HELIANTUS SCHOOL

WARSAW BAŹANCIA 16

LECTURE NOTES
IN NUMERICAL METHODS
OF
LINEAR ALGEBRA

T. STYŠ Ph.D. University of Warsaw

Warsaw, May 2006

PREFACE

This text is to introduce students of science and engineering to computational linear algebra. As a prerequisite to the numerical methods basic knowledge of linear algebra and computing are required. Also, the text assumes a previous knowledge on *Mathematica* as the systems for doing mathematics with computers. So, it is taken for granted that the reader has access to computer facilities for solving some of examples and exercise questions.
The text contains classical methods for solving linear systems of equations with emphasis put on error analysis, algorithm design and their implementation in computer arithmetic. There is also a desire that the reader will find interesting theorems with examples solved by included *Mathematica* modules.
The text begins with the notions and theorems concerning norms and operations on vectors and matrices. In the chapter 2, direct methods for solving linear systems of equations based on Gauss elimination are described and supported by examples and *Mathematica* programs.
The chapter 3, contains standard methods for solving eigenvalue problems for quadratic matrices. It includes Jacobi method, power method, and QR method with examples, questions and *Mathematica* modules.

Iterative methods for solving linear systems of equations are presented in the chapter 4. It starts with the sufficient and necessary condition for convergence of linear stationary one step methods. The class of linear stationary one step methods includes iterative Jacobi and Gauss Seidel methods, Successive Over relaxation method (SOR), Alternating directions method (ADI) and Gradient method (CG).

STYŠ Tadeusz

Contents

1 Vectors and Matrices **1**
- 1.1 Vector and Matrix Norms . . . 1
- 1.2 Conditional Number of a Matrix . . . 3
- 1.3 Positive Definite Matrices . . . 5
- 1.4 Diagonally Dominant Matrices . . . 6
- 1.5 Monotone Matrices . . . 8
- 1.6 Matrices of Positive Type . . . 9
- 1.7 Exercises . . . 11

2 Systems of Linear Equations **13**
- 2.1 Gauss Elimination Method . . . 13
- 2.2 Partial Pivoting . . . 20
- 2.3 Principal Element Strategy . . . 25
- 2.4 LU-Decomposition. . . . 28
- 2.5 Root Square Method . . . 30
- 2.6 Gauss Elimination for Tri-diagonal Matrices. . . . 32
- 2.7 Gauss Elimination for Block Tri-diagonal Matrices . . . 34
- 2.8 Gauss Elimination for Pentediagonal Matrices. . . . 39
- 2.9 Exercises . . . 42

3 Eigenvalues and Eigenvectors of a Matrix **47**
- 3.1 Eigenvalue Problem . . . 47
- 3.2 Jacobi Method for Real and Symmetric Matrices . . . 52
- 3.3 Power Method . . . 60
- 3.4 The Householder Transformation and Hessenberg Matrices . . . 64
- 3.5 QR Method . . . 70
- 3.6 Exercises . . . 79

4 Iterative Methods for Systems of Linear Equations **81**
- 4.1 Stationary One Step Linear Methods . . . 81
- 4.2 Jacobi Iterative Method . . . 83
- 4.3 Gauss Seidel Iterative Method . . . 86
- 4.4 Successive Overrelaxation Method (SOR) . . . 91
- 4.5 Alternating Direction Implicit Method (ADI) . . . 95

4.6 Conjugate Gradient Method (CG) 98
4.7 Exercises 103
4.8 References 106

Chapter 1

Vectors and Matrices

1.1 Vector and Matrix Norms

Let $x = (x_1, x_2, ..., x_n) \in R^n$ be a vector. Below, we shall consider the following three vector norms:

1. $\| x \|_S = \sqrt{| x_1 |^2 + | x_2 |^2 \cdots + | x_n |^2}$,
2. $\| x \|_1 = | x_1 | + | x_2 | \cdots + | x_n |$,
3. $\| x \|_\infty = \max\limits_{1 \le i \le n} | x_i |$.

The above vector norms satisfy the inequalities:

$$\begin{aligned} &\| x \|_S \le \| x \|_1 \le \sqrt{n} \| x \|_S, \\ &\| x \|_\infty \le \| x \|_S \le \sqrt{n} \| x \|_\infty, \\ &\| x \|_\infty \le \| x \|_1 \le n \| x \|_\infty, \end{aligned} \tag{1.1}$$

Let us note that if $\overline{x}$ is an approximate vector to a vector x then
the absolute error

$$\epsilon_x = \| \overline{x} - x \|$$

and *the relative error*

$$\delta_x = \frac{\| \overline{x} - x \|}{\| x \|}, \qquad x \neq 0.$$

Evidently, the relative error measured in the $\infty-$ norm expresses the number of correct significant digits of the largest component of the approximate $\overline{x}$. For instance, if

$$\frac{\| \overline{x} - x \|_\infty}{\| x \|_\infty} \approx 10^{-6}$$

then $\overline{x}$ should have 6 correct significant digits. If the norms $\| - \|_S$ or $\| - \|_1$ are used then all components of $\overline{x}$ may be biased by the error $\delta_x = 10^{-p}$.

Therefore $\overline{x}$ may have p correct significant digits.
A norm of a matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n-1} & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn-1} & a_{nn} \end{bmatrix}$$

is determined by the following relation:

$$\| A \| = \sup_{x \neq 0} \frac{\| Ax \|}{\| x \|}.$$

This means that the norm $\| A \|$ is the smallest constant for which the inequality

$$\| Ax \| \leq \| A \| \| x \|$$

holds for every $x \in R^n$.
One can show that the subordinated matrix norms to the three vector norms are:

1. (a) $\| A \|_S = \max\limits_{1 \leq i \leq n} \sqrt{\lambda_i(AA^T)}$, is the spectral norm of A, where $\lambda_i(AA^T)$ is the $i-th$ eigenvalue of the matrix AA^T,

   (b) $\| A \|_\infty = \max\limits_{1 \leq i \leq n} \sum\limits_{j=1}^{n} | a_{ij} |$ is the $\infty-$ norm of A,

   (c) $\| A \|_1 = \max\limits_{1 \leq j \leq n} \sum\limits_{j=1}^{n} | a_{ij} |$ is the first norm of A,

The above matrix norms satisfy the following inequalities:

1. (a) i. $\frac{1}{\sqrt{n}} \| A \|_\infty \leq \| A \|_S \leq \sqrt{n} \| A \|_\infty$
   ii. $\frac{1}{\sqrt{n}} \| A \|_1 \leq \| A \|_S \leq \sqrt{n} \| A \|_1$.

Let us note that if A is a symmetric matrix then the spectral norm of A is equal to the spectral radius $\rho(A)$, i.e.

$$\| A \|_S = \rho(A),$$

where $\rho(A) = \max\limits_{1 \leq i \leq n} | \lambda_i(A) |$, $\lambda_i(A)$ is $i-th$ eigenvalue of A.

1.2 Conditional Number of a Matrix

A resistance of a matrix A against perturbation of input data and round-off errors of partial results is measured by its conditional number

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

In case of a symmetric matrix A, where the spectral norm is involved, the conditional number of the matrix A is given by the formula

$$Cond(A) = \rho(A)\rho(A^{-1}).$$

A large conditional number $Cond(A)$ strongly affects final results of any algorithm that involves the matrix A.
For example, let the matrix

$$A = \begin{bmatrix} 1 & 0.99999 \\ 0.99999 & 1 \end{bmatrix}.$$

One may find that the eigenvalues of A and A^{-1} are:

$$\lambda_1(A) \approx 2, \quad \lambda_2(A) \approx 0 \quad \text{and} \quad \rho(A) = 2,$$

$$\lambda_1(A^{-1}) \approx 100000, \quad \lambda_2(A^{-1}) \approx 0.5 \ \text{ and } \ \rho(A^{-1}) = 100000.$$

Hence, the conditional number

$$Cond(A) = 200000.$$

Now, let us solve the following system of linear equations:

$$\begin{array}{rcrcl} x_1 & + & 0.99999x_2 & = & 2.99999 \\ 0.99999x_1 & - & x_2 & = & 0.99998 \end{array}$$

The solution of the above system of equations is:

$$x_1 = 2 \ \text{ and } \ x_2 = 1.$$

Changing the coefficient at x_1 in the first equation, by $\epsilon = 0.00001$, we obtain the following system of linear equations:

$$\begin{array}{rcrcl} 0.99999x_1 & + & 0.99999x_2 & = & 2.99999 \\ 0.99999x_1 & - & x_2 & = & 0.99998 \end{array}$$

We observe that the solution of the this system of equations

$$\overline{x}_1 = 100003, \quad \overline{x}_2 = -100001.$$

differs considerably from the solution of the original system of equations, in spite of very small change in the coefficient at x_1. This is due to the large

conditional number ($Cond(A) = 200000$) of the matrix A.
Solving numerically a system of linear equations

$$Ax = b,$$

we find an approximate solution $\overline{x}$ which is the exact solution of the system of equations

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

Having $\overline{x}$, we can compute the residual absolute error

$$r_b = Ax - A\overline{x} = b - \overline{b},$$

and the residual relative error

$$\delta_b = \frac{||b - \overline{b}||}{||b||}, \qquad b \neq 0.$$

The relative error

$$\delta_x = \frac{||x - \overline{x}||}{||x||}, \qquad x \neq 0,$$

satisfies the following inequality

$$\frac{\delta_b}{Cond(A)} \leq \delta_x \leq Cond(A)\ \delta_b.$$

Indeed, we have

$$A(x - \overline{x}) = b - \overline{b}, \qquad x - \overline{x} = A^{-1}(b - \overline{b}),$$

and

$$||A||\ ||x - \overline{x}|| \geq ||b - \overline{b}||, \qquad ||x - \overline{x}|| \leq ||A^{-1}||\ ||b - \overline{b}||.$$

Hence, we get

$$\frac{||b - \overline{b}||}{||A||\ ||x||} \leq \frac{||x - \overline{x}}{||x||} \leq \frac{||A^{-1}||\ ||b - \overline{b}||}{||x||}, \qquad x \neq 0.$$

Clearly, the solution x satisfies the inequality

$$\frac{||b||}{||A||} \leq ||x|| \leq ||A^{-1}||\ ||b||.$$

Combining the above inequalities, we obtain

$$\frac{1}{||A^{-1}||\ ||A||}\frac{||b - \overline{b}||}{||b||} \leq \frac{||x - \overline{x}||}{||x||} \leq ||A^{-1}||\ ||A||\ \frac{||b - \overline{b}||}{||b||}, \qquad x \neq 0,\ \ b \neq 0.$$

1.3 Positive Definite Matrices

A class of positive definite matrices plays an important role in different areas of mathematics (statistics, numerical analysis, differential equations, mechanics, algebra, geometry, etc.). Here, we shall consider positive definite matrices from the numerical point of view. As it is known, the most effective numerical methods for linear systems of equations are associated with positive definite matrices. Let us present the following definition:

Definition 1.1 *A matrix A is said to be positive definite if and only if the following conditions hold:*

1. *A is a symmetric matrix, i.e. $A^T = A$,*
2. *there exists a constant $\gamma > 0$ such that*

$$(A, x, x) = \sum_{i,j=1}^{n} a_{ij} x_i x_j \geq \gamma \sum_{i=1}^{n} x_i^2 = \gamma \, (x, x)$$

for every real vector $x = (x_1, x_2, ..., x_n) \in R^n$.

Example 1.1 *The matrix*

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

is positive definite.

Evidently, A is a symmetric matrix, i.e. $A^T = A$ and

$$(Ax, x) = 4(x_1^2 - x_1 x_2 + x_2^2) \geq 2(x_1^2 + x_2^2) = 2(x, x)$$

for every $x = (x_1, x_2) \in R^2$. So that $\gamma = 2$.

The following theorem holds:

Theorem 1.1 *An matrix A is positive definite if and only if all its eigenvalues are real and positive, i.e. $\lambda_1 > 0,\ \lambda_2 > 0, ..., \lambda_n > 0$.*

Proof. At first, let us assume that A is a positive definite matrix. Then, by condition 1, A is a symmetric matrix. Therefore all eigenvalues of A are real and eigenvectors $x^{(1)}, x^{(2)}, ..., x^{(n)}$ of A are orthonormal in the real space R^n. Evidently, for each eigenvector $x^{(k)},\ k = 1, 2, .., n$; we have

$$0 < (Ax^{(k)}, x^{(k)}) = \lambda_k (x^{(k)}, x^{(k)}) = \lambda_k, \quad k = 1, 2, ..., n.$$

Now, let us assume that all eigenvalues of A are real and positive, i.e.

$$0 < \lambda_1 < \lambda_2 <, ..., \lambda_n.$$

Then, each vector $x \neq 0$ can be presented in the form of the following linear combination:

$$x = \alpha_1 x^{(1)} + \alpha_2 x^{(2)} + \ldots + \alpha_n x^{(n)},$$

where $(x, x) = \alpha_1^2 + \alpha_2^2 + \ldots + \alpha_n^2 > 0$.
We thus have

$$\begin{aligned}(Ax, x) &= (A\sum_{j=1}^{n} \alpha_j x^{(j)}, \sum_{j=1}^{n} \alpha_j x^{(j)}) \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} \alpha_j \alpha_k (Ax^{(j)}, x^{(k)}) \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} \alpha_j \alpha_k \lambda_j (x^{(j)}, x^{(k)}) \\ &= \sum_{j=1}^{n} \lambda_j \alpha_j^2 \geq \lambda_1 \sum_{j=1}^{n} \alpha_j^2 = \lambda_1 (x, x).\end{aligned}$$

Hence $\gamma = \lambda_1$. End of the proof.

1.4 Diagonally Dominant Matrices

Below, we shall show that the class of diagonally dominant matrices is a sub class of the class of positive definite matrices.

Definition 1.2 *A matrix A is said to be diagonally positive dominant if and only if the following conditions hold:*

1. *(a)* $a_{ii} \geq \sum_{j=1, j\neq i}^{n} \mid a_{ij} \mid, \quad i = 1, 2, ..., n,$

   (b) there exists at least one natural i for which

   $$a_{ii} > \sum_{j=1, j\neq i}^{n} \mid a_{ij} \mid,$$

   (c) if condition (b) is satisfied for all $i = 1, 2, .., n$, then A is called a strongly diagonally dominant matrix.

Example 1.2 *Evidently, the matrix*

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

satisfies the conditions (a) *and* (b) *of definition and therefore,* A *is a diagonally positive dominant matrix.*

Let us note that conditions (a) and (b) are not difficult to check. Thus, we may easily find whether a matrix is or is not diagonally dominant. We may also use these conditions to determine whether a matrix is positive definite applying the following theorem:

Theorem 1.2 *Every non-singular symmetric and diagonally dominant matrix* A *is positive definite.*

Proof. It is sufficient to show that all eigenvalues of matrix A are real and positive. Then, by the theorem , A is a positive definite matrix. By assumption, A is a non-singular and symmetric matrix, therefore its eigenvalues are real and different from zero.

Now, we shall show that A does not have a negative eigenvalue. Evidently, for every negative $\lambda < 0$, the matrix $A - \lambda E$ is strongly diagonally dominant. Therefore, the homogeneous system of linear equations

$$(A - \lambda E)y = 0$$

has only one solution, i.e. $y = 0$. Indeed, let

$$\max_{1 \le i \le n} \mid y_i \mid = \mid y_k \mid .$$

Of course, without any additional restrictions, we may assume that $y_k \ge 0$. Because

$$\begin{aligned} 0 &= a_{k1}y_1 + a_{k2}y_2 + \cdots + (a_{kk} - \lambda)y_k + \cdots + a_{kn}y_n \\ &\ge [(a_{kk} - \lambda) - \sum_{j=1, j \ne i}^{n} |a_{kj}]y_k \ge 0, \end{aligned}$$

we get

$$[a_{kk} - \lambda - \sum_{j=1, j \ne i}^{n} \mid a_{kj}]y_k = 0.$$

However, since $A - \lambda E$, $(\lambda < 0)$ is a strongly diagonally dominant matrix,

$$a_{kk} - \lambda - \sum_{j=1, j \ne i}^{n} \mid a_{kj} \mid > 0.$$

Hence $y_k = 0$ and $y_1 = y_2 = \cdots = y_n = 0$. Thus, the matrix $A - \lambda E$ is non-singular for every negative $\lambda < 0$. This means that the matrix A has all positive eigenvalues. Finally, by the theorem , A is a positive definite matrix.

1.5 Monotone Matrices

Let us write the inequality $A \geq 0$ if all entries of the matrix A are non-negative, i.e. $a_{ij} \geq 0, \quad i, j = 1, 2..., n$. The monotone matrices are then defined as follows:

Definition 1.3 *A matrix A is said to be monotone if and only if the following implication his true:*

$$Ax \geq 0 \quad \textit{implies the inequality } x \geq 0.$$

The following theorem holds (cf. [18], [20]):

Theorem 1.3 *A is a monotone matrix if and only if A is non-singular and the inverse matrix to A satisfies the inequality $A^{-1} \geq 0$.*

Proof. At first, let us assume that A is a monotone matrix in the sense of definition. Then, the homogeneous system of linear equations $Ax = 0$ has only one solution $x = 0$. Indeed, by assumption

$$\text{the inequality } A(\pm x) \geq 0 \text{ implies the inequality } \pm x \geq 0.$$

Hence $x = 0$ and therefore A is a non-singular matrix.
Let z be a column of the inverse matrix A^{-1}. Then

$$Az = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix} \geq 0, \qquad z \geq 0.$$

Thus, the inverse matrix $A^{-1} \geq 0$.
Now, let us assume that A is a non-singular matrix and the inverse matrix $A^{-1} \geq 0$. Then, for $Ax \geq 0$, we have

$$x = A^{-1}Ax \geq 0.$$

Example 1.3 *The matrix*

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

is monotone.

Indeed, we have

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

1.6 Matrices of Positive Type

In general, it is not easy to determine, by definition, whether a matrix A is monotone or not. However, there are so called "matrices of positive type" which create a sub class of the class of all monotone matrices. The matrices of positive type are easy to investigate following the conditions of the definition:

Definition 1.4 *(cf. [18]). A matrix A is said to be of positive type if and only if the following conditions hold:*

1. *(a)* $a_{ij} \leq 0$ *for* $i \neq j$,

   (b) $\sum_{j=1}^{n} a_{ij} \geq 0$,

   (c) there exists a non-empty subset $J(A)$ of the set $\{1, 2, ..., n\}$ such $\sum_{j=1}^{n} a_{ij} > 0$ *for* $i \in J(A)$,

   (d) for every $k \in J(A)$ there exists $l \in J(A)$ and a sequence non-zero entries of the form $a_{kk_1}, a_{k_1k_2}, a_{k_2k_3}, ..., a_{k_rl}$

Let us note that condition (d) can be replaced by the condition:

A is an irreducible matrix (cf. [18]).

Example 1.4 *Let us consider the following matrix:*

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

Evidently, matrix A is monotone, since it satisfies conditions (a) and (b). Also, the conditions (c) and (d) hold for the set $J(A) = \{1, n\}$ while the sequence $a_{kk_1}, a_{k_1k_2}, ..., a_{k_rl} = -1, -1, .. - 1$.

The relation between matrices that are monotone and those of positive type is established in the following theorem (cf. [18]):

Theorem 1.4 *Every matrix of positive type is a monotone matrix.*

Proof. Let us assume that A is a matrix of positive type. Then, by conditions (a) and (b)

$$a_{ii} \geq 0, \quad i = 1, 2, ..., n.$$

If $a_{kk} = 0$ for a $k \in \{1, 2, .., n\}$ then, (also by (a) and (b)),

$$a_{kj} = 0 \quad j = 1, 2, ..., n.$$

However, in this case, condition (d) could not be satisfied. Therefore $a_{kk} > 0$ for all $k = 1, 2, ..., n$.
Now, we shall show that: the inequality $Ax \geq 0$ implies the inequality $x \geq 0$
Indeed, by conditions (a), (b), (c) and (d)

$$x_i \geq \sum_{j=1, j\neq i}^{n} \mid \frac{a_{ij}}{a_{ii}} \mid x_j, \tag{1.2}$$

$$\sum_{j=1, j\neq i}^{n} \mid \frac{a_{ij}}{a_{ii}} \mid < 1 \ \text{ for } \ i \in J(A), \tag{1.3}$$

and

$$\sum_{j=1, j\neq i}^{n} \mid \frac{a_{ij}}{a_{ii}} = 1 \ \text{ for } \ i not \in J(A). \tag{1.4}$$

Let

$$\min_{1\leq j\leq n} x_j = x_i = \alpha < 0.$$

Then, by the above inequalities, we get the following contrary inequality

$$\alpha = x_i \geq \sum_{j=1, j\neq i}^{n} \mid \frac{a_{ij}}{a_{ii}} \mid \alpha > \alpha \ \text{ for } \ i \in J(A).$$

So that $x_i \geq 0$ if $i \in J(A)$.
If $i \in J(A)$ and $k \neq i$ then, by condition (d), $a_{ik} < 0$ and $x_k = \alpha$ for all k such that $a_{ik} < 0$ or it contradicts the inequality (1.2). By condition (d), there exists k_1 such that $a_{kk_1} < 0$. Also, in this case when $x_{k_1} = \alpha$, we have $a_{kk_1} < 0$ or, it contradicts the inequality (1.2). Proceeding in this way, we may find a sequence

$$a_{kk_1}, a_{k_1k_2}, a_{k_2k_3}, ..., a_{k_rl}, \quad l \in J(A),$$

of non-zero entries of A. But then, for $l \in J(A)$, we arrive at a contradiction with inequality (1.2). Therefore, must be $\alpha \geq 0$. Hence $x_k \geq 0$ for all $k = 1, 2, ..., n$.

Example 1.5 *As we know, the following matrix is of positive type:*

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

Therefore, by the theorem, A is a monotone matrix.

1.7 Exercises

Assignment Questions

Question 1. Consider the following matrix

$$A = \begin{Bmatrix} 5 & -2 & -1 & 0 & 0 & \cdots & 0 & 0 \\ -2 & 5 & -2 & -1 & 0 & \cdots & 0 & 0 \\ -1 & -2 & 5 & -2 & -1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -2 & 5 \end{Bmatrix}$$

Show that

(a) A is a diagonally dominant matrix

(b) A is a positive definite matrix

(c) A is of a positive type matrix

(d) A is a monotone matrix

(e) A is a Stieltjes matrix

Question 2.

(2a) Show that if A is a monotone matrix then the inverse matrix A^{-1} exists and $A^{-1} \geq 0$.

(2b) Let the monotone matrices A and B satisfy the inequality $A \geq B$. Show that the inverse matrices A^{-1} and B^{-1} satisfy the inequality

$$B^{-1} \geq A^{-1} \geq 0.$$

(2c) Consider the following matrix

$$A = \begin{Bmatrix} 4 & -1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ -1 & 4 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 4 & -1 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & -1 & 4 \end{Bmatrix}$$

Show that A is a Stjeltjes matrix. Find the a constant $\gamma > 0$ such that $(Ax, x) \geq \gamma(x, x)$ for all real $x = (x_1, x_2, ..., x_n)$.

Quote the definitions and theorems used in the solution.

æ

Chapter 2

Systems of Linear Equations

2.1 Gauss Elimination Method

Introduction. Gauss elimination is a universal direct method. In general, this method can be successfully applied to any linear system of equations, provided that all arithmetic operations involved in the algorithm are not biased by round-off errors. However, it can hardly be the case, since any implementation of the method in a finite arithmetic yields round-off errors. Thus, restrictions are imposed on the class of equations because of computations in a finite arithmetic. For small systems of equations ($n \approx 100$, in 8-digit floating point arithmetic), Gauss method produces acceptable solutions if conditions of stability are satisfied. The number of equations can be considerably greater ($n >> 100$) if partial or full pivoting strategy is applied to a stable system of equations. For systems of equations with sparse matrices, Gauss elimination is also successfully applicable to large systems of linear equations. Applying Gauss elimination to large systems of equations, with multi-diagonal and diagonally dominant matrices, one can obtain a solution for the number of arithmetic operations proportional to the dimension n. This number of operations is significantly lower compared with the total number of arithmetic operations $\approx\ n^3$ that is required in Gauss elimination when it is applied to a system of equations with a full and non-singular matrix A. Although, Gauss method in its general form is costly in terms of arithmetic operations, the method provides $LU - decomposition$ of the matrix A and the determinant $det(A)$, as partial results in computing of the solution x.

Gauss elimination. We shall write a linear system of n equations in the following form:

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n = a_{1n+1}\\
&a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n = a_{2n+1}\\
&a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n = a_{3n+1}\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n = a_{nn+1}
\end{aligned} \tag{2.1}$$

or in the matrix form

$$Ax = a,$$

where the vectors are:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}, \qquad a = \begin{bmatrix} a_{1n+1} \\ a_{2n+1} \\ a_{3n+1} \\ \vdots \\ a_{nn+1} \end{bmatrix}$$

and the matrix is:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n-1} & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn-1} & a_{nn} \end{bmatrix}$$

Within this chapter, we shall assume that the matrix A is non-singular, so that $det(A) \neq 0$.
Let us demonstrate Gauss elimination solving the following sample system of four equations with four unknowns

$$\begin{array}{rcrl} 2x_1 + x_2 + 4x_3 - 3x_4 & = & 4 & \mid m_{21} = 2 \mid m_{31} = 3 \mid m_{41} = 4 \\ 4x_1 - 3x_2 + x_3 - 2x_4 & = & -7 & \\ 6x_1 + 4x_2 - 3x_3 - x_4 & = & 1 & \\ 8x_1 + 2x_2 + x_3 - 2x_4 & = & 7 & \end{array} \tag{2.2}$$

First step of elimination. At first step, we shall eliminate unknown x_1 from second, third and fourth equations. To eliminate x_1 from second equation, we multiply first equation by the coefficient

$$m_{21} = \frac{a_{21}}{a_{11}} = \frac{4}{2} = 2$$

and subtract the result from second equation. Then, we have

$$-5x_2 - 7x_3 + 4x_4 = -15.$$

To eliminate x_1 from third equation, we multiply first equation by the coefficient

$$m_{31} = \frac{a_{31}}{a_{11}} = \frac{6}{2} = 3$$

and subtract the result from third equation. Then, we have

$$x_2 - 15x_3 + 8x_4 = -11.$$

To eliminate x_1 from fourth equation, we multiply first equation by the coefficient

$$m_{41} = \frac{a_{41}}{a_{11}} = \frac{8}{2} = 4$$

and subtract the result from fourth equation. Then, we have

$$-2x_2 - 15x_3 + 10x_4 = -9.$$

After first step of elimination, we arrive at
First reduced system of equations

$$\begin{array}{rrrrl} 2x_1 & +x_2 & +4x_3 & -3x_4 = & 4 \\ & -5x_2 & -7x_3 & +4x_4 = & -15 \\ & x_2 & -15x_3 & +8x_4 = & -11 \\ & -2x_2 & -15x_3 & +10x_4 = & -9 \end{array} \quad \left| \; m_{32} = -\tfrac{1}{5} \; \right| \; m_{42} = \tfrac{2}{5} \tag{2.3}$$

Second step of elimination. At second step, we shall eliminate x_2 in (2.3) from third and fourth equations. To eliminate x_2 from third equation, we multiply second equation by the coefficient

$$m_{32} = \frac{a_{32}^{(1)}}{a_{22}^{(1)}} = \frac{1}{-5}$$

and subtract the result from third equation. Then, we have

$$-\frac{82}{5}x_3 + \frac{44}{5}x_4 = -14.$$

To eliminate x_2 from fourth equation, we multiply second equation by the coefficient

$$m_{42} = \frac{a_{42}^{(1)}}{a_{22}^{(1)}} = \frac{-2}{-5}$$

and subtract the result from fourth equation. Then, we have

$$-\frac{61}{5}x_3 + \frac{42}{5}x_4 = -3.$$

After second step of elimination, we arrive at
Second reduced system of equations

$$\begin{array}{rrrrrrrrl} 2x_1 & + & x_2 & + & 4x_3 & - & 3x_4 = & 4 & \\ & - & 5x_2 & - & 7x_3 & + & 4x_4 = & -15 & \\ & & & - & \frac{82}{5}x_3 & + & \frac{44}{5}x_4 = & -14 & \left| \; m_{43} = \frac{61}{82} \right. \\ & & & - & \frac{61}{5}x_3 & + & \frac{42}{5}x_4 = & -3 & \end{array} \tag{2.4}$$

Third step of elimination. At third step, we shall eliminate x_3 in (2.4) from fourth equation. To eliminate x_3 from fourth equation, we multiply third equation by the coefficient

$$m_{43} = \frac{a_{43}^{(2)}}{a_{33}^{(2)}} = \frac{61}{82}$$

and subtract from fourth equation.
Then, we have

$$\frac{76}{41}x_4 = \frac{304}{41}.$$

Finally, we have arrived at
Third reduced system of equations

$$\begin{array}{rcrcrcrcr} 2x_1 & + & x_2 & + & 4x_3 & - & 3x_4 & = & 4 \\ & - & 5x_2 & - & 7x_3 & + & 4x_4 & = & -15 \\ & & & - & \frac{82}{5}x_3 & + & \frac{44}{5}x_4 & = & -14 \\ & & & & & & \frac{76}{41}x_4 & = & \frac{304}{41} \end{array} \tag{2.5}$$

Let us observe that third reduced system of equations has upper-triangular form and its solution can be easily found by backward substitution. Indeed, from fourth equation

$$x_4 = \frac{\frac{304}{41}}{\frac{76}{41}} = 4,$$

from third equation

$$x_3 = -\frac{5}{82}(-14 - \frac{44}{5}4) = 3,$$

from second equation

$$x_2 = -\frac{1}{5}(-15 + 7 * 3 - 4 * 4) = 2,$$

and from first equation

$$x_1 = \frac{1}{2}(4 - 1 * 2 - 4 * 3 + 3 * 4) = 1.$$

Solving this example with the `Mathematica` program

```
n=4;
a={{2,1,4,-3,4},{4,-3,1,-2,-7},{6,4,-3,-1,1},{8,2,1,-2,7}};
     fi[a_,i_]:=ReplacePart[a,a[[i]]-      a[[s]]*a[[i,s]]/a[[s,s]],i];
     iter[a_,s_]:=Fold[fi,a,Range[s+1,n]];

     Do[a=iter[a,s],{s,1,n}];
     MatrixForm[a]
```

we obtain the upper triangular matrix

$$\begin{array}{ccccc} 2 & 1 & 4 & -3 & 4 \\ 0 & -5 & -7 & 4 & -15 \\ 0 & 0 & -\frac{82}{5} & \frac{44}{5} & -14 \\ 0 & 0 & 0 & \frac{76}{41} & \frac{304}{41} \end{array}$$

Then, we find the solution $x = \{1, 2, 3, 4\}$ using the following program

```
x=Table[0,{i,1,n}]; x[[n]]=a[[n,n+1]]/a[[n,n]];
Do[x[[n-i]]=(a[[n-i,n+1]]-
Sum[a[[n-i,j]]*x[[j]],{j,n-i+1,n}])/a[[n-i,n-i]],{i,1,n-1}];
```

Now, let us present Gauss elimination in the general form.
First step of elimination. At first step, we shall eliminate x_1 from second, third, . . .,n-th equations, provided that $a_{11} \neq 0$. To eliminate x_1, let us multiply first equation in (2.1) by the coefficient

$$m_{i1} = \frac{a_{i1}}{a_{11}}, \quad i = 2, 3, \ldots, n$$

and subtract first equation from i-th equation for $i = 2, 3, \ldots, n$. Then, we obtain
First reduced system of Gauss elimination

$$\begin{aligned} a_{11}^{(0)} x_1 + \quad & a_{12}^{(0)} x_2 + a_{13}^{(0)} x_3 + \cdots + a_{1n}^{(0)} x_n = a_{1n+1}^{(0)} \\ & a_{22}^{(1)} x_2 + a_{23}^{(1)} x_3 + \cdots + a_{2n}^{(1)} x_n = a_{2n+1}^{(1)} \\ & a_{32}^{(1)} x_2 + a_{33}^{(1)} x_3 + \cdots + a_{3n}^{(1)} x_n = a_{3n+1}^{(1)} \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & a_{n2}^{(1)} x_2 + a_{n3}^{(1)} x_3 + \cdots + a_{nn}^{(1)} x_n = a_{nn+1}^{(1)} \end{aligned} \tag{2.6}$$

where

$$a_{ik}^{(0)} = a_{ik}, \quad i = 1, 2, \ldots, n, \quad k = 1, 2, \ldots, n+1,$$

$$a_{ik}^{(1)} = a_{ik}^{(0)} - m_{i1} a_{1k}^{(0)}, \quad i = 2, 3, \ldots, n; \quad k = 2, 3, \ldots, n+1.$$

Second step of elimination. At second step, we shall eliminate x_2 from third, fourth,..., n-th equations in (2.6), provided that $a_{22}^{(1)} \neq 0$. Let us multiply second equation in (2.6) by the coefficient

$$m_{i2} = \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad i = 3, 4, \ldots, n$$

and subtract second equation from i-th equation for $i = 3, 4, \ldots, n$. Then, we obtain

Second reduced system of Gauss elimination

$$\begin{aligned} a_{11}^{(0)}x_1+ \quad & a_{12}^{(0)}x_2+ \quad a_{13}^{(0)}x_3 + \cdots + a_{1n}^{(0)}x_n = a_{1n+1}^{(0)} \\ & a_{22}^{(1)}x_2+ \quad a_{23}^{(1)}x_3 + \cdots + a_{2n}^{(1)}x_n = a_{2n+1}^{(1)} \\ & \qquad\quad a_{33}^{(2)}x_3 + \cdots + a_{3n}^{(2)}x_n = a_{3n+1}^{(2)} \\ & \qquad\quad a_{43}^{(2)}x_3 + \cdots + a_{4n}^{(2)}x_n = a_{4n+1}^{(2)} \\ & \qquad\quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & \qquad\quad a_{n3}^{(2)}x_3 + \cdots + a_{nn}^{(2)}x_n = a_{nn+1}^{(2)} \end{aligned} \tag{2.7}$$

where

$$a_{ik}^{(2)} = a_{ik}^{(1)} - m_{i2}a_{2k}^{(1)}, \quad i = 3, 4, \ldots, n, \quad k = 3, 4, \ldots, n+1.$$

We continue elimination of the unknowns $x_3, x_4 \ldots, x_{n-1}$, provided that $a_{33}^{(2)} \neq 0$, $a_{44}^{(3)} \neq 0$, $a_{55}^{(4)} \neq 0$, ... $a_{n-1n-1}^{(n-2)} \neq 0$. As the final step of elimination, we obtain

Last reduced system of Gauss elimination

$$\begin{aligned} a_{11}^{(0)}x_1+ \quad & a_{12}^{(0)}x_2+ \quad a_{13}^{(0)}x_3+ \quad \cdots \quad +a_{1n}^{(0)}x_n = a_{1n+1}^{(0)} \\ & a_{22}^{(1)}x_2+ \quad a_{23}^{(1)}x_3+ \quad \cdots \quad +a_{2n}^{(1)}x_n = a_{2n+1}^{(1)} \\ & \qquad\quad a_{33}^{(2)}x_3+ \quad \cdots \quad +a_{3n}^{(2)}x_n = a_{3n+1}^{(2)} \\ & \qquad\quad \ldots\ldots\ldots \quad \cdots \quad \ldots\ldots\ldots\ldots \\ & \qquad\qquad\qquad\qquad\quad a_{nn}^{(n-1)}x_n = a_{nn+1}^{(n-1)} \end{aligned} \tag{2.8}$$

where

$$a_{ik}^{(s)} = a_{ik}^{(s-1)} - m_{is}a_{sk}^{(s-1)}, \quad m_{is} = \frac{a_{is}^{(s-1)}}{a_{ss}^{(s-1)}}.$$

$$s = 1, 2 \ldots, n-1, \quad i = s+1, s+2, \ldots, n, \quad k = s+1, s+2, \ldots, n+1.$$

The system of equations (2.8) has upper-triangular form. Therefore, we can easily find its solution by backward substitution

$$\begin{aligned} x_n &= \frac{a_{nn+1}^{(n-1)}}{a_{nn}^{(n-1)}} \\ x_i &= \frac{1}{a_{ii}^{(i-1)}}[a_{in+1}^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)}x_j], \end{aligned} \tag{2.9}$$

for $i = n-1, n-2, \ldots, 1$.
Below, we give the above elimination step by step in `Mathematica`.
At step $s,\ s = 1, 2, ..., n$, we change $i{-}th$ row of the matrix $A,\ i = s+1, s+2, ..., n$, by replacing it with

$$i{-}th\ row - \frac{s{-}th\ row\ *\ i, s{-}th\ element}{s, s{-}th\ element}$$

When s is fixed, the following `Mathematica` function would change $i{-}th$ row:

```
oneRow[a_, i_]:=
      ReplacePart[a, a[[i]]-a[[s]]*a[[i,s]]/a[[s,s]], i];
```

Then, the $s-th$ iteration of Gausian elimination would require the use of `oneRow` with $i = s+1, s+2, ..., n$, which can be achieved using `Fold`:

```
iter[a_,s_]:=Fold[oneRow, a, Range[s+1,n]];
```

Let us take the sample numerical example, again:

```
n=4;
a={{2,1,4,-3,4},  {4,-3,1,-2,-7},
   {6,4,-3,-1,1},{8,2,1,-2,7}};

TableForm[a]
          2   1   4  -3   4
          4  -3   1  -2  -7
          6   4  -3  -1   1
          8   2   1  -2   7
```

Executing the following program

```
Do[oneRow[a_,i_]:=
   ReplacePart[a, a[[i]]-a[[s]]*a[[i,s]]/a[[s,s]], i];
   iter[a_,s_]:= Fold[oneRow, a, Range[s+1,n]];
   a = iter[a, s],{s, 1, n}];
```

we obtain the upper triangular matrix as in (2.5).
In general, the following module `gaussDirectElimination` solves a linear system of equations, provided that the diagonal elements $a_{ss}^{(s-1)} \neq 0,\ s = 1, 2, ..., n$.

```
gaussDirectElimination[a_]:=Module[{c,n,oneRow,iter,x },
c=a;
n=Length[a[[1]]]-1;
oneRow[c_,i_]:=
ReplacePart[c, c[[i]]-c[[s]]*c[[i,s]]/c[[s,s]], i];
iter[c_,s_]:=Fold[oneRow, c, Range[s+1,n]];
Do[c= iter[c, s],{s, 1, n}];
```

```
  x=Table[0,{i,1,n}]; x[[n]]=c[[n,n+1]]/c[[n,n]];
  Do[x[[n-i]]=(c[[n-i,n+1]]-Sum[c[[n-i,j]]*x[[j]],
        {j,n-i+1,n}])/c[[n-i,n-i]],{i,1,n-1}];
  x
]
```

Solving the sample axample, we input data matrix

```
a={{2,1,4,-3,4}, {4,-3,1,-2,-7}, {6,4,-3,-1,1},{8,2,1,-2,7}};
```

and invoke the module

```
gaussDirectElimination[a]
```

to obtain the solution $x = 1, 2, 3, 4$.
Let us note that, we can apply the general Gauss elimination to a system of linear equations if the pivotal elements

$$a_{11}^{(0)} \neq 0, \quad a_{22}^{(1)} \neq 0, \quad a_{33}^{(2)} \neq 0, \ \ldots, \ a_{nn}^{(n-1n-1)} \neq 0,$$

are different from zero. Such a system of equations can be solved by the `Mathematica` program given in the above example. However, this straight forward application of Gauss elimination might lead to a strong accumulation of round-off error. In applications, pivotal strategy is used to minimize accumulation of round-off errors. In the case when at least one pivotal element is equal to zero, say $a_{kk}^{(k-1)} = 0$, we can apply partial or full pivoting strategy.

2.2 Partial Pivoting

If the pivotal element $a_{kk}^{(k-1)} = 0$ then Gauss elimination cannot be continued without rearrangement of rows or columns of the matrix

$$A^{(k-1)} = \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & \cdots & a_{1k}^{(0)} & \cdots & a_{1n}^{(0)} & a_{1n+1}^{(0)} \\ & a_{22}^{91)} & a_{23}^{(1)} & \cdots & a_{2k}^{(1)} & \cdots & a_{2n}^{(1)} & a_{2n+1}^{(1)} \\ & & a_{33}^{(2)} & \cdots & a_{3k}^{(2)} & \cdots & a_{3n}^{(2)} & a_{3n+1}^{(2)} \\ & & & \ddots & \vdots & \cdots & \vdots & \vdots \\ zero & element & \rightarrow & & \mathbf{a_{kk}^{(k-1)}} & \cdots & a_{kn}^{(k-1)} & a_{kn+1}^{(k-1)} \\ & & & & \mathbf{a_{k+1k}^{(k-1)}} & \cdots & a_{k+1n}^{(k-1)} & a_{k+1n+1}^{(k-1)} \\ & & & & \vdots & \vdots & \vdots & \vdots \\ largest & element & \rightarrow & & \mathbf{a_{sk}^{(k-1)}} & \cdots & a_{sn}^{(k-1)} & a_{sn+1}^{(k-1)} \\ & & & & \vdots & \vdots & \vdots & \vdots \\ & & & & \mathbf{a_{nk}^{(k-1)}} & \cdots & a_{nn}^{(k-1)} & a_{nn+1}^{(k-1)} \end{bmatrix}$$

Also, if $a_{kk}^{(k-1)}$, $k = 1, 2, \ldots, n-1$, are small then pivoting strategy will minimize affect of round-off errors on the solution. Thus, to eliminate unknowns $x_k,\ x_{k+1}, \ldots, x_n$, $k = 1, 2, \ldots, n-1$, we find the greatest absolute value of the elements

$$\mathbf{a_{k+1k}^{(k-1)}}, \mathbf{a_{k+2k}^{(k-1)}}, \ldots, \mathbf{a_{nk}^{(k-1)}},$$

and interchange k-th equation with s-th equation to have the pivotal element $a_{sk}^{(k-1)}$ which has the greatest absolute value on the pivotal place in k-th row and k-th column. Then, we can continue elimination of remaining unknowns $x_k,\ x_{k+1}, \ldots, x_n$ taking pivotal entries with the greatest absolute values. The partial pivoting strategy always succeeds, if A is a non-singular matrix. Partial pivoting can be also done with interchange of relevant columns in matrix A. Then, the greatest absolute values of elements

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

allocated in k-th row of the matrix A should be found.
In order to illustrate the partial pivoting strategy, let us consider the following examples:

Example 2.1 *Solve the system of equations*

$$\begin{array}{rcrcrcrcl} 5x_1 & + & x_2 & + & 2x_3 & + & 3x_4 & = & 5 \\ 10x_1 & + & 2x_2 & - & 6x_3 & + & 9x_4 & = & 4 \\ 5x_1 & - & x_2 & + & x_3 & + & 4x_4 & = & 1 \\ 15x_1 & - & 3x_2 & - & 3x_3 & + & 9x_4 & = & 9 \end{array} \tag{2.10}$$

using

1. *(a) partial pivoting strategy for interchange of rows,*
   (b) partial pivoting strategy for interchange of columns.

Solution (a). According to partial pivoting, we interchange fourth equation with first equation to have the greatest pivotal element 15 on first pivotal place in first row and first column. Then, we obtain

$$\begin{array}{rcrcrcrcl} 15x_1 & - & 3x_2 & - & 3x_3 & + & 9x_4 & = & 9 \\ 10x_1 & + & 2x_2 & - & 6x_3 & + & 9x_4 & = & 4 \\ 5x_1 & - & x_2 & + & x_3 & + & 4x_4 & = & 1 \\ 5x_1 & + & x_2 & + & 2x_3 & + & 3x_4 & = & 5 \end{array} \tag{2.11}$$

Thus, first reduced system of equations is:

$$\begin{array}{rcrcrcrcr} 15x_1 & - & 3x_2 & - & 3x_3 & + & 9x_4 & = & 9 \\ & & 4x_2 & - & 4x_3 & + & 3x_4 & = & -2 \\ & & & & 2x_3 & + & x_4 & = & -2 \\ & & 2x_2 & + & 3x_3 & + & & = & 2 \end{array} \tag{2.12}$$

To get second reduced system, we do not need to make any interchange since $a_{22}^{(1)} = 4$ is the greatest entry in 2-nd column on the pivotal place. Thus, second reduced system is:

$$\begin{array}{rcrcrcrcr} 15x_1 & - & 3x_2 & - & 3x_3 & + & 9x_4 & = & 9 \\ & & 4x_2 & - & 4x_3 & + & 3x_4 & = & -2 \\ & & & & 2x_3 & + & x_4 & = & -2 \\ & & & & 5x_3 & - & \frac{3}{2}x_4 & = & 3 \end{array} \tag{2.13}$$

Obviously, third (last) reduced system of equations has upper-triangular form

$$\begin{array}{rcrcrcrcr} 15x_1 & - & 3x_2 & - & 3x_3 & + & 9x_4 & = & 9 \\ & & 4x_2 & - & 4x_3 & + & 3x_4 & = & -2 \\ & & & & 5x_3 & - & 1.5x_4 & = & 3 \\ & & & & & - & 1.6x_4 & = & 3.2 \end{array} \tag{2.14}$$

Hence, the solution is

$$\begin{aligned} x_4 &= -2, \\ x_3 &= \frac{1}{2}[-2-(-2)] = 0, \\ x_2 &= \frac{1}{4}[-2+4*0-3*(-2)] = 1, \\ x_1 &= \frac{1}{15}[9+3*1+3*0-9*(-2)] = 2. \end{aligned}$$

Solution (b). Let us come back to the original system of equations (2.10). Evidently, the entry $a_{11}^{(0)} = 5$ has greatest absolute value among all entries in first row of A. Therefore, there is no a need to interchange columns in A. Then, we obtain
First reduced system of equations:

$$\begin{array}{rcrcrcrcr} 5x_1 & + & 2x_2 & + & x_3 & + & 3x_4 & = & 5 \\ & & & - & 10x_3 & + & 3x_4 & = & -6 \\ & - & 2x_2 & - & x_3 & + & x_4 & = & -4 \\ & - & 6x_2 & - & 9x_3 & + & & = & -6 \end{array} \tag{2.15}$$

We shall interchange second and fourth equations in (2.15) to obtain the following system of equations:

$$\begin{array}{rcrcrcrcr} 5x_1 & + & x_2 & + & 2x_3 & + & 3x_4 & = & 5 \\ & - & 6x_2 & - & 9x_3 & & & = & -6 \\ & - & 2x_2 & - & x_3 & + & x_4 & = & -4 \\ & & & - & 10x_3 & + & 3x_4 & = & -6 \end{array} \tag{2.16}$$

Next, we shall interchange second and third columns in (2.16) to get the pivotal entry $a_{22}^{(1)} = -9$ on the pivotal place in second row and second column.

$$\begin{array}{rcrcrcrcr} 5x_1 & + & 2x_3 & + & x_2 & + & 3x_4 & = & 5 \\ & - & 9x_3 & - & 6x_2 & & & = & -6 \\ & - & x_3 & - & 2x_2 & + & x_4 & = & -4 \\ & - & 10x_3 & & & + & 3x_4 & = & -6 \end{array} \tag{2.17}$$

Now, we shall eliminate x_3 from third and fourth equations in (2.17) to obtain the second reduced system of equations:

$$\begin{array}{rcrcrcrcr} 5x_1 & + & 2x_3 & + & x_2 & + & 3x_4 & = & 5 \\ & - & 9x_3 & - & 6x_2 & & & = & -6 \\ & & & - & \frac{4}{3}x_2 & + & x_4 & = & -\frac{10}{3} \\ & & & & \frac{20}{3}x_2 & + & 3x_4 & = & \frac{2}{3} \end{array} \tag{2.18}$$

The pivotal entry $a_{33}^{(2)} = -\frac{4}{3}$ has the greatest absolute value in first row of the matrix

$$A^{(2)} = \begin{bmatrix} -\frac{4}{3} & 1 \\ \frac{20}{3} & 3 \end{bmatrix}.$$

Therefore, according to the partial pivoting, we shall not make any change of columns to eliminate x_2 from fourth equation in (2.18). Then, we obtain third (last) reduced system of equations:

$$\begin{array}{rcrcrcrcr} 5x_1 & + & 2x_3 & + & x_2 & + & 3x_4 & = & 5 \\ & - & 9x_3 & - & 6x_2 & & & = & -6 \\ & & & - & \frac{4}{3}x_2 & + & x_4 & = & -\frac{10}{3} \\ & & & & & & 8x_4 & = & -16 \end{array} \tag{2.19}$$

Hence, by backward substitution, we find the solution

$$\begin{aligned} x_4 &= -2, \\ x_2 &= -\frac{3}{4}[-\frac{10}{3} - 2 * (-2)] = 1, \\ x_3 &= -\frac{1}{9}[-6 + 6 * 1] = 0, \\ x_1 &= \frac{1}{5}[5 - 2 * 0 - 1 - 3(-2)] = 2. \end{aligned}$$

In the programming approach to the partial pivoting on rows, in `Mathematica`, we shall use the following algorithm

1. Set the matrix $m = [a|b]$ that includes right side vector b and the empty list {}.
2. Find the maximum element in the first column of m and denote it by mk.
3. Denote by k the position of mk in the first column of m.
4. Set $rowk = m[[k]]/mk$
5. Append $rowk$ to e.
6. Drop k-th row of m.
7. Replace each row of m by the $row - rowk * First[row]$
8. Drop first column of m
9. Return $\{m, e\}$.
10. Nest steps
    tt 1 - 7 n times, where n is the number of rows in m.

The `Mathematica` module `eliminatepivo` based on the partial pivoting reduces a system of algebraic equations to the upper triangular form and gives its solution.

```
eliminatepivo[a_, b_]:= Module[
                        {oneIter, mat, elimMatrix},

oneIter[{m_,e_}]:= Module[
{column1, mk, k, rowk, changeOneRow},

 column1=Map[First[#]&, m];
 mk= Max[column1];
 {{k}}= Position[column1, mk];
 rowk= m[[k]]/mk;

 changeOneRow[row_]:= row - rowk*First[row];

 {Map[Rest[#]&,
 Map[changeOneRow, Drop[m, {k}]]] ,
 Append[e, Rest[rowk]]}  ];

 mat = Transpose[ Append[Transpose[a], b]];

 {{}, elimMatrix}=
 Nest[oneIter, {mat, {}}, Length[First[mat]]-1];
```

```
        Fold[Prepend[#1, Last[#2] - Drop[#2, -1] . #1]&,
        Last[elimMatrix],
        Rest[Reverse[elimMatrix]]]
]
```

To solve the system of equations (2.11) using the module `eliminatepivo`, we input data

```
m={{5,1,2,3,},{10,2,-6,9},{5,-1,1,4},{15,-3,-3,9}};
b={5,4,1,9}
```

and invoke the module `eliminatepivo[m,b]`.

2.3 Principal Element Strategy

We apply principal element strategy to operate with possibly greatest pivotal entries in Gauss elimination. First, we find the greatest absolute value of entries in matrix A. Let

$$|a_{rs}| = \max_{1\leq i,j\leq n} |a_{ij}|.$$

Then, we interchange first row with r-th row and first column with s-th column to place the greatest value as first pivotal element. We should renumerate equations and unknowns, respectively. Next, we eliminate the relevant unknown from second, third,. . .,n-th of the newly renumerate equations to get first reduced system of equations. Secondly, we find the greatest absolute value among the entries $a_{ij}^{(1)},\ i,j = 2,3,\ldots,n$. Let

$$\mid a_{rs}^{(1)} \mid = \max_{2\leq i,j\leq n} \mid a_{ij}^{(1)} \mid .$$

Then, we interchange second row with the r-th row and second column with s-th column to place the greatest pivotal entry on second row and second column. We should renumerate equations and unknowns, again. After, elimination of x_2 from third, fourth,. . . , $n-th$ equations, we get second reduced system of equations. We repeat this process till upper-triangular systems appears.

Example 2.2 *Solve the system of equations*

$$\begin{array}{rcrcrcrcl} 5x_1 & + & x_2 & + & 2x_3 & + & 3x_4 & = & 5 \\ 10x_1 & + & 2x_2 & - & 6x_3 & + & 9x_4 & = & 4 \\ 5x_1 & - & x_2 & + & x_3 & + & 4x_4 & = & 1 \\ 15x_1 & - & 2x_2 & - & x_3 & + & 10x_4 & = & 8 \end{array} \tag{2.20}$$

using full pivoting strategy.

Solution. Evidently $a_{14} = 15$ is the greatest entry of the matrix

$$A = \begin{bmatrix} 5 & 1 & 2 & 3 \\ 10 & 2 & -6 & 9 \\ 5 & -1 & 1 & 4 \\ 15 & -2 & -1 & 10 \end{bmatrix}.$$

Therefore, we interchange first equation with fourth equation to get the greatest pivotal entry on first pivotal place. Then, we consider the system of equations

$$\begin{array}{rcrcrcrl} 15x_1 & - & 2x_2 & - & x_3 & + & 10x_4 & = 8 \\ 10x_1 & + & 2x_2 & - & 6x_3 & + & 9x_4 & = 4 \\ 5x_1 & - & x_2 & + & x_3 & + & 4x_4 & = 1 \\ 5x_1 & + & x_2 & + & 2x_3 & + & 3x_4 & = 5 \end{array} \tag{2.21}$$

After first step of elimination, we shall obtain the first reduced system of equations:

$$\begin{array}{rcrcrcrcr} 15x_1 & - & 2x_2 & - & x_3 & + & 10x_4 & = & 8 \\ & & \frac{10}{3}x_2 & - & \frac{16}{3}x_3 & + & \frac{7}{3}x_4 & = & -\frac{4}{3} \\ & - & \frac{1}{3}x_2 & - & \frac{4}{3}x_3 & + & \frac{2}{3}x_4 & = & -\frac{5}{3} \\ & & \frac{5}{3}x_2 & + & \frac{7}{3}x_3 & - & \frac{1}{3}x_4 & = & \frac{7}{3} \end{array} \tag{2.22}$$

Now, we shall find the greatest absolute entry in the matrix

$$\begin{bmatrix} \frac{10}{3} & -\frac{16}{3} & \frac{7}{3} \\ -\frac{1}{3} & -\frac{4}{3} & \frac{2}{3} \\ \frac{5}{3} & \frac{7}{3} & -\frac{1}{3} \end{bmatrix}.$$

The greatest absolute entry

$$| a_{23}^{(1)} | = | -\frac{16}{3} | .$$

We interchange second and third columns in the matrix

$$A^{(1)} = \begin{bmatrix} 15 & -2 & -1 & 10 \\ 0 & \frac{10}{3} & -\frac{16}{3} & \frac{7}{3} \\ 0 & -\frac{1}{3} & -\frac{4}{3} & \frac{2}{3} \\ 0 & \frac{5}{3} & \frac{7}{3} & -\frac{1}{3} \end{bmatrix}.$$

to get the greatest absolute pivotal entry on the pivotal place in second row and in second column. Then, x_2 takes the position of x_3 and x_3 takes the position of x_2. For second step of elimination, we consider the following system of equations:

$$\begin{array}{rcrcrcrcr} 15x_1 & - & x_3 & - & 2x_2 & + & 10x_4 & = & 8 \\ & - & \frac{16}{3}x_3 & + & \frac{10}{3}x_2 & + & \frac{7}{3}x_4 & = & -\frac{4}{3} \\ & - & \frac{4}{3}x_3 & - & \frac{1}{3}x_2 & + & \frac{2}{3}x_4 & = & -\frac{5}{3} \\ & & \frac{7}{3}x_3 & + & \frac{5}{3}x_2 & - & \frac{1}{3}x_4 & = & \frac{7}{3} \end{array} \tag{2.23}$$

Now, we shall eliminate x_3 from third and fourth equations in (2.23) to get second reduced system of equations:

$$\begin{array}{rcrcrcrcr} 15x_1 & - & x_3 & - & 2x_2 & + & 10x_4 & = & 8 \\ & - & \frac{16}{3}x_3 & + & \frac{10}{3}x_2 & + & \frac{7}{3}x_4 & = & -\frac{4}{3} \\ & & & - & \frac{7}{6}x_2 & + & \frac{1}{12}x_4 & = & -\frac{4}{3} \\ & & & + & \frac{25}{8}x_2 & + & \frac{11}{16}x_4 & = & \frac{7}{4} \end{array} \tag{2.24}$$

We us observe that $a_{43}^{(2)} = \frac{25}{8}$ is the greatest entry in the matrix

$$\begin{bmatrix} -\frac{7}{6} & \frac{1}{12} \\ \frac{25}{8} & \frac{11}{16} \end{bmatrix}$$

Thus, we shall interchange third and fourth equations in (2.24) to get the greatest pivotal entry on the pivotal place in third row and third column. Then, we consider the following system of equations:

$$\begin{array}{rcrcrcrcr} 15x_1 & - & x_3 & - & 2x_2 & + & 10x_4 & = & 8 \\ & - & \frac{16}{3}x_3 & + & \frac{10}{3}x_2 & + & \frac{7}{3}x_4 & = & -\frac{4}{3} \\ & & & + & \frac{25}{8}x_2 & + & \frac{11}{16}x_4 & = & \frac{7}{4} \\ & & & - & \frac{7}{6}x_2 & + & \frac{1}{12}x_4 & = & -\frac{4}{3} \end{array} \tag{2.25}$$

Finally, we shall eliminate x_2 from fourth equation in (2.25) to get the last reduced system of equations in the upper-triangular form

$$\begin{array}{rcrcrcl} 15x_1 & - & x_3 & - & 2x_2 & + & 10x_4 & = & 8 \\ & - & \frac{16}{3}x_3 & + & \frac{10}{3}x_2 & + & \frac{7}{3}x_4 & = & -\frac{4}{3} \\ & & & & \frac{25}{8}x_2 & + & \frac{11}{16}x_4 & = & \frac{7}{4} \\ & & & & & & \frac{17}{50}x_4 & = & -\frac{17}{25} \end{array} \tag{2.26}$$

Hence, the solution is:

$$\begin{aligned} x_4 &= -2 \\ x_2 &= -\frac{6}{7}[-\frac{4}{3} - \frac{1}{12}(-2)] = 1 \\ x_3 &= -\frac{3}{16}[-\frac{4}{3} - \frac{10}{3} - \frac{7}{3}(-2)] = 0 \\ x_1 &= \frac{1}{15}[8 + 2 - 10(-2)] = 2. \end{aligned}$$

2.4 LU-Decomposition.

Applying Gauss elimination method, we obtain the solution $x = (x_1, x_2, \ldots, x_n)$ of the system of equations

$$Ax = a \tag{2.27}$$

and the following factorized form of the matrix A:

$$A = LU,$$

where the lower triangular matrix

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ m_{21} & 1 & 0 & 0 & \cdots & 0 & 0 \\ m_{31} & m_{32} & 1 & 0 & \cdots & 0 & 0 \\ m_{41} & m_{42} & m_{43} & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ m_{n1} & m_{n2} & m_{n3} & m_{n4} & \cdots & m_{nn-1} & 1 \end{bmatrix},$$

and the upper-triangular matrix

$$U = \begin{bmatrix} a_{11}^{(0)} & a_{12}^{(0)} & a_{13}^{(0)} & a_{14}^{(0)} & \cdots & a_{1n-1}^{(0} & a_{1n}^{(0)} \\ 0 & a_{22}^{(1)} & a_{23}^{(1)} & a_{24}^{(1)} & \cdots & a_{2n-1}^{(1)} & a_{2n}^{(1)} \\ 0 & 0 & a_{33}^{(2)} & a_{34}^{(2)} & \cdots & a_{3n-1}^{(2)} & a_{3n}^{(2)} \\ 0 & 0 & 0 & a_{44}^{(3)} & \cdots & a_{4n-1}^{(3)} & a_{4n}^{(3)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & a_{nn}^{(n-1)} \end{bmatrix},$$

Indeed, the product of i-th row of the matrix L by k-th column of the matrix U is:

$$\sum_{s=1}^{n} L_{is}U_{sk} = \sum_{s=1}^{p} m_{is}a_{sk}^{(s-1)},$$

where $p = min\{i, k\}$.
Let us note that

$$a_{ik}^{(s)} = a_{ik}^{(s-1)} - m_{is}a_{sk}^{(s-1)},$$

By taking the sum of both hand sides, we obtain

$$a_{ik}^{(p)} = a_{ik} - \sum_{s=1}^{p-1} m_{is}a_{sk}^{(s-1)}.$$

Hence

$$a_{ik} = a_{ik}^{(p)} + \sum_{s=1}^{p} m_{is}a_{sk}^{(s-1)}.$$

One can check that the following equalities hold:

$$a_{ik}^{(p)} = a_{ik}^{(i)} \quad \text{if} \quad i \leq k$$

and

$$a_{ik}^{(p)} = 0 \quad \text{if} \quad i > k.$$

Thus, for $m_{ii} = 1$, we have

$$a_{ik} = \sum_{s=1}^{p} m_{is}a_{sk}^{(s-1)} = \sum_{s=1}^{n} L_{is}U_{sk}.$$

As an example of LU decomposition, we note that the matrix of the system of equations in (2.2) has the following LU decomposition:

$$A = \begin{bmatrix} 2 & 1 & 4 & -3 \\ 4 & -3 & 1 & -2 \\ 6 & 4 & -3 & -1 \\ 8 & 2 & 1 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 3 & -\frac{1}{5} & 1 & 0 \\ 4 & \frac{2}{5} & \frac{61}{82} & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 4 & -3 \\ 0 & -5 & -7 & 4 \\ 0 & 0 & -\frac{82}{5} & \frac{44}{5} \\ 0 & 0 & 0 & \frac{76}{41} \end{bmatrix}. = LU$$

2.5 Root Square Method

It is possible to present a symmetric matrix $A = \{A_{ij}\},\ i, j = 1, 2, ..., n$ as the square of a triangular matrix (cf. [6])

$$L = \begin{bmatrix} L_{11} & L_{12} & L_{13} & \cdots & L_{1n} \\ 0 & L_{22} & L_{23} & \cdots & L_{2n} \\ 0 & 0 & L_{33} & \cdots & L_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & L_{nn} \end{bmatrix}$$

so that $A = L^T L$. Indeed, we note that

$$A_{ij} = L_{1i}L_{1j} + L_{2i}L_{2j} + \cdots + L_{ii}L_{ij}, \quad i = 1, 2, ..., j-1$$

and

$$A_{ii} = L_{1i}^2 + L_{2i}^2 + \cdots + L_{ni}^2, \quad i = j.$$

Hence

$$\begin{aligned} &L_{11} = \sqrt{A_{11}}, && L_{1j} = \frac{A_{1j}}{L_{11}}, && j = 2, 3, ..., n, \\ &L_{ii} = \sqrt{A_{ii} - \sum_{k=1}^{i-1} L_{ki}^2}, && i = 2, 3, ..., n, \\ &L_{ij} = \frac{1}{L_{ii}}[A_{ij} - \sum_{k=1}^{i-1} L_{ki}L_{kj}], && j = i+1, i+2, ..., n, \\ &L_{ij} = 0, && j = 1, 2, ..., i-1. \end{aligned}$$

The LL-decomposition algorithm always succeeds if A is a positive definite matrix. However, the algorithm is also applicable when A is a symmetric matrix, provided that $L_{ii} \neq 0,\ i = 1, 2, ..., n$. In the case when complex numbers appear ($A_{ii} - \sum L_{ki}^2 < 0$) the algorithm produces complex entries of the triangular matrix L.

Having LL-decomposition of the matrix A, we can find solution x of the system of linear equations

$$Ax = b$$

by the substitution

$$L^T z = b \quad \text{and} \quad Lx = z.$$

So that

$$z_1 = \frac{b_1}{L_{11}}, \quad z_i = \frac{1}{L_{ii}}[b_i - \sum_{k=1}^{i-1} L_{ki}z_k], \quad i = 2, 3, ..., n,$$

and by backward substitution

$$x_n = \frac{z_n}{L_{nn}}, \quad x_i = \frac{1}{L_{ii}}[z_i - \sum_{k=i+1}^{n} L_{ik}x_k], \quad i = n-1, n-2, ..., 1.$$

Example 2.3 *Let us solve the following system of linear equations (cf. [6])*

$$\begin{array}{rcrcrcrcl} x_1 & + & 0.42x_2 & + & 0.54x_3 & + & 0.66x_4 & = & 0.3 \\ 0.42x_1 & + & 1x_2 & + & 0.32x_3 & + & 0.44x_4 & = & 0.5 \\ 0.54x_1 & + & 0.32x_2 & + & 1x_3 & + & 0.22x_4 & = & 0.7 \\ 0.66x_1 & + & 0.44x_2 & + & 0.22x_3 & + & 1x_4 & = & 0.9 \end{array}$$

using the following `Mathematica` *module* `choleva`

```
choleva[a_,b_]:=Module[{l,i,j,k,m,m1,n,x,z},
  n=Length[a[[1]]];
  l[1,1]:=l[1,1]=Sqrt[a[[1,1]]];
  l[1, j_]:=l[1,j]=a[[1,j]]/l[1,1];
  l[i_,i_]:=l[i,i]=Sqrt[a[[i,i]]- Sum[l[k,i]^2, {k, 1,i-1}]];
  l[i_, j_]:=l[i,j]=
  (a[[i,j]]-Sum[l[k,i] l[k,j], {k,1,i-1}])/l[i,i];
  m=Table[Join[Table[0,{i-1}],Table[l[i,j],{j,i,n}]],
  {i,1,n}];l[n,n];
  m1=Transpose[m];
  z[1]=b[[1]]/m1[[1,1]];
  z[i_]:=z[i]=(b[[i]]-Sum[m1[[i,j]]*z[j],{j,1,i-1}])/m1[[i,i]];
  x[n]=z[n]/m[[n,n]];
  x[i_]:=x[i]=(z[i]-Sum[m[[i,j]]*x[j],{j,i+1,n}])/m[[i,i]];
  Print["x = ",Table[x[i],{i,1,n}]];
  Print["Matrix L =",MatrixForm[m]]
  ]
```

We input the matrix a and the right side vector b

```
a={{1.,0.42,0.54,0.66},{0.42,1.,0.32,0.44},
   {0.54,0.32,1.,0.22},{0.66,0.44,0.22,1.}};
b={0.3,0.5,0.7,0.9};
```

Then, we invoke the module

```
choleva[a,b]
```

to obtain the solution

```
x = {-1.25779, 0.0434873, 1.03917, 1.48239}
```

and the upper triangular matrix

$$L = \begin{bmatrix} 1. & 0.42 & 0.54 & 0.66 \\ 0 & 0.907524 & 0.102697 & 0.179389 \\ 0 & 0 & 0.835376 & -0.185333 \\ 0 & 0 & 0 & 0.7056 \end{bmatrix}.$$

2.6 Gauss Elimination for Tri-diagonal Matrices.

Let us note that for a tri-diagonal matrix A, the system of linear equations (2.1) takes the following form:

$$\begin{array}{ccccccccccccc}
a_{11}x_1 & + & a_{12}x_2 & & & & & & & & & & = a_{1n+1} \\
a_{21}x_1 & + & a_{22}x_2 & + & a_{23}x_3 & & & & & & & & = a_{2n+1} \\
 & & a_{32}x_2 & + & a_{33}x_3 & + & a_{34}x_4 & & & & & & = a_{3n+1} \\
 & & \cdots\cdots & & \cdots\cdots & \cdots & \cdots\cdots & \cdots & \cdots & & \cdots & & \cdots \\
 & & & & a_{ii-1}x_{i-1} & + & a_{ii}x_i & + & a_{ii+1}x_{i+1} & & = a_{in+1} & & \\
 & & & & \cdots\cdots & \cdots & \cdots\cdots & \cdots & \cdots & & \cdots & & \\
 & & & & & & a_{nn-1}x_{n-1} & + & a_{nn}x_n & & = a_{nn+1} & &
\end{array}$$

Applying Gauss elimination to the above tri-diagonal system of equations, we obtain the last reduced system of equations:

$$\begin{array}{ccccccccccccc}
a_{11}x_1 & + & a_{12}x_2 & & & & & & & & = a_{1n+1} & & \\
 & & a_{22}^{(1)}x_2 & + & a_{23}x_3 & & & & & & = a_{2n+1}^{(1)} & & \\
 & & & & a_{33}^{(2)}x_3 & + & a_{34}x_4 & & & & = a_{3n+1}^{(2)} & & \\
 & & & & \cdots\cdots & \cdots & \cdots\cdots & \cdots & \cdots & & \cdots & & \\
 & & & & & & a_{ii}^{(i-1)}x_i & + & a_{ii+1}x_{i+1} & & = a_{in+1}^{(i-1)} & & \\
 & & & & & & \cdots\cdots & \cdots & \cdots & & \cdots & & \\
 & & & & & & & & a_{nn}^{(n-1)}x_n & & = a_{nn+1}^{(n-1)} & &
\end{array} \tag{2.28}$$

where (see (2.8) and (2.9))

$$a_{ii}^{(i-1)} = a_{ii} - \frac{a_{ii-1}}{a_{i-1i-1}^{(i-2)}}a_{i-1i},$$

$$a_{in+1}^{(i-1)} = a_{in+1} - \frac{a_{i-1n+1}^{(i-2)}}{a_{i-1i-1}^{(i-2)}}a_{ii-1},$$

for $i = 2, 3, \ldots, n$.
From formulae (2.9), we have

$$x_n = \frac{a_{nn+1}^{(n-1)}}{a_{nn}^{(n-1)}},$$

and

$$x_i = \frac{1}{a_{ii}^{(i-1)}}[a_{in+1}^{(i-1)} - a_{ii+1}x_{i+1}],$$

for $i = n-1, n-2, \ldots, 1$.
Let us denote by

$$\alpha_i = \frac{a_{ii+1}}{a_{ii}^{(i-1)}}, \quad i = 1, 2, ..., n-1 \qquad \beta_i = \frac{a_{in+1}^{(i-1)}}{a_{ii}^{(i-1)}} \quad i = 1, 2, ..., n.$$

Then, we have

$$\alpha_1 = \frac{a_{12}}{a_{11}}, \quad \beta_1 = \frac{a_{1n+1}}{a_{11}}$$

and

$$\alpha_i = \frac{a_{ii+1}}{a_{ii} - \alpha_{i-1}a_{ii-1}}, \quad \beta_i = \frac{a_{in+1} - \beta_{i-1}a_{ii-1}}{a_{ii} - \alpha_{i-1}a_{ii-1}}.$$

We obtain the following algorithm for solving a system of equations with a tri-diagonal matrix.

Algorithm.

$$\begin{aligned}
&Set\ :\\
&\qquad \alpha_1 = \frac{a_{12}}{a_{11}}, \qquad \beta_1 = \frac{a_{1n+1}}{a_{11}}\\
&for\ \ i = 2, 3, \ldots, n-1, \qquad evaluate\ :\\
&\qquad \alpha_i = \frac{a_{ii+1}}{a_{ii} - \alpha_{i-1}a_{ii-1}}\\
&for\ \ i = 2, 3, \ldots, n, \qquad evaluate\ :\\
&\qquad \beta_i = \frac{a_{in+1} - \beta_{i-1}a_{ii-1}}{a_{ii} - \alpha_{i-1}a_{ii-1}}\\
&set\ :\qquad x_n = \beta_n\\
&evaluate\ :\ \ for\ \ i = n-1, n-2, \ldots, 1\\
&\qquad x_i = \beta_i - \alpha_i x_{i+1}.
\end{aligned} \tag{2.29}$$

The above algorithm is stable with respect to round-off errors if the tri-diagonal matrix A satisfies the following conditions:

$$\begin{aligned}
&a_{11} > \mid a_{12} \mid, \qquad a_{nn} > \mid a_{nn-1} \mid,\\
&a_{ii} \geq \mid a_{ii-1} \mid + \mid a_{ii+1} \mid, \quad i = 2, 3, \ldots, n-1.
\end{aligned}$$

Then all $\alpha's$ coefficients are less than one.
Indeed, we have

$$\mid \alpha_1 \mid < 1$$

$$\mid \alpha_i \mid \leq \mid \frac{a_{ii+1}}{a_{ii} - \alpha_{i-1}a_{ii-1}} \mid \leq \frac{\mid a_{ii+1} \mid}{\mid a_{ii+1} \mid + (1 - \mid \alpha_{i-1} \mid) \mid a_{ii-1} \mid} < 1.$$

for $\quad i = 2, 3, \ldots, n-1.$

The solution x can be obtained by this algorithm with total number of $8n - 6$ arithmetic operations.

In order to implement the algorithm in `Mathematica`, we input data matrix and the right side vector as the lists

$$a = \{a_1, a_2, ..., a_n\}, \quad b = \{b_1, b_2, ..., b_{n-1}, 0\}$$
$$c = \{0, c_2, c_3, ..., c_n\}, \quad f = \{f_1, f_2, ..., f_n\}$$

and invoke the following module

```
soltri[a_,b_,c_,f_]:=Module[{al,be,n,x},
       n=Length[a];
       al[1]=b[[1]]/a[[1]];
       al[i_]:=al[i]=b[[i]]/(a[[i]]-al[i-1]*c[[i]]);
       be[1]=f[[1]]/a[[1]];
       be[i_]:=be[i]=(f[[i]]-be[i-1]*c[[i]])/
              (a[[i]]-al[i-1]*c[[i]]);
       x[n]=be[n];
       x[i_]:=x[i]=be[i]-al[i]*x[i+1];
       Table[x[i],{i,1,n}]
       ]
```

Example 2.4 *Let us consider the following system of linear equations:*

$$\begin{array}{rcrcrcrcrcr}
2x_1 & - & x_2 & & & & & & & & = f_1 \\
-x_1 & + & 2x_2 & - & x_3 & & & & & & = f_2 \\
 & - & x_2 & + & 2x_3 & - & x_4 & & & & = f_3 \\
 & & & & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
 & & & - & x_{i-1} & + & 2x_i & - & x_{i+1} & & = f_i \\
 & & & & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\
 & & & & & - & x_{n-1} & + & 2x_n & & = f_n
\end{array}$$

To solve this system of equations when $n = 8$, we input data

$$a = \{2., 2., 2., 2., 2., 2., 2., 2.\}, \quad b = \{-1., -1., -1., -1., -1., -1., -1., 0\}$$
$$c = \{0, -1., -1., -1., -1., -1., -1., -1.\}, \quad f = \{0., 2., -2., 2., -2., 2., -2., 0.\}$$

and invoke the module

```
soltri[a,b,c,f]
```

to obtain the solution $x = \{1., 2., 1., 2., 1., 2., 1., 2.\}$.

2.7 Gauss Elimination for Block Tri-diagonal Matrices

Let us consider the system of linear equations with tri-diagonal block matrix (cf. [7])

$$AX = B, \tag{2.30}$$

where the right side vector and unknown vector are:

$$B = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ \vdots \\ B_n \end{bmatrix}, \quad B_i = \begin{bmatrix} B_{i1} \\ B_{i2} \\ B_{i3} \\ \vdots \\ B_{in} \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{bmatrix}, \quad X_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \\ \vdots \\ x_{im} \end{bmatrix},$$

and the block matrix

$$A = \begin{bmatrix} A_{11} & A_{12} & 0 & 0 & \cdots & 0 & 0 \\ A_{21} & A_{22} & A_{23} & 0 & \cdots & 0 & 0 \\ 0 & A_{32} & A_{33} & A_{34} & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & A_{nn-1} & A_{nn} \end{bmatrix},$$

$$A_{ij} = \begin{bmatrix} a_{ij}^{11} & a_{ij}^{12} & \cdots & a_{ij}^{1m} \\ a_{ij}^{21} & a_{ij}^{22} & \cdots & a_{ij}^{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{ij}^{m1} & a_{ij}^{m2} & \cdots & a_{ij}^{mm} \end{bmatrix}, \quad i, j = 1, 2, ..., n.$$

Clearly, the system of equations (2.30) takes the following form in the block notation:

$$\begin{array}{llllllll} A_{11}X_1 & +A_{12}X_2 & & & & & & = B_1 \\ A_{21}X_1 & +A_{22}X_2 & +A_{23}X_3 & & & & & = B_2 \\ & A_{32}X_2 & +A_{33}X_3 & +A_{34}X_4 & & & & = B_3 \\ & \cdots & \cdots & \cdots\cdots & \cdots & \cdots\cdots & \cdots\cdots & \cdots \\ & & & & & A_{n-1n-1}X_{n-1} & +A_{n-1n}X_n & = B_{n-1} \\ & & & & & A_{nn-1}X_{n-1} & +A_{nn}X_n & = B_n \end{array} \tag{2.31}$$

Applying non-pivoting strategy, we can find the solution X, provided that the pivotal matrices are non-zero entries.
In order to get the first reduced system of equations, we multiply from the left first row-block in (2.31) by the matrix

$$M_{21} = A_{21}A_{11}^{-1}$$

providing that the inverse matrix A_{11}^{-1} exists. Then, we obtain
the first reduced system of equations:

$$\begin{array}{llllllll} A_{11}X_1 & +A_{12}X_2 & & & & & & = B_1 \\ & A_{22}^{(1)}X_2 & +A_{23}X_3 & & & & & = B_2^{(1)} \\ & A_{32}X_2 & +A_{33}X_3 & +A_{34}X_4 & & & & = B_3 \\ & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & & & A_{n-1n-1}X_{n-1} & +A_{n-1n}X_n & = B_{n-1} \\ & & & & & A_{nn-1}X_{n-1} & +A_{nn}X_n & = B_n \end{array} \tag{2.32}$$

where

$$A_{22}^{(1)} = A_{22} - A_{21}A_{11}^{-1}A_{12},$$
$$B_2^{(1)} = B_2 - A_{21}A_{11}^{-1}B_1.$$

Next, multiplying from the left second row-block in (2.32) by the matrix

$$M_{32} = A_{32}A_{22}^{-(1)}$$

we obtain
the second reduced system of equations:

$$\begin{array}{lllllll}
A_{11}X_1 & +A_{12}X_2 & & & & & = B_1 \\
 & A_{22}^{(1)}X_2 & +A_{23}X_3 & & & & = B_2^{(1)} \\
 & & A_{33}^{(2)}X_3 & +A_{34}X_4 & & & = B_3^{(2)} \\
 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
 & & & & A_{n-1n-1}X_{n-1} & +A_{n-1n}X_n & = B_{n-1} \\
 & & & & A_{nn-1}X_{n-1} & +A_{nn}X_n & = B_n
\end{array} \tag{2.33}$$

where

$$A_{33}^{(2)} = A_{33} - A_{32}A_{22}^{-(1)}A_{23},$$
$$B_3^{(2)} = B_3 - A_{32}A_{22}^{-(1)}B_2^{(1)}.$$

We continue the block elimination process if the pivotal matrices $A_{ii}^{(i-1)}$, $i = 1, 2, ..., n;$ are non-singular. As the final step of elimination, we obtain
the last reduced system of equations:

$$\begin{array}{lllllll}
A_{11}X_1 & +A_{12}X_2 & & & & & = B_1 \\
 & A_{22}^{(1)}X_2 & +A_{23}X_3 & & & & = B_2^{(1)} \\
 & & A_{33}^{(2)}X_3 & +A_{34}X_4 & & & = B_3^{(2)} \\
 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
 & & & & A_{n-1n-1}^{(n-2)}X_{n-1} & +A_{n-1n}X_n & = B_{n-1}^{(n-2)} \\
 & & & & & A_{nn}^{(n-1)}X_n & = B_n^{(n-1)}
\end{array} \tag{2.34}$$

where

$$A_{ii}^{(i-1)} = A_{ii} - A_{ii-1}A_{i-1i-1}^{-(i-2)}A_{i-1i},$$
$$B_i^{(i-1)} = B_i - A_{ii-1}A_{i-1i-1}^{-(i-2)}B_{i-1}^{-(i-2)},$$

$i = 2, 3, ..., n.$
Hence, the solution is

$$\begin{aligned}
&X_n = A_{nn}^{-(n-1)}B_n^{(n-1)}, \\
&X_i = A_{ii}^{-(i-1)}[B_i^{(i-1)} - A_{ii+1}X_{i+1}], \quad \text{i=n-1,n-2,...,1.}
\end{aligned} \tag{2.35}$$

The following theorem holds:

Theorem 2.1 *(cf. [5], p.112) If the block matrices* A_{ii}, $i = 1, 2, ..., n$ *are non-singular and satisfy the inequalities*

$$\| A_{ii}^{-1} \|_1 (\| A_{ii-1} \|_1 + \| A_{ii+1} \|) < 1$$

for $i = 1, 2, ..., n$. *when* $A_{10} = A_{nn+1} = 0$,
then the Gauss block elimination can be completed and the solution X *of the system of equations (2.32) is determined by the formula (2.35).*

Example 2.5 *Let us solve the system of equations*

$$\begin{array}{rrrrrrr}
4x_1 & -x_2 & +x_3 & & & & = 5 \\
-x_1 & +4x_2 & & +x_4 & & & = 11 \\
x_1 & & +4x_3 & -x_4 & +x_5 & & = 14 \\
& x_2 & -x_3 & +4x_4 & & +x_6 & = 21 \\
& & x_3 & & +4x_5 & -x_6 & = 17 \\
& & & x_4 & -x_5 & +4x_6 & = 23
\end{array}$$

by Gauss block elimination.

Solution. We can write the above system of equations in the block form

$$\begin{array}{rcrcrl}
A_{11}X_1 & + & A_{12}X_2 & & & = B_1 \\
A_{21}X_1 & + & A_{22}X_2 & + & A_{23}X_3 & = B_2 \\
& & A_{32}X_2 & + & A_{33}X_3 & = B_3
\end{array} \tag{2.36}$$

where the vector $B = (B_1, B_2, B_3)^T$ has the components

$$B_1 = \begin{bmatrix} 5 \\ 11 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 14 \\ 21 \end{bmatrix}, \quad B_3 = \begin{bmatrix} 17 \\ 23 \end{bmatrix},$$

and the unknown vector $X = (X_1, X_2, X_3)^T$ has the components

$$X_1 = \begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \quad X_2 = \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}, \quad X_3 = \begin{bmatrix} x_{31} \\ x_{32} \end{bmatrix},$$

and the matrices

$$A_{11} = A_{22} = A_{33} = \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix}, \quad A_{12} = A_{21} = A_{23} = A_{32} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Multiplying from the left first row-block in (2.36) by the matrix

$$M_{21} = A_{21}A_{11}^{-1} = \frac{1}{15} \begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix},$$

and then subtracting the result from the second row-block, we obtain *the first reduced system of equations:*

$$\begin{array}{rcrcrcl} A_{11}X_1 & + & A_{12}X_2 & & & = & B_1 \\ & & A_{22}^{(1)}X_2 & + & A_{23}X_3 & = & B_2^{(1)} \\ & & A_{32}X_2 & + & A_{33}X_3 & = & B_3 \end{array} \tag{2.37}$$

where

$$A_{22}^{(1)} = A_{22} - A_{21}A_{11}^{-1}A_{12} = \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix} - \frac{1}{15}\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix} = \frac{1}{15}\begin{bmatrix} 56 & -16 \\ -16 & 56 \end{bmatrix},$$

$$B_2^{(1)} = B_2 - A_{21}A_{11}^{-1}B_1 = \begin{bmatrix} 14 \\ 21 \end{bmatrix} - \frac{1}{15}\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}\begin{bmatrix} 5 \\ 11 \end{bmatrix} = \frac{1}{15}\begin{bmatrix} 179 \\ 266 \end{bmatrix}.$$

To eliminate the unknown X_2 from third equation in (2.37), we multiply second row-block by the matrix

$$M_{32} = A_{32}A_{22}^{-(1)} = \frac{1}{24}\begin{bmatrix} 7 & 2 \\ 2 & 7 \end{bmatrix}$$

and we subtract the result from third row-block in (2.37). Then, we obtain *the second reduced system of equations:*

$$\begin{array}{rcrcrcl} A_{11}X_1 & + & A_{12}X_2 & & & = & B_1 \\ & & A_{22}^{(1)}X_2 & + & A_{23}X_3 & = & B_2^{(1)} \\ & & & + & A_{33}^{(2)}X_3 & = & B_3^{(2)} \end{array} \tag{2.38}$$

where

$$A_{33}^{(2)} = A_{33} - A_{32}A_{22}^{-(1)}A_{23} = \begin{bmatrix} 4 & -1 \\ -1 & 4 \end{bmatrix} - \frac{1}{24}\begin{bmatrix} 7 & 2 \\ 2 & 7 \end{bmatrix} = \frac{1}{24}\begin{bmatrix} 89 & -26 \\ -26 & 89 \end{bmatrix},$$

$$B_3^{(2)} = B_3 - A_{22}^{-(1)}B_2^{(1)} = \begin{bmatrix} 17 \\ 23 \end{bmatrix} - \frac{1}{24}\begin{bmatrix} 7 & 2 \\ 2 & 7 \end{bmatrix}\frac{1}{15}\begin{bmatrix} 179 \\ 266 \end{bmatrix} = \frac{1}{24}\begin{bmatrix} 289 \\ 404 \end{bmatrix}.$$

Hence, by formula (2.35), we obtain

$$X_3 = A_{33}^{-(2)}B_3^{(2)} = \frac{24}{7245}\begin{bmatrix} 89 & 26 \\ 26 & 89 \end{bmatrix}\frac{1}{24}\begin{bmatrix} 289 \\ 404 \end{bmatrix} = \frac{1}{7245}\begin{bmatrix} 36225 \\ 43470 \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \end{bmatrix},$$

$$X_2 = A_{22}^{-(1)}[B_2^{(1)} - A_{23}X_3] = \frac{1}{24}\begin{bmatrix} 7 & 2 \\ 2 & 7 \end{bmatrix}\left\{\frac{1}{15}\begin{bmatrix} 179 \\ 266 \end{bmatrix} - \begin{bmatrix} 5 \\ 6 \end{bmatrix}\right\} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}.$$

$$X_1 = A_{11}^{-1}[B_1 - A_{12}X_2] = \frac{1}{15}\begin{bmatrix} 4 & 1 \\ 1 & 4 \end{bmatrix}\left\{\begin{bmatrix} 5 \\ 11 \end{bmatrix} - \begin{bmatrix} 3 \\ 4 \end{bmatrix}\right\} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

2.8 Gauss Elimination for Pentediagonal Matrices.

Let us write the system of linear equations (2.1) in the case when A is a pentadiagonal matrix

$$\begin{aligned}
&a_{11}x_1 + a_{12}x_2 + a_{13}x_3 && = a_{1n+1}\\
&a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 && = a_{2n+1}\\
&a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 + a_{35}x_5 && = a_{3n+1}\\
&\quad a_{42}x_2 + a_{43}x_3 + a_{44}x_4 + a_{45}x_5 + a_{46}x_6 && = a_{4n+1}\\
& && \ddots \quad \vdots\\
& && \cdots + a_{nn}x_n \quad = a_{nn+1}
\end{aligned}$$

This system of equations has the following pentadiagonal structure:

$$\begin{bmatrix}
* & * & * & & & & & & & \\
* & * & * & * & & & & & & \\
* & * & * & * & * & & & & & \\
 & * & * & * & * & * & & & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\
 & & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\
 & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & & & & & * & * & * & * & * \\
 & & & & & & * & * & * & * \\
 & & & & & & & * & * & * \\
 & & & & & & & & * & *
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ \vdots \\ \vdots \\ x_{n-3} \\ x_{n-2} \\ x_{n-1} \\ x_n \end{bmatrix}
=
\begin{bmatrix} * \\ * \\ * \\ * \\ \vdots \\ \vdots \\ \vdots \\ * \\ * \\ * \\ * \end{bmatrix}$$

where $* = a_{ik} = a_{ik}^{(0)}, \quad i = 1, 2, ..., n,\ k = 1, 2, ..., n+1.$
Eliminating unknown x_1, we obtain the first reduced system of equations which has the following structure

$$\begin{bmatrix}
* & * & * & & & & & & & \\
 & *^{(1)} & *^{(1)} & * & & & & & & \\
 & *^{(1)} & *^{(1)} & * & * & & & & & \\
 & * & * & * & * & * & & & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\
 & & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\
 & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & & & & & * & * & * & * & * \\
 & & & & & & * & * & * & * \\
 & & & & & & & * & * & * \\
 & & & & & & & & * & *
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ \vdots \\ \vdots \\ x_{n-3} \\ x_{n-2} \\ x_{n-1} \\ x_n \end{bmatrix}
=
\begin{bmatrix} * \\ *^{(1)} \\ *^{(1)} \\ * \\ \vdots \\ \vdots \\ \vdots \\ * \\ * \\ * \\ * \end{bmatrix}$$

where

$$*_{ik}^{(1)} = a_{ik} - m_{i1}a_{1k}, \qquad m_{i1} = \frac{a_{i1}}{a_{11}}, \qquad i = 2, 3; \quad k = 2, 3, n+1.$$

After second step of elimination, we obtain the second reduced system of equations which has the following structure

$$\begin{bmatrix}
* & * & * & & & & & & & \\
 & *^{(1)} & *^{(1)} & * & & & & & & \\
 & & *^{(2)} & *^{(2)} & * & & & & & \\
 & & *^{(2)} & *^{(2)} & * & * & & & & \\
 & & & \ddots & \ddots & \ddots & \ddots & & & \\
 & & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\
 & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & & & & & * & * & * & * & * \\
 & & & & & & * & * & * & * & * \\
 & & & & & & & * & * & * & * \\
 & & & & & & & & * & * & *
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ \vdots \\ \vdots \\ x_{n-3} \\ x_{n-2} \\ x_{n-1} \\ x_n \end{bmatrix}
=
\begin{bmatrix} * \\ *^{(1)} \\ *^{(2)} \\ *^{(2)} \\ \vdots \\ \vdots \\ \vdots \\ * \\ * \\ * \\ * \end{bmatrix}$$

where

$$*_{ik}^{(2)} = \begin{cases}
a_{ik}^{(1)} - m_{i2}a_{2k}^{(1)}, & m_{i2} = \dfrac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad i = 3, \ k = 3, n+1, \\
a_{ik} - m_{i2}a_{2k}, & m_{i2} = \dfrac{a_{i2}^{(1)}}{a_{22}^{(1)}}, \quad i = 3, \ k = 4, \\
a_{ik} - m_{i2}a_{2k}^{(1)}, & m_{i2} = \dfrac{a_{i2}}{a_{22}^{(1)}}, \quad i = 4, \ k = 3, 4, n+1
\end{cases}$$

Continuing elimination of successive unknowns, we obtain the following upper triangular system of equations:

$$\begin{bmatrix}
* & * & * & & & & & & \\
 & *^{(1)} & *^{(1)} & * & & & & & \\
 & & *^{(2)} & *^{(2)} & * & & & & \\
 & & & *^{(3)} & *^{(3)} & * & & & \\
 & & & & \ddots & \ddots & \ddots & & \\
 & & & & & \ddots & \ddots & \ddots & \\
 & & & & & & *^{(n-3)} & *^{(n-3)} & * \\
 & & & & & & & *^{(n-2)} & *^{(n-2)} \\
 & & & & & & & & *^{(n-1)}
\end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ \vdots \\ x_{n-2} \\ x_{n-1} \\ x_n \end{bmatrix}
=
\begin{bmatrix} * \\ *^{(1)} \\ *^{(2)} \\ *^{(3)} \\ \vdots \\ \vdots \\ *^{(n-3)} \\ *^{(n-2)} \\ *^{(n-1)} \end{bmatrix}$$

In general, the coefficients are determined by the formulas

$$*_{ik}^{(s)} = \begin{cases} a_{ik}^{(s-1)} - m_{is}a_{sk}^{(s-1)}, & m_{is} = \dfrac{a_{is}^{(s-1)}}{a_{ss}^{(s-1)}}, \quad i = s+1,\ k = s+1, n+1, \\ a_{ik} - m_{is}a_{sk}, & m_{is} = \dfrac{a_{is}^{(s-1)}}{a_{ss}^{(s-1)}}, \quad i = s+1,\ k = s+2, \\ a_{ik} - m_{is}a_{sk}^{(s-1)}, & m_{is} = \dfrac{a_{is}}{a_{ss}^{(s-1)}}, \quad i = s+2,\ k = s+1, s+2, n+1 \end{cases}$$

for $s = 2, 3, ..., n-2$, and for $s = n-1$, we have

$$*_{nn}^{(n-1)} = a_{nn}^{(n-2)} - m_{nn-1}a_{n-1n}^{(n-2)}, \quad m_{nn-1} = \frac{a_{nn-1}^{(n-2)}}{a_{n-1n-1}^{(n-2)}}.$$

and

$$*_{nn+1}^{(n-1)} = a_{nn+1}^{(n-2)} - m_{nn-1}a_{n-1n+1}^{(n-2)}.$$

Hence, by backward substitution, we find the solution

$$\begin{aligned} x_n &= \frac{a_{nn+1}^{(n-1)}}{a_{nn}^{(n-1)}} \\ x_{n-1} &= \frac{1}{a_{n-1n-1}^{(n-2)}}[a_{n-1n+1}^{(n-2)} - a_{n-1n}^{(n-2)}x_n] \\ x_{n-2} &= \frac{1}{a_{n-2n-2}^{(n-3)}}[a_{n-2n+1}^{(n-3)} - a_{n-2n-1}^{(n-3)}x_{n-1} - a_{n-2n}x_n] \\ x_{n-3} &= \frac{1}{a_{n-3n-3}^{(n-4)}}[a_{n-3n+1}^{(n-4)} - a_{n-3n-2}^{(n-4)}x_{n-2} - a_{n-3n-1}x_{n-1}] \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_s &= \frac{1}{a_{ss}^{(s-1)}}[a_{sn+1}^{(s-1)} - a_{ss+1}^{(s-1)}x_{s+1} - a_{ss+2}x_{s+2}] \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ x_2 &= \frac{1}{a_{22}^{(1)}}[a_{2n+1}^{(1)} - a_{23}^{(1)}x_3 - a_{24}x_4] \\ x_1 &= \frac{1}{a_{11}}[a_{1n+1} - a_{12}x_2 - a_{13}x_3] \end{aligned} \tag{2.39}$$

The module `solvefive` in `Mathematica` solves a system of linear equations with a pentadiagonal matrix A. The input entries of the pentadiagonal matrix are to be stored on the following list

$$\begin{aligned} a = \{&\{a_{31}, a_{42}, ..., a_{nn-1}\}, \{a_{21}, a_{32}, ..., a_{nn-1}\}, \\ &\{a_{11}, a_{22}, ..., a_{nn}\}, \{a_{12}, a_{23}, ..., a_{n-1n}\}, \\ &\{a_{13}, a_{24}, ..., a_{n-2,n}\}, \{a_{n+1,1}, a_{n+1,2}, ..., a_{n+1,n}\}\}. \end{aligned}$$

```
solvefive[a_]:=
Module[{a1,a2,d,a3,a4,f,x,x1,y,n},
{a1,a2,d,a3,a4,f}=Table[a[[i]],{i,1,6}];
    n=Length[d];
    Do[x=a2[[i-1]]/d[[i-1]];
    d[[i]]=d[[i]]-x*a3[[i-1]];
    a3[[i]]=a3[[i]]-x*a4[[i-1]];
    f[[i]]=f[[i]]-x*f[[i-1]];
    x=a1[[i-1]]/d[[i-1]];
    a2[[i]]=a2[[i]]-x*a3[[i-1]];
    d[[i+1]]=d[[i+1]]-x*a4[[i-1]];
    f[[i+1]]=f[[i+1]]-x*f[[i-1]],{i,2,n-1}];
    x1=a2[[n-1]]/d[[n-1]];
    d[[n]]=d[[n]]-x1*a3[[n-1]];
    y[n]=(f[[n]]-x1*f[[n-1]])/d[[n]];
    y[n-1]=(f[[n-1]]-a3[[n-1]]*y[n])/d[[n-1]];
    y[i_]:=y[i]=
    (f[[i]]-a4[[i]]*y[i+2]-a3[[i]]*y[i+1])/d[[i]];
    Table[y[i],{i,1,n}]
   ]
```

Entering the input data

```
a={{1.,1.,1.,1.,1.,0.},
   {-16.,-16.,-16.,-16.,-16.,-16.,-12.},
   {24.,30.,30.,30.,30.,30.,30.,24.},
   {-12.,-16.,-16.,-16.,-16.,-16.,-16.},
   {0.,1.,1.,1.,1.,1.},
   {12.,-1.,0.,0.,0.,0.,-1.,12.}};
```

we obtain the solution $x = \{1, 1, 1, 1, 1, 1, 1, 1\}$ by executing the command

```
solvefive[a].
```

2.9 Exercises

Question 2.1 *Solve the following system of equations:*

$$\begin{array}{rcrcrcrcl} 5x_1 & + & x_2 & + & 2x_3 & + & 5x_4 & = & 1 \\ 10x_1 & + & 2x_2 & - & 6x_3 & + & 9x_4 & = & 4 \\ 3x_1 & - & 2x_2 & + & 4x_3 & + & x_4 & = & 2 \\ 15x_1 & - & 2x_2 & - & x_3 & + & 10x_4 & = & 8 \end{array} \tag{2.40}$$

using

1. *(a) partial pivoting,*

(b) full pivoting.

Question 2.2 *Using a calculator, solve the following system of equations:*

$$\begin{aligned} 0.000003x_1 &+ 0.001x_2 &= 6 \\ 10x_1 &+ 3333.333x_2 &= 19999999 \end{aligned}$$

by

1. *(a) Gauss elimination without any pivoting,*
   (b) Gauss elimination with partial pivoting,
   (c) Gauss elimination with full pivoting.

Note that the exact solution : $x_1 = 1000000, \quad x_2 = 3000.$
Explain why Gauss elimination fails to get the accurate solution.

Question 2.3 .
(a). *Solve the following system of equations:*

$$\begin{aligned} 3x_1 &+ x_2 &+ 2x_3 &+ 3x_4 &= 10 \\ 6x_1 &+ 4x_2 &- 6x_3 &+ 9x_4 &= 25 \\ 9x_1 &- 6x_2 &+ 4x_3 &+ 8x_4 &= 20 \\ 15x_1 &- 8x_2 &- x_3 &+ 10x_4 &= 32 \end{aligned} \tag{2.41}$$

(b). *Find LU-decomposition of the matrix*

$$A = \begin{bmatrix} 3 & 1 & 2 & 3 \\ 6 & 4 & -6 & 9 \\ 9 & -6 & 4 & 8 \\ 15 & -8 & -1 & 10 \end{bmatrix}.$$

Question 2.4 *Solve the following system of linear equations by the root square method*

$$\begin{aligned} 4x_1 &+ 2x_2 &+ x_3 &+ x_4 &= 0 \\ 2x_1 &+ 6x_2 &+ x_3 &- x_4 &= 4 \\ x_1 &+ x_2 &+ 5x_3 &+ 2x_4 &= 27 \\ x_1 &- x_2 &+ 2x_3 &+ 7x_4 &= 19 \end{aligned}$$

Question 2.5 .

1. *(a) Find the upper-triangular form of the system of linear equations using Gauss elimination method*

$$\begin{aligned} 2x_1 &+ 3x_2 &- x_3 &= 1 \\ 4x_1 &+ 2x_2 &+ x_3 &= 2 \\ 6x_1 &+ x_2 &+ 4x_3 &= 7 \end{aligned}$$

Solve the above system of equations.

(b) Find the LU-decomposition of the matrix

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

Calculate the determinant of the matrix A using the LU-decomposition.

Question 2.6 *Consider the following system of equations:*

$$\begin{array}{ccccccccccc}
4x_1 & + & x_2 & & & & & & & & = 1 \\
x_1 & + & 4x_2 & + & x_3 & & & & & & = 4 \\
 & & x_2 & + & 4x_3 & + & x_4 & & & & = 9 \\
 & & & & \cdots & \cdots & \cdots & \cdots & \cdots & & \cdots \\
 & & & & x_{i-1} & + & 4x_i & + & x_{i+1} & & = i^2 \\
 & & & & \cdots & \cdots & \cdots & \cdots & \cdots & & \cdots \\
 & & & & & & x_{n-1} & + & 4x_n & & = n^2
\end{array}$$

Write an algorithm based on Gauss elimination to solve the above system of equations. Find the solution when $n = 10$.

Question 2.7 .

Derive the algorithm based on Square Root Method for solving the system of equations $Ax = F$*, where the tri-diagonal matrix*

$$A = \left\{ \begin{array}{ccccccccc}
a & b & 0 & 0 & 0 & \cdots & 0 & & 0 \\
b & a & b & 0 & 0 & \cdots & 0 & & 0 \\
0 & b & a & b & 0 & \cdots & 0 & & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & a \geq 2b > 0 \\
0 & 0 & 0 & 0 & 0 & \cdots & b & & a
\end{array} \right\}$$

(b) *Use the algorithm, which you have found in* (a)*, to solve the system of equatins*

$$\begin{aligned}
4x_1 - x_2 &= 3 \\
-x_1 + 4x_2 - x_3 &= 2 \\
-x_2 + 4x_3 - x_4 &= 2 \\
-x_3 + 4x_4 - x_5 &= 2 \\
-x_4 + 4x_5 &= 3
\end{aligned}$$

Question 2.8 .

Consider the system of equations

$$\begin{aligned}
3x_1 - x_2 &= F_1 \\
-x_1 + 3x_2 - x_3 &= F_2 \\
\ldots\ldots\ldots\ldots\ldots &\ldots \ldots\ldots \\
-x_{i-1} + 3x_i - x_{i+1} &= F_i, \qquad i = 2, 3, ..., n-1, \\
\ldots\ldots\ldots\ldots\ldots &\ldots \ldots\ldots \\
-x_{n-2} + 3x_{n-1} - x_n &= F_{n-1} \\
-x_4 + 3x_5 &= F_n
\end{aligned}$$

(a) Derive the algorithm based on Gause Elimination Method for solving the system of equations and show that the algorithm is numerically stable.

(b) Use the algorithm which you have found in (a) to solve the system of equatins

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

Chapter 3

Eigenvalues and Eigenvectors of a Matrix

3.1 Eigenvalue Problem

In this chapter, we shall consider the following eigenvalue problem:

Find all real or complex values of λ *and corresponding non-zero vectors* $x = (x_1, x_2, \ldots, x_n)^T \neq 0$, *such that*

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n-1} & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn-1} & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}. \tag{3.1}$$

Clearly, this system of equations possesses non-zero solutions if and only if the homogeneous system of equations

$$\begin{array}{rcrcrcrcl} (a_{11}-\lambda)x_1 & + & a_{12}x_2 & + & a_{13}x_3 & + \cdots + & a_{1n}x_n & = & 0 \\ a_{21}x_1 & + & (a_{22}-\lambda)x_2 & + & a_{23}x_3 & + \cdots + & a_{2n}x_n & = & 0 \\ \cdots\cdots & & \cdots\cdots\cdots & & \cdots\cdots & \cdots\cdots & \cdots\cdots & \cdots & \\ a_{n1}x_1 & + & a_{n2}x_2 & + & a_{n3}x_3 & + \cdots + & (a_{nn}-\lambda)x_n & = & 0 \end{array} \tag{3.2}$$

has non-zero solutions. It is well known, the homogeneous system of equations (3.2) has non-zero solutions if and only if the determinant

$$\Delta_n(\lambda) = \begin{vmatrix} a_{11}-\lambda & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22}-\lambda & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33}-\lambda & \cdots & a_{3n-1} & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn-1} & a_{nn}-\lambda \end{vmatrix} = 0 \tag{3.3}$$

Let us note that

$$\Delta_n(\lambda) = (-1)^n\lambda^n + a_{n-1}\lambda^{n-1} + a_{n-2}\lambda^{n-2} + \cdots + a_1\lambda + a_0,$$

is the polynomial of degree n with the leading term $(-1)^n\lambda^n$. This polynomial is called *characteristic polynomial of the matrix*

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n-1} & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn-1} & a_{nn} \end{bmatrix}$$

Thus, all eigenvalues of the matrix A are roots of the characteristic polynomial $\Delta_n(\lambda)$. Let a non-zero eigenvector $X^{(k)}$ corresponds to the root λ_k, so that

$$AX^{(k)} = \lambda_k X^{(k)}, \qquad k = 1, 2, ..., n.$$

For a real and symmetric matrix A, there exists exactly n orthonormal eigenvectors $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$, i.e.

$$(X^{(k)}, X^{(l)}) = \begin{cases} 1 & if \quad k = l, \\ 0 & if \quad k \neq l \end{cases}$$

where $X^{(k)} = [x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}]$ and $(X^{(k)}, X^{(l)}) = \sum_{i=1}^{n} x_i^{(k)} x_i^{(l)}$.

A matrix A for which there exists an orthonormal base of its eigenvectors is diagonalizable by the orthonormal matrix

$$X = \begin{bmatrix} x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(n)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(n)} \\ \cdots & \cdots & \cdots & \cdots \\ x_n^{(1)} & x_n^{(2)} & \cdots & x_n^{(n)} \end{bmatrix},$$

This means that

$$X^T A X = \Lambda,$$

where X^T denotes transposed matrix to X and

$$\Lambda = diagonal(\lambda_1, \lambda_2, \ldots, \lambda_n) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

Let us note that any matrix A can be transformed either to a diagonal form or to a Jordan form (cf. [6]). In the case when a matrix A (symmetric or not) possesses all distinct eigenvalues, so that $\lambda_k \neq \lambda_l$, for $k \neq l$, then there exists a non-singular matrix T such that

$$T^{-1} A T = \Lambda.$$

and A is a diagonalizable matrix.

Example 3.1 *Let us find all eigenvalues and corresponding eigenvectors for the matrix*

$$A = \begin{bmatrix} 20 & 6 & 8 \\ 6 & 20 & 0 \\ 8 & 0 & 20 \end{bmatrix}.$$

Solution Evidently, the characteristic polynomial of A is

$$\Delta_3(\lambda) = \begin{vmatrix} 20-\lambda & 6 & 8 \\ 6 & 20-\lambda & 0 \\ 8 & 0 & 20-\lambda \end{vmatrix} = -\lambda^3 + 60\lambda^2 - 1100\lambda + 6000.$$

The eigenvalues of the matrix A are the roots of the equation

$$\Delta_3(\lambda) = 0.$$

and these roots are:

$$\lambda_1 = 10, \quad \lambda_2 = 20 \quad \text{and} \quad \lambda_3 = 30.$$

In order to find eigenvectors corresponding to the eigenvalues $\lambda_1, \quad \lambda_2$ and λ_3, we shall solve the following homogeneous system of linear equations:

$$\begin{array}{rcrcrcl} (20-\lambda)x_1 & + & 6x_2 & + & 8x_3 & = & 0 \\ 6x_1 & + & (20-\lambda)x_2 & & & = & 0 \\ 8x_1 & & & + & (20-\lambda)x_3 & = & 0 \end{array} \tag{3.4}$$

when $\lambda_1 = 10, \quad \lambda_2 = 20$ and $\lambda_3 = 30$.
Thus, for $\lambda_1 = 10$, the homogeneous system of equations

$$\begin{array}{rcrcrcl} 10x_1 & + & 6x_2 & + & 8x_3 & = & 0 \\ 6x_1 & + & 10x_2 & & & = & 0 \\ 8x_1 & & & + & 10x_3 & = & 0 \end{array} \tag{3.5}$$

has the normalized solution

$$X^{(1)} = [-\frac{1}{\sqrt{2}}, \frac{3}{5\sqrt{2}}, \frac{4}{5\sqrt{2}}].$$

For $\lambda_2 = 20$, the homogeneous system of equations

$$\begin{array}{rcrcl} & & 6x_2 & + & 8x_3 & = & 0 \\ 6x_1 & & & & & = & 0 \\ 8x_1 & & & & & = & 0 \end{array} \tag{3.6}$$

has the normalized solution

$$X^{(2)} = [0, -\frac{4}{5}, \frac{3}{5}].$$

For $\lambda_3 = 30$, the homogeneous system of equations

$$\begin{array}{rcrcrcl} -10x_1 & + & 6x_2 & + & 8x_3 & = & 0 \\ 6x_1 & - & 10x_2 & & & = & 0 \\ 8x_1 & & & - & 10x_3 & = & 0 \end{array} \tag{3.7}$$

has the normalized solution

$$X^{(3)} = [\frac{1}{\sqrt{2}}, \frac{3}{5\sqrt{2}}, \frac{4}{5\sqrt{2}}].$$

One can check that, $X^{(1)}, X^{(2)}$ and $X^{(3)}$ are orthonormal eigenvectors, so that, the orthonormal matrix

$$X = \begin{bmatrix} -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{3}{5\sqrt{2}} & -\frac{4}{5} & \frac{3}{5\sqrt{2}} \\ \frac{4}{5\sqrt{2}} & \frac{3}{5} & \frac{4}{5\sqrt{2}} \end{bmatrix}$$

transforms the matrix A to the following diagonal matrix:

$$X^T A X = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 30 \end{bmatrix} = \Lambda.$$

Example 3.2 *Let us find eigenvalues and eigenvectors of the tri-diagonal matrix*

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 2 \end{bmatrix}_{(n \times n)}$$

Solution. The eigenvalues λ_k and corresponding eigenvectors $X^{(k)}$, $k = 1, 2, \ldots, n$, satisfy the following system of linear equations:

$$\begin{array}{rcrcrcl} 2x_1 & - & x_2 & & & = & \lambda x_1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ -x_{k-1} & + & 2x_k & - & x_{k+1} & = & \lambda x_k \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \\ -x_{n-1} & + & 2x_n & & & = & \lambda x_n \end{array} \tag{3.8}$$

In order to find all non-zero solutions of equations (3.8), we shall substitute to (3.8)

$$x_k = \alpha^k, \quad k = 1, 2, \ldots, n,$$

where α is a parameter.
Then, we obtain

$$\alpha^2 - (2 - \lambda)\alpha + 1 = 0.$$

All non-zero bounded solutions of (3.8) correspond to the complex roots of the quadratic equation for $0 < \lambda < 4$. Thus, the non-zero bounded solutions of the system of equations (3.8) are:

$$x_k^{(1)} = sin\ k\psi \qquad x_k^{(2)} = cos\ k\psi, \quad k = 1, 2, \ldots, n-1.$$

From the first equation in (3.8), we have

$$\lambda = 2(1 - cos\ \psi).$$

From the last equation in (3.8), we have

$$sin(n+1)\psi = 0.$$

Hence

$$\psi_k = \frac{k\pi}{n+1}, \quad k = 1, 2, \ldots, n.$$

So that, the eigenvalues of the matrix A

$$\lambda_k = 2(1 - cos\ \psi_k) = 4sin^2\frac{\psi_k}{2} = 4sin^2\frac{k\pi}{2(n+1)}, \quad k = 1, 2, \ldots, n$$

and the eigenvectors

$$X^{(k)} = [x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}],$$

where

$$x_s^{(k)} = sin\ \frac{sk\pi}{n+1}, \ \text{ for } \ k, s = 1, 2, \ldots, n.$$

The eigenvectors $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ are orthogonal. Indeed, it can be proved in an elementary way that

$$(X^{(k)}, X^{(l)}) = \sum_{s=1}^{n} x_s^{(k)} x_s^{(l)} = \sum_{s=1}^{n} sin\frac{sk\pi}{n+1} sin\frac{sl\pi}{n+1} = \begin{cases} \frac{n+1}{2} & if \quad k = l, \\ 0 & if \quad k \neq l. \end{cases}$$

Thus, the orthonormal set of eigenvectors of the matrix A is

$$\overline{X}^{(k)} = \sqrt{\frac{2}{n+1}} X^{(k)} = \sqrt{\frac{2}{n+1}} \begin{bmatrix} sin\frac{k\pi}{n+1} \\ sin\frac{2k\pi}{n+1} \\ sin\frac{3k\pi}{n+1} \\ \vdots \\ sin\frac{nk\pi}{n+1} \end{bmatrix}, \quad k = 1, 2, \ldots, n.$$

Let us note that any Hermitian matrix A [1] possesses all real eigenvalues. If A is a real symmetric matrix then A has also a real orthonormal base of eigenvectors. Indeed, we have

$$AX = \lambda X$$

and the scalar

$$X^*AX = \lambda X^*X.$$

Since $A^* = A$, we get

$$(X^*AX)^* = X^*A^*X^{**} = X^*AX$$

Then, the scalars X^*X and X^*AX are real. Therefore, λ must be also real.

3.2 Jacobi Method for Real and Symmetric Matrices

The idea of Jacobi method is to find an orthonormal matrix V (i.e. $V^{-1} = V^T$) such that

$$V^TAV = \Lambda,$$

where Λ is a diagonal matrix, V^T is transposed matrix to V and V^{-1} is the inverse matrix to V. As we know, such unitary matrix V exists for any real and symmetric matrix A. Evidently, if A is a diagonal matrix then $V = E$ is a unite matrix. Let A be a non-diagonal matrix. Then, we may choose k and l such that

$$\mid a_{kl} \mid = \max_{i,j=1,2,\ldots,n;\ i\neq j} \mid a_{ij} \mid > 0.$$

Now, let us consider the orthogonal matrix

$$C^{(1)} = \begin{array}{c} \begin{matrix} & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \\ & & & & & & & & & & \end{matrix} \\ \left[\begin{array}{ccccccccccc} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & \cos\psi & 0 & \cdots & 0 & -\sin\psi & & & \\ & & & 0 & 1 & \cdots & 0 & 0 & & & \\ & & & \vdots & & \ddots & & \vdots & & & \\ & & & 0 & 0 & \cdots & 1 & 0 & & & \\ & & & \sin\psi & 0 & \cdots & 0 & \cos\psi & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{array}\right] \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{l} \\ \\ \\ \leftarrow row_k \\ \\ \\ \\ \leftarrow row_l \\ k < l \\ \\ \\ \end{array} \tag{3.9}$$

$$\begin{matrix} column & & column \\ k & & l \\ \downarrow & & \downarrow \end{matrix}$$

[1] A is a Hermitian matrix if $A^* = A$, where A^* denotes transposed to A with conjugate entries of A

We can determine the angle ψ in such a way to nullify the entry $a_{kl}^{(1)}$ of the matrix $C^{(1)T}AC^{(1)}$. Namely, let $\overline{a}_{kl}$ be entry of the matrix $AC^{(1)}$. Then, we find

$$\begin{array}{l} \overline{a}_{kl} = -\sin\psi\ a_{kk} + \cos\psi\ a_{kl} \\ \overline{a}_{ll} = -\sin\psi\ a_{lk} + \cos\psi\ a_{ll} \\ a_{kl}^{(1)} = \cos\psi\ \overline{a}_{kl} + \sin\psi\ \overline{a}_{ll} \end{array} \tag{3.10}$$

Hence, we have

$$a_{kl}^{(1)} = \cos\psi(-\sin\psi\ a_{kk} + \cos\psi\ a_{kl}) + \sin\psi(-\sin\psi\ a_{kl} + \cos\psi\ a_{ll}) = a_{kl}((\cos\psi)^2 - (\sin\psi)^2) + \cos\psi\sin\psi(a_{ll} - a_{kk}).$$

and $a_{kl}^{(1)} = 0$ if the angle ψ satisfies the following equation:

$$a_{kl}(\cos\psi)^2 + (a_{ll} - a_{kk})\cos\psi\ \sin\psi - a_{kl}(\sin\psi)^2 = 0.$$

So that

$$a_{kl}\cos 2\psi - \frac{1}{2}(a_{kk} - a_{ll})\sin 2\psi = 0.$$

and

$$tan2\psi = \begin{cases} \dfrac{2a_{kl}}{a_{kk} - a_{ll}} & if \quad a_{kk} \neq a_{ll}, \\ \infty & if \quad a_{kk} = a_{ll}. \end{cases}$$

Hence, we get

$$\psi_k = \begin{cases} \dfrac{1}{2} arctan \dfrac{2a_{kl}}{a_{kk} - a_{ll}} & if \quad a_{kk} \neq a_{ll}, \\ \dfrac{\pi}{4} & if \quad a_{kk} = a_{ll}, \end{cases}$$

We can transform matrix A to an almost diagonal form by the orthogonal mappings $C^{(1)}, C^{(2)}, \ldots, C^{(r)}$. of the form (3.9). Then, we shall show that the sequence of matrices

$$\begin{array}{l} A^{(0)} = A, \\ A^{(1)} = C^{(1)T}AC^{(1)}, \\ A^{(2)} = C^{(2)T}C^{(1)T}AC^{(1)}C^{(2)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ A^{(r)} = C^{(r)T}C^{(r-1)T}\ldots C^{(1)T}AC^{(1)}C^{(2)}\ldots C^{(r)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{array} \tag{3.11}$$

converges to a diagonal matrix Λ.
Indeed, let

$$A^{(q)} = \{a_{ij}^{(q)}\}, \quad i, j = 1, 2, \ldots, n; \quad q = 0, 1, \ldots, r.$$

The matrices $A^{(q+1)}, \quad q = 0, 1, \ldots, r-1;$ are determined by the condition

$$a_{k_q l_q}^{(q+1)} = 0, \tag{3.12}$$

where

$$a_{k_q l_q}^{(q+1)} = \max_{i,j=1,2,\ldots,n;\ i\neq j} \mid a_{ij}^{(q)} \mid . \tag{3.13}$$

In order to prove that the sequence (3.11) converges to a diagonal matrix Λ, it is sufficient to show that the non-diagonal entries of $A^{(r)}$ tend to zero when $r \to \infty$, so that

$$\lim_{r\to\infty} \mu^{(r)} = 0,$$

where

$$\mu^{(r)} = \sum_{i,j=1,\ i\neq j}^{n} [a_{ij}^{(r)}]^2. \tag{3.14}$$

Let

$$S(A) = \sum_{i,j=1}^{n} a_{ij}^2 \ .$$

Then, the following equality holds:

$$S(A) = Sp(A^T A) = Sp(A^2)$$

for a symmetric matrix A, where $Sp(A) = \sum_{i=1}^{n} a_{ii}$ is the trace of the matrix A. For two symmetric matrices B and $M = A^T B A$, we have

$$S(M) = Sp(M^2) = Sp((A^T B A)^2) = Sp(A^{-1} B A) = Sp(B^2) = Sp(B). \tag{3.15}$$

Next, let

$$\overline{A} = \begin{bmatrix} a_{kk} & a_{kl} \\ a_{lk} & a_{ll} \end{bmatrix}, \quad \overline{M} = \begin{bmatrix} m_{kk} & m_{kl} \\ m_{lk} & m_{ll} \end{bmatrix}, \quad \overline{C} = \begin{bmatrix} cos\psi & sin\psi \\ -sin\psi & cos\psi \end{bmatrix}.$$

Since

$$\overline{M} = \overline{C}^T \overline{A} \overline{C},$$

by (3.15)

$$S(\overline{A}) = S(\overline{M}). \tag{3.16}$$

Also, by (3.15), we obtain

$$\mu(M) - \mu(A) = S(M) - \sum_{i=1}^{n} m_{ii} - [S(A) - \sum_{i=1}^{n} a_{ii}^2] = \sum_{i=1}^{n} a_{ii}^2 - \sum_{i=1}^{n} m_{ii}^2.$$

Hence, we have

$$m_{ij} = a_{ij}, \quad \text{for} \quad i \neq k, \ \ j \neq l, \quad i, j = 1, 2, \ldots, n.$$

Therefore

$$\mu(M) - \mu(A) = a_{kk}^2 + a_{ll}^2 - m_{kk}^2 - m_{ll}^2 = S(\overline{A}) - 2a_{kl}^2 - S(\overline{M}) + 2m_{kl}^2.$$

By (3.16), we get equality

$$\mu(M) - \mu(A) = 2(m_{kl}^2 - a_{kl}^2) \tag{3.17}$$

Now, we shall compute the difference $\mu^{(q+1)} - \mu^{(q)}$. Namely, by (3.12) and (3.17), we have

$$\mu^{(q+1)} - \mu^{(q)} = \mu(C^{(q+1)}) - \mu(C^{(q)}) = 2[(c_{k_q l_q}^{(q+1)})^2 - (c_{k_q l_q}^{(q)})^2] = -2(c_{k_q l_q}^{(q)})^2.$$

Therefore

$$\mu^{(q+1)} = \mu^{(q)} - 2(a_{k_q l_q}^{(q)})^2, \;\; q = 0, 1, \ldots, r-1.$$

Because the entry $a_{k_q l_q}^{(q)}$ has been chosen to have the greatest absolute value (see 3.13), therefore

$$(a_{k_q l_q}^{(q)})^2 \geq \frac{\mu^q}{n(n+1)},$$

and

$$\mu^{(q+1)} \leq \mu^{(q)} - \frac{2\mu^{(q)}}{n(n+1)}, \;\; q = 0, 1, \ldots, r-1,$$

Then, we have

$$0 < \mu^{(r+1)} \leq \mu^{(r)}[1 - \frac{2}{n(n+1)}]^r.$$

Hence, we obtain the limit

$$\lim_{r \to \infty} \mu^{(r)} = 0.$$

This means that the sequence of matrices (3.11) converges to the diagonal matrix

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

The diagonal entries $\lambda_1, \lambda_2, \ldots, \lambda_n$ are eigenvalues of the matrix A, so that

$$AV = V\Lambda,$$

where

$$V = \lim_{r \to \infty} C^{(0)} C^{(1)} \cdots C^{(r)}.$$

Since a product of orthonormal matrices is also an orthonormal matrix, therefore $C^{(\infty)}$ is the orthonormal matrix of eigenvectors of A. In order to stop Jacobi iterations, we can use the condition

$$\mu(C^{(0)} C^{(1)} \cdots C^{(r)}) \leq \epsilon,$$

where ϵ is a given accuracy.

Example 3.3 *Let us use Jacobi method to find all eigenvalues and eigenvectors of the matrix*

$$A = \begin{bmatrix} 20 & 6 & 8 \\ 6 & 20 & 0 \\ 8 & 0 & 20 \end{bmatrix}.$$

Solution . The greatest entry of A out of diagonal

$$\max_{i,j=1,2,3;\, i\neq j} \mid a_{ij} \mid = a_{13} = 8.$$

Hence, $k = 1,\ \ l = 3$ and

$$C^{(1)} = \begin{bmatrix} cos\psi & 0 & -sin\psi \\ 0 & 1 & 0 \\ sin\psi & 0 & cos\psi \end{bmatrix},$$

where by (3.10) $\psi = \frac{\pi}{4}$, since $a_{11} = a_{33} = 20$.
First Jacobi iteration for $k = 1,\ l = 3,\ a_{13}^{(0)} = 8$ and $\psi = 0.785$

$$A^{(1)} = C^{(1)T} A C^{(1)} = \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 20 & 6 & 8 \\ 6 & 20 & 0 \\ 8 & 0 & 20 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

$$C^{(1)} = \begin{bmatrix} 0.707 & 0.000 & -0.707 \\ 0.000 & 1.000 & 0.000 \\ 0.707 & 0.000 & 0.707 \end{bmatrix},$$

Hence, we have

$$A^{(1)} = C^{(1)T} A C^{(1)} = \begin{bmatrix} 28.000 & 4.243 & 0.000 \\ 4.243 & 20.000 & -4.243 \\ 0.000 & -4.243 & 12.000 \end{bmatrix},$$

and

$$\begin{matrix} & X^{(1)} & X^{(2)} & X^{(3)} \end{matrix}$$

$$V^{(1)} = C^{(1)} = \begin{bmatrix} 0.707 & 0.000 & -0.707 \\ 0.000 & 1.000 & 0.000 \\ 0.707 & 0.000 & 0.707 \end{bmatrix}.$$

Second Jacobi iteration for $k = 1,\ l = 2,\ a_{12}^{(1)} = 4.243$ and $\psi = 0.407$

$$C^{(2)} = \begin{bmatrix} 0.918 & -0.396 & 0.000 \\ 0.396 & 0.918 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{bmatrix},$$

Hence, we have

$$A^{(2)} = C^{(2)T} A^{(1)} C^{(2)} = \begin{bmatrix} 29.831 & 0.000 & -1.681 \\ 0.000 & 18.169 & -3.895 \\ -1.681 & -3.895 & 12.000 \end{bmatrix}.$$

and

$$V^{(2)} = C^{(1)} C^{(2)} = \begin{array}{c} \begin{matrix} X^{(1)} & X^{(2)} & X^{(3)} \end{matrix} \\ \begin{bmatrix} 0.649 & -0.280 & -0.707 \\ 0.396 & 0.918 & 0.000 \\ 0.649 & -0.280 & 0.707 \end{bmatrix} \end{array}.$$

Third Jacobi iteration for $k = 2,\ l = 3,\ a_{23}^{(2)} = -3.895$ and $\psi = -0.451$

$$C^{(3)} = \begin{bmatrix} 1.000 & 0.000 & 0.000 \\ 0.000 & 0.900 & 0.435 \\ 0.000 & -0.435 & 0.900 \end{bmatrix},$$

Hence, we have

$$A^{(3)} = C^{(3)T} A^{(2)} C^{(3)} = \begin{bmatrix} 29.831 & 0.732 & -1.513 \\ 0.732 & 20.053 & 0.000 \\ -1.513 & 0.000 & 10.116 \end{bmatrix},$$

and

$$V^{(3)} = C^{(1)} C^{(2)} C^{(3)} = \begin{array}{c} \begin{matrix} X^{(1)} & X^{(2)} & X^{(3)} \end{matrix} \\ \begin{bmatrix} 0.649 & 0.056 & -0.759 \\ 0.396 & 0.827 & 0.400 \\ 0.649 & -0.560 & 0.515 \end{bmatrix} \end{array}.$$

Fourth Jacobi iteration for $k = 1,\ l = 3,\ a_{13}^{(3)} = -1.513$, and $\psi = -0.076$

$$C^{(4)} = \begin{bmatrix} 0.997 & 0.000 & 0.076 \\ 0.000 & 1.000 & 0.000 \\ -0.076 & 0.000 & 0.997 \end{bmatrix},$$

Hence, we have

$$A^{(4)} = C^{(4)T} A^{(3)} C^{(4)} = \begin{bmatrix} 29.946 & 0.730 & 0.000 \\ 0.730 & 20.053 & 0.056 \\ 0.000 & 0.056 & 10.000 \end{bmatrix},$$

and

$$V^{(4)} = C^{(1)} C^{(2)} C^{(3)} C^{(4)} = \begin{array}{c} \begin{matrix} X^{(1)} & X^{(2)} & X^{(3)} \end{matrix} \\ \begin{bmatrix} 0.705 & 0.056 & -0.707 \\ 0.365 & 0.827 & 0.429 \\ 0.608 & -0.560 & 0.562 \end{bmatrix} \end{array}.$$

Fifth Jacobi iteration for $k = 1,\ l = 2,\ a_{12}^{(4)} = 0.73$ and $\psi = 0.073$

$$C^{(5)} = \begin{bmatrix} 0.997 & -0.073 & 0.000 \\ 0.073 & 0.997 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{bmatrix},$$

Hence, we have

$$A^{(5)} = C^{(5)T} A^{(4)} C^{(5)} = \begin{bmatrix} 30.000 & 0.000 & 0.004 \\ 0.000 & 20.000 & 0.056 \\ 0.004 & 0.056 & 10.000 \end{bmatrix},$$

and

$$V^{(5)} = C^{(1)}C^{(2)}C^{(3)}C^{(4)}C^{(5)} = \begin{array}{c} \begin{matrix} X^{(1)} & X^{(2)} & X^{(3)} \end{matrix} \\ \begin{bmatrix} 0.707 & 0.004 & -0.707 \\ 0.424 & 0.798 & 0.429 \\ 0.566 & -0.603 & 0.562 \end{bmatrix} \end{array}.$$

Sixth Jacobi iteration for $k = 2,\ l = 3,\ a_{23}^{(5)} = 0.056,$ and $\psi = 0.006$

$$C^{(6)} = \begin{bmatrix} 1.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & -0.006 \\ 0.000 & 0.006 & 1.000 \end{bmatrix},$$

Hence, we have

$$A^{(6)} = C^{(6)T} A^{(5)} C^{(6)} = \begin{bmatrix} 30.000 & 0.000 & 0.004 \\ 0.000 & 20.000 & 0.000 \\ 0.004 & 0.000 & 10.000 \end{bmatrix},$$

and

$$V^{(6)} = C^{(1)}C^{(2)}C^{(3)}C^{(4)}C^{(5)}C^{(6)} = \begin{array}{c} \begin{matrix} X^{(1)} & X^{(2)} & X^{(3)} \end{matrix} \\ \begin{bmatrix} 0.707 & 0.000 & -0.707 \\ 0.424 & 0.800 & 0.424 \\ 0.566 & -0.600 & 0.566 \end{bmatrix} \end{array}.$$

Seventh Jacobi iteration for $k = 1,\ \ l = 3,\ \ a_{13}^{(6)} = 0.004$, and $\psi = 0$

$$C^{(6)} = \begin{bmatrix} 1.000 & 0.000 & 0.000 \\ 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 1.000 \end{bmatrix},$$

Hence, we have

$$A^{(7)} = C^{(7)T} A^{(6)} C^{(7)} = \begin{bmatrix} 30.000 & 0.000 & 0.000 \\ 0.000 & 20.000 & 0.000 \\ 0.00 & 0.000 & 10.000 \end{bmatrix},$$

and

$$V^{(7)} = C^{(1)}C^{(2)}C^{(3)}C^{(4)}C^{(5)}C^{(6)}C^{(7)} = \begin{array}{c} \begin{matrix} X^{(1)} & X^{(2)} & X^{(3)} \end{matrix} \\ \begin{bmatrix} 0.707 & 0.000 & -0.707 \\ 0.424 & 0.800 & 0.424 \\ 0.566 & -0.600 & 0.566 \end{bmatrix} \end{array}.$$

Finally, the matrix A has the following eigenvalues

$$\begin{aligned} \lambda_1 &= 30.00 \\ \lambda_2 &= 20.00 \\ \lambda_3 &= 10.00 \end{aligned}$$

and eigenvectors

$X^{(1)}$	$X^{(2)}$	$X^{(3)}$
0.707	0.000	−0.707
0.424	0.800	0.424
0.566	−0.600	0.566

The following module in `Mathematica` finds eigenvalues and eigenvector of a symmetric matrix a by Jacobi method of iterations

Program 3.1 *Mathematica module that solves an eigenvalue problem by iterative Jacobi's method.*

```
 jacobi[a_]:=Module[{m,n,ckl,v},
 n=Length[a[[1]]]; v=IdentityMatrix[n];m=a;
 (* Module ckl finds orthogonal matrix *)
 ckl[m_]:=Module[{b,p,k,l,psi,c,s},
 b=Abs[m];
 Do[b[[i,i]]=0,{i,1,n}];
 p=Position[b,Max[b]];  p=First[p];
 k=p[[1]]; l=p[[2]];
 e=IdentityMatrix[n];
 psi=If[m[[k,k]]-m[[l,l]]==0,Pi/4,ArcTan[2*m[[k,l]]/
 (m[[k,k]]-m[[l,l]])]/2];
 c=N[Cos[psi]]; s=N[Sin[psi]];
 e[[k,k]]=c; e[[k,l]]=-s; e[[l,k]]=s; e[[l,l]]=c;
 e
];

Do[e=ckl[m];m=Transpose[e].m.e;v=v.e,{7}];
Print["Diagonal matrix of eigenvalues "];
```

```
 Print[Chop[m]//TableForm];
 Print["Matrix of eigenvectors"];
 Print[Chop[v]//TableForm];
];
```

Solving the example by the above program, we invoke the module `jacobi`

```
(* Data segment *);
a={{20.,6.,8.},{6.,20.,0.},{8.,0.,20}};
jacobi[a];
```

Then, we find the diagonalmatrix of eigenvalues and orthonormal matrix of eigenvectors.

```
Diagonal matrix of eigenvalues
30.     0    0
 0     20.   0
 0      0   10.
Matrix of eigenvectors
0.707107     0    -0.707107
0.424264     0.8   0.424264
0.565685    -0.6   0.565685
```

3.3 Power Method

Let $\lambda_1,\ \lambda_2, \cdots, \lambda_n$ be eigenvalues of a matrix A (real or complex). We shall consider λ_1 as dominant eigenvalue of A, so that

1. (a) $\lambda_1 = \lambda_2 = \cdots = \lambda_r$ for certain $1 \leq r \leq n$, i.e. λ_1 can be repeating eigenvalue of A,
   (b) $\mid \lambda_1 \mid = \mid \lambda_2 \mid = \cdots = \mid \lambda_r \mid > \mid \lambda_{r+1} \mid \geq \mid \lambda_{r+2} \mid \geq \cdots \geq \mid \lambda_n \mid$.

In order to find a dominant eigenvalue λ_1 of the matrix A and corresponding eigenvector $X^{(1)}$, we can apply the power method, provided that the eigenvectors $X^{(1)}, X^{(2)}, \ldots, X^{(n)}$ of A are linearly independent in the real space R^n, (or in the complex space C^n). Therefore, every vector $Y \in R^n$ can be written as the following linear combination:

$$Y = a_1 X^{(1)} + a_2 X^{(2)} + \cdots + a_n X^{(n)}.$$

Now, let us choose a starting vector Y to determine the iterations:

$$\begin{aligned}
AY &= a_1\lambda_1 X^{(1)} + a_2\lambda_2 X^{(2)} + \cdots + a_n\lambda_n X^{(n)},\\
A^2Y &= a_1\lambda_1^2 X^{(1)} + a_2\lambda_2^2 X^{(2)} + \cdots + a_n\lambda_n^2 X^{(n)},\\
A^3Y &= a_1\lambda_1^3 X^{(1)} + a_2\lambda_2^3 X^{(2)} + \cdots + a_n\lambda_n^2 X^{(n)},\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
A^kY &= a_1\lambda_1^k X^{(1)} + a_2\lambda_2^k X^{(k)} + \cdots + a_n\lambda_n^k X^{(n)}.
\end{aligned} \tag{3.18}$$

Hence, we have

$$A^kY = \lambda_1^k[a_1X^{(1)} + a_2(\frac{\lambda_2}{\lambda_1})^kX^{(2)} + a_3(\frac{\lambda_3}{\lambda_1})^kX^{(3)} + \cdots + a_n(\frac{\lambda_n}{\lambda_1})^kX^{(n)}]. \quad (3.19)$$

Since

$$|\frac{\lambda_i}{\lambda_1}| < 1, \quad i = r+1, r+2, \ldots, n,$$

we get

$$\frac{\lambda_i}{\lambda_1} \to 0 \text{ when } \quad k \to \infty, \quad i = r+1, r+2, \ldots, n.$$

Thus, if $a_1 \neq 0$ then

$$A^kY \approx \lambda_1 a_1 X^{(1)},$$

and the vector A^kY approximates the eigenvector $X^{(1)}$. It can happen that the component $a_1 = 0$. Then, we can change the starting vector Y to get non-zero component a_1. (In practice, it is reasonable to choose $Y = [1, 1, \cdots, 1]$). However, round-off errors yield a non-zero term $\lambda_1 \epsilon X^{(1)}$, so that, in the presence of round-off errors $A^kY \to X^{(1)}$ when $k \to \infty$. In the absence of round-off errors, when $a_1 = 0$, we can get the next eigenvalue λ_2 and corresponding eigenvector $X^{(2)}$. For a distinct dominant eigenvalue λ_1, from (3.18), we obtain the following formula:

$$\begin{aligned}\frac{[A^{k+1}Y]_i}{[A^kY]_i} &= \lambda_1 \frac{[a_1X^{(1)} + a_2(\frac{\lambda_2}{\lambda_1})^{k+1}X^{(2)} + \cdots + a_n(\frac{\lambda_n}{\lambda_1})^{k+1}X^{(n)}]_i}{[a_1X^{(1)} + a_2(\frac{\lambda_2}{\lambda_1})^kX^{(2)} + \cdots + a_n(\frac{\lambda_n}{\lambda_1})^kX^{(n)}]_i} \\ &= \lambda_1 + O((\frac{\lambda_2}{\lambda_1})^{k+1}),\end{aligned}$$

where $[X^{(k)}]_i$ denotes $i-th$ component of the vector $X^{(k)}$.
Hence, the approximate value of λ_1 is

$$\lambda_1^* = \frac{[A^{k+1}Y]_i}{[A^kY]_i}. \quad (3.20)$$

for $i = 1, 2, \ldots, n$.

We shall use the dominant component to evaluate

$$\lambda_1^* = \textit{diminant components } \frac{[A^{k+1}Y]_j}{[A^kY]_j}.$$

Similar formula can be obtained for a repeating dominant eigenvalue λ_1.

Example 3.4 *Let us find the dominant eigenvalue λ_1 and corresponding eigenvector $X^{(1)}$ for the matrix*

$$A = \begin{bmatrix} 2 & 0 & 3 \\ 1 & 4 & 2 \\ 7 & 0 & 6 \end{bmatrix}.$$

using power method.

Solution. One can find, in elementary way, that the eigenvalues and eigenvectors of A are:

$$\lambda_1 = 9, \quad \lambda_2 = 4, \quad \lambda_3 = -1$$

$$X^{(1)} = \begin{bmatrix} 0.359700 \\ 0.407661 \\ 0.839309 \end{bmatrix}, \quad X^{(2)} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad X^{(3)} = \begin{bmatrix} -0.7001401 \\ 0.140028 \\ -0.7001401 \end{bmatrix}.$$

Choosing the starting vector $Y = [1, 1, 1]$, we obtain
The first iteration:

$$Y_1 = AY = \begin{bmatrix} 2 & 0 & 3 \\ 1 & 4 & 2 \\ 7 & 0 & 6 \end{bmatrix}. \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 5 \\ 7 \\ 13 \end{bmatrix}.$$

Thus, the dominant component of Y_1 is: $\lambda_1 \approx \lambda_1^{(1)} = 13$
The second iteration:

$$Y_2 = AY_1 = \begin{bmatrix} 2 & 0 & 3 \\ 1 & 4 & 2 \\ 7 & 0 & 6 \end{bmatrix}. \begin{bmatrix} 5 \\ 7 \\ 13 \end{bmatrix} = \begin{bmatrix} 49 \\ 59 \\ 113 \end{bmatrix}.$$

The dominant component of Y_2 is : $\lambda_1 \approx \lambda_1^{(2)} = 9.8$
The third iteration:

$$Y_3 = AY_2 = \begin{bmatrix} 2 & 0 & 3 \\ 1 & 4 & 2 \\ 7 & 0 & 6 \end{bmatrix}. \begin{bmatrix} 49 \\ 59 \\ 113 \end{bmatrix} = \begin{bmatrix} 437 \\ 511 \\ 1021 \end{bmatrix}.$$

The dominant component is : $\lambda_1 \approx \lambda_1^{(3)} = 9.035399$
The fourth iteration :

$$Y_4 = AY_3 = \begin{bmatrix} 2 & 0 & 3 \\ 1 & 4 & 2 \\ 7 & 0 & 6 \end{bmatrix}. \begin{bmatrix} 437 \\ 511 \\ 1021 \end{bmatrix} = \begin{bmatrix} 3937 \\ 4523 \\ 9185 \end{bmatrix}.$$

The dominant component is: $\lambda_1 \approx \lambda_1^{(4)} = 9.00915$
The fifth iteration :

$$Y_5 = AY_4 = \begin{bmatrix} 2 & 0 & 3 \\ 1 & 4 & 2 \\ 7 & 0 & 6 \end{bmatrix} . \begin{bmatrix} 3937 \\ 4523 \\ 9185 \end{bmatrix} = \begin{bmatrix} 35429 \\ 40399 \\ 82669 \end{bmatrix}.$$

The dominant component is $\lambda_1 \approx \lambda_1^{(5)} = 9.00044$.
In computations, *the scaled power method* is used to avoid large numbers. Then, we consider the following vectors as an approximation of $X^{(1)}$:

$$Y_k = \frac{A^k Y}{\sqrt{[A^1 Y]_1^2 + [A^k Y]_2^2 + \cdots + [A^k Y]_n^2}}, \quad k = 1, 2, \ldots;$$

The following module in `Mathematica` finds the dominant eigenvalue and corresponding eigenvector of a given matrix A.

Program 3.2 *Mathematica module that finds a dominant eigenvalue by power method.*

```
power[a_,iter_]:=Module[{lambda,n,s,x,vector1,vector2},
n=Length[a];vector2=Table[1,{n}];
Do[{vector1=a.vector2; vector2=a.vector1;
   vector3=Table[vector2[[i]]/vector1[[]],{i,1,n}];
   lambda=Max[Abs[vector3]];
   x=Sqrt[Sum[vector1[[i]]^2,{i,1,n}]];
   vector1=vector1/x;
   x=Sqrt[Sum[vector2[[i]]^2,{i,1,n}]];
   vector2=vector2/x},
{iter}];
Print["Eigenvalue lambda = ",N[lambda,4]];
Print["Eigenvector vector2 = ",vector2];
{lambda,vector2}
]
```

In order to evaluate the dominant eigenvalue of the matrix

```
a={{2.,0.,3.},{1.,4.,2.},{7.,0.,6.}};
```

we enter the number of required iterations `iter=3` and execute the command `power[a,iter]` .
Then, we obtain the following output

```
Eigenvalue lambda = 9.
Eigenvector vector2 = {0.359539, 0.408585, 0.838922}

{8.99995,{0.35954, 0.40859, 0.83892}}
```

3.4 The Householder Transformation and Hessenberg Matrices

Let us state the definition of the Hessenberg matrix

Definition 3.1 *A matrix $B = \{B_{ij}\},\ i, j = 1, 2, ..., n,$ is upper Hessenberg if $B_{ij} = 0$ for $i > j + 1$, that is, B has the following diagram:*

$$B = \begin{bmatrix} * & * & * & \cdots & * & * \\ * & * & * & \cdots & * & * \\ 0 & * & * & \cdots & * & * \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & * & * \\ 0 & 0 & 0 & \cdots & * & * \end{bmatrix}.$$

To reduce a matrix A to the Hessenberg matrix B, we apply the Householder transformation given below.

Definition 3.2 *The matrix*

$$H = I - 2x\ x^*.$$

is called Householder transformation. [2] *where I is n-th order identity matrix and the unitary vector*

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad x^* = (x_1^*, x_2^*, ..., x_n^*) \in C^n, \quad ||x||_2 = \sqrt{x^* x} = 1.$$

Let us note that the Householder transformation is Hermitian orthogonal matrix. Indeed, we have

$$H^* = I^* - (2x\ x^*)^* = I - 2x\ x^* = H,$$

and

$$H^* H = H^2 = (I - 2x\ x^*)(I - 2x\ x^*) = I - 2x\ x^* - 2x\ x^* + 4x\ x^*\ x\ x^* = I,$$

[2] Householder transformation is also applicable to a matrix with complex entries, then x_i^* means conjugate to $x_i,\ i = 1, 2, ..., n$.

since $x\ x^*\ x\ x^* = x\ x^*$.
We shall consider the following Householder transformation:

$$R = I - 2\frac{u}{||u||_2}\frac{u^*}{||u||_2}, \qquad ||u||_2 = \sqrt{u^*u},$$

where

$$u = \begin{cases} x + ||x||_2 e_1 \ \ if \quad x + ||x||_2 e_1 \neq 0, \\ x - ||x||_2 e_1 \ \ if \quad x + ||x||_2 e_1 = 0, \end{cases}$$

for any non zero real vector x with components $x_i,\ i = 1, 2, ..., n$, and $e_1 = (1, 0, 0, ..., 0)^T$.
The transformation R maps a non zero vector x to the vector $\pm||x||_2^2 e_1$, that is

$$Rx = \pm||x||_2^2 e_1. \tag{3.21}$$

Indeed, we note that

$$\begin{aligned} Rx = x - \frac{2}{||u||_2^2}uu^*x = \ & x - \frac{2}{||u||_2^2}(x \pm ||x||_2 e_1)(x \pm ||x||_2 e_1)^* x \\ & = x - \frac{2}{||u||_2^2}(x \pm ||x||_2 e_1)(||x||_2^2 \pm ||x||_2 x_1). \end{aligned}$$

Since

$$||u||_2^2 = (x \pm ||x||_2 e_1)^*(x \pm ||x||_2 e_1) = 2(||x||_2^2 \pm ||x||_2 x_1),$$

we have

$$Rx = x - (x \pm ||x||_2 e_1) = \pm||x||_2 e_1 = \begin{cases} -||x||_2 e_1, & x + ||x||_2 e_1 \neq 0, \\ ||x||_2 e_1, & x + ||x||_2 e_1 = 0. \end{cases}$$

In order to transform a real matrix $A = (a_{ij}),\ i, j = 1, 2, ..., n$, to the upper Hessenberg form B, we apply the following algorithm:
Algorithm

1. Set

$$A_1 = A, \ \ x = (a_{21}, a_{31}, ..., a_{n1})^*,$$

$$u = x \pm ||x||_2 e_1 = \begin{cases} x + ||x||_2 e_1 & x + ||x||_2 e_1 \neq 0, \\ x - ||x||_2 e_1 & x + ||x||_2 e_1 = 0. \end{cases}$$

$$R_{n-1} = I_{n-1} - \frac{2}{||u||_2^2}uu^*,$$

$$V_1 = \begin{bmatrix} I_1 & 0 \\ 0 & R_{n-1,} \end{bmatrix},$$

where I_1 is the identity matrix of order 1, and R_{n-1} is the Householder transformation of order $n-1$.
To get zeros in the first column down, we compute the matrix

$$A_2 = V_1 A_1 V_1^* = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \cdots & a_{2n}^{(1)} \\ 0 & a_{32}^{(1)} & a_{33}^{(1)} & \cdots & a_{3n}^{(1)} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & a_{n2}^{(1)} & a_{n3}^{(1)} & \cdots & a_{nn}^{(1)} \end{bmatrix}$$

2. Set

$$x = (a_{32}, a_{42}, ..., a_{n2})^*, u = x \pm ||x||_2 e_1,$$

$$R_{n-2} = I_{n-2} - \frac{2}{||u||_2^2} uu^*,$$

$$V_2 = \begin{bmatrix} I_2 & 0 \\ 0 & R_{n-2}, \end{bmatrix},$$

where I_2 is the identity matrix of order 2, and R_{n-2} is the Householder transformation of order $n-2$.
To get zeros in the second column down, we compute the matrix

$$A_3 = V_2 A_1 V_2^* = \begin{bmatrix} a_{11}^{(2)} & a_{12}^{(2)} & a_{13}^{(2)} & \cdots & a_{1n}^{(2)} \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ 0 & 0 & a_{43}^{(2)} & \cdots & a_{4n}^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}$$

3. Set, for $k = 3, 4, ..., n-2$,

$$x = (a_{k+1k}, a_{k+2k}, ..., a_{nk})^*, \quad u = x \pm ||x||_2 e_1,$$

$$R_{n-k} = I_{n-k} - \frac{2}{||u||_2^2} uu^*,$$

$$V_k = \begin{bmatrix} I_k & 0 \\ 0 & R_{n-k}, \end{bmatrix},$$

where I_k is the identity matrix of order k, and R_{n-k} is the Householder transformation of order $n-k$.
To get zeros in the k-th column down, we compute the matrix

$$A_{k+1} = V_k A_k V_k^* = \begin{bmatrix} a_{11}^{(k)} & a_{12}^{(k)} & a_{13}^{(k)} & \cdots & a_{1n-1}^{(k)} & a_{1n}^{(k)} \\ a_{21}^{(k)} & a_{22}^{(k)} & a_{23}^{(k)} & \cdots & a_{2n-1}^{(k)} & a_{2n}^{(k)} \\ 0 & a_{32}^{(k)} & a_{33}^{(k)} & \cdots & a_{3n-1}^{(k)} & a_{3n}^{(k)} \\ 0 & 0 & a_{43}^{(k)} & \cdots & a_{4n-1}^{(k)} & a_{4n}^{(k)} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & a_{nn-1}^{(k)} & a_{nn}^{(k)} \end{bmatrix}$$

Clearly, we get the Hessenberg matrix $B = V_{n-1} A_{n-1} V_{n-1}^*$.

Example 3.5 *Let us reduce the matrix*

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

to the upper Hessenberg form.
Following the algorithm, we find

1.

$$x = \begin{bmatrix} -2 \\ 1 \\ 2 \end{bmatrix}, \quad ||x||_2 = 3, \quad u = x + ||x||_2 e_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \quad ||u||^2 = 6,$$

$$R_3 = I_1 - \frac{2}{||u||^2} u\, u^* = \begin{bmatrix} 0.6667 & -0.3333 & -0.6667 \\ -0.3333 & 0.6667 & -0.6667 \\ -0.6667 & -0.6667 & -0.3333 \end{bmatrix},$$

$$V_1 = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.6667 & -0.3333 & -0.6667 \\ 0.0000 & -0.3333 & 0.6667 & -0.6667 \\ 0.0000 & -0.6667 & -0.6667 & -0.3333 \end{bmatrix},$$

and

$$A_2 = V_1\, A_1 V_1 = \begin{bmatrix} 1.0000 & -2.6667 & -1.6667 & -6.3333 \\ -3.0000 & -1.4444 & 0.5556 & 4.7778 \\ 0.0000 & 0.5556 & -2.4444 & 3.7778 \\ 0.0000 & 4.7778 & 3.7778 & 12.8889 \end{bmatrix}.$$

2.

$$x = \begin{bmatrix} 0.5556 \\ 4.7778 \end{bmatrix}, \quad ||x||_2 = 1.6178,$$

$$u = x + ||x||_2 e_1 = \begin{bmatrix} 5.3655 \\ 4.7778 \end{bmatrix}, \quad ||u||^2 = 51.6160,$$

$$R_2 = I_2 - \frac{2}{||u||^2} u\, u^* = \begin{bmatrix} -0.1155 & -0.9933 \\ -0.9933 & 0.1155 \end{bmatrix},$$

$$V_2 = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & -0.1155 & -0.9933 \\ 0.0000 & 0.0000 & -0.9933 & 0.1155 \end{bmatrix},$$

and the upper Hessenberg form of the matrix A is

$$B = A_3 = V_2\, A_2 V_2 = \begin{bmatrix} 1.0000 & -2.6667 & 6.4834 & 0.9240 \\ -3.0000 & -1.4444 & -4.8100 & 0.000 \\ 0.0000 & -4.8100 & 13.5512 & 1.9178 \\ 0.0000 & 0.0000 & 1.9178 & -3.1067 \end{bmatrix}.$$

We can solve the example using the following module in `Mathematica`

Program 3.3 *Mathematica module that finds Householder transformation.*

```
householder[{a_,k_}]:=Module[{n,e,s,t,u,u2,uu,v,ik,ink,rnk,x,x2},
n=Length[a];
x=Take[Map[#[[k]]&,a],{k+1,n}];
e=Prepend[Table[0,{n-k-1}],1];
x2=N[Sqrt[x.x]];
u=If[(x+x2*e).(x+x2*e)==0,x-x2*e,x+x2*e];
u2=u.u; ik=IdentityMatrix[k];
ink=IdentityMatrix[n-k];
s=Length[u];
uu=Table[u[[i]]*u[[j]],{i,1,s},{j,1,s}];
rnk=ink-2*uu/u2;
v=IdentityMatrix[n];
Do[v[[i,j]]=rnk[[i-k,j-k]],{i,k+1,n},{j,k+1,n}];
t=k+1;
{v.a.v,t}
]
```

To find Hessenberg form of the input data matrix

```
a={{1.,3.,4.,5.},{-2.,2.,5.,6.},
   {1.,5.,3.,7.},{2.,6.,7.,4.}};
```

we execute the commands

```
n=4;
b=Nest[householder,{a,1},n-1];
Chop[b[[1]],10^-4]//TableForm
```

Then, we obtain the following Hessenberg matrix

```
 1.     -2.66667   6.48345    -0.924007
-3.     -1.44444  -4.80997     0
 0      -4.80997  13.5512     -1.91782
 0       0        -1.91782    -3.10672
```

Let us use the Householder transformation for a matrix deflation, that is, to reduce an eigenvalue problem of dimension n to an eigenvalue problem of dimension $n-1$.

Matrix Deflation. Let A be a matrix of order n for which an eigenvalue λ and an eigenvector x, with the norm $||x||_2 = 1$, are known. Using the deflation, one can reduce the matrix A to a matrix C of order $n-1$, whose eigenvalues are the same as the remaining eigenvalues of A. To find such a matrix C, we consider the Householder transformation R of the eigenvector x, so that, by (3.21), we have

$$Rx = -e_1, \qquad e_1 = (1, 0, ..., 0)^T,$$

and

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = -R^T e_1 = - \begin{bmatrix} R_{11} \\ R_{1,2} \\ \vdots \\ R_{1n} \end{bmatrix}.$$

Hence, the matrix

$$R^T = \begin{bmatrix} -x_1 & R_{21} & R_{31} & \cdots & R_{n1} \\ -x_2 & R_{22} & R_{32} & \cdots & R_{n2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -x_n & R_{2n} & R_{3n} & \cdots & R_{nn} \end{bmatrix} = [-x, V],$$

where the matrix V with n rows and $(n-1)$ columns is

$$V = \begin{bmatrix} R_{21} & R_{31} & \cdots & R_{n1} \\ R_{22} & R_{32} & \cdots & R_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ R_{2n} & R_{3n} & \cdots & R_{nn} \end{bmatrix}_{n(n-1)}.$$

Clearly, we have

$$AR^T = [-Ax, AV] = [-\lambda x, AV],$$

and

$$RAR^T = \begin{bmatrix} -x^T \\ V^T \end{bmatrix} [-\lambda x, AV] = \begin{bmatrix} \lambda & -x^T AV \\ -\lambda V^T x & V^T AV \end{bmatrix}$$

Because $V^T x = 0$, we have

$$RAR^T = \begin{bmatrix} \lambda & -x^T AV \\ 0 & V^T AV \end{bmatrix},$$

Thus, the matrix $C = RAR^T$ of dimension $n-1$ is similar to A, and therefore C has the same eigenvalues as the matrix A, except λ.

3.5 QR Method

In order to compute all the eigenvalues of a square matrix A, the **QR** method is widely recommended. This method consists of two the following parts:

- In the first part, the Householder transformation is used to reduce the matrix A to the Hessenberg matrix B,
- In the second part, the **QR** decomposition is used to factorize the Hessenberg matrix $B = QR$ with an orthogonal matrix Q and an upper triangular matrix R.

Now, let us consider the **QR** decomposition of a Hessenberg matrix

$$B = QR, \tag{3.22}$$

where Q is an orthogonal matrix and R is an upper triangular matrix.

We can obtain such a decomposition, multiplying matrix B by plane rotation matrices $C^{(2,1)}, C^{(3,2)}, ..., C^{(n,n-1)}$, where

$$C^{(k,l)} = \begin{array}{c} \begin{matrix} & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \end{matrix} \\ \begin{matrix} \textit{column} \\ k \\ \downarrow \end{matrix} \end{array} \begin{array}{c} \begin{matrix} & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \\ & & & & & & & & & \end{matrix} \\ \begin{matrix} \textit{column} \\ l \\ \downarrow \end{matrix} \end{array} \begin{array}{l} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array}$$

$$\begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & \cos\psi & 0 & \cdots & 0 & -\sin\psi & & \\ & & & 0 & 1 & \cdots & 0 & 0 & & \\ & & & \vdots & & \ddots & & \vdots & & \\ & & & 0 & 0 & \cdots & 1 & 0 & & \\ & & & \sin\psi & 0 & \cdots & 0 & \cos\psi & & \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{bmatrix} \begin{array}{l} \\ \\ \\ \leftarrow row_k \\ \\ \\ \leftarrow row_l \\ k < l \\ \\ \\ \end{array} \tag{3.23}$$

Solving the equation

$$[C^{(k+1,k)}B]_{k+1,k} = B_{kk}\sin\psi + B_{k+1k}\cos\psi = 0,$$

we compute the angle

$$\psi = \begin{cases} Arctan(-\dfrac{B_{k+1k}}{B_{kk}}), & B_{kk} \neq 0, \\ \dfrac{\pi}{2}, & B_{kk} = 0. \end{cases} \tag{3.24}$$

One can show that the multiplication of the Hessenberg matrix B by a plane rotation matrix $C^{(k,k-1)}$ preserves the Hessenberg form of the matrix B.
Thus, the matrix

$$R = [C^{(2,1)}, C^{(3,2)}, ..., C^{(n,n-1)}]B = Q^T B,$$

is upper triangular.
Hence, we have

$$B = QR,$$

where the matrix $Q = C^{(2,1)}C^{(3,2)}, ..., C^{(n,n-1)}$, is orthogonal, since the product of orthogonal matrices $C^{(k,k-1)}, \quad k = 2, 3, ..., n$, is also orthogonal.
Let us note that the Householder transformation as well as orthogonal plane rotation matrices preserve the eigenvalues, so that, the matrices B and R have the same eigenvalues as the original matrix A. Assuming that the matrix A

is reduced to the Hessenberg matrix B, and the decomposition $B = QR$ is known.
Thus, the QR method is given by the following iterative process:
The sequence of matrices

$$A^{(0)},\ A^{(1)}, ..., A^{(m)}, ...,$$

is built according to the following recursive rule:

1. Set $\quad A^{(0)} = B,$
2. For $m = 0, 1, ...,$ compute the orthogonal matrix
$$Q^{(m)} = C^{(2,1)}C^{(3,2)}, ..., C^{(n,n-1)},$$
where $C^{(k,k-1)}$ depends on the matrix $A^{(m)}$. This dependence is given, by formula (3.24), to compute the angle ψ.
3. Use the orthogonal matrix $Q^{(m)}$, to find decomposition
$$A^{(m)} = Q^{(m)}R^{(m)},$$
with the upper triangular matrix $R^{(m)}$
4. Compute the matrix
$$A^{(m+1)} = R^{(m)}Q^{(m)}.$$

The matrices $A^{(m)},\ m = 0, 1, ...,$ are similar to the matrix B. Indeed, we have

$$A^{(m+1)} = R^{(m)}Q^{(m)} = [Q^{(m)}]^{-1}Q^{(m)}RQ^{(m)} = [Q^{(m)}]^{-1}A^{(m)}Q^{(m)}.$$

Therefore, all the matrices $A^{(m)},\ m = 0, 1, ...,$ have the same eigenvalues as the matrix B.
Clearly, at m-th iteration, the **QR** decomposition of the matrix $A^{(m)}$ is needed to compute the next term $A^{(m+1)}$. As we know now, such a decomposition can be found with use of the orthogonal plane rotation matrices.
Thus, we arrive to the following algorithm:

QR Algorithm

Step 1 Transform the matrix

$$A = \begin{Bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \cdots & a_{1n-1} & a_{1n} & \\ a_{21} & a_{22} & a_{23} & a_{24} & \cdots & a_{2n-1} & a_{2n} & \\ a_{31} & a_{32} & a_{33} & a_{34} & \cdots & a_{3n-1} & a_{3n} & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & a_{n4} & \cdots & a_{nn-1} & a_{nn} & \end{Bmatrix}$$

by Householder transformation to the Hessenberg's matrix

$$B = \begin{Bmatrix} b_{11} & b_{12} & b_{13} & b_{14} & \cdots & b_{1n-1} & b_{1n} \\ b_{21} & b_{22} & b_{23} & b_{24} & \cdots & b_{2n-1} & b_{2n} \\ 0 & b_{32} & b_{33} & b_{34} & \cdots & b_{3n-1} & b_{3n} \\ 0 & 0 & b_{43} & b_{44} & \cdots & b_{4n-1} & b_{4n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & b_{nn-1} & b_{nn} \end{Bmatrix}$$

End of step 1.

1. Compute the sequence of matrices

$$A^{(0)},\ A^{(1)},\ A^{(2)}, ..., A^{(m)}, ...$$

as follows:

Step 2

(a) Set $A^{(0)} = B$

(b) Find the QR factorization of the matrix $A^{(0)}$, that is $A^{(0)} = Q^{(0)}R^{(0)}$ following the scheme

- Compute the angle

$$\psi = \begin{cases} ArcTan(-\dfrac{A_{21}^{(0)}}{A_{11}^{(0)}}), & A_{11}^{(0)} \neq 0 \\ \dfrac{\pi}{2}, & A_{11}^{(0)} = 0. \end{cases}$$

and compute the orthonormal matrix

$$C^{(21)} = \begin{Bmatrix} cos\psi & -sin\psi & 0 & 0 & 0 & \cdots & 0 & 0 \\ sin\psi & cos\psi & & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{Bmatrix}$$

- Compute the matrix $A^{(01)} = C^{(21)}A^{(0)} = C^{(21)}B$.
To eliminate the element $A_{21}^{(0)}$ in $A^{(0)}$.

- Compute the angle

$$\psi = \begin{cases} ArcTan(-\dfrac{A_{32}^{(01)}}{A_{22}^{(01)}}), & A_{22}^{(01)} \neq 0 \\ \dfrac{\pi}{2}, & A_{22}^{(01)} = 0. \end{cases}$$

and compute the orthonormal matrix

$$C^{(32)} = \begin{Bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & cos\psi & -sin\psi & 0 & 0 & \cdots & 0 & 0 \\ 0 & sin\psi & cos\psi & & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{Bmatrix}$$

- Compute the matrix $A^{(11)} = C^{(32)}A^{(01)} = C^{(32)}C^{21)}B$
  To eliminate the element $A_{32}^{(01)}$ in $A^{(01)}$,
- Compute the angle

$$\psi = \begin{cases} ArcTan(-\dfrac{A_{32}^{(01)}}{A_{33}^{(01)}}), & A_{33}^{(01)} \neq 0 \\ \dfrac{\pi}{2}, & A_{33}^{(01)} = 0. \end{cases}$$

and compute the orthonormal matrix

$$C^{(43)} = \begin{Bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos\psi & -\sin\psi & 0 & \cdots & 0 & 0 \\ 0 & 0 & \sin\psi & \cos\psi & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{Bmatrix}$$

- Compute the matrix $A^{(21)} = C^{(43)}A^{(01)} = C^{(43)}C^{(32)}C^{21)}B$

- Continue the process of elimination of the elements $A_{21}^{(0)}, A_{32}^{(0)}, A_{43}^{(0)}, ..., A_{n,n-1}^{(0)}$ in $A^{(0)}$ to obtain the upper triangular matrix

$$R^{(0)} = Q^{(0)}A^{(0)}$$

where $Q^{(0)} = [C^{n,n-1}C^{n-1,n-2}...C^{21}]^T$.
Then, set $A^{(1)} = R^{(0)}Q^{(0)}$. We note that the matrix

$$A^{(1)} = [Q^{(0)}]^T \; B \; Q^{(0)}$$

is similar to the matrix B and therefore both matrices $A^{(1)}$ and B have the same eigenvalues. End of step 2.

In order to compute next matrix $A^{(2)}$ in the sequence, we replace the matrix $A^{(0)}$ by the matrix $A^{(1)}$ in part 2. Then, we repeat part 2 for $A^{(1)}$ to obtain the QR factorization of the matrix $A^{(1)}$, that is

$$R^{(1)} = Q^{(1)}A^{(1)}$$

Then, we set

$$A^{(2)} = R^{(1)}Q^{(1)} = [Q^{(1)}]^T B Q^{(1)}.$$

Thus, the matrix $A^{(2)}$ is similar to the matrix B and both matrices have the same eigenvalues.

(c) We continue replacement of the matrix $A^{(0)}$ in part 2 by successive matrices $A^{(1)},\ A^{(2)},\ A^{(3)}, ..., A^{(m)}$ until certain m.

Under conditions stated in theorem 3.1 , the sequence $A^{(1)},\ A^{(2)}, A^{(3)}, ..., A^{(m)}, ...$ converges to an upper triangular matrix and its diagonal elements are eigenvalues of the matrix B

Theorem 3.1 *If the following assumptions are satisfied:*

- *The real matrix A is diagonalizable, that is, there exists a non singular matrix T such that*

$$A = T \; Diagonal(\lambda_1, \lambda_2, ..., \lambda_n) \; T^{-1},$$

- *the eigenvalues $\lambda_k,\ k = 1, 2, ..., n,$ have different absolute values*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|.$$

Then, the sequence of matrices $A^{(m)},\ m = 0, 1, ...,$ converges to an upper triangular matrix R, so that

$$\lim_{m\to\infty} A^{(m)} = R,$$

and the diagonal entries $R_{s,s},\ s = 1, 2, ..., n$ of the matrix R are the eigenvalues of the matrix A, that is

$$\lambda_s = R_{s,s}, \qquad s = 1, 2, ..., n.$$

In order to accelerate convergence of the sequence $A^{(m)},\ m = 0, 1, ...,$ one can use the following algorithm with shift

QR Algorithm with shift. The algorithm with shift is a modification of the **QR** algorithm.

Let $\alpha_m,\ m = 0, 1, ...,$ be a sequence of shift numbers. The shift numbers are chosen to accelerate convergence. In the **QR** algorithm with shift, the sequence of matrices

$$A^{(0)},\ A^{(1)},\ ...,\ A^{(m)}, ...,$$

is constructed according to the following recursive rule:

1. Set $A^{(0)} = B$,
2. For $m = 0, 1, ...,$ compute orthogonal matrix $Q^{(m)}$ and upper triangular matrix $R^{(m)}$, to find the decomposition
$$A^{(m)} - \alpha_m I = Q^{(m)} R^{(m)},$$
3. Compute the matrix
$$A^{(m+1)} = R^{(m)} Q^{(m)} + \alpha_m I.$$

If we have an estimate of the eigenvalues $\lambda_k,\ k = 1, 2, ..., n$ we are able to find good shift numbers.

One of a strategy to choose the shift numbers $\alpha_m,\ m = 1, 2, ...,$ is to put $\alpha_m = A^{(m)}_{nn}$, to get λ_n, since $A^{(m)}_{ss} \to \lambda_s,\ s = 1, 2, ..., n,$ when $m \to \infty$.

For a broad range of matrices, the sequence $A^{(0)},\ A^{(1)},\ ...,\ A^{(m)}, ...,$ which is produced by the **QR** algorithm with shift converges to an upper triangular matrix R, that is

$$\lim_{m\to\infty} A^{(m)} = R,$$

Then, the diagonal entries of the matrix R are eigenvalues of the original matrix A.

Example 3.6 *Let us apply the* **QR** *algorithm to find the eigenvalues of the Hessenberg matrix B that is found in the example 1.*
Thus, we have $n = 4$, and

$$B = \begin{bmatrix} 0.0000 & 0.5126 & 0.4954 & -0.7013 \\ -5.8523 & 4.8248 & 0.7287 & -1.3189 \\ 0.0000 & -0.8535 & 2.5479 & -0.5786 \\ 0.0000 & 0.0000 & -0.4931 & 2.6273 \end{bmatrix}.$$

Following the algorithm, we set

$$A^{(0)} = B,$$

and, we compute the plane rotation matrices:
Since $B[1,1] = 0$, *we find* $\psi = \frac{\pi}{2}$ *and*

$$C^{(21)} = \begin{bmatrix} \cos\psi & -\sin\psi & 0 & 0 \\ \sin\psi & \cos\psi & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

We compute

$$\psi = Arctan(-\frac{[C^{(2,1)}B]_{21}}{[C^{(2,1)}B]_{11}} = 1.0299$$

and

$$C^{(32)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\psi & -\sin\psi & 0 \\ 0 & \sin\psi & \cos\psi & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0.5149 & -0.8573 & 0 \\ 0 & 0.8573 & 0.5149 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

We compute

$$\psi = Arctan(-\frac{[C^{(32)}C^{(2,1)}B]_{32}}{[C^{(32)}C^{(2,1)}B]_{22}} = 0.2767$$

and

$$C^{(43)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos\psi & -\sin\psi \\ 0 & 0 & \sin\psi & \cos\psi \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9620 & -0.2732 \\ 0 & 0 & 0.2732 & 0.9620 \end{bmatrix},$$

The orthogonal matrix

$$Q^{(0)} = C^{(21)}C^{(32)}C^{(43)} = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0.5149 & 0 & -0.8573 & 0 \\ 0.8247 & 0 & 0.4953 & -0.2732 \\ 0.2342 & 0 & 0.1406 & 0.9620 \end{bmatrix}.$$

The upper triangular matrix

$$R^{(0)} = Q^T B = \begin{bmatrix} 5.8523 & -4.8248 & -0.7287 & 1.3189 \\ 0 & 0.9956 & -1.9292 & 0.1349 \\ 0 & 0 & 1.8052 & -1.5826 \\ 0 & 0 & 0 & 2.2818 \end{bmatrix},$$

The matrix

$$A^{(1)} = R^{(0)}Q^{(0)} = \begin{bmatrix} 4.8248 & 3.6378 & 4.1050 & 2.5367 \\ -0.9956 & 1.6538 & -0.9924 & -0.1415 \\ 0 & -1.5475 & 1.3264 & -1.2685 \\ 0 & 0 & -0.6233 & 2.1950 \end{bmatrix}.$$

In order to get the eigenvalues four decimal places accurate, we need to execute about 30 iterations so that, the final approximate eigenvalues are diagonal entries of the matrix

$$A^{(30)} = R^{(29)}Q^{(29)} = \begin{bmatrix} 4.0000 & 0.3058 & 0.5249 & 5.1008 \\ 0 & 3.0000 & 0.0765 & 3.8528 \\ 0 & 0 & 2.0000 & 2.0646 \\ 0 & 0 & 0 & 1.0000 \end{bmatrix}.$$

Hence, the eigenvalues $\lambda_1 = 1,\ \lambda_2 = 2,\ \lambda_3 = 3$, and $\lambda_4 = 4$. These eigenvalues, we can obtain using the following module in `Mathematica`

Program 3.4 *Mathematica module that solves an eigenvalue problem by* **QR** *method.*

```
rq[a_,iter]:=Module[{rq1,r1},
 a1=a;
 rq1[a1_]:=Module[{b,n,v,r},
    b=a1; n=Length[a1]; v=IdentityMatrix[n];
    ckl[{b_,v_,k_}]:=Module[{pa,c,s,e,t,n},
       n=Length[b]; e=IdentityMatrix[n];
       pa=If[b[[k,k]]==0,Pi/2,ArcTan[-b[[k+1,k]]/b[[k,k]]]];
       c=Cos[pa];  s=Sin[pa];
       e[[k,k]]=c; e[[k,k+1]]=-s; e[[k+1,k]]=s; e[[k+1,k+1]]=c;
       t=k+1;{e.b,e.v,t}
     ];

     r=FixedPoint[ckl,{a1,v,1},n-1];
     r[[1]].Transpose[r[[2]]]
    ];
     r1=FixedPoint[rq1,a1,iter]

 ]
```

To find the eigenvalues of the Hessenberg matrix

```
a={{0.0,0.5126,0.4954,-0.7013},
{-5.8523,4.8248,0.7287,-1.3189},
{0.0,-0.8535,2.5479,-0.5786},
{0.0,0.0,-0.4931,2.6273}};
```

we enter the matrix a, the number of iterations `iter=30` and execute the commands

```
iter=30;
MatrixForm[N[Chop[qr[a,iter],10^-3],4]]
```

Then, we obtain the following output

```
4.          0.3058      0.5249      5.101
0           3.          0.07649     3.853
0           0           2.          2.065
0           0           0             1.
```

3.6 Exercises

Question 3.1 *Find all eigenvalues and all orthonormal eigenvectors of the matrix*

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

Question 3.2 *Find an orthonormal matrix V that transforms the matrix*

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

to a diagonal form.

Question 3.3 *Find all eigenvalues and eigenvectors of the following $n \times n$ matrix:*

$$A = \begin{bmatrix} 4 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 4 & -1 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 4 \end{bmatrix}_{(n \times n)}$$

Question 3.4 *Solve the following eigenvalue problem:*

$$Ax = \lambda x$$

by Jacobi method with accuracy $\epsilon = 0.05$, where

$$A = \begin{bmatrix} 2 & 0 & 3 \\ 1 & 4 & 2 \\ 9 & 0 & 6 \end{bmatrix}.$$

Question 3.5 *Find the dominant eigenvalue and corresponding eigenvector of the matrix*

$$A = \begin{bmatrix} 5 & 0 & 0 & 3 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 10 & 0 \\ 10 & 0 & 0 & 12 \end{bmatrix}.$$

Question 3.6 *Use the Householder transformation to reduce the following matrix to the Hessenberg form:*

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

Question 3.7 *Use the QR method to find all eigenvalues and all orthonormal eigenvectors of the matrix*

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

Question 3.8 *(a) Let*

$$B = \begin{Bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ 0 & b_{32} & b_{33} \end{Bmatrix}$$

be a Hessenberg's matrix and let $A^{(0)}, A^{(1)}, A^{(2)}, ...$ be the sequence of matrices determined by QR factorization of the matrix B. Write the algorithm to compute the term $A^{(1)}$ and $A^{(2)}$

(b) Let the Hessenberg's matrix

$$B = \begin{Bmatrix} 4 & 0 & 1 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{Bmatrix}$$

Compute the terms $A^{(0)}$, $A^{(1)}$ and $A^{(2)}$ of the QR sequence $A^{(0)}, A^{(1)}, A^{(2)}, ...$

Chapter 4

Iterative Methods for Systems of Linear Equations

4.1 Stationary One Step Linear Methods

In this section, we shall consider a class of one step linear stationary iterative methods of the following form (cf. [8], [14],[19], [22]):

$$x^{(m+1)} = Gx^{(m)} + F, \quad m = 0, 1, 2, \ldots; \tag{4.1}$$

where $x^{(0)}$ is a starting vector, in general, arbitrarily chosen, G is an iterative matrix, and F is a given vector.

Definition 4.1 *An iterative method of class (4.1) is said to be consistent with the system of linear equations*

$$Ax = b, \tag{4.2}$$

if and only if the exact solution $x^{(d)}$ of the system of equations (4.2) is a stationary point of the iterative method i.e.,

$$x^{(d)} = Gx^{(d)} + F, \tag{4.3}$$

where the vectors

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

and the matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n-1} & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn-1} & a_{nn} \end{bmatrix}$$

In order to determine a consistent iterative method, we should give relationship between matricesA, G and vectors b, F. Later on, we shall establish such relationship for Jacobi Iterative Method (JI), Gauss Seidel Iterative Method (GS), Successive Overrelaxation Method (SOR) and Alternating Direction Implicit Method (ADI).
Now, let us state the necessary and sufficient condition for convergence of a stationary one step linear iterative method.

Definition 4.2 *An iterative method of the class (4.1) consistent with the system of linear equations (4.2) is convergent if and only if for every starting vector* $x^{(0)}$ *the sequence*

$$\{x^{(m)}\}, \qquad m = 0, 1, \ldots;$$

determined by the iterative method is convergent to the exact solution of the system of linear equations (4.2), i.e.,

$$\lim x^{(m)} = x^{(d)} \quad \text{and} \quad Ax^{(d)} = b.$$

Let

$$\epsilon^{(m)} = x^{(m)} - x^{(d)}.$$

be the error of $m - th$ iteration. Then, from (4.1) and (4.3), we have

$$\epsilon^{(m+1)} = G^m \epsilon^{(0)}. \tag{4.4}$$

Hence, we obtain the following sufficient and necessary condition of convergence:
An iterative method of the class (4.1) is convergent, i.e.,

$$\epsilon^{(m)} \to 0 \ \ when \ \ m \to \infty$$

if and only if

$$\rho(G) < 1, \tag{4.5}$$

where $\rho(G) = \max_{1 \le i \le n} \mid \lambda_i \mid$, is the spectral radius of the iterative matrix G, and $\lambda_i, \ \ i = 1, 2, \ldots, n,$ are eigenvalues of G.
Rates of convergence. In order to estimate the rate of convergence of an iterative method, we may use the *Average Rate of Convergence* $R_m(G)$ or the *Asymptotic Rate of Convergence* $R_\infty(G)$.(cf. [19],[22]). The rates $R_m(G)$ and $R_\infty(G)$ are defined as follows:
From formula (4.4)

$$\parallel \epsilon^{(m)} \parallel \le \parallel G^m \parallel \parallel \epsilon^{(0)} \parallel .$$

So that, the norm $\parallel G^m \parallel$ determines the rate of approaching $x^{(m)}$ to $x^{(d)}$ when $m \to \infty$. Usually, we finish an iterative process if the error $\epsilon^{(m)}$ is a small fraction of the initial error $\epsilon^{(0)}$, i.e.,

$$\parallel \epsilon^{(m)} \parallel \le \mu \parallel \epsilon^{(0)} \parallel .$$

The above inequality holds if

$$m \geq -\frac{log\mu}{-\frac{1}{m}log \parallel G^m \parallel}.$$

Hence, *the Average Rate of Convergence is:*

$$R_m(G) = -\frac{1}{m}log \parallel G^m \parallel,$$

and the *Asymptotic Rate of Convergence is:*

$$R_\infty(G) = \lim_{m\to\infty} R_m(G) = -log\rho(G).$$

Now, we note that to reduce the initial error μ times

$$m \approx -\frac{log\mu}{R_\infty(G)} = \frac{log\mu}{log\rho(G)}.$$

iterations are needed.
Below, we shall give some of well known stationary one-step iterative methods.

4.2 Jacobi Iterative Method

Let A be a non-singular matrix. and let the diagonal entries $a_{ii} \neq 0$, for $i = 1, 2, \ldots n$. Clearly, the matrix A can be written in the following form:

$$A = L + D + U,$$

where the lower-triangular matrix

$$L = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn-1} & 0 \end{bmatrix}$$

the upper-triangular matrix

$$U = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ 0 & 0 & 0 & \cdots & a_{3n-1} & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}$$

and the diagonal matrix

$$D = \begin{bmatrix} a_{11} & 0 & 0 & \cdots & 0 & 0 \\ 0 & a_{22} & 0 & \cdots & 0 & 0 \\ 0 & 0 & a_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{nn} \end{bmatrix}$$

Now, let us write the linear system of equations (4.2) in the following equivalent form:

$$Dx = -(L+U)x + b,$$

or

$$x = -D^{-1}(L+U)x + D^{-1}b.$$

Then, Jacobi iterative method is:

$$x^{(m+1)} = -D^{-1}(L+U)x^{(m)} + D^{-1}b, \quad m = 0, 1, \ldots; \tag{4.6}$$

or

$$x^{(m+1)} = -G_J x^{(m)} + F_J, \quad m = 0, 1, \ldots;$$

where $x^{(0)}$ is the starting vector, $G_J = D^{-1}(L+U)$ is Jacobi's iterative matrix, and F_j is the following vector $F_J = D^{-1}b$.
In terms of coordinates, Jacobi iterative method takes the following form:

$$x_i^{(m+1)} = \frac{1}{a_{ii}} \sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(m)} + \frac{b_i}{a_{ii}},$$

for $i = 1, 2, \ldots, n$; and $m = 0, 1, \ldots;$
In order to stop Jacobi iterations, we can use the condition

$$\mid x_i^{(m+1)} - x_i^{(m)} \mid \leq \epsilon, \tag{4.7}$$

where ϵ is a given accuracy and m is the least integer for which condition (4.7) holds.

Example 4.1 *Let us solve the following system of equations*

$$\begin{array}{rcrcrcrcr} 10x_1 & - & x_2 & - & x_3 & - & x_4 & = & 34 \\ -x_1 & + & 10x_2 & - & x_3 & - & x_4 & = & 23 \\ -x_1 & - & x_2 & + & 10x_3 & - & x_4 & = & 12 \\ -x_1 & - & x_2 & - & x_3 & + & 10x_4 & = & 1 \end{array}$$

by Jacobi iterative method using condition (4.7) to stop the iterations when $\epsilon = 0.0005$.

Solution. Let $x^{(0)} = (0, 0, 0, 0)$ be the starting vector. Following Jacobi iterations, we find
The first iteration:

$$x_1^{(1)} = \tfrac{1}{10}[34 + x_2^{(0)} + x_3^{(0)} + x_4^{(0)}] = 3.4$$

$$x_2^{(1)} = \tfrac{1}{10}[23 + x_1^{(0)} + x_3^{(0)} + x_4^{(0)}] = 2.3$$

$$x_3^{(1)} = \tfrac{1}{10}[12 + x_1^{(0)} + x_2^{(0)} + x_4^{(0)}] = 1.2$$

$$x_4^{(1)} = \tfrac{1}{10}[1 + x_1^{(0)} + x_2^{(0)} + x_3^{(0)}] = 0.1$$

The second iteration:

$$x_1^{(2)} = \tfrac{1}{10}[34 + x_2^{(1)} + x_3^{(1)} + x_4^{(1)}] = \tfrac{1}{10}[34 + 2.3 + 1.2 + 0.1] = 3.76$$

$$x_2^{(2)} = \tfrac{1}{10}[23 + x_1^{(1)} + x_3^{(1)} + x_4^{(1)}] = \tfrac{1}{10}[23 + 3.4 + 1.2 + 0.1] = 2.77$$

$$x_3^{(2)} = \tfrac{1}{10}[12 + x_1^{(1)} + x_2^{(1)} + x_4^{(1)}] = \tfrac{1}{10}[12 + 3.4 + 2.3 + 0.1] = 1.78$$

$$x_4^{(2)} = \tfrac{1}{10}[1 + x_1^{(1)} + x_2^{(1)} + x_3^{(1)}] = \tfrac{1}{10}[1 + 3.4 + 2.3 + 1.2] = 0.79$$

The third iteration:

$$x_1^{(3)} = \tfrac{1}{10}[34 + x_2^{(2)} + x_3^{(2)} + x_4^{(2)}] = \tfrac{1}{10}[34 + 2.77 + 1.78 + 0.79] = 3.934$$

$$x_2^{(3)} = \tfrac{1}{10}[23 + x_1^{(2)} + x_3^{(2)} + x_4^{(2)}] = \tfrac{1}{10}[23 + 3.76 + 1.78 + 0.79] = 2.933$$

$$x_3^{(3)} = \tfrac{1}{10}[12 + x_1^{(2)} + x_2^{(2)} + x_4^{(2)}] = \tfrac{1}{10}[12 + 3.76 + 2.77 + 0.79] = 1.932$$

$$x_4^{(3)} = \tfrac{1}{10}[1 + x_1^{(2)} + x_2^{(2)} + x_3^{(2)}] = \tfrac{1}{10}[1 + 3.76 + 2.77 + 1.78] = 0.931$$

The fourth iteration:

$$x_1^{(4)} = \tfrac{1}{10}[34 + x_2^{(3)} + x_3^{(3)} + x_4^{(3)}] = \tfrac{1}{10}[34 + 2.933 + 1.932 + 0.931] = 3.9796$$

$$x_2^{(4)} = \tfrac{1}{10}[23 + x_1^{(3)} + x_3^{(3)} + x_4^{(3)}] = \tfrac{1}{10}[23 + 3.934 + 1.932 + 0.931] = 2.9797$$

$$x_3^{(4)} = \tfrac{1}{10}[12 + x_1^{(3)} + x_2^{(3)} + x_4^{(3)}] = \tfrac{1}{10}[12 + 3.3.934 + 2.933 + 0.931] = 1.9798$$

$$x_4^{(4)} = \tfrac{1}{10}[1 + x_1^{(3)} + x_2^{(3)} + x_3^{(3)}] = \tfrac{1}{10}[1 + 3.934 + 2.933 + 1.932] = 0.9799$$

We can obtain more accurate approximate solution using the following module

Program 4.1 *Mathematica module that solves a linear system of equations by Jacobi iterative method*

```
jacobiItation[a_,x0_]:=Module[{b,d,d1,i,k,l,n,x,u},
n=Length[First[a]]-1;
b=Map[#[[n+1]]&,a];
l=Table[0,{i,1,n},{k,1,n}]; u=l;d=l;
Do[l[[i,k]]=a[[i,k]],{i,2,n},{k,1,i-1}];
Do[u[[i,k]]=a[[i,k]],{i,1,n-1},{k,i+1,n}];
Do[d[[i,i]]=a[[i,i]],{i,1,n}];
d1=Inverse[d];
x=x0;
Do[x=d1.(-(l+u).x+b),{iter}];
x
```

Entering input data `iter=8;` and

```
a={{10.,-1.,-1.,-1.,34.},{-1.,10,-1.,-1.,23.},
   {-1.,-1.,10,-1.,12.},{-1.,-1.,-1.,10.,1.}};
x0={0,0,0,0};
```

we invoke the module

```
jacobiIteration[a,x0];
```

Then, we obtain the approximate solution

```
{3.9999852376, 2.9999852377, 1.9999852378, 0.9999852379}
```

This solution satisfies the condition (4.7) for $m = 7$, so that

$$| x_i^{(8)} - x_i^{(7)} | < 0.0003 < \epsilon, \quad i = 1, 2, 3, 4.$$

4.3 Gauss Seidel Iterative Method

Gauss Seidel iterative method is a simple modification of Jacobi iterative method. Namely, we may use already evaluated $x_j^{(m+1)}$, for $j = 1, 2, \ldots, i-1$ to determine $x_j^{(m+1)}$, for $j = i, i+1, \ldots, n$.
Then, we have

$$Dx^{(m+1)} = -Lx^{(m+1)} - Ux^{(m)} + b, \quad m = 0, 1, \ldots;$$

and the Gauss Seidel iterative method takes the following form:

$$x^{(m+1)} = G_S x^{(m)} + F_S, \quad m = 0, 1, \ldots; \tag{4.8}$$

where the iterative matrix $G_S = -(D+L)^{-1}U$ and the vector $F_S = (D+L)^{-1}b$.
In the terms of coordinates Gauss Seidel iterations are:

$$x_i^{(m+1)} = -\frac{1}{a_{ii}}[\sum_{j=1}^{i-1} a_{ij}x_j^{(m+1)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(m)} - b_i],$$

for $i = 1, 2, \ldots, n; \quad m = 0, 1, \ldots;$
Let us now state the sufficient condition for convergence of Jacobi and Gauss Seidel methods.
The following theorem holds:

Theorem 4.1 *(cf. [17]). If the matrix A is positive definite then Jacobi and Gauss Seidel iterative methods are convergent.*

Proof. By the assumption, A is a symmetric matrix. Therefore

$$A = U + D + U^T$$

and hence the iterative matrix of Gauss Seidel method is:

$$G_S = -(D + U)^{-1}U^T.$$

Let $-\lambda$ be an eigenvalue of the matrix G_S corresponding to the eigenvector v. Then, the following equality holds:

$$(D + U)^{-1}U^T v = \lambda v,$$

or

$$U^T v = \lambda(D + U)v.$$

In general, the iterative matrix G_S can have complex eigenvalues, so that

$$v^* U^T v = \lambda v^*(D + U)v, \tag{4.9}$$

where v^* is conjugate to v.
Adding the term $v^*(D + U)v$ to both sides of (4.9), we obtain

$$v^* A v = (1 + \lambda)v^*(D + U)v. \tag{4.10}$$

Since A is a symmetric matrix, therefore

$$\overline{(v^* A v)} = v^* A v.$$

Hence, by (4.9)

$$\begin{array}{l}(1 + \overline{\lambda})v^*(D + U)^T v = (1 + \lambda)v^*(D + U)v = \\ (1 + \lambda)[v^* D v + v^* U v] = \\ (1 + \lambda)[v^* D v + \overline{\lambda} v^*(D + U)^T v].\end{array} \tag{4.11}$$

Grouping like terms in (4.11), we arrive at the equality

$$(1- \mid \lambda \mid^2)v^*(D+U)^T v = (1+\lambda)v^* Dv. \tag{4.12}$$

Multiplying (4.12) by $1+\lambda$, we obtain

$$(1- \mid \lambda \mid^2)v^* Av = \mid 1+\lambda \mid^2 v^* Dv. \tag{4.13}$$

By the assumption, A is a positive definite matrix, therefore D is also a positive definite matrix. Moreover, $\lambda = 1$ cannot be an eigenvalue of $G_S = -(D+U)^{-1}U^T$ since A is a non-singular matrix and $v \neq 0$ (4.10). Therefore, we have

$$1- \mid \lambda \mid > 0,$$

and hence all eigenvalues of G_S satisfy the inequality

$$\mid \lambda \mid < 1.$$

This means that the spectral radius $\rho(G_S) < 1$. By the necessary and sufficient condition of convergence, the Gauss Seidel method is convergent.
In order to prove that Jacobi method is also convergent, when A is a positive definite matrix, we may use the following relations between $\rho(G_J)$ and $\rho(G_S)$.

1. (a) $\rho(G_J) = \rho(G_S) = 0$,
   (b) $\rho(G_j) = \rho(G_S) = 1$,
   (c) $\rho(G_S) < \rho(G_J) < 1$,
   (d) $1 < \rho(G_J) < \rho(G_S)$.

From the above relations, it follows that Jacobi iterative method is convergent if and only if Gauss Seidel method is convergent. One can show that Gauss Seidel method is asymptotically twice faster than Jacobi method, i.e.,

$$R_\infty(G_S) = 2R_\infty(G_J).$$

and the number of iterations needed to reduce μ times the initial error $\epsilon^{(0)}$ by Gauss Seidel iterations is:

$$m \approx -\frac{log\mu}{2R_\infty(G_J)}.$$

We can improve the accuracy of the final result of Jacobi iterative and Gauss Seidel iterative methods using the following formula:

$$x_i^* = x_i^{(m+2)} - \frac{(x_i^{(m+2)} - x_{i+1}^{(m)})^2}{x_i^{(m+2)} - 2x_i^{m+1)} + x_i^{(m)}}. \tag{4.14}$$

Example 4.2 *Let us solve the following system of linear equations*

$$\begin{array}{rcrcrcrcl} 10x_1 & - & x_2 & - & x_3 & - & x_4 & = & 34 \\ -x_1 & + & 10x_2 & - & x_3 & - & x_4 & = & 23 \\ -x_1 & - & x_2 & + & 10x_3 & - & x_4 & = & 12 \\ -x_1 & - & x_2 & - & x_3 & + & 10x_4 & = & 1 \end{array} \tag{4.15}$$

by Gauss Seidel iterative method using condition (4.7) to stop the iterations when $\epsilon = 0.01$.

Solution. Let $x^{(0)} = (0, 0, 0, 0)$ be the starting vector. Following Gauss Seidel iterations, we find
The first iteration:

$$x_1^{(1)} = \tfrac{1}{10}[34 + x_2^{(0)} + x_3^{(0)} + x_4^{(0)}] = 3.4$$

$$x_2^{(1)} = \tfrac{1}{10}[23 + x_1^{(1)} + x_3^{(0)} + x_4^{(0)}] = \tfrac{1}{10}[23 + 3.4 + 0 + 0 + 0] = 2.64$$

$$x_3^{(1)} = \tfrac{1}{10}[12 + x_1^{(1)} + x_2^{(1)} + x_4^{(0)}] = \tfrac{1}{10}[12 + 3.4 + 2.64 + 0 + 0] = 1.804$$

$$x_4^{(1)} = \tfrac{1}{10}[1 + x_1^{(1)} + x_2^{(1)} + x_3^{(1)}] = \tfrac{1}{10}[1 + 3.4 + 2.64 + 1.804 + 0] = 0.8844$$

The second iteration:

$$x_1^{(2)} = \tfrac{1}{10}[34 + x_2^{(1)} + x_3^{(1)} + x_4^{(1)}] = \tfrac{1}{10}[34 + 2.64 + 1.804 + 0.8844] = 3.9328$$

$$x_2^{(2)} = \tfrac{1}{10}[23 + x_1^{(2)} + x_3^{(1)} + x_4^{(1)}] = \tfrac{1}{10}[23 + 3.9328 + 1.804 + 0.8844] = 2.9621$$

$$x_3^{(2)} = \tfrac{1}{10}[12 + x_1^{(2)} + x_2^{(2)} + x_4^{(1)}] = \tfrac{1}{10}[12 + 3.9328 + 2.9621 + 0.8844] = 1.9779$$

$$x_4^{(2)} = \tfrac{1}{10}[1 + x_1^{(2)} + x_2^{(2)} + x_3^{(2)}] = \tfrac{1}{10}[1 + 3.9328 + 2.9621 + 1.979] = 0.9873$$

The third iteration:

$$x_1^{(3)} = \tfrac{1}{10}[34 + x_2^{(2)} + x_3^{(2)} + x_4^{(2)}] = \tfrac{1}{10}[34 + 2.9621 + 1.9779 + 0.9873] = 3.9927$$

$$x_2^{(3)} = \tfrac{1}{10}[23 + x_1^{(3)} + x_3^{(2)} + x_4^{(2)}] = \tfrac{1}{10}[23 + 3.9927 + 1.9779 + 0.9873] = 2.9958$$

$$x_3^{(3)} = \tfrac{1}{10}[12 + x_1^{(3)} + x_2^{(3)} + x_4^{(2)}] = \tfrac{1}{10}[12 + 3.9927 + 2.9958 + 0.9873] = 1.9976$$

$$x_4^{(3)} = \tfrac{1}{10}[1 + x_1^{(3)} + x_2^{(3)} + x_3^{(3)}] = \tfrac{1}{10}[1 + 3.9927 + 2.9958 + 1.9976] = 0.9986$$

The fourth iteration:

$$x_1^{(4)} = \frac{1}{10}[34 + x_2^{(3)} + x_3^{(3)} + x_4^{(3)}] = \frac{1}{10}[34 + 2.9958 + 1.9976 + 0.9986] = 3.9992$$

$$x_2^{(4)} = \frac{1}{10}[23 + x_1^{(4)} + x_3^{(3)} + x_4^{(3)}] = \frac{1}{10}[23 + 3.9992 + 1.9976 + 0.9986] = 2.9995$$

$$x_3^{(4)} = \frac{1}{10}[12 + x_1^{(4)} + x_2^{(4)} + x_4^{(3)}] = \frac{1}{10}[12 + 3.9992 + 2.9995 + 0.9986] = 1.9997$$

$$x_4^{(4)} = \frac{1}{10}[1 + x_1^{(4)} + x_2^{(4)} + x_3^{(4)}] = \frac{1}{10}[1 + 3.9992 + 2.9995 + 1.9997] = 0.9998$$

Evidently, condition (4.7) is satisfied for $m = 3$, so that

$$\mid x_i^{(4)} - x_i^{(3)} \mid \leq 0.008 < \epsilon, \quad i = 1, 2, 3, 4.$$

We can solve a system of linear equations by Gauss Seidel method of iterations using the following `Mathematica` module:

Program 4.2 *Mathematica module that solves a linear system of equations by Gauss Seidel iterative method*

```
        gaussSeidel[a_,x0_]:=Module[{b,d,d1,i,k,l,n,x,u},
        n=Length[a[[1]]]-1;
        b=Map[#[[n+1]]&,a];
        l=Table[0,{i,1,n},{k,1,n}]; u=l;d=l;
        Do[l[[i,k]]=a[[i,k]],{i,2,n},{k,1,i-1}];
        Do[u[[i,k]]=a[[i,k]],{i,1,n-1},{k,i+1,n}];
        Do[d[[i,i]]=a[[i,i]],{i,1,n}];
        d1=Inverse[d+l];
        x=x0;
        Do[x=d1.(-u.x+b),{4}];
        x
];
```

In order to repeat the solution of the above example, using the module `seidel`, we enter input data matrix and the starting vector

```
a={{10.,-1.,-1.,-1.,34.},{-1.,10,-1.,-1.,23.},
   {-1.,-1.,10,-1.,12.},{-1.,-1.,-1.,10.,1.}};
x0={0,0,0,0};
```

Then, we execute the instruction

```
N[gaussSeidel[a,x0],4]
```

to obtain the approximate solution `3.999, 3., 2., 0.9998.`
Let us note that, we have got the same numerical solution of the system of equations (4.15) by Gauss Seidel iterative method for four iterations, and by Jacobi iterative method for 8 iterations. Still, we can improve the results using formula (4.14). Namely, we obtain the four digit accurate solution using only three iterations: the second, the third and the fourth, i.e.,

$$x_1^{(*)} = x_1^{(4)} - \frac{[x_1^{(4)} - x_1^{(3)}]^2}{x_1^{(4)} - 2x_1^{(3)} + x_1^{(2)}} = 3.9992 - \frac{[3.9992 - 3.9927]^2}{3.9992 - 2 * 3.9927 + 3.9328} = 4.0000$$
$$x_2^{(*)} = x_2^{(4)} - \frac{[x_2^{(4)} - x_2^{(3)}]^2}{x_2^{(4)} - 2x_2^{(3)} + x_2^{(2)}} = 2.9995 - \frac{(2.9995 - 2.9958)^2}{2.9995 - 2 * 2.9958 + 2.9621} = 3.0000$$
$$x_3^{(*)} = x_3^{(4)} - \frac{[x_3^{(4)} - x_3^{(3)}]^2}{x_3^{(4)} - 2x_3^{(3)} + x_3^{(2)}} = 1.9997 - \frac{(1.9997 - 1.9976)^2}{1.9997 - 2 * 1.9976 + 1.9779} = 2.0000$$
$$x_4^{(*)} = x_4^{(4)} - \frac{[x_4^{(4)} - x_4^{(3)}]^2}{x_4^{(4)} - 2x_4^{(3)} + x_4^{(2)}} = 0.9998 - \frac{(0.9998 - 0.9986)^2}{0.9998 - 2 * 0.9986 + 0.9873} = 1.0000$$

4.4 Successive Overrelaxation Method (SOR)

Let us rewrite the system of equations (4.2) in the following form:

$$x = (-D^{-1}L - D^{-1}U)x + D^{-1}b. \tag{4.16}$$

Then, the successive overrelaxation iterations take the following form:

$$\begin{aligned} &x^{(m+1)} = w[-D^{-1}Lx^{(m+1)} - D^{-1}Ux^{(m)} + D^{-1}b] + (1-w)x^{(m)}, \\ &m = 0, 1, \ldots; \end{aligned} \tag{4.17}$$

where w is a parameter, $x^{(0)}$ is a starting vector.
In terms of coordinates, SOR iterations are:

$$x_i^{(m+1)} = -w[\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(m+1)} + \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(m)} - \frac{b_i}{a_{ii}}] + (1-w)x_i^{(m)}, \tag{4.18}$$

$i = 1, 2, \ldots;\ \ m = 0, 1, \ldots;$
Clearly, SOR method is one step linear stationary method, since from (4.17), we get

$$x^{(m+1)} = G_w x^{(m)} + F_w, \quad m = 0, 1, \ldots; \tag{4.19}$$

where the iterative matrix

$$G_w = (E + wD^{-1}L)^{-1}[(1-w)E - wD^{-1}U]$$

and the vector

$$F_w = wD^{-1}b$$

.
Let us note that for $w = 1$, SOR iterations (4.19) are the same as Gauss Seidel iterations (4.8). The rate of convergence of SOR method depends on the value of the parameter w. Naturally, the fastest convergence of SOR iterations will be for optimal value of $w = w_{opt}$ for which the spectral radius $\rho(G_w)$ attains its minimum. The optimal value of the parameter w can be determined by the following formula (cf. [22]):

$$w_{opt} = \frac{2}{1 + \sqrt{1 - \rho^2(G_J)}}, \tag{4.20}$$

where $\rho(G_J)$ is the spectral the matrix G_J.
The conditions of convergence of SOR method are given in the following theorem:

Theorem 4.2 *Let A be a symmetric matrix with the positive diagonal entries $a_{ii} > 0, \quad i = 1, 2, \ldots$; Then, SOR method converges if and only if A is a positive definite matrix and $0 < w < 2$.*

Thus, SOR method as well as Jacobi and Gauss Seidel methods are convergent for positive definite matrices. Among them SOR method has the greatest rate of convergence, i.e.,

$$R_\infty(G_{w_{opt}}) = 2\sqrt{R_\infty(G_J)},$$

where $R_\infty(G_J)$ is the asymptotic rate of convergence of Jacobi method. However, in order to use SOR method with the optimal parameter w_{opt}, we have to know the radius of convergence $\rho(G_J)$ of Jacobi method. In some interesting cases (for example when approximating elliptic equations), $\rho(G_J)$ is known and then, we may apply SOR method successfully.

Example 4.3 *Let us solve the following system of linear equations*

$$\begin{array}{rcrcrcrcr} 10x_1 & - & x_2 & - & x_3 & - & x_4 & = & 34 \\ -x_1 & + & 10x_2 & - & x_3 & - & x_4 & = & 23 \\ -x_1 & - & x_2 & + & 10x_3 & - & x_4 & = & 12 \\ -x_1 & - & x_2 & - & x_3 & + & 10x_4 & = & 1 \end{array}$$

by SOR method using condition (4.7) to stop the iterations when $\epsilon = 0.05$.

Solution. One can find that the spectral radius $\rho(G_J) = 0.3$, where Jacobi iterative matrix

$$G_J = E - D^{-1}A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \frac{1}{10} \begin{bmatrix} 10 & -1 & -1 & -1 \\ -1 & 10 & -1 & -1 \\ -1 & -1 & 10 & -1 \\ -1 & -1 & -1 & 10 \end{bmatrix} =$$

$$\frac{1}{10}\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

Therefore, by formula (4.20) the optimal parameter

$$w_{opt} = \frac{2}{1+\sqrt{1-0.3^2}} = 1.023573302.$$

Let $x^{(0)} = (0,0,0,0)$ be the starting vector. Following SOR iterations, we find
The first SOR iteration:

$$x_1^{(1)} = \frac{w}{10}[34 + x_2^{(0)} + x_3^{(0)} + x_4^{(0)}] + (1-w)x_1^{(0)} =$$
$$0.1023573302(34+0+0+0)+0 = 3.4801$$
$$x_2^{(1)} = \frac{w}{10}[23 + x_1^{(1)} + x_3^{(0)} + x_4^{(0)}] + (1-w)x_2^{(0)} =$$
$$0.1023573302(23+3.4801+0+0)+0 = 2.7104$$
$$x_3^{(1)} = \frac{w}{10}[12 + x_1^{(1)} + x_2^{(1)} + x_4^{(0)}] + (1-w)x_3^{(0)} =$$
$$0.10235702(12+3.4801+2.7104+0)+0 = 1.8619$$
$$x_4^{(1)} = \frac{w}{10}[1 + x_1^{(1)} + x_2^{(1)} + x_3^{(1)}] + (1-w)x_4^{(0)} =$$
$$0.1023573302(1+3.4801+2.7104+1.8698) = 0.9259$$

The second SOR iteration:

$$x_1^{(2)} = \frac{w}{10}[34 + x_2^{(1)} + x_3^{(1)} + x_4^{(1)}] + (1-w)x_1^{(1)} =$$
$$0.1023573302(34+2.7104+1.8619+0.9266) - 0.023573302 * 3.4801 = 3.9610$$
$$x_2^{(2)} = \frac{w}{10}[23 + x_1^{(2)} + x_3^{(1)} + x_4^{(1)}] + (1-w)x_2^{(1)} =$$
$$0.1023573302(23+3.9610+1.8619+0.9266) - 0.023573302 * 2.7104 = 2.9812$$
$$x_3^{(2)} = \frac{w}{10}[12 + x_1^{(2)} + x_2^{(2)} + x_4^{(1)}] + (1-w)x_3^{(1)} =$$
$$0.10235702(12+3.9610+2.9812+0.9266) - 0.023573302 * 1.8619 = 1.9898$$
$$x_4^{(2)} = \frac{w}{10}[1 + x_1^{(2)} + x_2^{(2)} + x_3^{(2)}] + (1-w)x_4^{(1)} =$$
$$0.1023573302(1+3.9610+2.9812+1.9898) - 0.023573302 * 0.9266 = 0.9948$$

The third SOR iteration:

$$x_1^{(3)} = \frac{w}{10}[34 + x_2^{(2)} + x_3^{(2)} + x_4^{(2)}] + (1-w)x_1^{(2)} =$$

$$0.1023573302(34 + 2.9812 + 1.9898 + 0.9948) - 0.023573302 * 3.9610 = 3.9974$$

$$x_2^{(3)} = \frac{w}{10}[23 + x_1^{(3)} + x_3^{(2)} + x_4^{(2)}] + (1-w)x_2^{(2)} =$$

$$0.1023573302(23 + 3.9974 + 1.9898 + 0.9948) - 0.023573302 * 2.9812 = 2.9986$$

$$x_3^{(3)} = \frac{w}{10}[12 + x_1^{(3)} + x_2^{(3)} + x_4^{(2)}] + (1-w)x_3^{(2)} =$$

$$0.10235702(12 + 3.9974 + 2.9986 + 0.9948) - 0.023573302 * 1.9898 = 1.9993$$

$$x_4^{(3)} = \frac{w}{10}[1 + x_1^{(3)} + x_2^{(3)} + x_3^{(3)}] + (1-w)x_4^{(2)} =$$

$$0.1023573302(1 + 3.9974 + 2.9986 + 1.9993) - 0.023573302 * 0.9948 = 0.9996$$

The condition (4.7) is satisfied for $m = 2$, so that

$$\mid x_i^{(3)} - x_i^{(2)} \mid \leq 0.0364 < \epsilon$$

$i = 1, 2, 3, 4$.
Comparing the results of the three methods, we observe that SOR method produces the most accurate results at each iteration.
We can solve the above example using the following `Mathematica` module

Program 4.3 *Mathematica module that solves a system of linear equations by Gauus-Seidel iterative method*

```
sor[a_,x0_]:=
Module[{b,d,d1,i,k,l,n,w,fw,gw,id,u,x},
n=Length[a[[1]]]-1;
b=Map[#[[n+1]]&,a];
w=2/(1+Sqrt[1-0.3^2]);
l=Table[0,{i,1,n},{k,1,n}]; u=l;d=l;
Do[l[[i,k]]=a[[i,k]],{i,2,n},{k,1,i-1}];
Do[u[[i,k]]=a[[i,k]],{i,1,n-1},{k,i+1,n}];
Do[d[[i,i]]=a[[i,i]],{i,1,n}];
d1=Inverse[d]; id=IdentityMatrix[n];
d2=Inverse[id+w*d1.l];
gw=d2.((1-w)*id-w*d1.u);
fw=w*d2.d1.b;
x=x0;
```

```
    Do[x=gw.x+fw,{3}];
    x
  ];
```

Entering data matrix a and starting vector x_0

```
a={{10.,-1.,-1.,-1.,34.},{-1.,10,-1.,-1.,23.},
   {-1.,-1.,10,-1.,12.},{-1.,-1.,-1.,10.,1.}};
   x0={0,0,0,0};
```

we obtain the approximate solution 3.9974, 2.9986, 1.9993, 0.99964, by execution of the command `N[sor[a,x0],5]`.

4.5 Alternating Direction Implicit Method (ADI)

The ADI iterative method in its first version was published by D.W. Peaceman and H.H. Rachford in 1955. Here, we shall present a stationary variant of ADI method, (cf. [22]).
[1] Let us assume that A is a positive definite matrix. We split A in three components as follows:

$$A = L + D + U,$$

where D is diagonal matrix.
By ADI method, the sequence $\{x^{(m)}\},\ m = 0, 1, \ldots;$ of successive iterations is determined in two steps. Namely, for a given vector $x^{(m)}$ the next two terms $x^{(m+\frac{1}{2})}$ and $x^{(m+1)}$ are computed by the following recursive formulas:

$$\begin{aligned} &(L_1 + \beta E)x^{(m+\frac{1}{2})} = b - (U_1 - \beta E)x^{(m)}, \\ &(U_1 + \beta E)x^{(m+1)} = b - (L_1 - \beta E)x^{(m+\frac{1}{2})}, \quad m = 0, 1, ...; \end{aligned} \tag{4.21}$$

where $L_1 = L + \frac{1}{2}D, \quad U_1 = U + \frac{1}{2}D$, and β is a parameter.
Eliminating $x^{(m+\frac{1}{2})}$ from equations (4.21), we obtain

$$x^{(m+1)} = G(A, \beta)x^{(m)} + F(A, \beta), \qquad m = 0, 1, ...;$$

where the iterative matrix of ADI method

$$G(A, \beta) = (U_1 + \beta E)^{-1}(L_1 - \beta E)(L_1 + \beta E)^{-1}(U_1 - \beta E),$$

and the vector

$$F(A, \beta) = (U_1 + \beta E)^{-1}b - (U_1 + \beta E)^{-1}(L_1 - \beta E)^{-1}(L_1 + \beta E)^{-1}b.$$

[1] See a broad description of ADI method in [19], [22]

Thus, ADI method is also one step stationary method, and it is convergent if the spectral radius $\rho(G(A,\beta)) < 1$.
On the other hand, the spectral radius

$$\rho(G(A,\beta)) = \max_{1\le k\le n}[\frac{\lambda_k-\beta}{\lambda_k+\beta}]^2,$$

where λ_k is the eigenvalue of the matrix L_1, and

$$[\frac{\lambda_k-\beta}{\lambda_k+\beta}]^2$$

is the eigenvalue of the iterative matrix $G(A,\beta)$.
By the assumption, A is a positive definite matrix. Therefore, L_1 and U_1 are also a positive definite matrices and all their eigenvalues are positive, so that

$$0 < a \le \lambda_k \le b, \qquad k = 1, 2, ..., n;$$

Then, the spectral radius

$$\rho(G(A,\beta)) < 1$$

for every $\beta > 0$. Therefore ADI method is convergent for every $\beta > 0$.
In order to reach the greatest rate of convergence, we can choose an optimal value of the parameter β to obtain the smallest value of $\rho(G(A,\beta))$. Thus, we shall find such β_{opt} for which

$$\rho(A,\beta_{opt}) = \min_{\beta>0}\max_{a\le\lambda\le b}[\frac{\lambda-\beta}{\lambda+\beta}]^2.$$

One can find that

$$\max_{a\le\lambda\le b}|\frac{\lambda-\beta}{\lambda+\beta}| = \max\{\frac{\beta-a}{\beta+a}, \frac{b-\beta}{b+\beta}\} = \begin{cases} \dfrac{b-\beta}{b+\beta} & if \quad a \le \beta \le \sqrt{ab}, \\ \dfrac{\beta-a}{\beta+a} & if \quad \sqrt{ab} \le \beta \le b. \end{cases}$$

Hence, the optimal value of parameter $\beta_{opt} = \sqrt{ab}$.

Example 4.4 *Let us solve the following system of linear equations:*

$$\begin{array}{rcrcrcrcr} 10x_1 & - & x_2 & - & x_3 & - & x_4 & = & 34 \\ -x_1 & + & 10x_2 & - & x_3 & - & x_4 & = & 23 \\ -x_1 & - & x_2 & + & 10x_3 & - & x_4 & = & 12 \\ -x_1 & - & x_2 & - & x_3 & + & 10x_4 & = & 1 \end{array}$$

by ADI method using condition (4.7) to stop the iterations when $\epsilon = 0.005$.

Solution. Let us note that the matrix

$$A = \begin{bmatrix} 10 & -1 & -1 & -1 \\ -1 & 10 & -1 & -1 \\ -1 & -1 & 10 & -1 \\ -1 & -1 & -1 & 10 \end{bmatrix}$$

can be written as follows:

$$A = L + D + U,$$

where

$$L = \begin{bmatrix} 4 & -1 & 0 & 0 \\ -1 & 4 & -1 & 0 \\ 0 & -1 & 4 & -1 \\ 0 & 0 & -1 & 4 \end{bmatrix}, \quad D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix},$$

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

Evidently, the matrices L, D and U satisfy conditions (a) and (b) and the assumptions of theorem 1. Therefore, ADI method (4.21) is convergent. In the example, we shall use the optimal value of the parameters when

$$\beta_{opt} = \beta_{opt} = \sqrt{ab}$$

Namely, one can find that the matrices

$$L_1 == L + \frac{1}{2}D = \begin{bmatrix} 5 & -1 & 0 & 0 \\ -1 & 5 & -1 & 0 \\ 0 & -1 & 5 & -1 \\ 0 & 0 & -1 & 5 \end{bmatrix}$$

$$U_1 == U + \frac{1}{2}D = \begin{bmatrix} 5 & 0 & -1 & -1 \\ 0 & 5 & 0 & -1 \\ -1 & 0 & 5 & 0 \\ -1 & -1 & 0 & 5 \end{bmatrix}$$

have the same eigenvalues

$$\nu_1 = \lambda_1 = 3.381966, \quad \nu_2 = \lambda_2 = 4.381966,$$

$$\nu_3 = \lambda_3 = 5.618033, \quad \nu_4 = \lambda_4 = 6.618033.$$

Hence $a = 3.381966$, $b = 6.6180340$ and $\beta_{opt} = \sqrt{ab} = 4.7309579$.
Let $x^{(0)} = (0, 0, 0, 0)$ be the starting vector. Following ADI iterations (4.21), we arrive at the following results:

	ADI Iterations			
$x^{(m)}$	x_1	x_2	x_3	x_4
$x^{(0)}$	0.0000	0.0000	0.0000	0.0000
$x^{(0.5)}$	3.7934	2.9137	1.5596	0.2630
$x^{(1)}$	3.9858	2.9325	1.9261	0.9667
$x^{(1.5)}$	3.9891	2.9972	1.9995	0.9924
$x^{(2)}$	3.9999	2.9989	1.9990	1.0000
The error : $\max[x_i^{(2)} - x_i^{(0)}] = 0.0011$				

.

Let us note that ADI method produces the most accurate results at each iteration as compared with the other methods (Jacobi, GS, SOR). However, ADI method converges well when optimal parameters are used. Then, two linear systems of equations have to be solved at each iteration. This makes ADI method less effective. Although, in some cases when the matrices L_1 and U_1 have simple structure (for instance, L_1 and U_1 are tri-diagonal matrices), ADI method can produce a satisfactory solution using relatively small number of arithmetic operations. In order to reduce the initial error μ times

$$m \approx \frac{\log(\mu)}{\log(\rho(G_A))}$$

arithmetic operations are needed.

4.6 Conjugate Gradient Method (CG)

The conjugate gradient method is applicable to a linear system of n equations

$$Ax = b, \tag{4.22}$$

with a positive definite matrix A.
This method produces a solution of the system (4.22) in at most n iterations, provided that computations are done in exact arithmetic. An implementation of the method on a computer may affect infinite iterative process.
The method is based on a set $v^{(1)}, v^{(2)}, ..., v^{(n)}$ of A-orthogonal vectors in the sense of the following inner product

$$(Av^{(i)}, v^{(j)}) = \sum_{k=1}^{n}\sum_{s=1}^{n} a_{ks} v_k^{(i)} v_s^{(j)},$$

where the vector $v^{(i)} = (v_1^{(i)}, v_2^{(i)}, \ldots, v_n^{(i)})$ is in the real space R^n. So that

$$(Av^{(i)}, v^{(j)}) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

Let us note that if $v^{(1)}, v^{(2)}, ..., v^{(n)}$ are A-orthogonal vectors then the exact solution $x^{(d)} = (x_1^{(d)}, x_2^{(d)}, \ldots, x_n^{(d)})$ can be presented as follows:

$$x^{(d)} = x^{(1)} + \sum_{k=1}^{n} \alpha_k v^{(k)}, \tag{4.23}$$

where $x^{(1)}$ is a starting vector arbitrarily chosen, and the coefficients

$$r^{(1)} = b - Ax^{(1)}.$$

$$\alpha_k = \frac{(A(x^{(d)} - x^{(1)}), v^{(k)})}{(Av^{(k)}, v^{(k)})} = \frac{(v^{(k)}, r^{(1)})}{(Av^{(k)}, v^{(k)})}, \quad k = 1, 2, \ldots, n,$$

Indeed, we have

$$\begin{aligned} (A(x^{(1)} + \sum_{k=1}^{n} \alpha_k v^{(k)}), v^{(s)}) &= \\ Ax^{(1)} + \alpha_s &= (b_s, v^{(s)}), \end{aligned}$$

for $s = 1, 2, ..., n$.
Hence, we obtain

$$A(x^{(1)} + \sum_{k=1}^{n} \alpha_k v^{(k)}) = b.$$

Because the system of equations (4.22) has a unique solution, therefore (4.23) holds.

The main problem in the CG method is to find an A-orthogonal set of n vectors. In order to obtain an A-orthogonal set of n vectors in the real space R^n, we can use Gram-Schmidt like procedure. Namely, let us choose a linearly independent set of n vectors $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$ and let put

$$v^{(1)} = u^{(1)}, \quad v^{(i+1)} = u^{(i+1)} - \sum_{k=1}^{i} \beta_{i+1k} v^{(k)}, \tag{4.24}$$

where the coefficients

$$\beta_{i+1k} = \frac{(Au^{(i+1)}, v^{(k)})}{(Av^{(k)}, v^{(k)})}, \quad k = 1, 2, \ldots, i; \quad i = 1, 2, \ldots, n.$$

There are many ways to choose linearly independent vectors $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$. One way is to set

$$u^{(i)} = r^{(i)}, \quad i = 1, 2, \ldots, n,$$

where the residual vector

$$r^{(i)} = b - Ax^{(i)}, \quad i = 1, 2, .., n,$$

with

$$x^{(i+1)} = x^{(i)} + \alpha_i v^{(i)}, \quad i = 1, 2, ..., n-1,$$

and with starting vector $x^{(1)}$.
For the above choice of vectors $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$, the CG algorithm is:

Choose a vector $x^{(1)}$,

then evaluate :

$$\begin{aligned} & & v^{(1)} &= r^{(1)} = b - Ax^{(1)}, \\ & For\ i = 1, 2, ..., n, & & \\ & & \alpha_i &= \frac{(v^{(i)}, r^{(i)})}{(Av^{(i)}, v^{(i)})}, \\ & & x^{(i+1)} &= x^{(i)} + \alpha_i v^{(i)}, \\ & & r^{(i+1)} &= r^{(i)} - \alpha_i A v^{(i)}, \\ & & \beta_i &= \frac{(Ar^{(i+1)}, v^{(i)})}{(Av^{(i)}, v^{(i)})}, \\ & & v^{(i+1)} &= r^{(i+1)} + \beta_i v^{(i)}. \end{aligned} \tag{4.25}$$

Implementing the above algorithm in exact arithmetic, we obtain the solution $x^{(d)} = x^{(n+1)}$. As we have mentioned, the CG iterative process can be infinite if round-off errors are involved in the calculations. To stop CG iterations, in such a case, we can use the following conditions:

1. (a)
$$\| r^{(m)} \|^2 = (r^{(m)}, r^{(m)}) \approx 0,$$

   (b) determine the maximum number of iterations m.
   For well conditioned matrices the maximum number of iterations is $m \approx 2$n.

The CG method when it is applied to a matrix with n^2 entries requires $O(n^3)$ arithmetic operations. So that, the method is equivalent to Gauss elimination in terms of the number of operations. However, the CG method is very efficient when it is applied to sparse matrix. Below, we give the module `conjugateGradient` that solves a system of linear equations with a positive definite matrix.

Program 4.4 *Mathematica module that solves a system of linear equations by SOR method*

```
Options[conjugateGradient]=
   {x0value   -> zeroVector,
    r0value   -> bVector,
    maxIter   -> twon,
    tollerance-> 10^-8};

conjugateGradient[a_, b_, opts___]:=Module[
    {n, r, v, al, be, oneiter, norm, x0, r0,
    iters, eps},

    n=Length[b];
    twon =2 n;
    zeroVector = Table[0, {n}];
    bVector = b;

    x0= x0value/.{opts}/.Options[conjugateGradient];
    r0= r0value/.{opts}/.Options[conjugateGradient];
    iters= maxIter/.{opts}/.Options[conjugateGradient];
    eps= tollerance/.{opts}/.Options[conjugateGradient];

    r[0]=v[0]=r0;
    x[0]=x0;
    al[i_]:=al[i]=r[i].r[i]/(r[i].a.v[i]);
    r[i_]:=r[i]=r[i-1]-al[i-1]  a.v[i-1];
    be[i_]:=be[i]=-r[i+1].a.v[i]/(v[i].a.v[i]);
    v[i_]:=v[i]= r[i] + be[i-1] v[i-1];
    x[i_]:=x[i]=x[i-1]+al[i-1] v[i-1];

   oneiter[{k_, residuals_, solution_}]:=
          {k+1, r[k+1],x[k+1]};

          norm[w_]:=Apply[Plus,w^2];
          N[FixedPoint[oneiter,{0, b, x0}, iters,
          SameTest->((norm[#2[[2]]]<eps)&)]]
    ]
```

Example 4.5 *Let us solve the same system of linear equations as in example 1.*

$$\begin{array}{rcrcrcrcr} 10x_1 & - & x_2 & - & x_3 & - & x_4 & = & 34 \\ -x_1 & + & 10x_2 & - & x_3 & - & x_4 & = & 23 \\ -x_1 & - & x_2 & + & 10x_3 & - & x_4 & = & 12 \\ -x_1 & - & x_2 & - & x_3 & + & 10x_4 & = & 1 \end{array}$$

by CG method using the condition either 1 or 2 to stop the iterations.

Solving this example with `Mathematica` module, we enter data

```
n=4;
a={{10,-1,-1,-1},{-1,10,-1,-1},
{-1,-1,10,-1},{-1,-1,-1,10}};
b={34,23,12,1};
```

and invoke the module `conjugateGradient[a,b]` to obtain the output {{2., {0, 0, 0, 0}, {4., 3., 2., 1
where the number of iterations $k = 2$, the residual vector $r = (0, 0, 0, 0)$, and
the solution $x = (4., 3., 2., 1.)$
Also, we find the solution x following the algorithm step by step Let

$$x^{(1)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad A = \begin{bmatrix} 10 & -1 & -1 & -1 \\ -1 & 10 & -1 & -1 \\ -1 & -1 & 10 & -1 \\ -1 & -1 & -1 & 10 \end{bmatrix}.$$

Then, we find

The first CG iteration

$$v^{(1)} = r^{(1)} = \begin{bmatrix} 34 \\ 23 \\ 12 \\ 1 \end{bmatrix},$$

$$\alpha_1 = \frac{(v^{(1)}, r^{(1)})}{(Av^{(1)}, v^{(1)})} = 0.12016,$$

$$x^{(2)} = x^{(1)} + \alpha_1 v^{(1)} = \begin{bmatrix} 4.0854 \\ 2.7636 \\ 1.4419 \\ 0.1202 \end{bmatrix},$$

$$r^{(2)} = r^{(1)} - \alpha_1 A v^{(1)} = \begin{bmatrix} -2.5279 \\ 1.0112 \\ 4.5502 \\ 8.0893 \end{bmatrix},$$

$$\beta_1 = \frac{(Ar^{(2)}, v^{(1)})}{(Av^{(1)}, v^{(1)})} = 0.05112,$$

$$v^{(2)} = r^{(2)} + \beta_1 v^{(1)} = \begin{bmatrix} -0.7897 \\ 2.1870 \\ 5.1637 \\ 8.1404 \end{bmatrix},$$

The second CG iteration:

$$\alpha_2 = \frac{(v^{(2)}, r^{(2)})}{(Av^{(2)}, v^{(2)})} = 0.10808,$$

$$x^{(3)} = x^{(2)} + \alpha_2 v^{(2)} = \begin{bmatrix} 4.0000 \\ 3.0000 \\ 2.0000 \\ 1.0000 \end{bmatrix},$$

$$r^{(3)} = r^{(2)} - \alpha_2 A v^{(2)} = \begin{bmatrix} 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \end{bmatrix},$$

$$\beta_2 = \frac{(Ar^{(3)}, v^{(2)})}{(Av^{(2)}, v^{(2)})} = 0.0000,$$

$$v^{(3)} = r^{(3)} + \beta_2 v^{(2)} = \begin{bmatrix} 0.0000 \\ 0.0000 \\ 0.0000 \\ 0.0000 \end{bmatrix}.$$

Since $r^{(3)} = 0$, by the second iteration, we get the exact solution:

$$x^{(d)} = x^{(3)} = \begin{bmatrix} 4 \\ 3 \\ 2 \\ 1 \end{bmatrix}.$$

4.7 Exercises

Question 4.1 *Solve the following system of linear equations:*

$$\begin{array}{rcrcrcl} 10x_1 & - & x_2 & - & x_3 & = & 35 \\ -x_1 & + & 10x_2 & - & x_3 & = & 24 \\ -x_1 & - & x_2 & + & 10x_3 & = & 13 \end{array} \tag{4.26}$$

by

1. *(a) Jacobi method,*
   (b) Gauss Seidel method,
   (c) SOR method,
   (d) ADI method,

using $\epsilon = 0.05$ *to stop the iterations.*

Question 4.2 *Investigate the convergence of Jacobi and Gauss Seidel iterative methods for the following system of linear equations:*

$$\begin{array}{rcrcrcr} 2x_1 & - & x_2 & - & x_3 & = & 3 \\ -x_1 & + & 2x_2 & - & x_3 & = & 0 \\ -x_1 & - & x_2 & + & 2x_3 & = & -3 \end{array}$$

Question 4.3 *Solve the system of equations*

$$\begin{array}{rcrcrcr} 3x_1 & - & x_2 & - & x_3 & = & -1 \\ -x_1 & + & 3x_2 & - & x_3 & = & 2 \\ -x_1 & - & x_2 & + & 3x_3 & = & 6 \end{array}$$

by CG method.

Question 4.4 *(a) State the necessary and sufficient conditions for convergence of the stationary linear one step iterative methods. Show that the iterative method*

$$x^{(m+1)} = G\, x^{(m)} + F, \qquad m = 0, 1, 2, ...$$

satisfies the necessary and sufficient condition when

$$G = \frac{1}{10} \begin{Bmatrix} 3 & 1 \\ 1 & 3 \end{Bmatrix}$$

(b) State a sufficient condition for convergence of the Jacobi and Gauss Seidel iterative methods. Show the the Jacobi and Gauss Seidel iterative methods are convergent when they are applied to the system of equations

$$\begin{array}{rcrcrcr} 5x_1 & - & 2x_2 & - & x_3 & = & 10 \\ -2x_1 & + & 10x_2 & - & x_3 & = & 13 \\ -x_1 & - & x_2 & + & 5x_3 & = & 0 \end{array} \tag{4.27}$$

(c) Solve the system of equations (4.27) by Jacobi and Gauss Seidel iterative methods using starting vector $x^{(0)} = \{2.5, 1.5, 1\}$ *with the accuracy* $\epsilon = 0.01$

Question 4.5 *(a) State the algorithm of the Conjugate Gradient Method for solving a linear system of equations* $Ax = b$ *where* $A = \{a_{ij}\}$, $i, j = 1, 2, ..., n.$ *is a symmetric non-singular matrix*

(b) Solve the system of equations

$$\begin{aligned} 6x_1 &- 3x_2 - x_3 = 7 \\ -3x_1 &+ 8x_2 - x_3 = 0 \\ -x_1 &- x_2 + 6x_3 = 9 \end{aligned} \tag{4.28}$$

by the Conjugate Gradient Method.

4.8 References

[**1**] Conte, S. D., & de Boor, C., (1983), Elementary Numerical Analysis, Algorithmic Approach, McGraw-Hill.

[**5**] Duff, I.S., Erisman, A.M., & Reid, J.K., (1986), Direct Methods for Sparse Matrices, Clarendon Press-Oxford.

[**6**] Faddeev, D.K., & Faddeeva, W.N., (1963), Numerical Methods in Linear Algebra, State Publisher, Ser.,Phys. & Maths.

[**7**] Golub, G., & Van Loan, C.F., (1983), Matrix Computations, John Hopkins University Press.

[**8**] Hageman, L.A., & Young, D.M., (1981), Applied Iterative Methods, Academic Press.

[**13**] Moris, J.L., (1983), Computational Methods in Numerical Analysis, John Wiley & Sons.

[**17**] Ralston, A., (1965), A First Course in Numerical Analysis, McGraw-Hill.

[**18**] Stoer, J. & Bulrisch R. (1980), Introduction to Numerical Analysis, Springer-Verlag

[**19**] Varga, R.S. (1962), Matrix Iterative Analysis, Printice-Hall, INC.

[**21**] Wolfram, S., (1992), Mathematica a System for Doing Mathematics by Computers, Addison-Wesley Publishing Company.

[**22**] Young, D.M. (1971), Iterative Solution of Large Linear Systems, Academic Press.