_{k+1} = A_k$ and we may pick $P_k = E_k = I$. □

Remark: Obviously, the matrix M can be computed as

$$M = E_{n-1}P_{n-1} \cdots E_2P_2E_1P_1,$$

but this expression is of no use. Indeed, what we need is M^{-1}; when no permutations are needed, it turns out that M^{-1} can be obtained immediately from the matrices E_k's, in fact, from their inverses, and no multiplications are necessary.

Remark: Instead of looking for an invertible matrix M so that MA is upper-triangular, we can look for an invertible matrix M so that MA is a diagonal matrix. Only a simple change to Gaussian elimination is needed. At every Stage k, after the pivot has been found and pivoting been performed, if necessary, in addition to adding suitable multiples of the kth row to the rows *below* row k in order to zero the entries in column k for $i = k+1, \ldots, n$, also add suitable multiples of the kth row to the rows *above* row k in order to zero the entries in column k for $i = 1, \ldots, k-1$. Such steps are also achieved by multiplying on the left by elementary matrices $E_{i,k;\beta_{i,k}}$, except that $i < k$, so that these matrices are not lower-triangular matrices. Nevertheless, at the end of the process, we find that $A_n = MA$, is a diagonal matrix.

This method is called the *Gauss-Jordan factorization*. Because it is more expensive than Gaussian elimination, this method is not used much in practice. However, Gauss-Jordan factorization can be used to compute the inverse of a matrix A. Indeed, we find the jth column of A^{-1} by solving the system $Ax^{(j)} = e_j$ (where e_j is the jth canonical basis vector of $\mathbb{R}^n$). By applying Gauss-Jordan, we are led to a system of the form $D_jx^{(j)} = M_je_j$, where D_j is a diagonal matrix, and we can immediately compute $x^{(j)}$.

It remains to discuss the choice of the pivot, and also conditions that guarantee that no permutations are needed during the Gaussian elimination process. We begin by stating a necessary and sufficient condition for an invertible matrix to have an LU-factorization (*i.e.*, Gaussian elimination does not require pivoting).

7.4 LU-Factorization

Definition 7.1. We say that an invertible matrix A has an *LU-factorization* if it can be written as $A = LU$, where U is upper-triangular invertible and L is lower-triangular, with $L_{i\,i} = 1$ for $i = 1, \ldots, n$.

A lower-triangular matrix with diagonal entries equal to 1 is called a *unit lower-triangular* matrix. Given an $n \times n$ matrix $A = (a_{ij})$, for any k with $1 \leq k \leq n$, let $A(1:k, 1:k)$ denote the submatrix of A whose entries are a_{ij}, where $1 \leq i, j \leq k$.[1] For example, if A is the 5×5 matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{pmatrix},$$

then

$$A(1:3, 1:3) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}.$$

Proposition 7.2. *Let A be an invertible $n \times n$-matrix. Then A has an LU-factorization $A = LU$ iff every matrix $A(1:k, 1:k)$ is invertible for $k = 1, \ldots, n$. Furthermore, when A has an LU-factorization, we have*

$$\det(A(1:k, 1:k)) = \pi_1 \pi_2 \cdots \pi_k, \quad k = 1, \ldots, n,$$

where π_k is the pivot obtained after $k-1$ elimination steps. Therefore, the kth pivot is given by

$$\pi_k = \begin{cases} a_{11} = \det(A(1:1, 1:1)) & \text{if } k = 1 \\ \dfrac{\det(A(1:k, 1:k))}{\det(A(1:k-1, 1:k-1))} & \text{if } k = 2, \ldots, n. \end{cases}$$

Proof. First assume that $A = LU$ is an LU-factorization of A. We can write

$$A = \begin{pmatrix} A(1:k, 1:k) & A_2 \\ A_3 & A_4 \end{pmatrix} = \begin{pmatrix} L_1 & 0 \\ L_3 & L_4 \end{pmatrix} \begin{pmatrix} U_1 & U_2 \\ 0 & U_4 \end{pmatrix} = \begin{pmatrix} L_1 U_1 & L_1 U_2 \\ L_3 U_1 & L_3 U_2 + L_4 U_4 \end{pmatrix},$$

where L_1, L_4 are unit lower-triangular and U_1, U_4 are upper-triangular. (Note, $A(1:k, 1:k)$, L_1, and U_1 are $k \times k$ matrices; A_2 and U_2 are $k \times (n-k)$ matrices; A_3 and L_3 are $(n-k) \times k$ matrices; A_4, L_4, and U_4 are $(n-k) \times (n-k)$ matrices.) Thus,

$$A(1:k, 1:k) = L_1 U_1,$$

and since U is invertible, U_1 is also invertible (the determinant of U is the product of the diagonal entries in U, which is the product of the diagonal entries in U_1 and U_4). As L_1 is invertible (since its diagonal entries are equal to 1), we see that $A(1:k, 1:k)$ is invertible for $k = 1, \ldots, n$.

Conversely, assume that $A(1:k, 1:k)$ is invertible for $k = 1, \ldots, n$. We just need to show that Gaussian elimination does not need pivoting. We prove by induction on k that the kth step does not need pivoting.

[1] We are using `Matlab`'s notation.

This holds for $k = 1$, since $A(1:1,1:1) = (a_{11})$, so $a_{11} \neq 0$. Assume that no pivoting was necessary for the first $k-1$ steps ($2 \leq k \leq n-1$). In this case, we have

$$E_{k-1} \cdots E_2 E_1 A = A_k,$$

where $L = E_{k-1} \cdots E_2 E_1$ is a unit lower-triangular matrix and $A_k(1:k,1:k)$ is upper-triangular, so that $LA = A_k$ can be written as

$$\begin{pmatrix} L_1 & 0 \\ L_3 & L_4 \end{pmatrix} \begin{pmatrix} A(1:k,1:k) & A_2 \\ A_3 & A_4 \end{pmatrix} = \begin{pmatrix} U_1 & B_2 \\ 0 & B_4 \end{pmatrix},$$

where L_1 is unit lower-triangular and U_1 is upper-triangular. (Once again $A(1:k,1:k)$, L_1, and U_1 are $k \times k$ matrices; A_2 and B_2 are $k \times (n-k)$ matrices; A_3 and L_3 are $(n-k) \times k$ matrices; A_4, L_4, and B_4 are $(n-k) \times (n-k)$ matrices.) But then,

$$L_1 A(1:k,1:k)) = U_1,$$

where L_1 is invertible (in fact, $\det(L_1) = 1$), and since by hypothesis $A(1:k,1:k)$ is invertible, U_1 is also invertible, which implies that $(U_1)_{kk} \neq 0$, since U_1 is upper-triangular. Therefore, no pivoting is needed in Step k, establishing the induction step. Since $\det(L_1) = 1$, we also have

$$\begin{aligned} \det(U_1) = \det(L_1 A(1:k,1:k)) &= \det(L_1)\det(A(1:k,1:k)) \\ &= \det(A(1:k,1:k)), \end{aligned}$$

and since U_1 is upper-triangular and has the pivots $\pi_1, \ldots, \pi_k$ on its diagonal, we get

$$\det(A(1:k,1:k)) = \pi_1 \pi_2 \cdots \pi_k, \quad k = 1, \ldots, n,$$

as claimed. □

Remark: The use of determinants in the first part of the proof of Proposition 7.2 can be avoided if we use the fact that a triangular matrix is invertible iff all its diagonal entries are nonzero.

Corollary 7.3. *(LU-Factorization) Let A be an invertible $n \times n$-matrix. If every matrix $A(1:k,1:k)$ is invertible for $k = 1, \ldots, n$, then Gaussian elimination requires no pivoting and yields an LU-factorization $A = LU$.*

Proof. We proved in Proposition 7.2 that in this case Gaussian elimination requires no pivoting. Then since every elementary matrix $E_{i,k;\beta}$ is lower-triangular (since we always arrange that the pivot π_k occurs above the rows that it operates on), since $E_{i,k;\beta}^{-1} = E_{i,k;-\beta}$ and the $E_k s$ are products of $E_{i,k;\beta_{i,k}} s$, from

$$E_{n-1} \cdots E_2 E_1 A = U,$$

where U is an upper-triangular matrix, we get

$$A = LU,$$

where $L = E_1^{-1}E_2^{-1}\cdots E_{n-1}^{-1}$ is a lower-triangular matrix. Furthermore, as the diagonal entries of each $E_{i,k;\beta}$ are 1, the diagonal entries of each E_k are also 1. □

Example 7.1. The reader should verify that

$$\begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{pmatrix}$$

is an LU-factorization.

One of the main reasons why the existence of an LU-factorization for a matrix A is interesting is that if we need to solve *several* linear systems $Ax = b$ corresponding to the same matrix A, we can do this cheaply by solving the two triangular systems

$$Lw = b, \quad \text{and} \quad Ux = w.$$

There is a certain asymmetry in the LU-decomposition $A = LU$ of an invertible matrix A. Indeed, the diagonal entries of L are all 1, but this is generally false for U. This asymmetry can be eliminated as follows: if

$$D = \mathrm{diag}(u_{11}, u_{22}, \ldots, u_{nn})$$

is the diagonal matrix consisting of the diagonal entries in U (the pivots), then we if let $U' = D^{-1}U$, we can write

$$A = LDU',$$

where L is lower- triangular, U' is upper-triangular, all diagonal entries of both L and U' are 1, and D is a diagonal matrix of pivots. Such a decomposition leads to the following definition.

Definition 7.2. We say that an invertible $n \times n$ matrix A has an LDU-factorization if it can be written as $A = LDU'$, where L is lower- triangular, U' is upper-triangular, all diagonal entries of both L and U' are 1, and D is a diagonal matrix.

We will see shortly than if A is real symmetric, then $U' = L^\top$.

As we will see a bit later, real symmetric positive definite matrices satisfy the condition of Proposition 7.2. *Therefore, linear systems involving real symmetric positive definite matrices can be solved by Gaussian elimination without pivoting.* Actually, it is possible to do better: this is the Cholesky factorization.

If a square invertible matrix A has an LU-factorization, then it is possible to find L and U while performing Gaussian elimination. Recall that at Step k, we pick a pivot $\pi_k = a_{ik}^{(k)} \neq 0$

in the portion consisting of the entries of index $j \geq k$ of the k-th column of the matrix A_k obtained so far, we swap rows i and k if necessary (the pivoting step), and then we zero the entries of index $j = k+1, \ldots, n$ in column k. Schematically, we have the following steps:

$$\begin{pmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{pivot}}{\Longrightarrow} \begin{pmatrix} \times & \times & \times & \times & \times \\ 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{elim}}{\Longrightarrow} \begin{pmatrix} \times & \times & \times & \times & \times \\ 0 & \times & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times & \times \\ 0 & \mathbf{0} & \times & \times & \times \end{pmatrix}.$$

More precisely, after permuting row k and row i (the pivoting step), if the entries in column k below row k are $\alpha_{k+1k}, \ldots, \alpha_{nk}$, then we add $-\alpha_{jk}/\pi_k$ times row k to row j; this process is illustrated below:

$$\begin{pmatrix} a_{kk}^{(k)} \\ a_{k+1k}^{(k)} \\ \vdots \\ a_{ik}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{pmatrix} \overset{\text{pivot}}{\Longrightarrow} \begin{pmatrix} a_{ik}^{(k)} \\ a_{k+1k}^{(k)} \\ \vdots \\ a_{kk}^{(k)} \\ \vdots \\ a_{nk}^{(k)} \end{pmatrix} = \begin{pmatrix} \pi_k \\ \alpha_{k+1k} \\ \vdots \\ \alpha_{ik} \\ \vdots \\ \alpha_{nk} \end{pmatrix} \overset{\text{elim}}{\Longrightarrow} \begin{pmatrix} \pi_k \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{pmatrix}.$$

Then if we write $\ell_{jk} = \alpha_{jk}/\pi_k$ for $j = k+1, \ldots, n$, the kth column of L is

$$\begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \ell_{k+1k} \\ \vdots \\ \ell_{nk} \end{pmatrix}.$$

Observe that the signs of the multipliers $-\alpha_{jk}/\pi_k$ have been flipped. Thus, we obtain the unit lower triangular matrix

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{pmatrix}.$$

It is easy to see (and this is proven in Theorem 7.5) that the inverse of L is obtained from

L by flipping the signs of the ℓ_{ij}:

$$L^{-1} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\ell_{21} & 1 & 0 & \cdots & 0 \\ -\ell_{31} & -\ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ -\ell_{n1} & -\ell_{n2} & -\ell_{n3} & \cdots & 1 \end{pmatrix}.$$

Furthermore, if the result of Gaussian elimination (without pivoting) is $U = E_{n-1} \cdots E_1 A$, then

$$E_k = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -\ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\ell_{nk} & 0 & \cdots & 1 \end{pmatrix} \quad \text{and} \quad E_k^{-1} = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nk} & 0 & \cdots & 1 \end{pmatrix},$$

so the kth column of E_k is the kth column of L^{-1}.

Here is an example illustrating the method.

Example 7.2. Given

$$A = A_1 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix},$$

we have the following sequence of steps: The first pivot is $\pi_1 = 1$ in row 1, and we substract row 1 from rows 2, 3, and 4. We get

$$A_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & -2 & -1 & -1 \end{pmatrix} \quad L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_2 = -2$ in row 2, and we subtract row 2 from row 4 (and add 0 times row 2 to row 3). We get

$$A_3 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_3 = -2$ in row 3, and since the fourth entry in column 3 is already a zero, we add 0 times row 3 to row 4. We get

$$A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

The procedure is finished, and we have

$$L = L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \qquad U = A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix}.$$

It is easy to check that indeed

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

We now show how to extend the above method to deal with pivoting efficiently. This is the $PA = LU$ factorization.

7.5 $PA = LU$ Factorization

The following easy proposition shows that, in principle, A can be premultiplied by some permutation matrix P, so that PA can be converted to upper-triangular form without using any pivoting. Permutations are discussed in some detail in Section 6.1, but for now we just need this definition. For the precise connection between the notion of permutation (as discussed in Section 6.1) and permutation matrices, see Problem 7.16.

Definition 7.3. A *permutation matrix* is a square matrix that has a single 1 in every row and every column and zeros everywhere else.

It is shown in Section 6.1 that every permutation matrix is a product of transposition matrices (the $P(i, k)$s), and that P is invertible with inverse $P^\top$.

Proposition 7.4. *Let A be an invertible $n \times n$-matrix. There is some permutation matrix P so that $(PA)(1 : k, 1 : k)$ is invertible for $k = 1, \ldots, n$.*

Proof. The case $n = 1$ is trivial, and so is the case $n = 2$ (we swap the rows if necessary). If $n \geq 3$, we proceed by induction. Since A is invertible, its columns are linearly independent; in particular, its first $n - 1$ columns are also linearly independent. Delete the last column of

A. Since the remaining $n-1$ columns are linearly independent, there are also $n-1$ linearly independent rows in the corresponding $n \times (n-1)$ matrix. Thus, there is a permutation of these n rows so that the $(n-1) \times (n-1)$ matrix consisting of the first $n-1$ rows is invertible. But then there is a corresponding permutation matrix P_1, so that the first $n-1$ rows and columns of P_1A form an invertible matrix A'. Applying the induction hypothesis to the $(n-1) \times (n-1)$ matrix A', we see that there some permutation matrix P_2 (leaving the nth row fixed), so that $(P_2P_1A)(1:k, 1:k)$ is invertible, for $k = 1, \ldots, n-1$. Since A is invertible in the first place and P_1 and P_2 are invertible, P_1P_2A is also invertible, and we are done. □

Remark: One can also prove Proposition 7.4 using a clever reordering of the Gaussian elimination steps suggested by Trefethen and Bau [68] (Lecture 21). Indeed, we know that if A is invertible, then there are permutation matrices P_i and products of elementary matrices E_i, so that

$$A_n = E_{n-1}P_{n-1} \cdots E_2P_2E_1P_1A,$$

where $U = A_n$ is upper-triangular. For example, when $n = 4$, we have $E_3P_3E_2P_2E_1P_1A = U$. We can define new matrices E_1', E_2', E_3' which are still products of elementary matrices so that we have

$$E_3'E_2'E_1'P_3P_2P_1A = U.$$

Indeed, if we let $E_3' = E_3$, $E_2' = P_3E_2P_3^{-1}$, and $E_1' = P_3P_2E_1P_2^{-1}P_3^{-1}$, we easily verify that each E_k' is a product of elementary matrices and that

$$\begin{aligned} E_3'E_2'E_1'P_3P_2P_1 &= E_3(P_3E_2P_3^{-1})(P_3P_2E_1P_2^{-1}P_3^{-1})P_3P_2P_1 \\ &= E_3P_3E_2P_2E_1P_1. \end{aligned}$$

It can also be proven that E_1', E_2', E_3' are lower triangular (see Theorem 7.5).

In general, we let

$$E_k' = P_{n-1} \cdots P_{k+1}E_kP_{k+1}^{-1} \cdots P_{n-1}^{-1},$$

and we have

$$E_{n-1}' \cdots E_1'P_{n-1} \cdots P_1A = U,$$

where each E_j' is a lower triangular matrix (see Theorem 7.5).

It is remarkable that if pivoting steps are necessary during Gaussian elimination, a very simple modification of the algorithm for finding an LU-factorization yields the matrices L, U, and P, such that $PA = LU$. To describe this new method, since the diagonal entries of L are 1s, it is convenient to write

$$L = I + \Lambda.$$

Then in assembling the matrix Λ while performing Gaussian elimination with pivoting, we make the same transposition on the rows of Λ (really Λ_{k-1}) that we make on the rows of A (really A_k) during a pivoting step involving row k and row i. We also assemble P by starting with the identity matrix and applying to P the same row transpositions that we apply to A and Λ. Here is an example illustrating this method.

Example 7.3. Given

$$A = A_1 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix},$$

we have the following sequence of steps: We initialize $\Lambda_0 = 0$ and $P_0 = I_4$. The first pivot is $\pi_1 = 1$ in row 1, and we subtract row 1 from rows 2, 3, and 4. We get

$$A_2 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & -2 & -1 & -1 \end{pmatrix} \quad \Lambda_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_2 = -2$ in row 3, so we permute row 2 and 3; we also apply this permutation to Λ and P:

$$A'_3 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & -2 & -1 & -1 \end{pmatrix} \quad \Lambda'_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next we subtract row 2 from row 4 (and add 0 times row 2 to row 3). We get

$$A_3 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad \Lambda_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The next pivot is $\pi_3 = -2$ in row 3, and since the fourth entry in column 3 is already a zero, we add 0 times row 3 to row 4. We get

$$A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} \quad \Lambda_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The procedure is finished, and we have

$$L = \Lambda_3 + I = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \quad U = A_4 = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix}$$

$$P = P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

It is easy to check that indeed

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & -2 & -1 & 1 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix}$$

and

$$PA = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & 0 \\ 1 & -1 & 0 & -1 \end{pmatrix}.$$

Using the idea in the remark before the above example, we can prove the theorem below which shows the correctness of the algorithm for computing P, L and U using a simple adaptation of Gaussian elimination.

We are not aware of a detailed proof of Theorem 7.5 in the standard texts. Although Golub and Van Loan [30] state a version of this theorem as their Theorem 3.1.4, they say that "The proof is a messy subscripting argument." Meyer [48] also provides a sketch of proof (see the end of Section 3.10). In view of this situation, we offer a complete proof. It does involve a lot of subscripts and superscripts, but in our opinion, it contains some techniques that go far beyond symbol manipulation.

Theorem 7.5. *For every invertible $n \times n$-matrix A, the following hold:*

(1) There is some permutation matrix P, some upper-triangular matrix U, and some unit lower-triangular matrix L, so that $PA = LU$ (recall, $L_{ii} = 1$ for $i = 1, \ldots, n$). Furthermore, if $P = I$, then L and U are unique and they are produced as a result of Gaussian elimination without pivoting.

(2) If $E_{n-1} \ldots E_1 A = U$ is the result of Gaussian elimination without pivoting, write as usual $A_k = E_{k-1} \ldots E_1 A$ (with $A_k = (a_{ij}^{(k)})$), and let $\ell_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$, with $1 \leq k \leq n-1$ and $k+1 \leq i \leq n$. Then

$$L = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ \ell_{21} & 1 & 0 & \cdots & 0 \\ \ell_{31} & \ell_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \ell_{n1} & \ell_{n2} & \ell_{n3} & \cdots & 1 \end{pmatrix},$$

where the kth column of L is the kth column of E_k^{-1}, for $k = 1, \ldots, n-1$.

(3) If $E_{n-1} P_{n-1} \cdots E_1 P_1 A = U$ is the result of Gaussian elimination with some pivoting, write $A_k = E_{k-1} P_{k-1} \cdots E_1 P_1 A$, and define E_j^k, with $1 \leq j \leq n-1$ and $j \leq k \leq n-1$,

such that, for $j = 1, \ldots, n-2$,

$$\begin{aligned} E_j^j &= E_j \\ E_j^k &= P_k E_j^{k-1} P_k, \quad \textit{for } k = j+1, \ldots, n-1, \end{aligned}$$

and

$$E_{n-1}^{n-1} = E_{n-1}.$$

Then,

$$\begin{aligned} E_j^k &= P_k P_{k-1} \cdots P_{j+1} E_j P_{j+1} \cdots P_{k-1} P_k \\ U &= E_{n-1}^{n-1} \cdots E_1^{n-1} P_{n-1} \cdots P_1 A, \end{aligned}$$

and if we set

$$\begin{aligned} P &= P_{n-1} \cdots P_1 \\ L &= (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}, \end{aligned}$$

then

$$PA = LU. \tag{$\dagger_1$}$$

Furthermore,

$$(E_j^k)^{-1} = I + \mathcal{E}_j^k, \quad 1 \leq j \leq n-1,\ j \leq k \leq n-1,$$

where $\mathcal{E}_j^k$ *is a lower triangular matrix of the form*

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix},$$

we have

$$E_j^k = I - \mathcal{E}_j^k,$$

and

$$\mathcal{E}_j^k = P_k \mathcal{E}_j^{k-1}, \quad 1 \leq j \leq n-2,\ j+1 \leq k \leq n-1,$$

where $P_k = I$ *or else* $P_k = P(k,i)$ *for some* i *such that* $k+1 \leq i \leq n$*; if* $P_k \neq I$*, this means that* $(E_j^k)^{-1}$ *is obtained from* $(E_j^{k-1})^{-1}$ *by permuting the entries on rows* i *and* k *in column* j*. Because the matrices* $(E_j^k)^{-1}$ *are all lower triangular, the matrix* L *is also lower triangular.*

In order to find L, define lower triangular $n \times n$ matrices Λ_k of the form

$$\Lambda_k = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ \lambda_{21}^{(k)} & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ \lambda_{31}^{(k)} & \lambda_{32}^{(k)} & \ddots & 0 & 0 & \vdots & \vdots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \vdots & \vdots & \vdots \\ \lambda_{k+11}^{(k)} & \lambda_{k+12}^{(k)} & \cdots & \lambda_{k+1k}^{(k)} & 0 & \cdots & \cdots & 0 \\ \lambda_{k+21}^{(k)} & \lambda_{k+22}^{(k)} & \cdots & \lambda_{k+2k}^{(k)} & 0 & \ddots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1}^{(k)} & \lambda_{n2}^{(k)} & \cdots & \lambda_{nk}^{(k)} & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

to assemble the columns of L iteratively as follows: let

$$(-\ell_{k+1k}^{(k)}, \ldots, -\ell_{nk}^{(k)})$$

be the last $n-k$ elements of the kth column of E_k, and define Λ_k inductively by setting

$$\Lambda_1 = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ \ell_{21}^{(1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1}^{(1)} & 0 & \cdots & 0 \end{pmatrix},$$

then for $k = 2, \ldots, n-1$, define

$$\Lambda'_k = P_k \Lambda_{k-1}, \tag{$\dagger_2$}$$

and $\Lambda_k = (I + \Lambda'_k)E_k^{-1} - I$, with

$$\Lambda_k = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ \lambda_{21}^{\prime(k-1)} & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ \lambda_{31}^{\prime(k-1)} & \lambda_{32}^{\prime(k-1)} & \ddots & 0 & 0 & \vdots & \vdots & 0 \\ \vdots & \vdots & \ddots & 0 & 0 & \vdots & \vdots & \vdots \\ \lambda_{k1}^{\prime(k-1)} & \lambda_{k2}^{\prime(k-1)} & \cdots & \lambda_{k\,(k-1)}^{\prime(k-1)} & 0 & \cdots & \cdots & 0 \\ \lambda_{k+11}^{\prime(k-1)} & \lambda_{k+12}^{\prime(k-1)} & \cdots & \lambda_{k+1\,(k-1)}^{\prime(k-1)} & \ell_{k+1k}^{(k)} & \ddots & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda_{n1}^{\prime(k-1)} & \lambda_{n2}^{\prime(k-1)} & \cdots & \lambda_{n\,k-1}^{\prime(k-1)} & \ell_{nk}^{(k)} & \cdots & \cdots & 0 \end{pmatrix},$$

with $P_k = I$ or $P_k = P(k,i)$ for some $i > k$. This means that in assembling L, row k and row i of Λ_{k-1} need to be permuted when a pivoting step permuting row k and row i of A_k is required. Then

$$I + \Lambda_k = (E_1^k)^{-1} \cdots (E_k^k)^{-1}$$
$$\Lambda_k = \mathcal{E}_1^k + \cdots + \mathcal{E}_k^k,$$

for $k = 1, \ldots, n-1$, and therefore

$$L = I + \Lambda_{n-1}.$$

The proof of Theorem 7.5, which is very technical, is given in Section 7.6.

We emphasize again that Part (3) of Theorem 7.5 shows the remarkable fact that in assembling the matrix L while performing Gaussian elimination with pivoting, the only change to the algorithm is to make the same transposition on the rows of Λ_{k-1} that we make on the rows of A (really A_k) during a pivoting step involving row k and row i. We can also assemble P by starting with the identity matrix and applying to P the same row transpositions that we apply to A and Λ. Here is an example illustrating this method.

Example 7.4. Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix}.$$

We set $P_0 = I_4$, and we can also set $\Lambda_0 = 0$. The first step is to permute row 1 and row 2, using the pivot 4. We also apply this permutation to P_0:

$$A'_1 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 1 & 2 & -3 & 4 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix} \quad P_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next we subtract 1/4 times row 1 from row 2, 1/2 times row 1 from row 3, and add 3/4 times row 1 to row 4, and start assembling Λ:

$$A_2 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 0 & -6 & 6 \\ 0 & -1 & -4 & 5 \\ 0 & 5 & 10 & -10 \end{pmatrix} \quad \Lambda_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \end{pmatrix} \quad P_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Next we permute row 2 and row 4, using the pivot 5. We also apply this permutation to Λ and P:

$$A'_3 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & -1 & -4 & 5 \\ 0 & 0 & -6 & 6 \end{pmatrix} \quad \Lambda'_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Next we add 1/5 times row 2 to row 3, and update Λ'_2:

$$A_3 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & -6 & 6 \end{pmatrix} \quad \Lambda_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/2 & -1/5 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \end{pmatrix} \quad P_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Next we permute row 3 and row 4, using the pivot -6. We also apply this permutation to Λ and P:

$$A'_4 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & -2 & 3 \end{pmatrix} \quad \Lambda'_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 1/2 & -1/5 & 0 & 0 \end{pmatrix} \quad P_3 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Finally we subtract $1/3$ times row 3 from row 4, and update Λ'_3:

$$A_4 = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \Lambda_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -3/4 & 0 & 0 & 0 \\ 1/4 & 0 & 0 & 0 \\ 1/2 & -1/5 & 1/3 & 0 \end{pmatrix} \quad P_3 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Consequently, adding the identity to Λ_3, we obtain

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

We check that

$$PA = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & -3 & 4 \\ 4 & 8 & 12 & -8 \\ 2 & 3 & 2 & 1 \\ -3 & -1 & 1 & -4 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 12 & -8 \\ -3 & -1 & 1 & -4 \\ 1 & 2 & -3 & 4 \\ 2 & 3 & 2 & 1 \end{pmatrix},$$

and that

$$LU = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -3/4 & 1 & 0 & 0 \\ 1/4 & 0 & 1 & 0 \\ 1/2 & -1/5 & 1/3 & 1 \end{pmatrix} \begin{pmatrix} 4 & 8 & 12 & -8 \\ 0 & 5 & 10 & -10 \\ 0 & 0 & -6 & 6 \\ 0 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 12 & -8 \\ -3 & -1 & 1 & -4 \\ 1 & 2 & -3 & 4 \\ 2 & 3 & 2 & 1 \end{pmatrix} = PA.$$

Note that if one willing to overwrite the lower triangular part of the evolving matrix A, one can store the evolving Λ there, since these entries will eventually be zero anyway! There is also no need to save explicitly the permutation matrix P. One could instead record the permutation steps in an extra column (record the vector $(\pi(1), \ldots, \pi(n))$ corresponding to the permutation π applied to the rows). We let the reader write such a bold and space-efficient version of LU-decomposition!

Remark: In `Matlab` the function `lu` returns the matrices P, L, U involved in the $PA = LU$ factorization using the call $[L, U, P] = \mathrm{lu}(A)$.

As a corollary of Theorem 7.5(1), we can show the following result.

Proposition 7.6. *If an invertible real symmetric matrix A has an LU-decomposition, then A has a factorization of the form*

$$A = LDL^\top,$$

where L is a lower-triangular matrix whose diagonal entries are equal to 1, and where D consists of the pivots. Furthermore, such a decomposition is unique.

Proof. If A has an LU-factorization, then it has an LDU factorization

$$A = LDU,$$

where L is lower-triangular, U is upper-triangular, and the diagonal entries of both L and U are equal to 1. Since A is symmetric, we have

$$LDU = A = A^\top = U^\top DL^\top,$$

with $U^\top$ lower-triangular and $DL^\top$ upper-triangular. By the uniqueness of LU-factorization (Part (1) of Theorem 7.5), we must have $L = U^\top$ (and $DU = DL^\top$), thus $U = L^\top$, as claimed. □

Remark: It can be shown that Gaussian elimination plus back-substitution requires $n^3/3 + O(n^2)$ additions, $n^3/3 + O(n^2)$ multiplications and $n^2/2 + O(n)$ divisions.

7.6 Proof of Theorem 7.5 ⊛

Proof. (1) The only part that has not been proven is the uniqueness part (when $P = I$). Assume that A is invertible and that $A = L_1U_1 = L_2U_2$, with L_1, L_2 unit lower-triangular and U_1, U_2 upper-triangular. Then we have

$$L_2^{-1}L_1 = U_2U_1^{-1}.$$

However, it is obvious that L_2^{-1} is lower-triangular and that U_1^{-1} is upper-triangular, and so $L_2^{-1}L_1$ is lower-triangular and $U_2U_1^{-1}$ is upper-triangular. Since the diagonal entries of L_1 and L_2 are 1, the above equality is only possible if $U_2U_1^{-1} = I$, that is, $U_1 = U_2$, and so $L_1 = L_2$.

(2) When $P = I$, we have $L = E_1^{-1}E_2^{-1}\cdots E_{n-1}^{-1}$, where E_k is the product of $n-k$ elementary matrices of the form $E_{i,k;-\ell_i}$, where $E_{i,k;-\ell_i}$ subtracts ℓ_i times row k from row i, with $\ell_{ik} = a_{ik}^{(k)}/a_{kk}^{(k)}$, $1 \le k \le n-1$, and $k+1 \le i \le n$. Then it is immediately verified that

$$E_k = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -\ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\ell_{nk} & 0 & \cdots & 1 \end{pmatrix},$$

and that

$$E_k^{-1} = \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nk} & 0 & \cdots & 1 \end{pmatrix}.$$

If we define L_k by

$$L_k = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \vdots & 0 \\ \ell_{21} & 1 & 0 & 0 & 0 & \vdots & 0 \\ \ell_{31} & \ell_{32} & \ddots & 0 & 0 & \vdots & 0 \\ \vdots & \vdots & \ddots & 1 & 0 & \vdots & 0 \\ \ell_{k+11} & \ell_{k+12} & \cdots & \ell_{k+1k} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \vdots & 0 \\ \ell_{n1} & \ell_{n2} & \cdots & \ell_{nk} & 0 & \cdots & 1 \end{pmatrix}$$

for $k = 1, \ldots, n-1$, we easily check that $L_1 = E_1^{-1}$, and that

$$L_k = L_{k-1}E_k^{-1}, \quad 2 \leq k \leq n-1,$$

because multiplication on the right by E_k^{-1} adds ℓ_i times column i to column k (of the matrix L_{k-1}) with $i > k$, and column i of L_{k-1} has only the nonzero entry 1 as its ith element. Since

$$L_k = E_1^{-1} \cdots E_k^{-1}, \quad 1 \leq k \leq n-1,$$

we conclude that $L = L_{n-1}$, proving our claim about the shape of L.

(3)

Step 1. Prove ($\dagger_1$).

First we prove by induction on k that

$$A_{k+1} = E_k^k \cdots E_1^k P_k \cdots P_1 A, \quad k = 1, \ldots, n-2.$$

For $k = 1$, we have $A_2 = E_1 P_1 A = E_1^1 P_1 A$, since $E_1^1 = E_1$, so our assertion holds trivially.

Now if $k \geq 2$,

$$A_{k+1} = E_k P_k A_k,$$

and by the induction hypothesis,

$$A_k = E_{k-1}^{k-1} \cdots E_2^{k-1} E_1^{k-1} P_{k-1} \cdots P_1 A.$$

Because P_k is either the identity or a transposition, $P_k^2 = I$, so by inserting occurrences of P_kP_k as indicated below we can write

$$\begin{aligned}
A_{k+1} &= E_kP_kA_k \\
&= E_kP_kE_{k-1}^{k-1}\cdots E_2^{k-1}E_1^{k-1}P_{k-1}\cdots P_1A \\
&= E_kP_kE_{k-1}^{k-1}(P_kP_k)\cdots(P_kP_k)E_2^{k-1}(P_kP_k)E_1^{k-1}(P_kP_k)P_{k-1}\cdots P_1A \\
&= E_k(P_kE_{k-1}^{k-1}P_k)\cdots(P_kE_2^{k-1}P_k)(P_kE_1^{k-1}P_k)P_kP_{k-1}\cdots P_1A.
\end{aligned}$$

Observe that P_k has been "moved" to the right of the elimination steps. However, by definition,

$$\begin{aligned}
E_j^k &= P_kE_j^{k-1}P_k, \quad j = 1,\ldots,k-1 \\
E_k^k &= E_k,
\end{aligned}$$

so we get

$$A_{k+1} = E_k^kE_{k-1}^k\cdots E_2^kE_1^kP_k\cdots P_1A,$$

establishing the induction hypothesis. For $k = n-2$, we get

$$U = A_{n-1} = E_{n-1}^{n-1}\cdots E_1^{n-1}P_{n-1}\cdots P_1A,$$

as claimed, and the factorization $PA = LU$ with

$$\begin{aligned}
P &= P_{n-1}\cdots P_1 \\
L &= (E_1^{n-1})^{-1}\cdots(E_{n-1}^{n-1})^{-1}
\end{aligned}$$

is clear.

Step 2. Prove that the matrices $(E_j^k)^{-1}$ are lower-triangular. To achieve this, we prove that the matrices $\mathcal{E}_j^k$ are strictly lower triangular matrices of a very special form.

Since for $j = 1,\ldots,n-2$, we have $E_j^j = E_j$,

$$E_j^k = P_kE_j^{k-1}P_k, \quad k = j+1,\ldots,n-1,$$

since $E_{n-1}^{n-1} = E_{n-1}$ and $P_k^{-1} = P_k$, we get $(E_j^j)^{-1} = E_j^{-1}$ for $j = 1,\ldots,n-1$, and for $j = 1,\ldots,n-2$, we have

$$(E_j^k)^{-1} = P_k(E_j^{k-1})^{-1}P_k, \quad k = j+1,\ldots,n-1.$$

Since

$$(E_j^{k-1})^{-1} = I + \mathcal{E}_j^{k-1}$$

and $P_k = P(k,i)$ is a transposition or $P_k = I$, so $P_k^2 = I$, and we get

$$\begin{aligned}
(E_j^k)^{-1} = P_k(E_j^{k-1})^{-1}P_k = P_k(I+\mathcal{E}_j^{k-1})P_k = P_k^2 + P_k\,\mathcal{E}_j^{k-1}\,P_k \\
= I + P_k\,\mathcal{E}_j^{k-1}\,P_k.
\end{aligned}$$

Therefore, we have

$$(E_j^k)^{-1} = I + P_k\,\mathcal{E}_j^{k-1}\,P_k, \quad 1 \leq j \leq n-2,\ j+1 \leq k \leq n-1.$$

We prove for $j = 1, \ldots, n-1$, that for $k = j, \ldots, n-1$, each $\mathcal{E}_j^k$ is a lower triangular matrix of the form

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix},$$

and that

$$\mathcal{E}_j^k = P_k\,\mathcal{E}_j^{k-1}, \quad 1 \leq j \leq n-2,\ j+1 \leq k \leq n-1,$$

with $P_k = I$ or $P_k = P(k,i)$ for some i such that $k+1 \leq i \leq n$.

For each j $(1 \leq j \leq n-1)$ we proceed by induction on $k = j, \ldots, n-1$. Since $(E_j^j)^{-1} = E_j^{-1}$ and since E_j^{-1} is of the above form, the base case holds.

For the induction step, we only need to consider the case where $P_k = P(k,i)$ is a transposition, since the case where $P_k = I$ is trivial. We have to figure out what $P_k\,\mathcal{E}_j^{k-1}\,P_k = P(k,i)\,\mathcal{E}_j^{k-1}\,P(k,i)$ is. However, since

$$\mathcal{E}_j^{k-1} = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k-1)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k-1)} & 0 & \cdots & 0 \end{pmatrix},$$

and because $k+1 \leq i \leq n$ and $j \leq k-1$, multiplying $\mathcal{E}_j^{k-1}$ on the right by $P(k,i)$ will permute *columns* i and k, which are columns of zeros, so

$$P(k,i)\,\mathcal{E}_j^{k-1}\,P(k,i) = P(k,i)\,\mathcal{E}_j^{k-1},$$

and thus,

$$(E_j^k)^{-1} = I + P(k,i)\,\mathcal{E}_j^{k-1}.$$

But since

$$(E_j^k)^{-1} = I + \mathcal{E}_j^k,$$

we deduce that

$$\mathcal{E}_j^k = P(k,i)\,\mathcal{E}_j^{k-1}.$$

We also know that multiplying $\mathcal{E}_j^{k-1}$ on the left by $P(k,i)$ will permute *rows* i and k, which shows that $\mathcal{E}_j^k$ has the desired form, as claimed. Since all $\mathcal{E}_j^k$ are strictly lower triangular, all $(E_j^k)^{-1} = I + \mathcal{E}_j^k$ are lower triangular, so the product

$$L = (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}$$

is also lower triangular.

Step 3. Express L as $L = I + \Lambda_{n-1}$, with $\Lambda_{n-1} = \mathcal{E}_1^1 + \cdots + \mathcal{E}_{n-1}^{n-1}$.

From Step 1 of Part (3), we know that

$$L = (E_1^{n-1})^{-1} \cdots (E_{n-1}^{n-1})^{-1}.$$

We prove by induction on k that

$$\begin{aligned} I + \Lambda_k &= (E_1^k)^{-1} \cdots (E_k^k)^{-1} \\ \Lambda_k &= \mathcal{E}_1^k + \cdots + \mathcal{E}_k^k, \end{aligned}$$

for $k = 1, \ldots, n-1$.

If $k = 1$, we have $E_1^1 = E_1$ and

$$E_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -\ell_{21}^{(1)} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -\ell_{n1}^{(1)} & 0 & \cdots & 1 \end{pmatrix}.$$

We also get

$$(E_1^{-1})^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \ell_{21}^{(1)} & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \ell_{n1}^{(1)} & 0 & \cdots & 1 \end{pmatrix} = I + \Lambda_1.$$

Since $(E_1^{-1})^{-1} = I + \mathcal{E}_1^1$, we find that we get $\Lambda_1 = \mathcal{E}_1^1$, and the base step holds.

Since $(E_j^k)^{-1} = I + \mathcal{E}_j^k$ with

$$\mathcal{E}_j^k = \begin{pmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & \ell_{j+1j}^{(k)} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \ell_{nj}^{(k)} & 0 & \cdots & 0 \end{pmatrix}$$

and $\mathcal{E}_i^k \mathcal{E}_j^k = 0$ if $i < j$, as in part (2) for the computation involving the products of L_k's, we get

$$(E_1^{k-1})^{-1} \cdots (E_{k-1}^{k-1})^{-1} = I + \mathcal{E}_1^{k-1} + \cdots + \mathcal{E}_{k-1}^{k-1}, \quad 2 \leq k \leq n. \qquad (*)$$

Similarly, from the fact that $\mathcal{E}_j^{k-1} P(k,i) = \mathcal{E}_j^{k-1}$ if $i \geq k+1$ and $j \leq k-1$ and since

$$(E_j^k)^{-1} = I + P_k \mathcal{E}_j^{k-1}, \quad 1 \leq j \leq n-2,\ j+1 \leq k \leq n-1,$$

we get

$$(E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1} = I + P_k(\mathcal{E}_1^{k-1} + \cdots + \mathcal{E}_{k-1}^{k-1}), \quad 2 \leq k \leq n-1. \tag{$**$}$$

By the induction hypothesis,

$$I + \Lambda_{k-1} = (E_1^{k-1})^{-1} \cdots (E_{k-1}^{k-1})^{-1},$$

and from $(*)$, we get

$$\Lambda_{k-1} = \mathcal{E}_1^{k-1} + \cdots + \mathcal{E}_{k-1}^{k-1}.$$

Using $(**)$, we deduce that

$$(E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1} = I + P_k \Lambda_{k-1}.$$

Since $E_k^k = E_k$, we obtain

$$(E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1}(E_k^k)^{-1} = (I + P_k \Lambda_{k-1}) E_k^{-1}.$$

However, by definition

$$I + \Lambda_k = (I + P_k \Lambda_{k-1}) E_k^{-1},$$

which proves that

$$I + \Lambda_k = (E_1^k)^{-1} \cdots (E_{k-1}^k)^{-1}(E_k^k)^{-1}, \tag{$\dagger$}$$

and finishes the induction step for the proof of this formula.

If we apply Equation $(*)$ again with $k+1$ in place of k, we have

$$(E_1^k)^{-1} \cdots (E_k^k)^{-1} = I + \mathcal{E}_1^k + \cdots + \mathcal{E}_k^k,$$

and together with $(\dagger)$, we obtain,

$$\Lambda_k = \mathcal{E}_1^k + \cdots + \mathcal{E}_k^k,$$

also finishing the induction step for the proof of this formula. For $k = n-1$ in $(\dagger)$, we obtain the desired equation: $L = I + \Lambda_{n-1}$. □

7.7 Dealing with Roundoff Errors; Pivoting Strategies

Let us now briefly comment on the choice of a pivot. Although theoretically, any pivot can be chosen, the possibility of roundoff errors implies that it is not a good idea to pick very small pivots. The following example illustrates this point. Consider the linear system

$$\begin{array}{rcrcl} 10^{-4}x & + & y & = & 1 \\ x & + & y & = & 2. \end{array}$$

Since 10^{-4} is nonzero, it can be taken as pivot, and we get

$$\begin{array}{rcl} 10^{-4}x + y & = & 1 \\ (1-10^4)y & = & 2-10^4. \end{array}$$

Thus, the exact solution is

$$x = \frac{10^4}{10^4-1}, \quad y = \frac{10^4-2}{10^4-1}.$$

However, if roundoff takes place on the fourth digit, then $10^4 - 1 = 9999$ and $10^4 - 2 = 9998$ will be rounded off both to 9990, and then the solution is $x = 0$ and $y = 1$, very far from the exact solution where $x \approx 1$ and $y \approx 1$. The problem is that we picked a very small pivot. If instead we permute the equations, the pivot is 1, and after elimination we get the system

$$\begin{array}{rcl} x + y & = & 2 \\ (1-10^{-4})y & = & 1 - 2 \times 10^{-4}. \end{array}$$

This time, $1 - 10^{-4} = 0.9999$ and $1 - 2 \times 10^{-4} = 0.9998$ are rounded off to 0.999 and the solution is $x = 1, y = 1$, much closer to the exact solution.

To remedy this problem, one may use the strategy of *partial pivoting*. This consists of choosing during Step k ($1 \leq k \leq n-1$) one of the entries $a_{i\,k}^{(k)}$ such that

$$|a_{i\,k}^{(k)}| = \max_{k \leq p \leq n} |a_{p\,k}^{(k)}|.$$

By maximizing the value of the pivot, we avoid dividing by undesirably small pivots.

Remark: A matrix, A, is called *strictly column diagonally dominant* iff

$$|a_{j\,j}| > \sum_{i=1,\, i \neq j}^{n} |a_{i\,j}|, \quad \text{for } j = 1, \ldots, n$$

(resp. *strictly row diagonally dominant* iff

$$|a_{i\,i}| > \sum_{j=1,\, j \neq i}^{n} |a_{i\,j}|, \quad \text{for } i = 1, \ldots, n.)$$

For example, the matrix

$$\begin{pmatrix} \frac{7}{2} & 1 & & & \\ 1 & 4 & 1 & & 0 \\ & \ddots & \ddots & \ddots & \\ 0 & & 1 & 4 & 1 \\ & & & 1 & \frac{7}{2} \end{pmatrix}$$

of the curve interpolation problem discussed in Section 7.1 is strictly column (and row) diagonally dominant.

It has been known for a long time (before 1900, say by Hadamard) that if a matrix A is strictly column diagonally dominant (resp. strictly row diagonally dominant), then it is invertible. It can also be shown that if A is strictly column diagonally dominant, then Gaussian elimination with partial pivoting does not actually require pivoting (see Problem 7.12).

Another strategy, called *complete pivoting*, consists in choosing some entry $a_{ij}^{(k)}$, where $k \leq i, j \leq n$, such that

$$|a_{ij}^{(k)}| = \max_{k \leq p,q \leq n} |a_{pq}^{(k)}|.$$

However, in this method, if the chosen pivot is not in column k, it is also necessary to permute columns. This is achieved by multiplying on the right by a permutation matrix. However, complete pivoting tends to be too expensive in practice, and partial pivoting is the method of choice.

A special case where the LU-factorization is particularly efficient is the case of tridiagonal matrices, which we now consider.

7.8 Gaussian Elimination of Tridiagonal Matrices

Consider the tridiagonal matrix

$$A = \begin{pmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ & \ddots & \ddots & \ddots & & \\ & & a_{n-2} & b_{n-2} & c_{n-2} & \\ & & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & & a_n & b_n \end{pmatrix}.$$

Define the sequence

$$\delta_0 = 1, \quad \delta_1 = b_1, \quad \delta_k = b_k \delta_{k-1} - a_k c_{k-1} \delta_{k-2}, \quad 2 \leq k \leq n.$$

Proposition 7.7. *If A is the tridiagonal matrix above, then $\delta_k = \det(A(1:k, 1:k))$ for $k = 1, \ldots, n$.*

Proof. By expanding $\det(A(1:k, 1:k))$ with respect to its last row, the proposition follows by induction on k. □

Theorem 7.8. *If A is the tridiagonal matrix above and $\delta_k \neq 0$ for $k = 1, \ldots, n$, then A has*

the following LU-factorization:

$$A = \begin{pmatrix} 1 & & & & & & \\ a_2\dfrac{\delta_0}{\delta_1} & 1 & & & & & \\ & a_3\dfrac{\delta_1}{\delta_2} & 1 & & & & \\ & & \ddots & \ddots & & & \\ & & & a_{n-1}\dfrac{\delta_{n-3}}{\delta_{n-2}} & 1 & & \\ & & & & a_n\dfrac{\delta_{n-2}}{\delta_{n-1}} & 1 \end{pmatrix} \begin{pmatrix} \dfrac{\delta_1}{\delta_0} & c_1 & & & & \\ & \dfrac{\delta_2}{\delta_1} & c_2 & & & \\ & & \dfrac{\delta_3}{\delta_2} & c_3 & & \\ & & & \ddots & \ddots & \\ & & & & \dfrac{\delta_{n-1}}{\delta_{n-2}} & c_{n-1} \\ & & & & & \dfrac{\delta_n}{\delta_{n-1}} \end{pmatrix}.$$

Proof. Since $\delta_k = \det(A(1:k, 1:k)) \neq 0$ for $k = 1, \ldots, n$, by Theorem 7.5 (and Proposition 7.2), we know that A has a unique LU-factorization. Therefore, it suffices to check that the proposed factorization works. We easily check that

$$\begin{aligned} (LU)_{k\,k+1} &= c_k, \quad 1 \leq k \leq n-1 \\ (LU)_{k\,k-1} &= a_k, \quad 2 \leq k \leq n \\ (LU)_{k\,l} &= 0, \quad |k-l| \geq 2 \\ (LU)_{1\,1} &= \frac{\delta_1}{\delta_0} = b_1 \\ (LU)_{k\,k} &= \frac{a_k c_{k-1}\delta_{k-2} + \delta_k}{\delta_{k-1}} = b_k, \quad 2 \leq k \leq n, \end{aligned}$$

since $\delta_k = b_k\delta_{k-1} - a_k c_{k-1}\delta_{k-2}$. □

It follows that there is a simple method to solve a linear system $Ax = d$ where A is tridiagonal (and $\delta_k \neq 0$ for $k = 1, \ldots, n$). For this, it is convenient to "squeeze" the diagonal matrix Δ defined such that $\Delta_{k\,k} = \delta_k/\delta_{k-1}$ into the factorization so that $A = (L\Delta)(\Delta^{-1}U)$, and if we let

$$z_1 = \frac{c_1}{b_1}, \quad z_k = c_k\,\frac{\delta_{k-1}}{\delta_k}, \quad 2 \leq k \leq n-1, \quad z_n = \frac{\delta_n}{\delta_{n-1}} = b_n - a_n z_{n-1},$$

$A = (L\Delta)(\Delta^{-1}U)$ is written as

$$A = \begin{pmatrix} \frac{c_1}{z_1} & & & & & \\ a_2 & \frac{c_2}{z_2} & & & & \\ & a_3 & \frac{c_3}{z_3} & & & \\ & & \ddots & \ddots & & \\ & & & a_{n-1} & \frac{c_{n-1}}{z_{n-1}} & \\ & & & & a_n & z_n \end{pmatrix} \begin{pmatrix} 1 & z_1 & & & & \\ & 1 & z_2 & & & \\ & & 1 & z_3 & & \\ & & & \ddots & \ddots & \\ & & & 1 & z_{n-2} & \\ & & & & 1 & z_{n-1} \\ & & & & & 1 \end{pmatrix}.$$

As a consequence, the system $Ax = d$ can be solved by constructing three sequences: First, the sequence

$$z_1 = \frac{c_1}{b_1}, \quad z_k = \frac{c_k}{b_k - a_k z_{k-1}}, \quad k = 2, \ldots, n-1, \quad z_n = b_n - a_n z_{n-1},$$

corresponding to the recurrence $\delta_k = b_k\delta_{k-1} - a_k c_{k-1}\delta_{k-2}$ and obtained by dividing both sides of this equation by δ_{k-1}, next

$$w_1 = \frac{d_1}{b_1}, \quad w_k = \frac{d_k - a_k w_{k-1}}{b_k - a_k z_{k-1}}, \quad k = 2, \ldots, n,$$

corresponding to solving the system $L\Delta w = d$, and finally

$$x_n = w_n, \quad x_k = w_k - z_k x_{k+1}, \quad k = n-1, n-2, \ldots, 1,$$

corresponding to solving the system $\Delta^{-1}Ux = w$.

Remark: It can be verified that this requires $3(n-1)$ additions, $3(n-1)$ multiplications, and $2n$ divisions, a total of $8n-6$ operations, which is much less that the $O(2n^3/3)$ required by Gaussian elimination in general.

We now consider the special case of symmetric positive definite matrices (SPD matrices).

7.9 SPD Matrices and the Cholesky Decomposition

Recall that an $n \times n$ real symmetric matrix A is *positive definite* iff

$$x^\top Ax > 0 \quad \text{for all } x \in \mathbb{R}^n \text{ with } x \neq 0.$$

Equivalently, A is symmetric positive definite iff all its eigenvalues are strictly positive. The following facts about a symmetric positive definite matrix A are easily established (some left as an exercise):

(1) The matrix A is invertible. (Indeed, if $Ax = 0$, then $x^\top Ax = 0$, which implies $x = 0$.)

(2) We have $a_{ii} > 0$ for $i = 1, \ldots, n$. (Just observe that for $x = e_i$, the ith canonical basis vector of $\mathbb{R}^n$, we have $e_i^\top A e_i = a_{ii} > 0$.)

(3) For every $n \times n$ real invertible matrix Z, the matrix $Z^\top AZ$ is real symmetric positive definite iff A is real symmetric positive definite.

(4) The set of $n \times n$ real symmetric positive definite matrices is convex. This means that if A and B are two $n \times n$ symmetric positive definite matrices, then for any $\lambda \in \mathbb{R}$ such that $0 \leq \lambda \leq 1$, the matrix $(1 - \lambda)A + \lambda B$ is also symmetric positive definite. Clearly since A and B are symmetric, $(1 - \lambda)A + \lambda B$ is also symmetric. For any nonzero $x \in \mathbb{R}^n$, we have $x^\top Ax > 0$ and $x^\top Bx > 0$, so

$$x^\top((1 - \lambda)A + \lambda B)x = (1 - \lambda)x^\top Ax + \lambda x^\top Bx > 0,$$

because $0 \leq \lambda \leq 1$, so $1-\lambda \geq 0$ and $\lambda \geq 0$, and $1-\lambda$ and λ can't be zero simultaneously.

(5) The set of $n \times n$ real symmetric positive definite matrices is a cone. This means that if A is symmetric positive definite and if $\lambda > 0$ is any real, then λA is symmetric positive definite. Clearly λA is symmetric, and for nonzero $x \in \mathbb{R}^n$, we have $x^\top Ax > 0$, and since $\lambda > 0$, we have $x^\top \lambda Ax = \lambda x^\top Ax > 0$.

Remark: Given a complex $m \times n$ matrix A, we define the matrix $\overline{A}$ as the $m \times n$ matrix $\overline{A} = (\overline{a_{ij}})$. Then we define A^* as the $n \times m$ matrix $A^* = (\overline{A})^\top = \overline{(A^\top)}$. The $n \times n$ complex matrix A is *Hermitian* if $A^* = A$. This is the complex analog of the notion of a real symmetric matrix. A Hermitian matrix A is *positive definite* if

$$z^* Az > 0 \quad \text{for all } z \in \mathbb{C}^n \text{ with } z \neq 0.$$

It is easily verified that Properties (1)-(5) hold for Hermitian positive definite matrices; replace $\top$ by $*$.

It is instructive to characterize when a 2×2 real symmetric matrix A is positive definite. Write

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

Then we have

$$\begin{pmatrix} x & y \end{pmatrix} \begin{pmatrix} a & c \\ c & b \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = ax^2 + 2cxy + by^2.$$

If the above expression is strictly positive for all nonzero vectors $\binom{x}{y}$, then for $x = 1, y = 0$ we get $a > 0$ and for $x = 0, y = 1$ we get $b > 0$. Then we can write

$$\begin{aligned} ax^2 + 2cxy + by^2 &= \left(\sqrt{a}x + \frac{c}{\sqrt{a}}y\right)^2 + by^2 - \frac{c^2}{a}y^2 \\ &= \left(\sqrt{a}x + \frac{c}{\sqrt{a}}y\right)^2 + \frac{1}{a}\left(ab - c^2\right)y^2. \end{aligned} \tag{$\dagger$}$$

Since $a > 0$, if $ab - c^2 \leq 0$, then we can choose $y > 0$ so that the second term is negative or zero, and we can set $x = -(c/a)y$ to make the first term zero, in which case $ax^2 + 2cxy + by^2 \leq 0$, so we must have $ab - c^2 > 0$.

Conversely, if $a > 0, b > 0$ and $ab > c^2$, then for any $(x, y) \neq (0, 0)$, if $y = 0$, then $x \neq 0$ and the first term of (†) is positive, and if $y \neq 0$, then the second term of (†) is positive. Therefore, the symmetric matrix A is positive definite iff

$$a > 0,\ b > 0,\ ab > c^2. \tag{$*$}$$

Note that $ab - c^2 = \det(A)$, so the third condition says that $\det(A) > 0$.

Observe that the condition $b > 0$ is redundant, since if $a > 0$ and $ab > c^2$, then we must have $b > 0$ (and similarly $b > 0$ and $ab > c^2$ implies that $a > 0$).

We can try to visualize the space of 2×2 real symmetric positive definite matrices in $\mathbb{R}^3$, by viewing (a, b, c) as the coordinates along the x, y, z axes. Then the locus determined by the strict inequalities in $(*)$ corresponds to the region on the side of the cone of equation $xy = z^2$ that does not contain the origin and for which $x > 0$ and $y > 0$. For $z = \delta$ fixed, the equation $xy = \delta^2$ define a hyperbola in the plane $z = \delta$. The cone of equation $xy = z^2$ consists of the lines through the origin that touch the hyperbola $xy = 1$ in the plane $z = 1$. We only consider the branch of this hyperbola for which $x > 0$ and $y > 0$. See Figure 7.6.

It is not hard to show that the inverse of a real symmetric positive definite matrix is also real symmetric positive definite, but the product of two real symmetric positive definite matrices may *not* be symmetric positive definite, as the following example shows:

$$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1/\sqrt{2} & -1\sqrt{2} \\ -1/\sqrt{2} & 3/\sqrt{2} \end{pmatrix} = \begin{pmatrix} 0 & 2/\sqrt{2} \\ -1/\sqrt{2} & 5/\sqrt{2} \end{pmatrix}.$$

According to the above criterion, the two matrices on the left-hand side are real symmetric positive definite, but the matrix on the right-hand side is not even symmetric, and

$$\begin{pmatrix} -6 & 1 \end{pmatrix} \begin{pmatrix} 0 & 2/\sqrt{2} \\ -1/\sqrt{2} & 5/\sqrt{2} \end{pmatrix} \begin{pmatrix} -6 \\ 1 \end{pmatrix} = \begin{pmatrix} -6 & 1 \end{pmatrix} \begin{pmatrix} 2/\sqrt{2} \\ 11/\sqrt{2} \end{pmatrix} = -1/\sqrt{5},$$

even though its eigenvalues are both real and positive.

Next we show that a real symmetric positive definite matrix has a special LU-factorization of the form $A = BB^\top$, where B is a lower-triangular matrix whose diagonal elements are strictly positive. This is the *Cholesky factorization*.

First we note that a symmetric positive definite matrix satisfies the condition of Proposition 7.2.

Proposition 7.9. *If A is a real symmetric positive definite matrix, then $A(1 : k, 1 : k)$ is symmetric positive definite and thus invertible for $k = 1, \ldots, n$.*

Proof. Since A is symmetric, each $A(1 : k, 1 : k)$ is also symmetric. If $w \in \mathbb{R}^k$, with $1 \leq k \leq n$, we let $x \in \mathbb{R}^n$ be the vector with $x_i = w_i$ for $i = 1, \ldots, k$ and $x_i = 0$ for

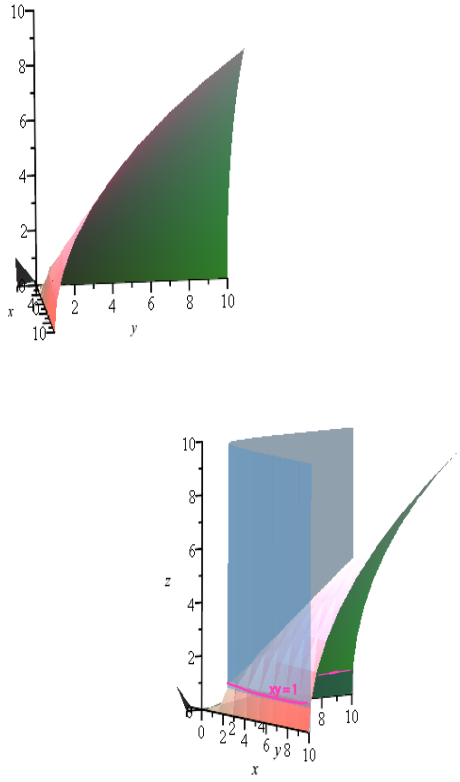

Figure 7.6: Two views of the surface $xy = z^2$ in $\mathbb{R}^3$. The intersection of the surface with a constant z plane results in a hyperbola. The region associated with the 2×2 symmetric positive definite matrices lies in "front" of the green side.

$i = k+1, \ldots, n$. Now since A is symmetric positive definite, we have $x^\top Ax > 0$ for all $x \in \mathbb{R}^n$ with $x \neq 0$. This holds in particular for all vectors x obtained from nonzero vectors $w \in \mathbb{R}^k$ as defined earlier, and clearly

$$x^\top Ax = w^\top A(1:k, 1:k)\, w,$$

which implies that $A(1:k, 1:k)$ is positive definite. Thus, by Fact 1 above, $A(1:k, 1:k)$ is also invertible. □

Proposition 7.9 also holds for a complex Hermitian positive definite matrix. Proposition 7.9 can be strengthened as follows: *A real symmetric (or complex Hermitian) matrix A is positive definite iff* $\det(A(1:k, 1:k)) > 0$ *for* $k = 1, \ldots, n$.

The above fact is known as *Sylvester's criterion*. We will prove it after establishing the Cholesky factorization.

Let A be an $n \times n$ real symmetric positive definite matrix and write

$$A = \begin{pmatrix} a_{11} & W^\top \\ W & C \end{pmatrix},$$

where C is an $(n-1)\times(n-1)$ symmetric matrix and W is an $(n-1)\times 1$ matrix. Since A is symmetric positive definite, $a_{11}>0$, and we can compute $\alpha=\sqrt{a_{11}}$. The trick is that we can factor A uniquely as

$$A=\begin{pmatrix} a_{11} & W^\top \\ W & C\end{pmatrix}=\begin{pmatrix} \alpha & 0\\ W/\alpha & I\end{pmatrix}\begin{pmatrix} 1 & 0\\ 0 & C-WW^\top/a_{11}\end{pmatrix}\begin{pmatrix} \alpha & W^\top/\alpha\\ 0 & I\end{pmatrix},$$

i.e., as $A=B_1A_1B_1^\top$, where B_1 is lower-triangular with positive diagonal entries. Thus, B_1 is invertible, and by Fact (3) above, A_1 is also symmetric positive definite.

Remark: The matrix $C-WW^\top/a_{11}$ is known as the *Schur complement* of the matrix (a_{11}).

Theorem 7.10. *(Cholesky factorization) Let A be a real symmetric positive definite matrix. Then there is some real lower-triangular matrix B so that $A=BB^\top$. Furthermore, B can be chosen so that its diagonal elements are strictly positive, in which case B is unique.*

Proof. We proceed by induction on the dimension n of A. For $n=1$, we must have $a_{11}>0$, and if we let $\alpha=\sqrt{a_{11}}$ and $B=(\alpha)$, the theorem holds trivially. If $n\geq 2$, as we explained above, again we must have $a_{11}>0$, and we can write

$$A=\begin{pmatrix} a_{11} & W^\top \\ W & C\end{pmatrix}=\begin{pmatrix} \alpha & 0\\ W/\alpha & I\end{pmatrix}\begin{pmatrix} 1 & 0\\ 0 & C-WW^\top/a_{11}\end{pmatrix}\begin{pmatrix} \alpha & W^\top/\alpha\\ 0 & I\end{pmatrix}=B_1A_1B_1^\top,$$

where $\alpha=\sqrt{a_{11}}$, the matrix B_1 is invertible and

$$A_1=\begin{pmatrix} 1 & 0\\ 0 & C-WW^\top/a_{11}\end{pmatrix}$$

is symmetric positive definite. However, this implies that $C-WW^\top/a_{11}$ is also symmetric positive definite (consider $x^\top A_1x$ for every $x\in\mathbb{R}^n$ with $x\neq 0$ and $x_1=0$). Thus, we can apply the induction hypothesis to $C-WW^\top/a_{11}$ (which is an $(n-1)\times(n-1)$ matrix), and we find a unique lower-triangular matrix L with positive diagonal entries so that

$$C-WW^\top/a_{11}=LL^\top.$$

But then we get

$$\begin{aligned} A &= \begin{pmatrix} \alpha & 0\\ W/\alpha & I\end{pmatrix}\begin{pmatrix} 1 & 0\\ 0 & C-WW^\top/a_{11}\end{pmatrix}\begin{pmatrix} \alpha & W^\top/\alpha\\ 0 & I\end{pmatrix}\\ &= \begin{pmatrix} \alpha & 0\\ W/\alpha & I\end{pmatrix}\begin{pmatrix} 1 & 0\\ 0 & LL^\top\end{pmatrix}\begin{pmatrix} \alpha & W^\top/\alpha\\ 0 & I\end{pmatrix}\\ &= \begin{pmatrix} \alpha & 0\\ W/\alpha & I\end{pmatrix}\begin{pmatrix} 1 & 0\\ 0 & L\end{pmatrix}\begin{pmatrix} 1 & 0\\ 0 & L^\top\end{pmatrix}\begin{pmatrix} \alpha & W^\top/\alpha\\ 0 & I\end{pmatrix}\\ &= \begin{pmatrix} \alpha & 0\\ W/\alpha & L\end{pmatrix}\begin{pmatrix} \alpha & W^\top/\alpha\\ 0 & L^\top\end{pmatrix}. \end{aligned}$$

Therefore, if we let

$$B = \begin{pmatrix} \alpha & 0 \\ W/\alpha & L \end{pmatrix},$$

we have a unique lower-triangular matrix with positive diagonal entries and $A = BB^\top$. $\square$

Remark: The uniqueness of the Cholesky decomposition can also be established using the uniqueness of an LU-decomposition. Indeed, if $A = B_1B_1^\top = B_2B_2^\top$ where B_1 and B_2 are lower triangular with positive diagonal entries, if we let Δ_1 (resp. Δ_2) be the diagonal matrix consisting of the diagonal entries of B_1 (resp. B_2) so that $(\Delta_k)_{ii} = (B_k)_{ii}$ for $k = 1, 2$, then we have two LU-decompositions

$$A = (B_1\Delta_1^{-1})(\Delta_1 B_1^\top) = (B_2\Delta_2^{-1})(\Delta_2 B_2^\top)$$

with $B_1\Delta_1^{-1}, B_2\Delta_2^{-1}$ unit lower triangular, and $\Delta_1 B_1^\top, \Delta_2 B_2^\top$ upper triangular. By uniquenes of LU-factorization (Theorem 7.5(1)), we have

$$B_1\Delta_1^{-1} = B_2\Delta_2^{-1}, \quad \Delta_1 B_1^\top = \Delta_2 B_2^\top,$$

and the second equation yields

$$B_1\Delta_1 = B_2\Delta_2. \tag{$*$}$$

The diagonal entries of $B_1\Delta_1$ are $(B_1)_{ii}^2$ and similarly the diagonal entries of $B_2\Delta_2$ are $(B_2)_{ii}^2$, so the above equation implies that

$$(B_1)_{ii}^2 = (B_2)_{ii}^2, \quad i = 1, \ldots, n.$$

Since the diagonal entries of both B_1 and B_2 are assumed to be positive, we must have

$$(B_1)_{ii} = (B_2)_{ii}, \quad i = 1, \ldots, n;$$

that is, $\Delta_1 = \Delta_2$, and since both are invertible, we conclude from $(*)$ that $B_1 = B_2$.

Theorem 7.10 also holds for complex Hermitian positive definite matrices. In this case, we have $A = BB^*$ for some unique lower triangular matrix B with positive diagonal entries.

The proof of Theorem 7.10 immediately yields an algorithm to compute B from A by solving for a lower triangular matrix B such that $A = BB^\top$ (where both A and B are real matrices). For $j = 1, \ldots, n$,

$$b_{j\,j} = \left(a_{j\,j} - \sum_{k=1}^{j-1} b_{j\,k}^2\right)^{1/2},$$

and for $i = j+1, \ldots, n$ (and $j = 1, \ldots, n-1$)

$$b_{i\,j} = \left(a_{i\,j} - \sum_{k=1}^{j-1} b_{i\,k}b_{j\,k}\right)/b_{j\,j}.$$

The above formulae are used to compute the jth column of B from top-down, using the first $j-1$ columns of B previously computed, and the matrix A. In the case of $n = 3$, $A = BB^\top$ yields

$$\begin{pmatrix} a_{11} & a_{12} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} b_{11} & 0 & 0 \\ b_{21} & b_{22} & 0 \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \begin{pmatrix} b_{11} & b_{21} & b_{31} \\ 0 & b_{22} & b_{32} \\ 0 & 0 & b_{33} \end{pmatrix}$$
$$= \begin{pmatrix} b_{11}^2 & b_{11}b_{21} & b_{11}b_{31} \\ b_{11}b_{21} & b_{21}^2 + b_{22}^2 & b_{21}b_{31} + b_{22}b_{32} \\ b_{11}b_{31} & b_{21}b_{31} + b_{22}b_{32} & b_{31}^2 + b_{32}^2 + b_{33}^2 \end{pmatrix}.$$

We work down the first column of A, compare entries, and discover that

$$\begin{aligned} a_{11} &= b_{11}^2 & b_{11} &= \sqrt{a_{11}} \\ a_{21} &= b_{11}b_{21} & b_{21} &= \frac{a_{21}}{b_{11}} \\ a_{31} &= b_{11}b_{31} & b_{31} &= \frac{a_{31}}{b_{11}}. \end{aligned}$$

Next we work down the second column of A using previously calculated expressions for b_{21} and b_{31} to find that

$$\begin{aligned} a_{22} &= b_{21}^2 + b_{22}^2 & b_{22} &= \left(a_{22} - b_{21}^2\right)^{\frac{1}{2}} \\ a_{32} &= b_{21}b_{31} + b_{22}b_{32} & b_{32} &= \frac{a_{32} - b_{21}b_{31}}{b_{22}}. \end{aligned}$$

Finally, we use the third column of A and the previously calculated expressions for b_{31} and b_{32} to determine b_{33} as

$$a_{33} = b_{31}^2 + b_{32}^2 + b_{33}^2 \qquad b_{33} = \left(a_{33} - b_{31}^2 - b_{32}^2\right)^{\frac{1}{2}}.$$

For another example, if

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

we find that

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

We leave it as an exercise to find similar formulae (involving conjugation) to factor a complex Hermitian positive definite matrix A as $A = BB^*$. The following `Matlab` program implements the Cholesky factorization.

```
function B = Cholesky(A)
n = size(A,1);
B = zeros(n,n);
for j = 1:n-1;
    if j == 1
       B(1,1) = sqrt(A(1,1));
       for i = 2:n
           B(i,1) = A(i,1)/B(1,1);
       end
    else
       B(j,j) = sqrt(A(j,j) - B(j,1:j-1)*B(j,1:j-1)');
       for i = j+1:n
           B(i,j) = (A(i,j) - B(i,1:j-1)*B(j,1:j-1)')/B(j,j);
       end
    end
end
B(n,n) = sqrt(A(n,n) - B(n,1:n-1)*B(n,1:n-1)');
end
```

If we run the above algorithm on the following matrix

$$A = \begin{pmatrix} 4 & 1 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 1 & 4 \end{pmatrix},$$

we obtain

$$B = \begin{pmatrix} 2.0000 & 0 & 0 & 0 & 0 \\ 0.5000 & 1.9365 & 0 & 0 & 0 \\ 0 & 0.5164 & 1.9322 & 0 & 0 \\ 0 & 0 & 0.5175 & 1.9319 & 0 \\ 0 & 0 & 0 & 0.5176 & 1.9319 \end{pmatrix}.$$

The Cholesky factorization can be used to solve linear systems $Ax = b$ where A is symmetric positive definite: Solve the two systems $Bw = b$ and $B^\top x = w$.

Remark: It can be shown that this methods requires $n^3/6 + O(n^2)$ additions, $n^3/6 + O(n^2)$ multiplications, $n^2/2 + O(n)$ divisions, and $O(n)$ square root extractions. Thus, the Cholesky method requires half of the number of operations required by Gaussian elimination (since Gaussian elimination requires $n^3/3 + O(n^2)$ additions, $n^3/3 + O(n^2)$ multiplications, and

$n^2/2 + O(n)$ divisions). It also requires half of the space (only B is needed, as opposed to both L and U). Furthermore, it can be shown that Cholesky's method is numerically stable (see Trefethen and Bau [68], Lecture 23). In `Matlab` the function `chol` returns the lower-triangular matrix B such that $A = BB^\top$ using the call $B = \text{chol}(A, \text{'lower'})$.

Remark: If $A = BB^\top$, where B is any invertible matrix, then A is symmetric positive definite.

Proof. Obviously, $BB^\top$ is symmetric, and since B is invertible, $B^\top$ is invertible, and from

$$x^\top Ax = x^\top BB^\top x = (B^\top x)^\top B^\top x,$$

it is clear that $x^\top Ax > 0$ if $x \neq 0$. □

We now give three more criteria for a symmetric matrix to be positive definite.

Proposition 7.11. *Let A be any $n \times n$ real symmetric matrix. The following conditions are equivalent:*

(a) A is positive definite.

(b) All principal minors of A are positive; that is: $\det(A(1:k, 1:k)) > 0$ for $k = 1, \ldots, n$ (Sylvester's criterion).

(c) A has an LU-factorization and all pivots are positive.

(d) A has an $LDL^\top$-factorization and all pivots in D are positive.

Proof. By Proposition 7.9, if A is symmetric positive definite, then each matrix $A(1:k, 1:k)$ is symmetric positive definite for $k = 1, \ldots, n$. By the Cholsesky decomposition, $A(1:k, 1:k) = Q^\top Q$ for some invertible matrix Q, so $\det(A(1:k, 1:k)) = \det(Q)^2 > 0$. This shows that (a) implies (b).

If $\det(A(1:k, 1:k)) > 0$ for $k = 1, \ldots, n$, then each $A(1:k, 1:k)$ is invertible. By Proposition 7.2, the matrix A has an LU-factorization, and since the pivots π_k are given by

$$\pi_k = \begin{cases} a_{11} = \det(A(1:1, 1:1)) & \text{if } k = 1 \\ \dfrac{\det(A(1:k, 1:k))}{\det(A(1:k-1, 1:k-1))} & \text{if } k = 2, \ldots, n, \end{cases}$$

we see that $\pi_k > 0$ for $k = 1, \ldots, n$. Thus (b) implies (c).

Assume A has an LU-factorization and that the pivots are all positive. Since A is symmetric, this implies that A has a factorization of the form

$$A = LDL^\top,$$

with L lower-triangular with 1s on its diagonal, and where D is a diagonal matrix with positive entries on the diagonal (the pivots). This shows that (c) implies (d).

Given a factorization $A = LDL^\top$ with all pivots in D positive, if we form the diagonal matrix

$$\sqrt{D} = \mathrm{diag}(\sqrt{\pi_1}, \ldots, \sqrt{\pi_n})$$

and if we let $B = L\sqrt{D}$, then we have

$$A = BB^\top,$$

with B lower-triangular and invertible. By the remark before Proposition 7.11, A is positive definite. Hence, (d) implies (a). □

Criterion (c) yields a simple computational test to check whether a symmetric matrix is positive definite. There is one more criterion for a symmetric matrix to be positive definite: its eigenvalues must be positive. We will have to learn about the spectral theorem for symmetric matrices to establish this criterion.

Proposition 7.11 also holds for complex Hermitian positive definite matrices, where in (d), the factorization $LDL^\top$ is replaced by LDL^*.

For more on the stability analysis and efficient implementation methods of Gaussian elimination, LU-factoring and Cholesky factoring, see Demmel [16], Trefethen and Bau [68], Ciarlet [14], Golub and Van Loan [30], Meyer [48], Strang [63, 64], and Kincaid and Cheney [39].

7.10 Reduced Row Echelon Form (RREF)

Gaussian elimination described in Section 7.2 can also be applied to rectangular matrices. This yields a method for determining whether a system $Ax = b$ is solvable and a description of all the solutions when the system is solvable, for any rectangular $m \times n$ matrix A.

It turns out that the discussion is simpler if we rescale all pivots to be 1, and for this we need a third kind of elementary matrix. For any $\lambda \neq 0$, let $E_{i,\lambda}$ be the $n \times n$ diagonal matrix

$$E_{i,\lambda} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & \lambda & & & \\ & & & & 1 & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix},$$

with $(E_{i,\lambda})_{ii} = \lambda$ $(1 \leq i \leq n)$. Note that $E_{i,\lambda}$ is also given by

$$E_{i,\lambda} = I + (\lambda - 1)e_{ii},$$

and that $E_{i,\lambda}$ is invertible with

$$E_{i,\lambda}^{-1} = E_{i,\lambda^{-1}}.$$

Now after $k-1$ elimination steps, if the bottom portion

$$(a_{kk}^{(k)}, a_{k+1k}^{(k)}, \ldots, a_{mk}^{(k)})$$

of the kth column of the current matrix A_k is nonzero so that a pivot π_k can be chosen, after a permutation of rows if necessary, we also divide row k by π_k to obtain the pivot 1, and not only do we zero all the entries $i = k+1, \ldots, m$ in column k, but also all the entries $i = 1, \ldots, k-1$, so that the only nonzero entry in column k is a 1 in row k. These row operations are achieved by multiplication on the left by elementary matrices.

If $a_{kk}^{(k)} = a_{k+1k}^{(k)} = \cdots = a_{mk}^{(k)} = 0$, we move on to column $k+1$.

When the kth column contains a pivot, the kth stage of the procedure for converting a matrix to *rref* consists of the following three steps illustrated below:

$$\begin{pmatrix} 1 & \times & 0 & \times & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{pivot}}{\Longrightarrow} \begin{pmatrix} 1 & \times & 0 & \times & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times & \times \\ 0 & 0 & 0 & a_{ik}^{(k)} & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{rescale}}{\Longrightarrow}$$

$$\begin{pmatrix} 1 & \times & 0 & \times & \times & \times & \times \\ 0 & 0 & 1 & \times & \times & \times & \times \\ 0 & 0 & 0 & 1 & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \\ 0 & 0 & 0 & \times & \times & \times & \times \end{pmatrix} \overset{\text{elim}}{\Longrightarrow} \begin{pmatrix} 1 & \times & 0 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 1 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{1} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{0} & \times & \times & \times \\ 0 & 0 & 0 & \mathbf{0} & \times & \times & \times \end{pmatrix}.$$

If the kth column does not contain a pivot, we simply move on to the next column.

The result is that after performing such elimination steps, we obtain a matrix that has a special shape known as a *reduced row echelon matrix*, for short *rref*.

Here is an example illustrating this process: Starting from the matrix

$$A_1 = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 1 & 1 & 5 & 2 & 7 \\ 1 & 2 & 8 & 4 & 12 \end{pmatrix},$$

we perform the following steps

$$A_1 \longrightarrow A_2 = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 1 & 2 \\ 0 & 2 & 6 & 3 & 7 \end{pmatrix},$$

by subtracting row 1 from row 2 and row 3;

$$A_2 \longrightarrow \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 2 & 6 & 3 & 7 \\ 0 & 1 & 3 & 1 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 3/2 & 7/2 \\ 0 & 1 & 3 & 1 & 2 \end{pmatrix} \longrightarrow A_3 = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 3/2 & 7/2 \\ 0 & 0 & 0 & -1/2 & -3/2 \end{pmatrix},$$

after choosing the pivot 2 and permuting row 2 and row 3, dividing row 2 by 2, and subtracting row 2 from row 3;

$$A_3 \longrightarrow \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 0 & 1 & 3 & 3/2 & 7/2 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix} \longrightarrow A_4 = \begin{pmatrix} 1 & 0 & 2 & 0 & 2 \\ 0 & 1 & 3 & 0 & -1 \\ 0 & 0 & 0 & 1 & 3 \end{pmatrix},$$

after dividing row 3 by $-1/2$, subtracting row 3 from row 1, and subtracting $(3/2) \times$ row 3 from row 2.

It is clear that columns $1, 2$ and 4 are linearly independent, that column 3 is a linear combination of columns 1 and 2, and that column 5 is a linear combination of columns $1, 2, 4$.

In general, the sequence of steps leading to a reduced echelon matrix is not unique. For example, we could have chosen 1 instead of 2 as the second pivot in matrix A_2. Nevertheless, *the reduced row echelon matrix obtained from any given matrix is unique*; that is, it does not depend on the the sequence of steps that are followed during the reduction process. This fact is not so easy to prove rigorously, but we will do it later.

If we want to solve a linear system of equations of the form $Ax = b$, we apply elementary row operations to both the matrix A and the right-hand side b. To do this conveniently, we form the *augmented matrix* (A, b), which is the $m \times (n+1)$ matrix obtained by adding b as an extra column to the matrix A. For example if

$$A = \begin{pmatrix} 1 & 0 & 2 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 2 & 8 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 5 \\ 7 \\ 12 \end{pmatrix},$$

then the augmented matrix is

$$(A, b) = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 1 & 1 & 5 & 2 & 7 \\ 1 & 2 & 8 & 4 & 12 \end{pmatrix}.$$

Now for any matrix M, since

$$M(A, b) = (MA, Mb),$$

performing elementary row operations on (A, b) is equivalent to simultaneously performing operations on both A and b. For example, consider the system

$$\begin{array}{rcrcrcrcl} x_1 & & & + & 2x_3 & + & x_4 & = & 5 \\ x_1 & + & x_2 & + & 5x_3 & + & 2x_4 & = & 7 \\ x_1 & + & 2x_2 & + & 8x_3 & + & 4x_4 & = & 12. \end{array}$$

Its augmented matrix is the matrix

$$(A, b) = \begin{pmatrix} 1 & 0 & 2 & 1 & 5 \\ 1 & 1 & 5 & 2 & 7 \\ 1 & 2 & 8 & 4 & 12 \end{pmatrix}$$

considered above, so the reduction steps applied to this matrix yield the system

$$\begin{array}{rcrcrcl} x_1 & & & + & 2x_3 & = & 2 \\ & & x_2 & + & 3x_3 & = & -1 \\ & & & & x_4 & = & 3. \end{array}$$

This reduced system has the same set of solutions as the original, and obviously x_3 can be chosen arbitrarily. Therefore, our system has infinitely many solutions given by

$$x_1 = 2 - 2x_3, \quad x_2 = -1 - 3x_3, \quad x_4 = 3,$$

where x_3 is arbitrary.

The following proposition shows that the set of solutions of a system $Ax = b$ is preserved by any sequence of row operations.

Proposition 7.12. *Given any $m \times n$ matrix A and any vector $b \in \mathbb{R}^m$, for any sequence of elementary row operations $E_1, \ldots, E_k$, if $P = E_k \cdots E_1$ and $(A', b') = P(A, b)$, then the solutions of $Ax = b$ are the same as the solutions of $A'x = b'$.*

Proof. Since each elementary row operation E_i is invertible, so is P, and since $(A', b') = P(A, b)$, then $A' = PA$ and $b' = Pb$. If x is a solution of the original system $Ax = b$, then multiplying both sides by P we get $PAx = Pb$; that is, $A'x = b'$, so x is a solution of the new system. Conversely, assume that x is a solution of the new system, that is $A'x = b'$. Then because $A' = PA$, $b' = Pb$, and P is invertible, we get

$$Ax = P^{-1}A'x = P^{-1}b' = b,$$

so x is a solution of the original system $Ax = b$. □

Another important fact is this:

Proposition 7.13. *Given an $m \times n$ matrix A, for any sequence of row operations $E_1, \ldots, E_k$, if $P = E_k \cdots E_1$ and $B = PA$, then the subspaces spanned by the rows of A and the rows of B are identical. Therefore, A and B have the same row rank. Furthermore, the matrices A and B also have the same (column) rank.*

Proof. Since $B = PA$, from a previous observation, the rows of B are linear combinations of the rows of A, so the span of the rows of B is a subspace of the span of the rows of A. Since P is invertible, $A = P^{-1}B$, so by the same reasoning the span of the rows of A is a subspace of the span of the rows of B. Therefore, the subspaces spanned by the rows of A and the rows of B are identical, which implies that A and B have the same row rank.

Proposition 7.12 implies that the systems $Ax = 0$ and $Bx = 0$ have the same solutions. Since Ax is a linear combinations of the columns of A and Bx is a linear combinations of the columns of B, the maximum number of linearly independent columns in A is equal to the maximum number of linearly independent columns in B; that is, A and B have the same rank. □

Remark: The subspaces spanned by the columns of A and B can be different! However, their dimension must be the same.

We will show in Section 7.14 that the row rank is equal to the column rank. This will also be proven in Proposition 10.13 Let us now define precisely what is a reduced row echelon matrix.

Definition 7.4. An $m \times n$ matrix A is a *reduced row echelon matrix* iff the following conditions hold:

(a) The first nonzero entry in every row is 1. This entry is called a *pivot*.

(b) The first nonzero entry of row $i+1$ is to the right of the first nonzero entry of row i.

(c) The entries above a pivot are zero.

If a matrix satisfies the above conditions, we also say that it is in *reduced row echelon form*, for short *rref*.

Note that Condition (b) implies that the entries below a pivot are also zero. For example, the matrix

$$A = \begin{pmatrix} 1 & 6 & 0 & 1 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

is a reduced row echelon matrix. In general, a matrix in *rref* has the following shape:

$$\begin{pmatrix} 1 & 0 & 0 & \times & \times & 0 & 0 & \times \\ 0 & 1 & 0 & \times & \times & 0 & 0 & \times \\ 0 & 0 & 1 & \times & \times & 0 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

if the last row consists of zeros, or

$$\begin{pmatrix} 1 & 0 & 0 & \times & \times & 0 & 0 & \times & 0 & \times \\ 0 & 1 & 0 & \times & \times & 0 & 0 & \times & 0 & \times \\ 0 & 0 & 1 & \times & \times & 0 & 0 & \times & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \times & 0 & \times \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \times & \times & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \times \end{pmatrix}$$

if the last row contains a pivot.

The following proposition shows that every matrix can be converted to a reduced row echelon form using row operations.

Proposition 7.14. *Given any $m \times n$ matrix A, there is a sequence of row operations $E_1, \ldots, E_k$ such that if $P = E_k \cdots E_1$, then $U = PA$ is a reduced row echelon matrix.*

Proof. We proceed by induction on m. If $m = 1$, then either all entries on this row are zero, so $A = 0$, or if a_j is the first nonzero entry in A, let $P = (a_j^{-1})$ (a 1×1 matrix); clearly, PA is a reduced row echelon matrix.

Let us now assume that $m \geq 2$. If $A = 0$, we are done, so let us assume that $A \neq 0$. Since $A \neq 0$, there is a leftmost column j which is nonzero, so pick any pivot $\pi = a_{ij}$ in the jth column, permute row i and row 1 if necessary, multiply the new first row by π^{-1}, and clear out the other entries in column j by subtracting suitable multiples of row 1. At the end of this process, we have a matrix A_1 that has the following shape:

$$A_1 = \begin{pmatrix} 0 & \cdots & 0 & 1 & * & \cdots & * \\ 0 & \cdots & 0 & 0 & * & \cdots & * \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & * & \cdots & * \end{pmatrix},$$

where $*$ stands for an arbitrary scalar, or more concisely

$$A_1 = \begin{pmatrix} 0 & 1 & B \\ 0 & 0 & D \end{pmatrix},$$

where D is a $(m-1) \times (n-j)$ matrix (and B is a $1 \times n - j$ matrix). If $j = n$, we are done. Otherwise, by the induction hypothesis applied to D, there is a sequence of row operations that converts D to a reduced row echelon matrix R', and these row operations do not affect the first row of A_1, which means that A_1 is reduced to a matrix of the form

$$R = \begin{pmatrix} 0 & 1 & B \\ 0 & 0 & R' \end{pmatrix}.$$

Because R' is a reduced row echelon matrix, the matrix R satisfies Conditions (a) and (b) of the reduced row echelon form. Finally, the entries above all pivots in R' can be cleared out by subtracting suitable multiples of the rows of R' containing a pivot. The resulting matrix also satisfies Condition (c), and the induction step is complete. □

Remark: There is a `Matlab` function named `rref` that converts any matrix to its reduced row echelon form.

If A is any matrix and if R is a reduced row echelon form of A, the second part of Proposition 7.13 can be sharpened a little, since the structure of a reduced row echelon matrix makes it clear that its rank is equal to the number of pivots.

Proposition 7.15. *The rank of a matrix A is equal to the number of pivots in its rref R.*

7.11 RREF, Free Variables, and Homogenous Linear Systems

Given a system of the form $Ax = b$, we can apply the reduction procedure to the augmented matrix (A, b) to obtain a reduced row echelon matrix (A', b') such that the system $A'x = b'$ has the same solutions as the original system $Ax = b$. The advantage of the reduced system $A'x = b'$ is that there is a simple test to check whether this system is solvable, and to find its solutions if it is solvable.

Indeed, if any row of the matrix A' is zero and if the corresponding entry in b' is nonzero, then it is a pivot and we have the "equation"

$$0 = 1,$$

which means that the system $A'x = b'$ has no solution. On the other hand, if there is no pivot in b', then for every row i in which $b'_i \neq 0$, there is some column j in A' where the entry on row i is 1 (a pivot). Consequently, we can assign arbitrary values to the variable x_k if column k does not contain a pivot, and then solve for the pivot variables.

For example, if we consider the reduced row echelon matrix

$$(A', b') = \begin{pmatrix} 1 & 6 & 0 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

there is no solution to $A'x = b'$ because the third equation is $0 = 1$. On the other hand, the reduced system

$$(A', b') = \begin{pmatrix} 1 & 6 & 0 & 1 & 1 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

has solutions. We can pick the variables x_2, x_4 corresponding to nonpivot columns arbitrarily, and then solve for x_3 (using the second equation) and x_1 (using the first equation).

The above reasoning proves the following theorem:

Theorem 7.16. *Given any system $Ax = b$ where A is a $m \times n$ matrix, if the augmented matrix (A, b) is a reduced row echelon matrix, then the system $Ax = b$ has a solution iff there is no pivot in b. In that case, an arbitrary value can be assigned to the variable x_j if column j does not contain a pivot.*

Definition 7.5. Nonpivot variables are often called *free variables*.

Putting Proposition 7.14 and Theorem 7.16 together we obtain a criterion to decide whether a system $Ax = b$ has a solution: Convert the augmented system (A, b) to a row reduced echelon matrix (A', b') and check whether b' has no pivot.

Remark: When writing a program implementing row reduction, we may stop when the last column of the matrix A is reached. In this case, the test whether the system $Ax = b$ is

solvable is that the row-reduced matrix A' has no zero row of index $i > r$ such that $b'_i \neq 0$ (where r is the number of pivots, and b' is the row-reduced right-hand side).

If we have a *homogeneous system* $Ax = 0$, which means that $b = 0$, of course $x = 0$ is always a solution, but Theorem 7.16 implies that if the system $Ax = 0$ has more variables than equations, then it has some nonzero solution (we call it a *nontrivial solution*).

Proposition 7.17. *Given any homogeneous system $Ax = 0$ of m equations in n variables, if $m < n$, then there is a nonzero vector $x \in \mathbb{R}^n$ such that $Ax = 0$.*

Proof. Convert the matrix A to a reduced row echelon matrix A'. We know that $Ax = 0$ iff $A'x = 0$. If r is the number of pivots of A', we must have $r \leq m$, so by Theorem 7.16 we may assign arbitrary values to $n - r > 0$ nonpivot variables and we get nontrivial solutions. □

Theorem 7.16 can also be used to characterize when a square matrix is invertible. First, note the following simple but important fact:

If a square $n \times n$ matrix A is a row reduced echelon matrix, then either A is the identity or the bottom row of A is zero.

Proposition 7.18. *Let A be a square matrix of dimension n. The following conditions are equivalent:*

(a) The matrix A can be reduced to the identity by a sequence of elementary row operations.

(b) The matrix A is a product of elementary matrices.

(c) The matrix A is invertible.

(d) The system of homogeneous equations $Ax = 0$ has only the trivial solution $x = 0$.

Proof. First we prove that (a) implies (b). If (a) can be reduced to the identity by a sequence of row operations $E_1, \ldots, E_p$, this means that $E_p \cdots E_1 A = I$. Since each E_i is invertible, we get

$$A = E_1^{-1} \cdots E_p^{-1},$$

where each E_i^{-1} is also an elementary row operation, so (b) holds. Now if (b) holds, since elementary row operations are invertible, A is invertible and (c) holds. If A is invertible, we already observed that the homogeneous system $Ax = 0$ has only the trivial solution $x = 0$, because from $Ax = 0$, we get $A^{-1}Ax = A^{-1}0$; that is, $x = 0$. It remains to prove that (d) implies (a) and for this we prove the contrapositive: if (a) does not hold, then (d) does not hold.

Using our basic observation about reducing square matrices, if A does not reduce to the identity, then A reduces to a row echelon matrix A' whose bottom row is zero. Say $A' = PA$, where P is a product of elementary row operations. Because the bottom row of A' is zero, the system $A'x = 0$ has at most $n - 1$ nontrivial equations, and by Proposition 7.17, this system has a nontrivial solution x. But then, $Ax = P^{-1}A'x = 0$ with $x \neq 0$, contradicting the fact that the system $Ax = 0$ is assumed to have only the trivial solution. Therefore, (d) implies (a) and the proof is complete. □

Proposition 7.18 yields a method for computing the inverse of an invertible matrix A: reduce A to the identity using elementary row operations, obtaining

$$E_p \cdots E_1 A = I.$$

Multiplying both sides by A^{-1} we get

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

From a practical point of view, we can build up the product $E_p \cdots E_1$ by reducing to row echelon form the augmented $n \times 2n$ matrix (A, I_n) obtained by adding the n columns of the identity matrix to A. This is just another way of performing the Gauss–Jordan procedure.

Here is an example: let us find the inverse of the matrix

$$A = \begin{pmatrix} 5 & 4 \\ 6 & 5 \end{pmatrix}.$$

We form the 2×4 block matrix

$$(A, I) = \begin{pmatrix} 5 & 4 & 1 & 0 \\ 6 & 5 & 0 & 1 \end{pmatrix}$$

and apply elementary row operations to reduce A to the identity. For example:

$$(A, I) = \begin{pmatrix} 5 & 4 & 1 & 0 \\ 6 & 5 & 0 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 5 & 4 & 1 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix}$$

by subtracting row 1 from row 2,

$$\begin{pmatrix} 5 & 4 & 1 & 0 \\ 1 & 1 & -1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 5 & -4 \\ 1 & 1 & -1 & 1 \end{pmatrix}$$

by subtracting $4\times$ row 2 from row 1,

$$\begin{pmatrix} 1 & 0 & 5 & -4 \\ 1 & 1 & -1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 5 & -4 \\ 0 & 1 & -6 & 5 \end{pmatrix} = (I, A^{-1}),$$

by subtracting row 1 from row 2. Thus

$$A^{-1} = \begin{pmatrix} 5 & -4 \\ -6 & 5 \end{pmatrix}.$$

Proposition 7.18 can also be used to give an elementary proof of the fact that if a square matrix A has a left inverse B (resp. a right inverse B), so that $BA = I$ (resp. $AB = I$), then A is invertible and $A^{-1} = B$. This is an interesting exercise, try it!

7.12 Uniqueness of RREF Form

For the sake of completeness, we prove that the reduced row echelon form of a matrix is unique. The neat proof given below is borrowed and adapted from W. Kahan.

Proposition 7.19. *Let A be any $m \times n$ matrix. If U and V are two reduced row echelon matrices obtained from A by applying two sequences of elementary row operations $E_1, \ldots, E_p$ and $F_1, \ldots, F_q$, so that*

$$U = E_p \cdots E_1 A \quad \textit{and} \quad V = F_q \cdots F_1 A,$$

then $U = V$ and $E_p \cdots E_1 = F_q \cdots F_1$. In other words, the reduced row echelon form of any matrix is unique.

Proof. Let

$$C = E_p \cdots E_1 F_1^{-1} \cdots F_q^{-1}$$

so that

$$U = CV \quad \text{and} \quad V = C^{-1}U.$$

We prove by induction on n that $U = V$ (and $C = I$).

Let ℓ_j denote the jth column of the identity matrix I_n, and let $u_j = U\ell_j$, $v_j = V\ell_j$, $c_j = C\ell_j$, and $a_j = A\ell_j$, be the jth column of U, V, C, and A respectively.

First I claim that $u_j = 0$ iff $v_j = 0$ iff $a_j = 0$.

Indeed, if $v_j = 0$, then (because $U = CV$) $u_j = Cv_j = 0$, and if $u_j = 0$, then $v_j = C^{-1}u_j = 0$. Since $U = E_p \cdots E_1 A$, we also get $a_j = 0$ iff $u_j = 0$.

Therefore, we may simplify our task by striking out columns of zeros from U, V, and A, since they will have corresponding indices. We still use n to denote the number of columns of A. Observe that because U and V are reduced row echelon matrices with no zero columns, we must have $u_1 = v_1 = \ell_1$.

Claim. If U and V are reduced row echelon matrices without zero columns such that $U = CV$, for all $k \geq 1$, if $k \leq n$, then ℓ_k occurs in U iff ℓ_k occurs in V, and if ℓ_k does occur in U, then

1. ℓ_k occurs for the same column index j_k in both U and V;

2. the first j_k columns of U and V match;

3. the subsequent columns in U and V (of column index $> j_k$) whose coordinates of index $k+1$ through m are all equal to 0 also match. Let n_k be the rightmost index of such a column, with $n_k = j_k$ if there is none.

4. the first n_k columns of C match the first n_k columns of I_n.

We prove this claim by induction on k.

For the base case $k = 1$, we already know that $u_1 = v_1 = \ell_1$. We also have

$$c_1 = C\ell_1 = Cv_1 = u_1 = \ell_1.$$

If $v_j = \lambda\ell_1$ for some $\lambda \in \mathbb{R}$, then

$$u_j = U\ell_j = CV\ell_j = Cv_j = \lambda C\ell_1 = \lambda c_1 = \lambda\ell_1 = v_j.$$

A similar argument using C^{-1} shows that if $u_j = \lambda\ell_1$, then $v_j = u_j$. Therefore, all the columns of U and V proportional to ℓ_1 match, which establishes the base case. Observe that if ℓ_2 appears in U, then it must appear in both U and V for the same index, and if not then $n_1 = n$ and $U = V$.

Next us now prove the induction step. If $n_k = n$, then $U = V$ and we are done. Otherwise, ℓ_{k+1} appears in both U and V, in which case, by (2) and (3) of the induction hypothesis, it appears in both U and V for the same index, say j_{k+1}. Thus, $u_{j_{k+1}} = v_{j_{k+1}} = \ell_{k+1}$. It follows that

$$c_{k+1} = C\ell_{k+1} = Cv_{j_{k+1}} = u_{j_{k+1}} = \ell_{k+1},$$

so the first j_{k+1} columns of C match the first j_{k+1} columns of I_n.

Consider any subsequent column v_j (with $j > j_{k+1}$) whose elements beyond the $(k+1)$th all vanish. Then v_j is a linear combination of columns of V to the left of v_j, so

$$u_j = Cv_j = v_j.$$

because the first $k+1$ columns of C match the first column of I_n. Similarly, any subsequent column u_j (with $j > j_{k+1}$) whose elements beyond the $(k+1)$th all vanish is equal to v_j. Therefore, all the subsequent columns in U and V (of index $> j_{k+1}$) whose elements beyond the $(k+1)$th all vanish also match, so the first n_{k+1} columns of C match the first n_{k+1} columns of C, which completes the induction hypothesis.

We can now prove that $U = V$ (recall that we may assume that U and V have no zero columns). We noted earlier that $u_1 = v_1 = \ell_1$, so there is a largest $k \leq n$ such that ℓ_k occurs in U. Then the previous claim implies that all the columns of U and V match, which means that $U = V$. □

The reduction to row echelon form also provides a method to describe the set of solutions of a linear system of the form $Ax = b$.

7.13 Solving Linear Systems Using RREF

First we have the following simple result.

Proposition 7.20. *Let A be any $m \times n$ matrix and let $b \in \mathbb{R}^m$ be any vector. If the system $Ax = b$ has a solution, then the set Z of all solutions of this system is the set*

$$Z = x_0 + \mathrm{Ker}\,(A) = \{x_0 + x \mid Ax = 0\},$$

where $x_0 \in \mathbb{R}^n$ is any solution of the system $Ax = b$, which means that $Ax_0 = b$ (x_0 is called a special solution), and where $\mathrm{Ker}\,(A) = \{x \in \mathbb{R}^n \mid Ax = 0\}$*, the set of solutions of the homogeneous system associated with $Ax = b$.*

Proof. Assume that the system $Ax = b$ is solvable and let x_0 and x_1 be any two solutions so that $Ax_0 = b$ and $Ax_1 = b$. Subtracting the first equation from the second, we get

$$A(x_1 - x_0) = 0,$$

which means that $x_1 - x_0 \in \mathrm{Ker}\,(A)$. Therefore, $Z \subseteq x_0 + \mathrm{Ker}\,(A)$, where x_0 is a special solution of $Ax = b$. Conversely, if $Ax_0 = b$, then for any $z \in \mathrm{Ker}\,(A)$, we have $Az = 0$, and so

$$A(x_0 + z) = Ax_0 + Az = b + 0 = b,$$

which shows that $x_0 + \mathrm{Ker}\,(A) \subseteq Z$. Therefore, $Z = x_0 + \mathrm{Ker}\,(A)$. □

Given a linear system $Ax = b$, reduce the augmented matrix (A, b) to its row echelon form (A', b'). As we showed before, the system $Ax = b$ has a solution iff b' contains no pivot. Assume that this is the case. Then, if (A', b') has r pivots, which means that A' has r pivots since b' has no pivot, we know that the first r columns of I_m appear in A'.

We can permute the columns of A' and renumber the variables in x correspondingly so that the first r columns of I_m match the first r columns of A', and then our reduced echelon matrix is of the form (R, b') with

$$R = \begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix}$$

and

$$b' = \begin{pmatrix} d \\ 0_{m-r} \end{pmatrix},$$

where F is a $r \times (n-r)$ matrix and $d \in \mathbb{R}^r$. Note that R has $m - r$ zero rows.

Then because

$$\begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix} \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix} = \begin{pmatrix} d \\ 0_{m-r} \end{pmatrix} = b',$$

we see that

$$x_0 = \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix}$$

is a special solution of $Rx = b'$, and thus to $Ax = b$. In other words, we get a special solution by assigning the first r components of b' to the pivot variables and setting the nonpivot variables (the *free variables*) to zero.

Here is an example of the preceding construction taken from Kumpel and Thorpe [40]. The linear system

$$\begin{aligned} x_1 - x_2 + x_3 + x_4 - 2x_5 &= -1 \\ -2x_1 + 2x_2 - x_3 + x_5 &= 2 \\ x_1 - x_2 + 2x_3 + 3x_4 - 5x_5 &= -1, \end{aligned}$$

is represented by the augmented matrix

$$(A, b) = \begin{pmatrix} 1 & -1 & 1 & 1 & -2 & -1 \\ -2 & 2 & -1 & 0 & 1 & 2 \\ 1 & -1 & 2 & 3 & -5 & -1 \end{pmatrix},$$

where A is a 3×5 matrix. The reader should find that the row echelon form of this system is

$$(A', b') = \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & -1 \\ 0 & 0 & 1 & 2 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The 3×5 matrix A' has rank 2. We permute the second and third columns (which is equivalent to interchanging variables x_2 and x_3) to form

$$R = \begin{pmatrix} I_2 & F \\ 0_{1,2} & 0_{1,3} \end{pmatrix}, \qquad F = \begin{pmatrix} -1 & -1 & 1 \\ 0 & 2 & -3 \end{pmatrix}.$$

Then a special solution to this linear system is given by

$$x_0 = \begin{pmatrix} d \\ 0_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0_3 \end{pmatrix}.$$

We can also find a basis of the kernel (nullspace) of A using F. If $x = (u, v)$ is in the kernel of A, with $u \in \mathbb{R}^r$ and $v \in \mathbb{R}^{n-r}$, then x is also in the kernel of R, which means that $Rx = 0$; that is,

$$\begin{pmatrix} I_r & F \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u + Fv \\ 0_{m-r} \end{pmatrix} = \begin{pmatrix} 0_r \\ 0_{m-r} \end{pmatrix}.$$

Therefore, $u = -Fv$, and $\mathrm{Ker}\,(A)$ consists of all vectors of the form

$$\begin{pmatrix} -Fv \\ v \end{pmatrix} = \begin{pmatrix} -F \\ I_{n-r} \end{pmatrix} v,$$

for any arbitrary $v \in \mathbb{R}^{n-r}$. It follows that the $n - r$ columns of the matrix

$$N = \begin{pmatrix} -F \\ I_{n-r} \end{pmatrix}$$

form a basis of the kernel of A. This is because N contains the identity matrix I_{n-r} as a submatrix, so the columns of N are linearly independent. In summary, if $N^1, \ldots, N^{n-r}$ are the columns of N, then the general solution of the equation $Ax = b$ is given by

$$x = \begin{pmatrix} d \\ 0_{n-r} \end{pmatrix} + x_{r+1}N^1 + \cdots + x_n N^{n-r},$$

where $x_{r+1}, \ldots, x_n$ are the free variables; that is, the nonpivot variables.

Going back to our example from Kumpel and Thorpe [40], we see that

$$N = \begin{pmatrix} -F \\ I_3 \end{pmatrix} = \begin{pmatrix} 1 & 1 & -1 \\ 0 & -2 & -3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and that the general solution is given by

$$x = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} + x_3 \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} + x_4 \begin{pmatrix} 1 \\ -2 \\ 0 \\ 1 \\ 0 \end{pmatrix} + x_5 \begin{pmatrix} -1 \\ -3 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

In the general case where the columns corresponding to pivots are mixed with the columns corresponding to free variables, we find the special solution as follows. Let $i_1 < \cdots < i_r$ be the indices of the columns corresponding to pivots. Assign b'_k to the pivot variable x_{i_k} for $k = 1, \ldots, r$, and set all other variables to 0. To find a basis of the kernel, we form the $n - r$ vectors N^k obtained as follows. Let $j_1 < \cdots < j_{n-r}$ be the indices of the columns corresponding to free variables. For every column j_k corresponding to a free variable ($1 \leq k \leq n - r$), form the vector N^k defined so that the entries $N^k_{i_1}, \ldots, N^k_{i_r}$ are equal to the negatives of the first r entries in column j_k (flip the sign of these entries); let $N^k_{j_k} = 1$, and set all other entries to zero. Schematically, if the column of index j_k (corresponding to the free variable x_{j_k}) is

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_r \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

then the vector N^k is given by

$$\begin{array}{c|c} 1 & 0 \\ \vdots & \vdots \\ i_1 - 1 & 0 \\ i_1 & -\alpha_1 \\ i_1 + 1 & 0 \\ \vdots & \vdots \\ i_r - 1 & 0 \\ i_r & -\alpha_r \\ i_r + 1 & 0 \\ \vdots & \vdots \\ j_k - 1 & 0 \\ j_k & 1 \\ j_k + 1 & 0 \\ \vdots & \vdots \\ n & 0 \end{array}.$$

The presence of the 1 in position j_k guarantees that $N^1, \ldots, N^{n-r}$ are linearly independent.

As an illustration of the above method, consider the problem of finding a basis of the subspace V of $n \times n$ matrices $A \in \mathrm{M}_n(\mathbb{R})$ satisfying the following properties:

1. The sum of the entries in every row has the same value (say c_1);

2. The sum of the entries in every column has the same value (say c_2).

It turns out that $c_1 = c_2$ and that the $2n-2$ equations corresponding to the above conditions are linearly independent. We leave the proof of these facts as an interesting exercise. It can be shown using the duality theorem (Theorem 10.4) that the dimension of the space V of matrices satisying the above equations is $n^2 - (2n - 2)$. Let us consider the case $n = 4$. There are 6 equations, and the space V has dimension 10. The equations are

$$\begin{aligned}
a_{11} + a_{12} + a_{13} + a_{14} - a_{21} - a_{22} - a_{23} - a_{24} &= 0 \\
a_{21} + a_{22} + a_{23} + a_{24} - a_{31} - a_{32} - a_{33} - a_{34} &= 0 \\
a_{31} + a_{32} + a_{33} + a_{34} - a_{41} - a_{42} - a_{43} - a_{44} &= 0 \\
a_{11} + a_{21} + a_{31} + a_{41} - a_{12} - a_{22} - a_{32} - a_{42} &= 0 \\
a_{12} + a_{22} + a_{32} + a_{42} - a_{13} - a_{23} - a_{33} - a_{43} &= 0 \\
a_{13} + a_{23} + a_{33} + a_{43} - a_{14} - a_{24} - a_{34} - a_{44} &= 0,
\end{aligned}$$

and the corresponding matrix is

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

The result of performing the reduction to row echelon form yields the following matrix in rref:

$$U = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & -1 & -1 & -1 & 2 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & -1 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \end{pmatrix}$$

The list *pivlist* of indices of the pivot variables and the list *freelist* of indices of the free variables is given by

$$\begin{aligned} \mathit{pivlist} &= (1, 2, 3, 4, 5, 9), \\ \mathit{freelist} &= (\mathbf{6}, \mathbf{7}, \mathbf{8}, \mathbf{10}, \mathbf{11}, \mathbf{12}, \mathbf{13}, \mathbf{14}, \mathbf{15}, \mathbf{16}). \end{aligned}$$

After applying the algorithm to find a basis of the kernel of U, we find the following 16×10 matrix

$$BK = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & -2 & -1 & -1 & -1 \\ -1 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & -1 & 0 & 0 & -1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 & 0 & -1 & 1 & 1 & 1 & 0 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} \end{pmatrix}.$$

The reader should check that that in each column j of BK, the lowest bold 1 belongs to the row whose index is the jth element in *freelist*, and that in each column j of BK, the

signs of the entries whose indices belong to *pivlist* are the flipped signs of the 6 entries in the column U corresponding to the jth index in *freelist*. We can now read off from BK the 4×4 matrices that form a basis of V: every column of BK corresponds to a matrix whose rows have been concatenated. We get the following 10 matrices:

$$M_1 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_2 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_3 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_4 = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_5 = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad M_6 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$M_7 = \begin{pmatrix} -2 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad M_8 = \begin{pmatrix} -1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \quad M_9 = \begin{pmatrix} -1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$M_{10} = \begin{pmatrix} -1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Recall that a *magic square* is a square matrix that satisfies the two conditions about the sum of the entries in each row and in each column to be the same number, and also the additional two constraints that the main descending and the main ascending diagonals add up to this common number. Furthermore, the entries are also required to be positive integers. For $n = 4$, the additional two equations are

$$a_{22} + a_{33} + a_{44} - a_{12} - a_{13} - a_{14} = 0$$
$$a_{41} + a_{32} + a_{23} - a_{11} - a_{12} - a_{13} = 0,$$

and the 8 equations stating that a matrix is a magic square are linearly independent. Again, by running row elimination, we get a basis of the "generalized magic squares" whose entries are not restricted to be positive integers. We find a basis of 8 matrices. For $n = 3$, we find a basis of 3 matrices.

A magic square is said to be *normal* if its entries are precisely the integers $1, 2 \ldots, n^2$. Then since the sum of these entries is

$$1 + 2 + 3 + \cdots + n^2 = \frac{n^2(n^2+1)}{2},$$

and since each row (and column) sums to the same number, this common value (the *magic sum*) is

$$\frac{n(n^2+1)}{2}.$$

It is easy to see that there are no normal magic squares for $n = 2$. For $n = 3$, the magic sum is 15, for $n = 4$, it is 34, and for $n = 5$, it is 65.

In the case $n = 3$, we have the additional condition that the rows and columns add up to 15, so we end up with a solution parametrized by two numbers x_1, x_2; namely,

$$\begin{pmatrix} x_1 + x_2 - 5 & 10 - x_2 & 10 - x_1 \\ 20 - 2x_1 - x_2 & 5 & 2x_1 + x_2 - 10 \\ x_1 & x_2 & 15 - x_1 - x_2 \end{pmatrix}.$$

Thus, in order to find a normal magic square, we have the additional inequality constraints

$$\begin{aligned} x_1 + x_2 &> 5 \\ x_1 &< 10 \\ x_2 &< 10 \\ 2x_1 + x_2 &< 20 \\ 2x_1 + x_2 &> 10 \\ x_1 &> 0 \\ x_2 &> 0 \\ x_1 + x_2 &< 15, \end{aligned}$$

and all 9 entries in the matrix must be distinct. After a tedious case analysis, we discover the remarkable fact that there is a unique normal magic square (up to rotations and reflections):

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

It turns out that there are 880 different normal magic squares for $n = 4$, and 275, 305, 224 normal magic squares for $n = 5$ (up to rotations and reflections). Even for $n = 4$, it takes a fair amount of work to enumerate them all! Finding the number of magic squares for $n > 5$ is an open problem!

7.14 Elementary Matrices and Columns Operations

Instead of performing elementary row operations on a matrix A, we can perform elementary columns operations, which means that we multiply A by elementary matrices on the *right*. As elementary row and column operations, $P(i, k)$, $E_{i,j;\beta}$, $E_{i,\lambda}$ perform the following actions:

1. As a row operation, $P(i,k)$ permutes row i and row k.
2. As a column operation, $P(i,k)$ permutes column i and column k.
3. The inverse of $P(i,k)$ is $P(i,k)$ itself.
4. As a row operation, $E_{i,j;\beta}$ adds β times row j to row i.
5. As a column operation, $E_{i,j;\beta}$ adds β times column i to column j (note the switch in the indices).
6. The inverse of $E_{i,j;\beta}$ is $E_{i,j;-\beta}$.
7. As a row operation, $E_{i,\lambda}$ multiplies row i by λ.
8. As a column operation, $E_{i,\lambda}$ multiplies column i by λ.
9. The inverse of $E_{i,\lambda}$ is $E_{i,\lambda^{-1}}$.

We can define the notion of a reduced column echelon matrix and show that every matrix can be reduced to a unique reduced column echelon form. Now given any $m \times n$ matrix A, if we first convert A to its reduced row echelon form R, it is easy to see that we can apply elementary column operations that will reduce R to a matrix of the form

$$\begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix},$$

where r is the number of pivots (obtained during the row reduction). Therefore, for every $m \times n$ matrix A, there exist two sequences of elementary matrices $E_1, \ldots, E_p$ and $F_1, \ldots, F_q$, such that

$$E_p \cdots E_1 A F_1 \cdots F_q = \begin{pmatrix} I_r & 0_{r,n-r} \\ 0_{m-r,r} & 0_{m-r,n-r} \end{pmatrix}.$$

The matrix on the right-hand side is called the *rank normal form* of A. Clearly, r is the rank of A. As a corollary we obtain the following important result whose proof is immediate.

Proposition 7.21. *A matrix A and its transpose $A^\top$ have the same rank.*

7.15 Transvections and Dilatations ⊛

In this section we characterize the linear isomorphisms of a vector space E that leave every vector in some hyperplane fixed. These maps turn out to be the linear maps that are represented in some suitable basis by elementary matrices of the form $E_{i,j;\beta}$ (transvections) or $E_{i,\lambda}$ (dilatations). Furthermore, the transvections generate the group $\mathbf{SL}(E)$, and the dilatations generate the group $\mathbf{GL}(E)$.

Let H be any hyperplane in E, and pick some (nonzero) vector $v \in E$ such that $v \notin H$, so that

$$E = H \oplus Kv.$$

Assume that $f\colon E \to E$ is a linear isomorphism such that $f(u) = u$ for all $u \in H$, and that f is not the identity. We have

$$f(v) = h + \alpha v, \quad \text{for some } h \in H \text{ and some } \alpha \in K,$$

with $\alpha \neq 0$, because otherwise we would have $f(v) = h = f(h)$ since $h \in H$, contradicting the injectivity of f ($v \neq h$ since $v \notin H$). For any $x \in E$, if we write

$$x = y + tv, \quad \text{for some } y \in H \text{ and some } t \in K,$$

then

$$f(x) = f(y) + f(tv) = y + tf(v) = y + th + t\alpha v,$$

and since $\alpha x = \alpha y + t\alpha v$, we get

$$\begin{aligned} f(x) - \alpha x &= (1-\alpha)y + th \\ f(x) - x &= t(h + (\alpha - 1)v). \end{aligned}$$

Observe that if E is finite-dimensional, by picking a basis of E consisting of v and basis vectors of H, then the matrix of f is a lower triangular matrix whose diagonal entries are all 1 except the first entry which is equal to α. Therefore, $\det(f) = \alpha$.

Case 1. $\alpha \neq 1$.

We have $f(x) = \alpha x$ iff $(1-\alpha)y + th = 0$ iff

$$y = \frac{t}{\alpha - 1} h.$$

Then if we let $w = h + (\alpha - 1)v$, for $y = (t/(\alpha - 1))h$, we have

$$x = y + tv = \frac{t}{\alpha - 1} h + tv = \frac{t}{\alpha - 1}(h + (\alpha - 1)v) = \frac{t}{\alpha - 1} w,$$

which shows that $f(x) = \alpha x$ iff $x \in Kw$. Note that $w \notin H$, since $\alpha \neq 1$ and $v \notin H$. Therefore,

$$E = H \oplus Kw,$$

and f is the identity on H and a magnification by α on the line $D = Kw$.

Definition 7.6. Given a vector space E, for any hyperplane H in E, any nonzero vector $u \in E$ such that $u \notin H$, and any scalar $\alpha \neq 0, 1$, a linear map f such that $f(x) = x$ for all $x \in H$ and $f(x) = \alpha x$ for every $x \in D = Ku$ is called a *dilatation of hyperplane H, direction D, and scale factor α.*

If π_H and π_D are the projections of E onto H and D, then we have

$$f(x) = \pi_H(x) + \alpha\pi_D(x).$$

The inverse of f is given by

$$f^{-1}(x) = \pi_H(x) + \alpha^{-1}\pi_D(x).$$

When $\alpha = -1$, we have $f^2 = \mathrm{id}$, and f is a symmetry about the hyperplane H in the direction D. This situation includes orthogonal reflections about H.

Case 2. $\alpha = 1$.

In this case,

$$f(x) - x = th,$$

that is, $f(x) - x \in Kh$ for all $x \in E$. Assume that the hyperplane H is given as the kernel of some linear form φ, and let $a = \varphi(v)$. We have $a \neq 0$, since $v \notin H$. For any $x \in E$, we have

$$\varphi(x - a^{-1}\varphi(x)v) = \varphi(x) - a^{-1}\varphi(x)\varphi(v) = \varphi(x) - \varphi(x) = 0,$$

which shows that $x - a^{-1}\varphi(x)v \in H$ for all $x \in E$. Since every vector in H is fixed by f, we get

$$\begin{aligned} x - a^{-1}\varphi(x)v &= f(x - a^{-1}\varphi(x)v) \\ &= f(x) - a^{-1}\varphi(x)f(v), \end{aligned}$$

so

$$f(x) = x + \varphi(x)(f(a^{-1}v) - a^{-1}v).$$

Since $f(z) - z \in Kh$ for all $z \in E$, we conclude that $u = f(a^{-1}v) - a^{-1}v = \beta h$ for some $\beta \in K$, so $\varphi(u) = 0$, and we have

$$f(x) = x + \varphi(x)u, \quad \varphi(u) = 0. \tag{$*$}$$

A linear map defined as above is denoted by $\tau_{\varphi,u}$.

Conversely for any linear map $f = \tau_{\varphi,u}$ given by Equation $(*)$, where φ is a nonzero linear form and u is some vector $u \in E$ such that $\varphi(u) = 0$, if $u = 0$, then f is the identity, so assume that $u \neq 0$. If so, we have $f(x) = x$ iff $\varphi(x) = 0$, that is, iff $x \in H$. We also claim that the inverse of f is obtained by changing u to $-u$. Actually, we check the slightly more general fact that

$$\tau_{\varphi,u} \circ \tau_{\varphi,w} = \tau_{\varphi,u+w}.$$

Indeed, using the fact that $\varphi(w) = 0$, we have

$$\begin{aligned} \tau_{\varphi,u}(\tau_{\varphi,w}(x)) &= \tau_{\varphi,w}(x) + \varphi(\tau_{\varphi,w}(x))u \\ &= \tau_{\varphi,w}(x) + (\varphi(x) + \varphi(x)\varphi(w))u \\ &= \tau_{\varphi,w}(x) + \varphi(x)u \\ &= x + \varphi(x)w + \varphi(x)u \\ &= x + \varphi(x)(u + w). \end{aligned}$$

For $v = -u$, we have $\tau_{\varphi,u+v} = \varphi_{\varphi,0} = \mathrm{id}$, so $\tau_{\varphi,u}^{-1} = \tau_{\varphi,-u}$, as claimed.

Therefore, we proved that every linear isomorphism of E that leaves every vector in some hyperplane H fixed and has the property that $f(x) - x \in H$ for all $x \in E$ is given by a map $\tau_{\varphi,u}$ as defined by Equation $(*)$, where φ is some nonzero linear form defining H and u is some vector in H. We have $\tau_{\varphi,u} = \mathrm{id}$ iff $u = 0$.

Definition 7.7. Given any hyperplane H in E, for any nonzero nonlinear form $\varphi \in E^*$ defining H (which means that $H = \mathrm{Ker}\,(\varphi)$) and any nonzero vector $u \in H$, the linear map $f = \tau_{\varphi,u}$ given by

$$\tau_{\varphi,u}(x) = x + \varphi(x)u, \quad \varphi(u) = 0,$$

for all $x \in E$ is called a *transvection of hyperplane H and direction u*. The map $f = \tau_{\varphi,u}$ leaves every vector in H fixed, and $f(x) - x \in Ku$ for all $x \in E$.

The above arguments show the following result.

Proposition 7.22. *Let $f\colon E \to E$ be a bijective linear map and assume that $f \neq \mathrm{id}$ and that $f(x) = x$ for all $x \in H$, where H is some hyperplane in E. If there is some nonzero vector $u \in E$ such that $u \notin H$ and $f(u) - u \in H$, then f is a transvection of hyperplane H; otherwise, f is a dilatation of hyperplane H.*

Proof. Using the notation as above, for some $v \notin H$, we have $f(v) = h + \alpha v$ with $\alpha \neq 0$, and write $u = y + tv$ with $y \in H$ and $t \neq 0$ since $u \notin H$. If $f(u) - u \in H$, from

$$f(u) - u = t(h + (\alpha - 1)v),$$

we get $(\alpha - 1)v \in H$, and since $v \notin H$, we must have $\alpha = 1$, and we proved that f is a transvection. Otherwise, $\alpha \neq 0, 1$, and we proved that f is a dilatation. □

If E is finite-dimensional, then $\alpha = \det(f)$, so we also have the following result.

Proposition 7.23. *Let $f\colon E \to E$ be a bijective linear map of a finite-dimensional vector space E and assume that $f \neq \mathrm{id}$ and that $f(x) = x$ for all $x \in H$, where H is some hyperplane in E. If $\det(f) = 1$, then f is a transvection of hyperplane H; otherwise, f is a dilatation of hyperplane H.*

Suppose that f is a dilatation of hyperplane H and direction u, and say $\det(f) = \alpha \neq 0, 1$. Pick a basis $(u, e_2, \ldots, e_n)$ of E where $(e_2, \ldots, e_n)$ is a basis of H. Then the matrix of f is of the form

$$\begin{pmatrix} \alpha & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which is an elementary matrix of the form $E_{1,\alpha}$. Conversely, it is clear that every elementary matrix of the form $E_{i,\alpha}$ with $\alpha \neq 0, 1$ is a dilatation.

Now, assume that f is a transvection of hyperplane H and direction $u \in H$. Pick some $v \notin H$, and pick some basis $(u, e_3, \ldots, e_n)$ of H, so that $(v, u, e_3, \ldots, e_n)$ is a basis of E. Since $f(v) - v \in Ku$, the matrix of f is of the form

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ \alpha & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix},$$

which is an elementary matrix of the form $E_{2,1;\alpha}$. Conversely, it is clear that every elementary matrix of the form $E_{i,j;\alpha}$ $(\alpha \neq 0)$ is a transvection.

The following proposition is an interesting exercise that requires good mastery of the elementary row operations $E_{i,j;\beta}$; see Problems 7.10 and 7.11.

Proposition 7.24. *Given any invertible $n \times n$ matrix A, there is a matrix S such that*

$$SA = \begin{pmatrix} I_{n-1} & 0 \\ 0 & \alpha \end{pmatrix} = E_{n,\alpha},$$

with $\alpha = \det(A)$, and where S is a product of elementary matrices of the form $E_{i,j;\beta}$; that is, S is a composition of transvections.

Surprisingly, every transvection is the composition of two dilatations!

Proposition 7.25. *If the field K is not of characteristic 2, then every transvection f of hyperplane H can be written as $f = d_2 \circ d_1$, where d_1, d_2 are dilatations of hyperplane H, where the direction of d_1 can be chosen arbitrarily.*

Proof. Pick some dilatation d_1 of hyperplane H and scale factor $\alpha \neq 0, 1$. Then, $d_2 = f \circ d_1^{-1}$ leaves every vector in H fixed, and $\det(d_2) = \alpha^{-1} \neq 1$. By Proposition 7.23, the linear map d_2 is a dilatation of hyperplane H, and we have $f = d_2 \circ d_1$, as claimed. □

Observe that in Proposition 7.25, we can pick $\alpha = -1$; that is, every transvection of hyperplane H is the compositions of two symmetries about the hyperplane H, one of which can be picked arbitrarily.

Remark: Proposition 7.25 holds as long as $K \neq \{0, 1\}$.

The following important result is now obtained.

Theorem 7.26. *Let E be any finite-dimensional vector space over a field K of characteristic not equal to 2. Then the group $\mathbf{SL}(E)$ is generated by the transvections, and the group $\mathbf{GL}(E)$ is generated by the dilatations.*

Proof. Consider any $f \in \mathbf{SL}(E)$, and let A be its matrix in any basis. By Proposition 7.24, there is a matrix S such that

$$SA = \begin{pmatrix} I_{n-1} & 0 \\ 0 & \alpha \end{pmatrix} = E_{n,\alpha},$$

with $\alpha = \det(A)$, and where S is a product of elementary matrices of the form $E_{i,j;\beta}$. Since $\det(A) = 1$, we have $\alpha = 1$, and the result is proven. Otherwise, if f is invertible but $f \notin \mathbf{SL}(E)$, the above equation shows $E_{n,\alpha}$ is a dilatation, S is a product of transvections, and by Proposition 7.25, every transvection is the composition of two dilatations. Thus, the second result is also proven. □

We conclude this section by proving that any two transvections are conjugate in $\mathbf{GL}(E)$. Let $\tau_{\varphi,u}$ ($u \neq 0$) be a transvection and let $g \in \mathbf{GL}(E)$ be any invertible linear map. We have

$$\begin{aligned}(g \circ \tau_{\varphi,u} \circ g^{-1})(x) &= g(g^{-1}(x) + \varphi(g^{-1}(x))u) \\ &= x + \varphi(g^{-1}(x))g(u).\end{aligned}$$

Let us find the hyperplane determined by the linear form $x \mapsto \varphi(g^{-1}(x))$. This is the set of vectors $x \in E$ such that $\varphi(g^{-1}(x)) = 0$, which holds iff $g^{-1}(x) \in H$ iff $x \in g(H)$. Therefore, $\mathrm{Ker}\,(\varphi \circ g^{-1}) = g(H) = H'$, and we have $g(u) \in g(H) = H'$, so $g \circ \tau_{\varphi,u} \circ g^{-1}$ is the transvection of hyperplane $H' = g(H)$ and direction $u' = g(u)$ (with $u' \in H'$).

Conversely, let $\tau_{\psi,u'}$ be some transvection ($u' \neq 0$). Pick some vectors v, v' such that $\varphi(v) = \psi(v') = 1$, so that

$$E = H \oplus Kv = H' \oplus Kv'.$$

There is a linear map $g \in \mathbf{GL}(E)$ such that $g(u) = u'$, $g(v) = v'$, and $g(H) = H'$. To define g, pick a basis $(v, u, e_2, \ldots, e_{n-1})$ where $(u, e_2, \ldots, e_{n-1})$ is a basis of H and pick a basis $(v', u', e'_2, \ldots, e'_{n-1})$ where $(u', e'_2, \ldots, e'_{n-1})$ is a basis of H'; then g is defined so that $g(v) = v'$, $g(u) = u'$, and $g(e_i) = g(e'_i)$, for $i = 2, \ldots, n-1$. If $n = 2$, then e_i and e'_i are missing. Then, we have

$$(g \circ \tau_{\varphi,u} \circ g^{-1})(x) = x + \varphi(g^{-1}(x))u'.$$

Now $\varphi \circ g^{-1}$ also determines the hyperplane $H' = g(H)$, so we have $\varphi \circ g^{-1} = \lambda\psi$ for some nonzero λ in K. Since $v' = g(v)$, we get

$$\varphi(v) = \varphi \circ g^{-1}(v') = \lambda\psi(v'),$$

and since $\varphi(v) = \psi(v') = 1$, we must have $\lambda = 1$. It follows that

$$(g \circ \tau_{\varphi,u} \circ g^{-1})(x) = x + \psi(x)u' = \tau_{\psi,u'}(x).$$

In summary, we proved almost all parts the following result.

Proposition 7.27. *Let E be any finite-dimensional vector space. For every transvection $\tau_{\varphi,u}$ ($u \neq 0$) and every linear map $g \in \mathbf{GL}(E)$, the map $g \circ \tau_{\varphi,u} \circ g^{-1}$ is the transvection of hyperplane $g(H)$ and direction $g(u)$ (that is, $g \circ \tau_{\varphi,u} \circ g^{-1} = \tau_{\varphi \circ g^{-1}, g(u)}$). For every other transvection $\tau_{\psi,u'}$ ($u' \neq 0$), there is some $g \in \mathbf{GL}(E)$ such $\tau_{\psi,u'} = g \circ \tau_{\varphi,u} \circ g^{-1}$; in other words any two transvections ($\neq \mathrm{id}$) are conjugate in $\mathbf{GL}(E)$. Moreover, if $n \geq 3$, then the linear isomorphism g as above can be chosen so that $g \in \mathbf{SL}(E)$.*

Proof. We just need to prove that if $n \geq 3$, then for any two transvections $\tau_{\varphi,u}$ and $\tau_{\psi,u'}$ ($u, u' \neq 0$), there is some $g \in \mathbf{SL}(E)$ such that $\tau_{\psi,u'} = g \circ \tau_{\varphi,u} \circ g^{-1}$. As before, we pick a basis $(v, u, e_2, \ldots, e_{n-1})$ where $(u, e_2, \ldots, e_{n-1})$ is a basis of H, we pick a basis $(v', u', e'_2, \ldots, e'_{n-1})$ where $(u', e'_2, \ldots, e'_{n-1})$ is a basis of H', and we define g as the unique linear map such that $g(v) = v'$, $g(u) = u'$, and $g(e_i) = e'_i$, for $i = 1, \ldots, n-1$. But in this case, both H and $H' = g(H)$ have dimension at least 2, so in any basis of H' including u', there is some basis vector e'_2 independent of u', and we can rescale e'_2 in such a way that the matrix of g over the two bases has determinant $+1$. □

7.16 Summary

The main concepts and results of this chapter are listed below:

- One does not solve (large) linear systems by computing determinants.
- *Upper-triangular* (*lower-triangular*) matrices.
- Solving by *back-substitution* (*forward-substitution*).
- *Gaussian elimination*.
- Permuting rows.
- The *pivot* of an elimination step; *pivoting*.
- *Transposition matrix*; *elementary matrix*.
- The *Gaussian elimination theorem* (Theorem 7.1).
- *Gauss-Jordan factorization*.
- *LU-factorization*; Necessary and sufficient condition for the existence of an LU-factorization (Proposition 7.2).
- *LDU-factorization*.
- "$PA = LU$ theorem" (Theorem 7.5).
- $LDL^\top$*-factorization* of a symmetric matrix.

- Avoiding small pivots: *partial pivoting*; *complete pivoting*.
- Gaussian elimination of tridiagonal matrices.
- LU-factorization of tridiagonal matrices.
- *Symmetric positive definite* matrices (SPD matrices).
- *Cholesky factorization* (Theorem 7.10).
- Criteria for a symmetric matrix to be positive definite; *Sylvester's criterion*.
- *Reduced row echelon form*.
- Reduction of a rectangular matrix to its row echelon form.
- Using the reduction to row echelon form to decide whether a system $Ax = b$ is solvable, and to find its solutions, using a *special* solution and a basis of the *homogeneous system* $Ax = 0$.
- *Magic squares*.
- *Transvections and dilatations*.

7.17 Problems

Problem 7.1. Solve the following linear systems by Gaussian elimination:

$$\begin{pmatrix} 2 & 3 & 1 \\ 1 & 2 & -1 \\ -3 & -5 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ 2 \\ -7 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 6 \\ 9 \\ 14 \end{pmatrix}.$$

Problem 7.2. Solve the following linear system by Gaussian elimination:

$$\begin{pmatrix} 1 & 2 & 1 & 1 \\ 2 & 3 & 2 & 3 \\ -1 & 0 & 1 & -1 \\ -2 & -1 & 4 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 7 \\ 14 \\ -1 \\ 2 \end{pmatrix}.$$

Problem 7.3. Consider the matrix

$$A = \begin{pmatrix} 1 & c & 0 \\ 2 & 4 & 1 \\ 3 & 5 & 1 \end{pmatrix}.$$

When applying Gaussian elimination, which value of c yields zero in the second pivot position? Which value of c yields zero in the third pivot position? In this case, what can your say about the matrix A?

Problem 7.4. Solve the system

$$\begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$$

using the LU-factorization of Example 7.1.

Problem 7.5. Apply **rref** to the matrix

$$A_2 = \begin{pmatrix} 1 & 2 & 1 & 1 \\ 2 & 3 & 2 & 3 \\ -1 & 0 & 1 & -1 \\ -2 & -1 & 3 & 0 \end{pmatrix}.$$

Problem 7.6. Apply **rref** to the matrix

$$\begin{pmatrix} 1 & 4 & 9 & 16 \\ 4 & 9 & 16 & 25 \\ 9 & 16 & 25 & 36 \\ 16 & 25 & 36 & 49 \end{pmatrix}.$$

Problem 7.7. (1) Prove that the dimension of the subspace of 2×2 matrices A, such that the sum of the entries of every row is the same (say c_1) and the sum of entries of every column is the same (say c_2) is 2.

(2) Prove that the dimension of the subspace of 2×2 matrices A, such that the sum of the entries of every row is the same (say c_1), the sum of entries of every column is the same (say c_2), and $c_1 = c_2$ is also 2. Prove that every such matrix is of the form

$$\begin{pmatrix} a & b \\ b & a \end{pmatrix},$$

and give a basis for this subspace.

(3) Prove that the dimension of the subspace of 3×3 matrices A, such that the sum of the entries of every row is the same (say c_1), the sum of entries of every column is the same (say c_2), and $c_1 = c_2$ is 5. Begin by showing that the above constraints are given by the set of equations

$$\begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \\ 0 & 1 & 1 & -1 & 0 & 0 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{21} \\ a_{22} \\ a_{23} \\ a_{31} \\ a_{32} \\ a_{33} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Prove that every matrix satisfying the above constraints is of the form

$$\begin{pmatrix} a+b-c & -a+c+e & -b+c+d \\ -a-b+c+d+e & a & b \\ c & d & e \end{pmatrix},$$

with $a, b, c, d, e \in \mathbb{R}$. Find a basis for this subspace. (Use the method to find a basis for the kernel of a matrix).

Problem 7.8. If A is an $n \times n$ symmetric matrix and B is any $n \times n$ invertible matrix, prove that A is positive definite iff $B^\top AB$ is positive definite.

Problem 7.9. (1) Consider the matrix

$$A_4 = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix}.$$

Find three matrices of the form $E_{2,1;\beta_1}, E_{3,2;\beta_2}, E_{4,3;\beta_3}$, such that

$$E_{4,3;\beta_3}E_{3,2;\beta_2}E_{2,1;\beta_1}A_4 = U_4$$

where U_4 is an upper triangular matrix. Compute

$$M = E_{4,3;\beta_3}E_{3,2;\beta_2}E_{2,1;\beta_1}$$

and check that

$$MA_4 = U_4 = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 0 & 3/2 & -1 & 0 \\ 0 & 0 & 4/3 & -1 \\ 0 & 0 & 0 & 5/4 \end{pmatrix}.$$

(2) Now consider the matrix

$$A_5 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}.$$

Find four matrices of the form $E_{2,1;\beta_1}, E_{3,2;\beta_2}, E_{4,3;\beta_3}, E_{5,4;\beta_4}$, such that

$$E_{5,4;\beta_4}E_{4,3;\beta_3}E_{3,2;\beta_2}E_{2,1;\beta_1}A_5 = U_5$$

where U_5 is an upper triangular matrix. Compute

$$M = E_{5,4;\beta_4}E_{4,3;\beta_3}E_{3,2;\beta_2}E_{2,1;\beta_1}$$

and check that

$$MA_5 = U_5 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 \\ 0 & 3/2 & -1 & 0 & 0 \\ 0 & 0 & 4/3 & -1 & 0 \\ 0 & 0 & 0 & 5/4 & -1 \\ 0 & 0 & 0 & 0 & 6/5 \end{pmatrix}.$$

(3) Write a `Matlab` program defining the function Ematrix(n, i, j, b) which is the $n \times n$ matrix that adds b times row j to row i. Also write some `Matlab` code that produces an $n \times n$ matrix A_n generalizing the matrices A_4 and A_5.

Use your program to figure out which five matrices $E_{i,j;\beta}$ reduce A_6 to the upper triangular matrix

$$U_6 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 3/2 & -1 & 0 & 0 & 0 \\ 0 & 0 & 4/3 & -1 & 0 & 0 \\ 0 & 0 & 0 & 5/4 & -1 & 0 \\ 0 & 0 & 0 & 0 & 6/5 & -1 \\ 0 & 0 & 0 & 0 & 0 & 7/6 \end{pmatrix}.$$

Also use your program to figure out which six matrices $E_{i,j;\beta}$ reduce A_7 to the upper triangular matrix

$$U_7 = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3/2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4/3 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5/4 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6/5 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 7/6 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 8/7 \end{pmatrix}.$$

(4) Find the lower triangular matrices L_6 and L_7 such that

$$L_6 U_6 = A_6$$

and

$$L_7 U_7 = A_7.$$

(5) It is natural to conjecture that there are $n-1$ matrices of the form $E_{i,j;\beta}$ that reduce A_n to the upper triangular matrix

$$U_n = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3/2 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4/3 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5/4 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6/5 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & (n+1)/n \end{pmatrix},$$

namely,

$$E_{2,1;1/2}, E_{3,2;2/3}, E_{4,3;3/4}, \cdots, E_{n,n-1;(n-1)/n}.$$

It is also natural to conjecture that the lower triangular matrix L_n such that

$$L_n U_n = A_n$$

is given by

$$L_n = E_{2,1;-1/2} E_{3,2;-2/3} E_{4,3;-3/4} \cdots E_{n,n-1;-(n-1)/n},$$

that is,

$$L_n = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1/2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2/3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -3/4 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -4/5 & 1 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & 0 & 0 & \cdots & -(n-1)/n & 1 \end{pmatrix}.$$

Prove the above conjectures.

(6) Prove that the last column of A_n^{-1} is

$$\begin{pmatrix} 1/(n+1) \\ 2/(n+1) \\ \vdots \\ n/(n+1) \end{pmatrix}.$$

Problem 7.10. (1) Let A be any invertible 2×2 matrix

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

Prove that there is an invertible matrix S such that

$$SA = \begin{pmatrix} 1 & 0 \\ 0 & ad - bc \end{pmatrix},$$

where S is the product of at most four elementary matrices of the form $E_{i,j;\beta}$.

Conclude that every matrix A in $\mathbf{SL}(2)$ (the group of invertible 2×2 matrices A with $\det(A) = +1$) is the product of at most four elementary matrices of the form $E_{i,j;\beta}$.

For any $a \neq 0, 1$, give an explicit factorization as above for

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

What is this decomposition for $a = -1$?

(2) Recall that a rotation matrix R (a member of the group $\mathbf{SO}(2)$) is a matrix of the form

$$R = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

Prove that if $\theta \neq k\pi$ (with $k \in \mathbb{Z}$), any rotation matrix can be written as a product

$$R = ULU,$$

where U is upper triangular and L is lower triangular of the form

$$U = \begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix}, \quad L = \begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix}.$$

Therefore, every plane rotation (except a flip about the origin when $\theta = \pi$) can be written as the composition of three shear transformations!

Problem 7.11. (1) Recall that $E_{i,d}$ is the diagonal matrix

$$E_{i,d} = \mathrm{diag}(1, \ldots, 1, d, 1, \ldots, 1),$$

whose diagonal entries are all $+1$, except the (i,i)th entry which is equal to d.

Given any $n \times n$ matrix A, for any pair (i,j) of distinct row indices ($1 \leq i, j \leq n$), prove that there exist two elementary matrices $E_1(i,j)$ and $E_2(i,j)$ of the form $E_{k,\ell;\beta}$, such that

$$E_{j,-1}E_1(i,j)E_2(i,j)E_1(i,j)A = P(i,j)A,$$

the matrix obtained from the matrix A by permuting row i and row j. Equivalently, we have

$$E_1(i,j)E_2(i,j)E_1(i,j)A = E_{j,-1}P(i,j)A,$$

the matrix obtained from A by permuting row i and row j and multiplying row j by -1.

Prove that for every $i = 2, \ldots, n$, there exist four elementary matrices $E_3(i,d), E_4(i,d)$, $E_5(i,d), E_6(i,d)$ of the form $E_{k,\ell;\beta}$, such that

$$E_6(i,d)E_5(i,d)E_4(i,d)E_3(i,d)E_{n,d} = E_{i,d}.$$

What happens when $d = -1$, that is, what kind of simplifications occur?

Prove that all permutation matrices can be written as products of elementary operations of the form $E_{k,\ell;\beta}$ and the operation $E_{n,-1}$.

(2) Prove that for every invertible $n \times n$ matrix A, there is a matrix S such that

$$SA = \begin{pmatrix} I_{n-1} & 0 \\ 0 & d \end{pmatrix} = E_{n,d},$$

with $d = \det(A)$, and where S is a product of elementary matrices of the form $E_{k,\ell;\beta}$.

In particular, every matrix in $\mathbf{SL}(n)$ (the group of invertible $n \times n$ matrices A with $\det(A) = +1$) can be written as a product of elementary matrices of the form $E_{k,\ell;\beta}$. Prove that at most $n(n+1) - 2$ such transformations are needed.

(3) Prove that every matrix in $\mathbf{SL}(n)$ can be written as a product of at most $(n-1)(\max\{n,3\}+1)$ elementary matrices of the form $E_{k,\ell;\beta}$.

Problem 7.12. A matrix A is called *strictly column diagonally dominant* iff

$$|a_{j\,j}| > \sum_{i=1,\, i\neq j}^{n} |a_{i\,j}|, \quad \text{for } j = 1, \ldots, n$$

Prove that if A is strictly column diagonally dominant, then Gaussian elimination with partial pivoting does not require pivoting, and A is invertible.

Problem 7.13. (1) Find a lower triangular matrix E such that

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

(2) What is the effect of the product (on the left) with

$$E_{4,3;-1}E_{3,2;-1}E_{4,3;-1}E_{2,1;-1}E_{3,2;-1}E_{4,3;-1}$$

on the matrix

$$Pa_3=\begin{pmatrix}1&0&0&0\\1&1&0&0\\1&2&1&0\\1&3&3&1\end{pmatrix}.$$

(3) Find the inverse of the matrix Pa_3.

(4) Consider the $(n+1)\times(n+1)$ Pascal matrix Pa_n whose ith row is given by the binomial coefficients

$$\binom{i-1}{j-1},$$

with $1\leq i\leq n+1$, $1\leq j\leq n+1$, and with the usual convention that

$$\binom{0}{0}=1,\quad \binom{i}{j}=0\quad\text{if}\quad j>i.$$

The matrix Pa_3 is shown in Question (c) and Pa_4 is shown below:

$$Pa_4=\begin{pmatrix}1&0&0&0&0\\1&1&0&0&0\\1&2&1&0&0\\1&3&3&1&0\\1&4&6&4&1\end{pmatrix}.$$

Find n elementary matrices $E_{i_k,j_k;\beta_k}$ such that

$$E_{i_n,j_n;\beta_n}\cdots E_{i_1,j_1;\beta_1}Pa_n=\begin{pmatrix}1&0\\0&Pa_{n-1}\end{pmatrix}.$$

Use the above to prove that the inverse of Pa_n is the lower triangular matrix whose ith row is given by the signed binomial coefficients

$$(-1)^{i+j-2}\binom{i-1}{j-1},$$

with $1 \leq i \leq n+1$, $1 \leq j \leq n+1$. For example,

$$Pa_4^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 \\ -1 & 3 & -3 & 1 & 0 \\ 1 & -4 & 6 & -4 & 1 \end{pmatrix}.$$

Hint. Given any $n \times n$ matrix A, multiplying A by the elementary matrix $E_{i,j;\beta}$ *on the right* yields the matrix $AE_{i,j;\beta}$ in which β times the ith column is added to the jth column.

Problem 7.14. (1) Implement the method for converting a rectangular matrix to reduced row echelon form in `Matlab`.

(2) Use the above method to find the inverse of an invertible $n \times n$ matrix A by applying it to the the $n \times 2n$ matrix $[A\, I]$ obtained by adding the n columns of the identity matrix to A.

(3) Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & \cdots & n \\ 2 & 3 & 4 & 5 & \cdots & n+1 \\ 3 & 4 & 5 & 6 & \cdots & n+2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ n & n+1 & n+2 & n+3 & \cdots & 2n-1 \end{pmatrix}.$$

Using your program, find the row reduced echelon form of A for $n = 4, \ldots, 20$.

Also run the `Matlab rref` function and compare results.

Your program probably disagrees with `rref` even for small values of n. The problem is that some pivots are very small and the normalization step (to make the pivot 1) causes roundoff errors. Use a tolerance parameter to fix this problem.

What can you conjecture about the rank of A?

(4) Prove that the matrix A has the following row reduced form:

$$R = \begin{pmatrix} 1 & 0 & -1 & -2 & \cdots & -(n-2) \\ 0 & 1 & 2 & 3 & \cdots & n-1 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Deduce from the above that A has rank 2.

Hint. Some well chosen sequence of row operations.

(5) Use your program to show that if you add any number greater than or equal to $(2/25)n^2$ to every diagonal entry of A you get an invertible matrix! In fact, running the `Matlab` fuction `chol` should tell you that these matrices are SPD (symmetric, positive definite).

Problem 7.15. Let A be an $n \times n$ complex Hermitian positive definite matrix. Prove that the lower-triangular matrix B with positive diagonal entries such that $A = BB^*$ is given by the following formulae: For $j = 1, \ldots, n$,

$$b_{j\,j} = \left(a_{j\,j} - \sum_{k=1}^{j-1} |b_{j\,k}|^2 \right)^{1/2},$$

and for $i = j+1, \ldots, n$ (and $j = 1, \ldots, n-1$)

$$b_{i\,j} = \left(a_{i\,j} - \sum_{k=1}^{j-1} b_{i\,k} b_{j\,k} \right) / b_{j\,j}.$$

Problem 7.16. (Permutations and permutation matrices) A permutation can be viewed as an operation permuting the rows of a matrix. For example, the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 \\ 3 & 4 & 2 & 1 \end{pmatrix}$$

corresponds to the matrix

$$P_\pi = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Observe that the matrix P_π has a single 1 on every row and every column, all other entries being zero, and that if we multiply any 4×4 matrix A by P_π on the left, then the rows of A are permuted according to the permutation π; that is, the $\pi(i)$th row of $P_\pi A$ is the ith row of A. For example,

$$P_\pi A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{pmatrix}.$$

Equivalently, the ith row of $P_\pi A$ is the $\pi^{-1}(i)$th row of A. In order for the matrix P_π to move the ith row of A to the $\pi(i)$th row, the $\pi(i)$th row of P_π must have a 1 in column i and zeros everywhere else; this means that the ith column of P_π contains the basis vector $e_{\pi(i)}$, the vector that has a 1 in position $\pi(i)$ and zeros everywhere else.

This is the general situation and it leads to the following definition.

Definition 7.8. Given any permutation $\pi \colon [n] \to [n]$, the *permutation matrix* $P_\pi = (p_{ij})$ representing π is the matrix given by

$$p_{ij} = \begin{cases} 1 & \text{if } i = \pi(j) \\ 0 & \text{if } i \neq \pi(j); \end{cases}$$

equivalently, the jth column of P_π is the basis vector $e_{\pi(j)}$. A *permutation matrix* P is any matrix of the form P_π (where P is an $n \times n$ matrix, and $\pi\colon [n] \to [n]$ is a permutation, for some $n \geq 1$).

Remark: There is a confusing point about the notation for permutation matrices. A permutation matrix P acts on a matrix A by multiplication on the left by permuting the rows of A. As we said before, this means that the $\pi(i)$th row of $P_\pi A$ is the ith row of A, or equivalently that the ith row of $P_\pi A$ is the $\pi^{-1}(i)$th row of A. But then observe that the row index of the entries of the ith row of PA is $\pi^{-1}(i)$, and not $\pi(i)$! See the following example:

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = \begin{pmatrix} a_{41} & a_{42} & a_{43} & a_{44} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{pmatrix},$$

where

$$\begin{aligned} \pi^{-1}(1) &= 4 \\ \pi^{-1}(2) &= 3 \\ \pi^{-1}(3) &= 1 \\ \pi^{-1}(4) &= 2. \end{aligned}$$

Prove the following results

(1) Given any two permutations $\pi_1, \pi_2\colon [n] \to [n]$, the permutation matrix $P_{\pi_2 \circ \pi_1}$ representing the composition of π_1 and π_2 is equal to the product $P_{\pi_2} P_{\pi_1}$ of the permutation matrices P_{π_1} and P_{π_2} representing π_1 and π_2; that is,

$$P_{\pi_2 \circ \pi_1} = P_{\pi_2} P_{\pi_1}.$$

(2) The matrix $P_{\pi_1^{-1}}$ representing the inverse of the permutation π_1 is the inverse $P_{\pi_1}^{-1}$ of the matrix P_{π_1} representing the permutation π_1; that is,

$$P_{\pi_1^{-1}} = P_{\pi_1}^{-1}.$$

Furthermore,

$$P_{\pi_1}^{-1} = (P_{\pi_1})^\top.$$

(3) Prove that if P is the matrix associated with a transposition, then $\det(P) = -1$.

(4) Prove that if P is a permutation matrix, then $\det(P) = \pm 1$.

(5) Use permutation matrices to give another proof of the fact that the parity of the number of transpositions used to express a permutation π depends only on π.