Numerical Method(Math-2073/53) Lecture note: Chapter-III

"Mathematics is the most beautiful and most powerful creation of the human spirit."

Stefan Banac

Dejen Ketema
Department of Mathematics
dejen.ketema@amu.edu.et

April, 2019

Contents

Chapter 1

Solving System of Equations

1.1 Introduction

In this chapter we consider numerical methods for solving a system of linear equations $Ax = b$. We assume that the given matrix A is real, $n \times n$, and nonsingular and that b is a given real vector in $\mathbb{R}^n$, and we seek a solution x that is necessarily also a vector in $\mathbb{R}^n$. Such problems arise frequently in virtually any branch of science, engineering, economics, or finance.

There is really no single technique that is best for all cases. Nonetheless, the many available numerical methods can generally be divided into two classes: direct methods and iterative methods. The present chapter is devoted to this two methods. In the absence of roundoff error, direct method would yield the exact solution within a finite number of steps.
An example of a problem in electrical engineering that requires a solution of a system of equations is shown in Fig.1.1. Using Kirchhoff's law, the currents $i_1, i_2, i_3, \& i_4$ can be determined by solving the following system of four equations:

$$\begin{aligned} 9i_1 - 4i_2 - 2i_3 &= 24 \\ -4i_1 + 17i_2 - 6i_3 - 3i_4 &= -16 \\ -2i_1 - 6i_2 + 14i_3 - 6i_4 &= 0 \\ -3i_2 - 6i_3 + 1li_4 &= 18 \end{aligned} \tag{1.1}$$

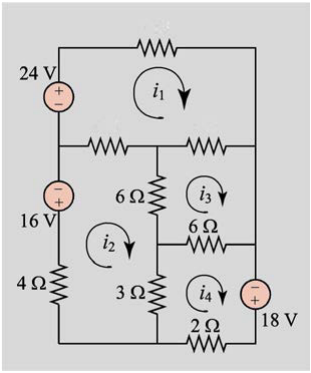

Figure 1.1: Electrical circuit.

Definition 1.1

linear equation in the variables $x_1,\ x_2,\ \cdots,\ x_n$ *is an equation of the form*

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b,$$

where the coefficients $a_1,\ a_2,\ \cdots,\ a_n$ *and* b *are constant real or complex numbers. The constant* a_i *is called the* ***coefficient*** *of* x_i*; and* b *is called the* ***constant term*** *of the equation.*

A **system of linear equations** (or linear system) is a finite collection of linear equations in same variables. For instance, a linear system of m equations in n variables $x_1, x_2, \cdots, x_n$ can be written as

$$\begin{cases} a_{11}x_1 & +a_{12}x_2 & +\cdots & +a_{1n}x_n & = b_1 \\ a_{21}x_1 & +a_{22}x_2 & +\cdots & +a_{2n}x_n & = b_2 \\ & & & \vdots & \\ a_{m1}x_1 & +a_{m2}x_2 & +\cdots & +a_{mn}x_n & = b_m \end{cases} \tag{1.2}$$

where $x_1, x_2, \ldots, x_n$ are the unknowns, $a_{11}, a_{12}, \ldots, a_{mn}$ are the coefficients of the system, and $b_1, b_2, \ldots, b_m$ the constant terms.

1.1.1 Vector equation

One extremely helpful view is that each unknown is a weight for a column vector in a linear combination.

$$x_1\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + x_2\begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + x_n\begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{1.3}$$

1.1.2 Matrix equation

The vector equation is equivalent to a matrix equation of the form

$$A\mathbf{x} = \mathbf{b}$$

where A is an $m \times n$ matrix, x is a column vector with n entries, and b is a column vector with m entries.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix} \tag{1.4}$$

A **solution** of a linear system (1.2) is a tuple $(s_1, s_2, \cdots, s_n)$ of numbers that makes each equation a true statement when the values $(s_1, s_2, \cdots, s_n)$ are substituted for$x_1, x_2, \cdots, x_n$, respectively. The set of all solutions of a linear system is called the **solution set** of the system.

Theorem 1.1

Any system of linear equations has one of the following exclusive conclusions.

(a) No solution.

(b) Unique solution.

(c) Infinitely many solutions.

A linear system is said to be **consistent** if it has at least one solution; and is said to be **inconsistent** if it has no solution.

1.1.3 Geometric interpretation

For a system involving two variables (x and y), each linear equation determines a line on the xy-plane. Because a solution to a linear system must satisfy all of the equations, the solution set is the intersection of these lines, and is hence either a line, a single point, or the empty set.

For three variables, each linear equation determines a plane in three-dimensional space, and the solution set is the intersection of these planes. Thus the solution set may be a plane, a line, a single point, or the empty set.

For n variables, each linear equation determines a hyperplane in n-dimensional space. The solution set is the intersection of these hyperplanes, which may be a flat of any dimension.

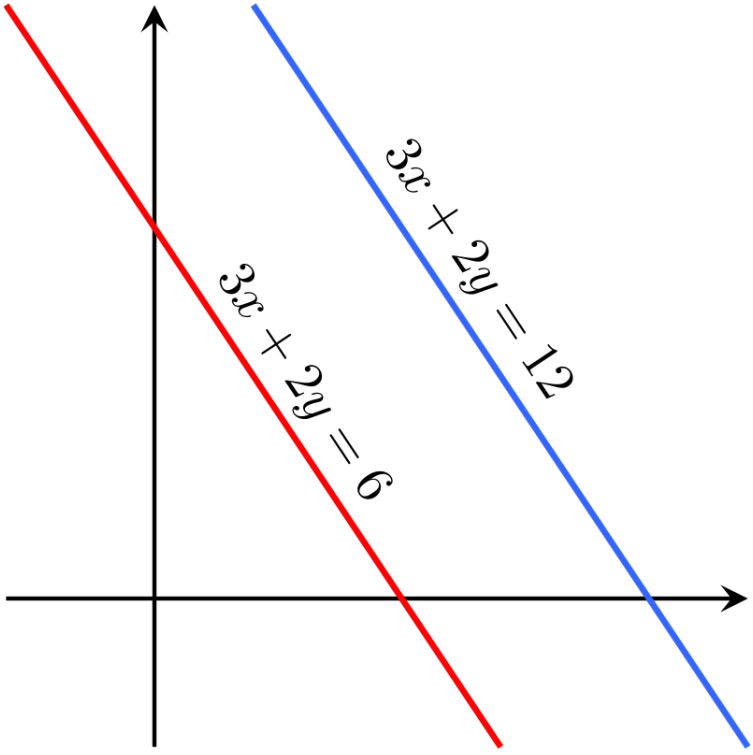

Figure 1.2: The equations $3x + 2y = 6$ and $3x + 2y = 12$ are (**inconsistent**).

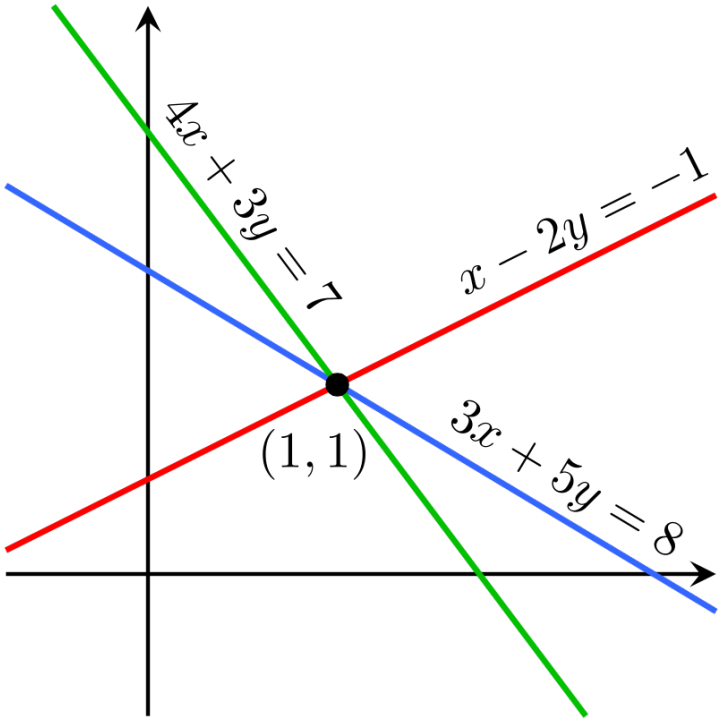

Figure 1.3: The equations $x - 2y = -1, 3x + 5y = 8$, and $4x + 3y = 7$ are linearly dependent(**consistent**).

1.2 Direct method

1.2.1 Cramer's rule

Consider a system of n linear equations for n unknowns, represented in matrix multiplication form as follows:

$$Ax = b$$

where the $n \times n$ matrix A has a nonzero determinant, and the vector $x = (x_1, \ldots, x_n)^{\mathrm{T}}$ is the column vector of the variables. Then the theorem states that in this case the system has a unique solution, whose individual values for the unknowns are given by:

$$x_i = \frac{\det(A_i)}{\det(A)} \qquad i = 1, \ldots, n$$

where A_i is the matrix formed by replacing the $i - th$ column of A by the column vector b.

A more general version of Cramer's rule considers the matrix equation

$$AX = B$$

where the $n \times n$ matrix A has a nonzero determinant, and X, B are $n \times m$ matrices. Given sequences $1 \leq i_1 < i_2 < \ldots < i_k \leq n$ and $1 \leq j_1 < j_2 < \ldots < j_k \leq m$, let $X_{I,J}$ be the $k \times k$ submatrix of X with rows in $I := (i_1, \ldots, i_k)$ and columns in $J := (j_1, \ldots, j_k)$. Let $A_B(I, J)$ be the $n \times n$ matrix formed by replacing the i_s column of A by the j_s column of B, for all $s = 1, \ldots, k$. Then

$$\det X_{I,J} = \frac{\det(A_B(I, J))}{\det(A)}.$$

In the case $k = 1$, this reduces to the normal Cramer's rule.

1.2.2 Explicit formulas for small systems

Consider the linear system

$$\begin{cases} a_1x + b_1y &= c_1 \\ a_2x + b_2y &= c_2 \end{cases}$$

which in matrix format is

$$\begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}.$$

Assume $a_1b_2 - b_1a_2$ nonzero. Then, with help of determinants, x and y can be found with Cramer's rule as

$$x = \frac{\begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{c_1b_2 - b_1c_2}{a_1b_2 - b_1a_2}, \quad y = \frac{\begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}} = \frac{a_1c_2 - c_1a_2}{a_1b_2 - b_1a_2}.$$

The rules for 3×3 matrices are similar. Given

$$\begin{cases} a_1x + b_1y + c_1z &= d_1 \\ a_2x + b_2y + c_2z &= d_2 \\ a_3x + b_3y + c_3z &= d_3 \end{cases}$$

which in matrix format is

$$\begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}.$$

Then the values of x, y and z can be found as follows:

$$x = \frac{\begin{vmatrix} d_1 & b_1 & c_1 \\ d_2 & b_2 & c_2 \\ d_3 & b_3 & c_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}}, \quad y = \frac{\begin{vmatrix} a_1 & d_1 & c_1 \\ a_2 & d_2 & c_2 \\ a_3 & d_3 & c_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}}, \text{ and } z = \frac{\begin{vmatrix} a_1 & b_1 & d_1 \\ a_2 & b_2 & d_2 \\ a_3 & b_3 & d_3 \end{vmatrix}}{\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}}.$$

Example 1.1

Solve the following system by Cramer's rule.
$$\begin{aligned} 2x_1 + 3x_2 + 4x_3 &= 19 \\ x_1 + 2x_2 + x_3 &= 4 \\ 3x_1 - x_2 + x_3 &= 9 \end{aligned}$$

Solution: *The coefficient matrix is* $A = \begin{bmatrix} 2 & 3 & 4 \\ 1 & 2 & 1 \\ 3 & -1 & 1 \end{bmatrix}$ *and column matrix* $b = \begin{bmatrix} 19 \\ 4 \\ 9 \end{bmatrix}$, *then* $det(A) = \begin{vmatrix} 2 & 3 & 4 \\ 1 & 2 & 1 \\ 3 & -1 & 1 \end{vmatrix} = 4 + 9 - 4 - 24 - 3 + 2 = -16 \neq 0$ *then the system has unique solution.*

$$A_1 = \begin{bmatrix} 19 & 3 & 4 \\ 4 & 2 & 1 \\ 9 & -1 & 1 \end{bmatrix} \;\&\; det(A_1) = \begin{vmatrix} 19 & 3 & 4 \\ 4 & 2 & 1 \\ 9 & -1 & 1 \end{vmatrix} = 38 + 27 - 16 - 72 - 12 + 19 = -16$$

$$A_2 = \begin{bmatrix} 2 & 19 & 4 \\ 1 & 4 & 1 \\ 3 & 9 & 1 \end{bmatrix} \;\&\; det(A_2 = \begin{vmatrix} 2 & 19 & 4 \\ 1 & 4 & 1 \\ 3 & 9 & 1 \end{vmatrix} = 8 + 57 + 36 - 48 - 19 - 18 = 16$$

$$A_3 = \begin{bmatrix} 2 & 3 & 19 \\ 1 & 2 & 4 \\ 3 & -1 & 9 \end{bmatrix} \;\&\; det(A_3) = \begin{vmatrix} 2 & 3 & 19 \\ 1 & 2 & 4 \\ 3 & -1 & 9 \end{vmatrix} = 36 + 36 - 19 - 114 - 27 + 8 = -80$$

$$\therefore\; x_1 = \frac{det(A_1)}{det(A)} = \frac{-16}{-16} = 1$$
$$x_2 = \frac{det(A_2)}{det(A)} = \frac{16}{-16} = -1$$
$$x_3 = \frac{det(A_3)}{det(A)} = \frac{-80}{-16} = 5.$$

This is the solution of the system.

Exercise 1.1

Use Cramer's Rule to solve each for each of the variables.

(a) $\begin{aligned} x - y &= 4 \\ -x + 2y &= -7 \end{aligned}$ (b) $\begin{aligned} -2x + y &= -2 \\ x - 2y &= -2 \end{aligned}$ (c) $\begin{aligned} 2x + y + z &= 1 \\ 3x + z &= 4 \\ x - y - z &= 2 \end{aligned}$

1.2.3 Inverse Matrix Method

Let $AX = b$ is a system of **n** linear equations with **n** unknowns and A is invertible, then the system has unique solution given by inversion method $X = A^{-1}b$.

$$A^{-1} = \frac{adj(A)}{\det(A)}$$

Note:- When A is not square or is singular, the system may not have a solution or may have more than one solution.

Example 1.2

Use the inverse of the coefficient matrix to solve the following system

$$\begin{aligned} 3x_1 + x_2 &= 6 \\ -x_1 + 2x_2 + 2x_3 &= -7 \\ 5x_1 - x_3 &= 10 \end{aligned}$$

***Solution**: Okay, let's first write down the matrix form of this system.*

$$\begin{bmatrix} 3 & 9 & 0 \\ -1 & 2 & 2 \\ 5 & 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -7 \\ 10 \end{bmatrix}$$

Now, we found the inverse of the coefficient matrix by using methods of finding Inverses and is the following;

Example

$$A = \begin{bmatrix} 3 & 1 & 0 \\ -1 & 2 & 2 \\ 5 & 0 & -1 \end{bmatrix} \Longrightarrow \quad C_A = \begin{bmatrix} -2 & 9 & -10 \\ 1 & 3 & 5 \\ 2 & -6 & 7 \end{bmatrix} \Longrightarrow \quad adj(A) = \begin{bmatrix} 2 & -1 & 2 \\ 9 & -3 & -6 \\ -10 & 5 & 7 \end{bmatrix}$$

and $det(A) = 3(-2) + 1(9) + 0(-10) = -6 + 9 = 3$, *then*

$$A^{-1} = 1/3 \begin{bmatrix} 2 & -1 & 2 \\ 9 & -3 & -6 \\ -10 & 5 & 7 \end{bmatrix} = \begin{bmatrix} 2/3 & -1/3 & 2/3 \\ 3 & -1 & -2 \\ -10/3 & 5/3 & 7/3 \end{bmatrix}$$

$$\therefore \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2/3 & -1/3 & 2/3 \\ 3 & -1 & -2 \\ -10/3 & 5/3 & 7/3 \end{bmatrix} \begin{bmatrix} 6 \\ -7 \\ 10 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 5 \\ -25/3 \end{bmatrix}$$

Now each of the entries of X *are* $x_1 = 1/3,\ x_2 = 5$ *and* $x_3 = -25/3$

1.2.4 Gaussian Elimination Method

In this section we show the following:

- How to solve linear equations when A is in upper triangular form. The algorithm is called backward substitution.
- How to transform a general system of linear equations into an upper triangular form, to which backward substitution can be applied. The algorithm is called Gaussian elimination.

A triangular matrix is a special kind of square matrix. A square matrix is called lower triangular if all the entries above the main diagonal are zero. Similarly, a square matrix is called upper triangular if all the entries below the main diagonal are zero. A triangular matrix is one that is either lower triangular or upper triangular. A matrix that is both upper and lower

triangular is called a diagonal matrix.

A matrix of the form

$$L = \begin{bmatrix} \ell_{1,1} & & & & 0 \\ \ell_{2,1} & \ell_{2,2} & & & \\ \ell_{3,1} & \ell_{3,2} & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ \ell_{n,1} & \ell_{n,2} & \dots & \ell_{n,n-1} & \ell_{n,n} \end{bmatrix}$$

is called a lower triangular matrix or left triangular matrix, and analogously a matrix of the form

$$U = \begin{bmatrix} u_{1,1} & u_{1,2} & u_{1,3} & \dots & u_{1,n} \\ & u_{2,2} & u_{2,3} & \dots & u_{2,n} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & u_{n-1,n} \\ 0 & & & & u_{n,n} \end{bmatrix}$$

is called an upper triangular matrix or right triangular matrix.

Forward and back substitution

A matrix equation in the form $\mathbf{Lx} = \mathbf{b}$ or $\mathbf{Ux} = \mathbf{b}$ is very easy to solve by an iterative process called **forward substitution** for lower triangular matrices and analogously **back substitution** for upper triangular matrices. The process is so called because for lower triangular matrices, one first computes x_1, then substitutes that forward into the next equation to solve for x_2, and repeats through to x_n. In an upper triangular matrix, one works backwards, first computing x_n, then substituting that back into the previous equation to solve for x_{n-1}, and repeating through x_1. Notice that this does not require inverting the matrix.

Forward substitution The matrix equation $Lx = b$ can be written as a system of linear equations

$$\begin{array}{ccccccc} \ell_{1,1}x_1 & & & & & = & b_1 \\ \ell_{2,1}x_1 & + & \ell_{2,2}x_2 & & & = & b_2 \\ \vdots & & \vdots & \ddots & & & \vdots \\ \ell_{m,1}x_1 & + & \ell_{m,2}x_2 & + \cdots + & \ell_{m,m}x_m & = & b_m \end{array}$$

Observe that the first equation ($\ell_{1,1}x_1 = b_1$ only involves x_1, and thus one can solve for x_1 directly. The second equation only involves x_1 and x_2, and thus can be solved once one substitutes in the already solved value for x_1. Continuing in this way, the k-th equation only involves $x_1, \dots, x_k$, and one can solve for x_k using the previously solved values for $x_1, \dots, x_{k-1}$.
The resulting formulas are:

$$\begin{aligned} x_1 &= \frac{b_1}{\ell_{1,1}}, \\ x_2 &= \frac{b_2 - \ell_{2,1}x_1}{\ell_{2,2}}, \\ &\vdots \\ x_m &= \frac{b_m - \sum_{i=1}^{m-1} \ell_{m,i}x_i}{\ell_{m,m}}. \end{aligned}$$

A matrix equation with an upper triangular matrix U can be solved in an analogous way, only working backwards.

Backward Substitution.

Given an upper triangular matrix A and a right-hand-side **b**,

$$\text{for } k = n : -1 : 1$$

$$x_k = b_k - \frac{\sum_{j=k+1}^{n} a_{kj} x_j}{a_{kk}}$$

$$\text{end}$$

Gauss elimination method is used to solve system of linear equations. In this method the linear system of equation is reduced to an upper triangular system by using successive elementary row operations. Finally we solve the value variables by using back ward substitution method. This method will be fail if any of the pivot element $a_{ii}, \quad i = 1, \ 2, \ \cdots, \ n$ becomes zero. In such case we re-write equation in such manner so that pivots are non zero. This procedure is called pivoting.
Consider system $AX = b$

Algorithm

Step 1: Form the augmented matrix $[A|b]$

Step 2: Transform $[A|b]$ to row echelon form $[U|d]$ using row operations.

Step 3: Solve the system $UX = d$ by back substitution.

The following row operations on the augmented matrix of a system produce the augmented matrix of an equivalent system, i.e., a system with the same solution as the original one.

- Interchange any two rows.
- Multiply each element of a row by a nonzero constant.
- Replace a row by the sum of itself and a constant multiple of another row of the matrix.

For these row operations, we will use the following notations.

- $R_i \leftrightarrow R_j$ means: Interchange row i and row j.
- αR_i means: Replace row i with α times row i.
- $R_i + \alpha R_j$ means: Replace row i with the sum of row i and α times row j.

Example 1.3

Solve the following system using Gauss elimination method.

$$\begin{aligned} 2x_1 - 3x_2 + x_3 &= 5 \\ 4x_1 + 14x_2 + 12x_3 &= 10 \\ 6x_1 + x_2 + 5x_3 &= 9 \end{aligned}$$

Solution: *The augmented matrix of the system is*

$$\begin{bmatrix} 2 & -3 & 1 & 5 \\ 4 & 14 & 12 & 10 \\ 6 & 1 & 5 & 9 \end{bmatrix}$$

Applying, elementary row operations on this matrix to change into its echelon form.

$$\begin{bmatrix} 2 & -3 & 1 & 5 \\ 4 & 14 & 12 & 10 \\ 6 & 1 & 5 & 9 \end{bmatrix} \begin{matrix} R_2 \longrightarrow R_2 - 2R_1 \\ R_3 \longrightarrow R_3 - 3R_1 \end{matrix} \begin{bmatrix} 2 & -3 & 1 & 5 \\ 0 & 20 & 10 & 0 \\ 0 & 10 & 2 & -6 \end{bmatrix}$$

$$R_3 \longrightarrow R_3 - 1/2R_2 \quad \begin{bmatrix} 2 & -3 & 1 & 5 \\ 0 & 20 & 10 & 0 \\ 0 & 0 & -3 & -6 \end{bmatrix}$$

Since $rank(A) = rank(A) = 3 = n$ *the solution exists and is unique.*

$$\begin{aligned} 2x_1 \quad -3x_2 \quad +x_3 \quad &= 5 \\ 20x_2 \quad +10x_3 \quad &= 0 \\ -3x_3 \quad &= -6 \end{aligned}$$

From this we get $x_3 = 2$. *And using back substitution we have* $x_2 = -1$ *and* $x_1 = 0$
Hence $(0, -1, 2)$ *is the solution of the system.*

Exercise 1.2

Solve the following system of four equations using the Gauss elimination method.

$$\begin{aligned} 4x_1 - 2x_2 - 3x_3 + 6x_4 &= 12 \\ -6x_1 + 7x_2 + 6.5x_3 - 6x_4 &= -6.5 \\ x_1 + 7.5x_2 + 6.25x_3 + 5.5x_4 &= 16 \\ -12x_1 + 22x_2 + 15.5x_3 - x4 &= 17 \end{aligned}$$

1.2.5 NAIVE GAUSSIAN ELIMINATION

Consider the system (1.2) in matrix form

$$Ax = b.$$

Let us denote the original system by $A^{(1)}x = b^{(1)}$. That is,

$$A = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_n^{(1)} \end{bmatrix} \tag{1.5}$$

The Gaussian elimination consists of reducing the system (1.5) to an equivalent system $Ux = d$, in which U is an upper triangular matrix. This new system can be easily solved by back substitution.

Algorithm:

Step 1: Assume $a_{11}^{(1)} \neq 0$. Define the row multipliers by

$$m_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}$$

Multiply the first row by m_{i1} and subtract from the i^{th} row $(i = 2, \cdots, n)$ to get

$$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i1} a_{1j}^{(1)}, \ j = 2, 3, \cdots, n$$
$$b_i^{(2)} = b_i^{(1)} - m_{i1} b_1^{(1)}.$$

Here, the first rows of A and b are left unchanged, and the entries of the first column of A below $a_{11}^{(1)}$ are set to zeros. The result of the transformed system is

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{bmatrix}$$

We continue in this way. At the k^{th} step we have

Step k: Assume $a_{kk}^{(k)} \neq 0$. Define the row multipliers by

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$$

Multiply the k^{th} row by m_{ik} and subtract from the i^{th} row $(i = k+1, \cdots, n)$ to get

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik} a_{kj}^{(k)}, \ j = k+1, \cdots, n$$
$$b_i^{(k+1)} = b_i^{(k)} - m_{ik} b_k^{(k)}.$$

At this step, the entries of column k below the diagonal element are set to zeros, and the rows 1 through k are left undisturbed. The result of the transformed system is

$$\begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1k}^{(1)} & a_{1,k+1}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{1k}^{(2)} & a_{1,k+1}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{kk}^{(k)} & a_{k,k+1}^{(k)} & \cdots & a_{kn}^{(k)} \\ 0 & 0 & \cdots & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{nn}^{(k+1)} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \\ x_{k+1} \\ \cdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_k^{(k)} \\ b_{k+1}^{(k+1)} \\ \vdots \\ b_n^{(k+1)} \end{bmatrix}$$

At $k = n - 1$, we obtain the final triangular system

$$\begin{aligned} a_{11}^{(1)} x_1 + a_{12}^{(1)} x_2 + \cdots + a_{1n}^{(1)} x_n &= b_1^{(1)} \\ a_{22}^{(2)} x_2 + \cdots + a_{2n}^{(2)} x_n &= b_2^{(2)} \\ \cdots &= \cdots \\ a_{n-1,n-1}^{(n-1)} x_{n-1} + a_{n-1,n}^{(n-1)} x_n &= b_{n-1}^{(n-1)} \\ a_{nn}^{(n)} x_n &= b_n^{(n)}. \end{aligned}$$

Using back substitution, we obtain the following solution of the system

$$\begin{aligned} x_n &= \frac{b_n^{(n)}}{a_{nn}^{(n)}} \\ x_{n-1} &= \frac{b_{n-1}^{(n-1)} - a_{n-1,n}^{(n-1)} x_n}{a_{n-1,n-1}^{(n-1)}} \\ x_i &= \frac{b_i^{(i)} - (a_{i,i+1}^{(i)} x_{i+1} + \cdots + a_{in}^{(i)} x_n)}{a_{ii}^{(i)}} \\ &= \frac{b_i^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(i)} x_j}{a_{ii}^{(i)}}, \; i = n-2, n-3, \cdots, 1. \end{aligned}$$

Remarks: In the Gaussian elimination algorithm described above, we used the equations in their natural order and we assumed at each step that the pivot element $a_{kk}^{(k)} \neq 0$. So the algorithm fails if the pivot element becomes zero during the elimination process. In order to avoid an accidental zero pivot, we use what is called Gaussian elimination with scaled partial pivoting.

Theorem 1.2

The total number of multiplications and divisions required to obtain the solution of an $n \times n$ linear system using naive Gaussian elimination is

$$\frac{n^3}{3} + n^2 - \frac{n}{3}.$$

Hence, for n large the total number of operations is approximately $n^3/3$.

Example 1.4

Solve the system of equations

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 \\ -1 & 1 & -5 & 3 \\ 3 & 1 & 7 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \begin{bmatrix} 10 \\ 31 \\ -2 \\ 18 \end{bmatrix}$$

Solution: *The augmented matrix along with the row multipliers m_{i1} are*

$$\begin{array}{rl} & \\ \textit{pivotal} & \textit{element} \rightarrow \\ m_{21} = 2 & \\ m_{31} = -1 & \\ m_{41} = 3 & \end{array} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 10 \\ 2 & 3 & 1 & 5 & 31 \\ -1 & 1 & -5 & 3 & -2 \\ 3 & 1 & 7 & -2 & 18 \end{array}\right].$$

Subtracting multiples of the first equation from the three others gives

$$\begin{array}{rl} & \\ \textit{pivotal} & \textit{element} \rightarrow \\ m_{32} = 2 & \\ m_{42} = -2 & \end{array} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 10 \\ 0 & 1 & -1 & 3 & 11 \\ 0 & 2 & -4 & 4 & 8 \\ 0 & -2 & 4 & -5 & -12 \end{array}\right].$$

Example

Subtracting multiples, of the second equation from the last two, gives

$$\begin{array}{ll} pivotal & element \rightarrow \\ & m_{43}=-1 \end{array} \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 10 \\ 0 & 1 & -1 & 3 & 11 \\ 0 & 0 & -2 & -2 & -14 \\ 0 & 0 & 2 & 1 & 10 \end{array}\right].$$

Subtracting multiples, of the third equation from the last one, gives the upper triangular system

$$\left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 10 \\ 0 & 1 & -1 & 3 & 11 \\ 0 & 0 & -2 & -2 & -14 \\ 0 & 0 & 0 & -1 & -4 \end{array}\right].$$

The process of the back substitution algorithm applied to the triangular system produces the solution

$$\begin{aligned} x_4 &= \frac{-4}{-1} = 4 \\ x_3 &= \frac{-14+2x_4}{-2} = \frac{-6}{-2} = 3 \\ x_2 &= 11 + x_3 - 3x_4 = 11 + 3 - 12 = 2 \\ x_1 &= 10 - x_2 - x_3 - x_4 = 10 - 2 - 3 - 4 = 1 \end{aligned}$$

Example: Matlab Solution

```
A=[1 1 1 1;2 3 1 5;-1 1 -5 3;3 1 7 -2];
 b=[10 31 -2 18]';
 ngaussel(A,b)


 The augmented matrix is
augm =
     1     1     1     1    10
     2     3     1     5    31
    -1     1    -5     3    -2
     3     1     7    -2    18
 The transformed upper triangular augmented matrix C is =
C =
     1     1     1     1    10
     0     1    -1     3    11
     0     0    -2    -2   -14
     0     0     0    -1    -4
 The vector solution is =
x =
     1
     2
     3
     4
```


1.2.6 GAUSS ELIMINATION WITH PIVOTING

Having a zero pivot element is not the only source of trouble that can arise when we apply naive Gaussian elimination. Under certain circumstances, the pivot element can become very small in magnitude compared to the rest of the elements in the pivot row. This can dramatically increase the round-off error, which can result in an inaccurate solution vector. To illustrate some of the effects of round-off error in the elimination process, we apply naive Gaussian elimination to the system

$$\begin{aligned} 0.0002x_1 + 1.471x_2 &= 1.473 \\ 0.2346x_1 - 1.317x_2 &= 1.029 \end{aligned}$$

using four-digit floating-point arithmetic with rounding. The exact solution of this system is $x_1 = 10.00$ and $x_2 = 1.000$. The multiplier for this system is

$$m_{21} = \frac{0.2346}{0.0002} = 1173.$$

Applying naive Gaussian elimination and performing the appropriate rounding gives

$$\begin{aligned} 0.0002x_1 + 1.471x_2 &= 1.473 \\ -1726x_2 &= -1727. \end{aligned}$$

Hence,

$$\begin{aligned} x_2 &= \frac{-1727}{-1726} = 1.001 \\ x_1 &= \frac{1.473 - (1.471)(1.001)}{0.0002} \\ &= \frac{1.473 - 1.472}{0.0002} \\ &= 5.000. \end{aligned}$$

As one can see, x_2 is a close approximation of the actual value. However, the relative error in the computed solution for x_1 is very large: 50%. The failure of naive Gaussian elimination in this example results from the fact that $|a_{11}| = 0.0002$ is small compared to $|a_{12}|$. Hence, a relatively small error due to round-off in the computed value, x_2, led to a relatively large error in the computed solution, x_1.

A useful strategy to avoid the problem of having a zero or very small pivot element is to use Gaussian elimination with scaled partial pivoting. In this method, the equations of the system (1.5) are used in a different order, and the pivot equation is selected by looking for the absolute largest coefficient of x_k relative to the size of the equation. The basic idea in elimination with partial pivoting is to avoid small pivots and control the size of the multipliers. The order in which the equations would be used as pivot equations is determined by the index vector that we call $d = [d_1, d_2, \cdots, d_n]$. At the beginning we set $d = [1, 2, \cdots, n]$. We then define the scale vector

$$c = [c_1, c_2, \cdots, c_n]$$

where

$$c_i = \max_{1 \le j \le n} |a_{ij}|, i = 1, 2, \cdots, n.$$

The elimination with scaled partial pivoting consists of choosing the pivot equation such that the ratio $|a_{i,1}|/c_i$ is greatest. To do that we define the ratio vector

$$r = [r_1, r_2, \cdots, r_n]$$

where

$$r_i = |a_{i1}|/c_i, i = 1, 2, \cdots, n.$$

If r_j is the largest element in r, we interchange d_1 and d_j in the index vector d to get the starting index vector

$$d = [d_j, d_2, \cdots, d_1, \cdots, d_n].$$

This means that row j is the pivot equation in step 1. The Gaussian elimination is then used to get an equivalent system of equation with zeros below and above the pivot element. Note that during the elimination process, only elements in the index vector d have been interchanged and not the equations. The process continues in this way until the end of step $(n-1)$ where a final index vector is obtained containing the order in which the pivot equations were selected. The solution of the system of equation is then obtained by performing a back substitution, reading the entries of the index vector from the last to the first.

Example 1.5

Solve the system of equation using Gaussian elimination with scaled partial pivoting

$$\begin{bmatrix} 1 & 3 & -2 & 4 \\ 2 & -3 & 3 & -1 \\ -1 & 7 & -4 & 2 \\ 3 & -1 & 6 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \begin{bmatrix} -11 \\ 6 \\ -9 \\ 15 \end{bmatrix}$$

Solution:

```
A=[1 3 -2 4;2 -3 3 -1;-1 7 -4 2;3 -1 6 2];
 b=[-11 6 -9 15]';
 gaussel(A,b)

 The augmented matrix is =
augm =
     1     3    -2     4   -11
     2    -3     3    -1     6
    -1     7    -4     2    -9
     3    -1     6     2    15
The scale vector =
c =
     4     3     7     6
The index vector =
d =    2     1     4     3
 The transformed upper triangular augmented matrix C is =
C =
    2.0000   -3.0000    3.0000   -1.0000  -14.0000
         0    4.5000   -3.5000    4.5000    6.0000
         0         0    4.2222         0    4.0000
         0         0         0   -4.0000   16.8889
 The vector solution is =
x =
    -2
     1
     4
    -1
```

Example 1.6

Solve the system of equation using Gaussian elimination with scaled partial pivoting

$$\begin{bmatrix} 0.0002 & 1.471 \\ 0.2346 & -1.317 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.473 \\ 1.029 \end{bmatrix}.$$

Solution: $x_1 = 10$ *and* $x_2 = 1$.

Ill-conditioning

A linear system is said to be ill-conditioned if the coefficient matrix tends to be singular, that is, small perturbations in the coefficient matrix will produce large changes in the solution of the system. For example, consider the following system in two equations:

$$\begin{bmatrix} 1.00 & 2.00 \\ 0.49 & 0.99 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3.00 \\ 1.47 \end{bmatrix}.$$

The exact solution of this system is $x = [30]'$. Now consider the system

$$\begin{bmatrix} 1.00 & 2.00 \\ 0.48 & 0.99 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3.001.47 \end{bmatrix}$$

obtained by changing the entry $a_{21} = 0.49$ of the previous system to 0.48.

The exact solution of this new system is $x = [11]'$ which differs significantly from the first one. Then it is realized that ill-conditioning is present. The difficulty arising from ill-conditioning cannot be solved by simple refinements in the Gaussian elimination procedure. To find out if a system of equations $Ax = b$ is ill-conditioned, one has to compute the so-called condition number of the matrix A.

Example 1.7

Using the four decimal places computer, solve the linear system

$$\begin{aligned} 0.729x_1 + 0.81x_2 + 0.9x_3 &= 0.6867 \\ x_1 + x_2 + x_3 &= 0.8338 \\ 1.331x_1 + 1.21x_2 + 1.1x_3 &= 1.0000 \end{aligned}$$

Its exact solution rounded to four decimal places, is $x_1 = 0.2245; x_2 = 0.2814 \& x_3 = 0.3279$

Solution without pivoting:

$$\left[\begin{array}{ccc|c} 0.729 & 0.81 & 0.9 & 0.6867 \\ 1 & 1 & 1 & 0.8338 \\ 1.331 & 1.21 & 1.1 & 1.0000 \end{array}\right] \Longrightarrow \begin{aligned} m_{21} &= \frac{a_{21}}{a_{11}} = 1.372 | R_2 \leftarrow R_2 - m_{21}R_1 \\ m_{31} &= \frac{a_{31}}{a_{11}} = 1.826 | R_3 \leftarrow R_3 - m_{31}R_1 \end{aligned}$$

$$\left[\begin{array}{ccc|c} 0.729 & 0.81 & 0.9 & 0.6867 \\ 0.0 & -0.1110 & -0.2350 & -0.1084 \\ 0.0 & -0.2690 & -0.430 & -0.2540 \end{array}\right] \Longrightarrow m_{32} = \frac{a_{21}^{(2)}}{a_{22}^{(2)}} = 2.423 | R_3 \leftarrow R_3 - m_{32}R_2$$

$$\left[\begin{array}{ccc|c} 0.729 & 0.81 & 0.9 & |0.6867 \\ 0.0 & .0.1110 & -0.2350 & -0.1084 \\ 0.0 & 0.0 & 0.0264 & 0.0087 \end{array}\right]$$

Using back substitution the solution becomes, $x_1 = 0.2251, x_2 = 0.2790, \& x_3 = 0.3295$.

Example

Solution with pivoting:
To interchange 0 the rows $i\&j$ we will use the notation $R_i \Leftrightarrow R_j$.

$$\begin{bmatrix} 0.729 & 0.81 & 0.9 & | & 0.6867 \\ 1 & 1 & 1 & | & 0.8338 \\ 1.331 & 1.21 & 1.1 & | & 1.0000 \end{bmatrix} \implies R_1 \Leftrightarrow R_3 \& \quad \begin{aligned} m_{21} &= \frac{a_{21}}{a_{11}} = 0.7513 | R_2 \leftarrow R_2 - m_{21} R_1 \\ m_{31} &= \frac{a_{31}}{a_{11}} = 0.5477 | R_3 \leftarrow R_3 - m_{31} R_1 \end{aligned}$$

$$\begin{bmatrix} 1.331 & 1.21 & 1.1 & | & 1.0000 \\ 0.0 & 0.0909 & 0.1736 & | & 0.0825 \\ 0.0 & 0.1473 & 0.2975 & | & 0.1390 \end{bmatrix} \implies m_{32} = \frac{a_{21}^{(2)}}{a_{22}^{(2)}} = 0.6171 | R_3 \leftarrow R_3 - m_{32} R_2$$

$$\begin{bmatrix} 1.331 & 1.21 & 1.1 & | & 1.0000 \\ 0.0 & 0.0909 & 0.1736 & | & 0.0825 \\ 0.0 & 0.0 & -0.010 & | & -0.0033 \end{bmatrix}$$

The solution using backward substitution is $x_1 = 0.2246, x_2 = 0.2812 \& x_3 = 0.3280$. Comparing solution with and without pivoting to the exact solution rounded to four decimal places, we observe that solution with pivoting is much accurate solution than the solution without pivoting.

1.2.7 Gauss-Jordan Elimination Method

The Gauss-Jordan elimination method to solve a system of linear equations is described in the following steps.

1. Write the augmented matrix of the system.
2. Use row operations to transform the augmented matrix in the form described below, which is called the reduced row echelon form (RREF).
   (a) The rows (if any) consisting entirely of zeros are grouped together at the bottom of the matrix.
   (b) In each row that does not consist entirely of zeros, the leftmost nonzero element is a 1 (called a leading 1 or a pivot).
   (c) Each column that contains a leading 1 has zeros in all other entries.
   (d) The leading 1 in any row is to the left of any leading 1's in the rows below it.
3. Stop process in step 2 if you obtain a row whose elements are all zeros except the last one on the right. In that case, the system is inconsistent and has no solutions. Otherwise, finish step 2 and read the solutions of the system from the final matrix.

Note: When doing step 2, row operations can be performed in any order. Try to choose row operations so that as few fractions as possible are carried through the computation. This makes calculation easier when working by hand.

Example 1.8

Given the following linear system with corresponding augmented matrix:

$$\begin{aligned} 3x_2 - 6x_3 + 6x_4 + 4x_5 &= -5 \\ 3x_1 - 7x_2 + 8x_3 - 5x_4 + 8x_5 &= 9 \\ 3x_1 - 9x_2 + 12x_3 - 9x_4 + 6x_5 &= 15 \end{aligned}$$

$$\begin{bmatrix} 0 & 3 & -6 & 6 & 4 & -5 \\ 3 & -7 & 8 & -5 & 8 & 9 \\ 3 & -9 & 12 & -9 & 6 & 15 \end{bmatrix}$$

To solve this system, the matrix has to be reduced into reduced echelon form.

Step 1: Switch row 1 and row 3. All leading zeros are now below non-zero leading entries.

$$\begin{bmatrix} 3 & -9 & 12 & -9 & 6 & 15 \\ 3 & -7 & 8 & -5 & 8 & 9 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix}$$

Step 2: Set row 2 to row 2 plus (-1) times row 1. In other words, subtract row 1 from row 2. This will eliminate the first entry of row 2.

$$\begin{bmatrix} 3 & -9 & 12 & -9 & 6 & 15 \\ 0 & 2 & -4 & 4 & 2 & -6 \\ 0 & 3 & -6 & 6 & 4 & -5 \end{bmatrix}$$

Step 3: Multiply row 2 by 3 and row 3 by 2. This will eliminate the first entry of row 3.

$$\begin{bmatrix} 3 & -9 & 12 & -9 & 6 & 15 \\ 0 & 6 & -12 & 12 & 6 & -18 \\ 0 & 6 & -12 & 12 & 8 & -10 \end{bmatrix}$$

Step 4: Set row 3 to row 3 plus (-1) times row 2. In other words, subtract row 2 from row 3. This will eliminate the second entry of row 3.

$$\begin{bmatrix} 3 & -9 & 12 & -9 & 6 & 15 \\ 0 & 6 & -12 & 12 & 6 & -18 \\ 0 & 0 & 0 & 0 & 2 & 8 \end{bmatrix}$$

Step 5: Multiply each row by the reciprocal of its first non-zero value. This will make every row start with a_1.

$$\begin{bmatrix} 1 & -3 & 4 & -3 & 2 & 5 \\ 0 & 1 & -2 & 2 & 1 & -3 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix}$$

The matrix is now in row echelon form: All nonzero rows are above any rows of all zeros (there are no zero rows), each leading entry of a row is in a column to the right of the leading entry of the row above it and all entries in a column below a leading entry are zeros.

Example

As can and will be shown later, from this form one can observe that the system has infinitely many solutions. To get those solutions, the matrix is further reduced into reduced echelon form.

Step 6: Set row 2 to row 2 plus (-1) times row 3 and row 1 to row 1 plus (-2) times row 3. This will eliminate the entries above the leading entry of row 3.

$$\begin{bmatrix} 1 & -3 & 4 & -3 & 0 & -3 \\ 0 & 1 & -2 & 2 & 0 & -7 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix}$$

Step 7: Set row 1 to row 1 plus 3 times row 2. This eliminates the entry above the leading entry of row 2.

$$\begin{bmatrix} 1 & 0 & -2 & 3 & 0 & -24 \\ 0 & 1 & -2 & 2 & 0 & -7 \\ 0 & 0 & 0 & 0 & 1 & 4 \end{bmatrix}$$

This is a reduced echelon form, since the leading entry in each nonzero row is 1 and each leading 1 is the only nonzero entry in its column.
From this the solution of the system can be read:

$$\begin{aligned} x_1 - 2x_3 + 3x_4 &= -24 \\ x_2 - 2x_3 + 2x_4 &= -7 \\ x_5 &= 4 \end{aligned}$$

Those equations can be solved for x_1, x_2 and x_5:

$$\begin{aligned} x_1 &= 2x_3 - 3x_4 - 24 \\ x_2 &= 2x_3 - 2x_4 - 7 \\ x_5 &= 4 \end{aligned}$$

This is the solution of the system. The variables x_3 and x_4 can take any value and are so called free variables. The solution is valid for any x_3 and x_4.

Exercise 1.3

Solve the following system by using the Gauss-Jordan elimination method.

1. $\begin{aligned} x + y + z &= 5 \\ 2x + 3y + 5z &= 8 \\ 4x + 5z &= 2 \end{aligned}$
2. $\begin{aligned} x + 2y - 3z &= 2 \\ 6x + 3y - 9z &= 6 \\ 7x + 14y - 21z &= 13 \end{aligned}$
3. $\begin{aligned} 4y + z &= 2 \\ 2x + 6y - 2z &= 3 \\ 4x + 8y - 5z &= 4 \end{aligned}$

4 *Ava invests a total of $\$10,000$ in three accounts, one paying 5 interest, another paying 8 interest, and the third paying 9 interest. The annual interest earned on the three investments last year was $\$770$. The amount invested at 9 was twice the amount invested at 5 . How much was invested at each rate?*

Exercise 1.4

1. *Solve the following linear system of equation by using Cramer's rule, Gaussian elimination method, and inverse method.*

$$(a)\ \begin{aligned} 2x_1 + 5x_2 + 3x_3 &= 9 \\ 3x_1 + x_2 + 2x_3 &= 3 \\ x_1 + 2x_2 - x_3 &= 6 \end{aligned} \qquad (b)\ \begin{aligned} x + z &= 1 \\ 2x + y + z &= 0 \\ x + y + 2z &= 1 \end{aligned} \qquad (c)\ \begin{aligned} x + 2y + z &= 3 \\ 2x + 5y - z &= -4 \\ 3x - 2y - z &= 5 \end{aligned}$$

2. *Use rank of matrix to determine the values of a, b and c so that the following system has:*
   a) no solution b) more than one solution c) a unique solution and solve it.

$$i)\ \begin{aligned} 1x + y - bz &= 1 \\ 2x + 3y + az &= 3 \\ x + ay + 3z &= 2 \end{aligned} \qquad ii)\ \begin{aligned} x + 2y - 3z &= a \\ 2x + 6y - 11z &= b \\ x - 2y + 7z &= c \end{aligned} \qquad iii)\ \begin{aligned} x - 2y + bz &= 3 \\ ax + 2z &= 2 \\ 5x + 2y &= 2 \end{aligned}$$

3. *An electrical network has two voltage sources and six resistors. By applying both Ohm's law and Kirchhoff's Current law, we get the following linear system of equations:*

$$\begin{bmatrix} R_1 + R_3 + R_4 & R_3 & R_4 \\ R_3 & R_2 + R_3 + R_5 & -R_5 \\ R_4 & -R_5 & R_4 + R_5 + R_6 \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \\ i_3 \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \\ 0 \end{bmatrix}.$$

   solve the linear system for the current i_1, i_2, and i_3 if

   (a) $R_1 = 1, R_2 = 2, R_3 = 1, R_4 = 2, R_5 = 1, R_6 = 6$, *and* $V_1 = 20, V_2 = 30$,

   (b) $R_1 = 1, R_2 = 1, R_3 = 1, R_4 = 2, R_5 = 2, R_6 = 4$, *and* $V_1 = 12.5, V_2 = 22.5$,

   (c) $R_1 = 2, R_2 = 2, R_3 = 4, R_4 = 1, R_5 = 4, R_6 = 3$, *and* $V_1 = 40, V_2 = 36$.

4. *Use both naive Gaussian elimination and Gaussian elimination with scaled partial pivoting to solve the following linear system using four-decimal floating point arithmetic*

$$\begin{aligned} 0.0003x_1 + 1.354x_2 &= 1.357 \\ 0.2322x_1 - 1.544x_2 &= 0.7780 \end{aligned}$$

5. *Solve the following set of four equations using the Gauss-Jordan elimination method.*

$$\begin{aligned} 4x_1 - 2x_2 - x_3 + 6x_4 &= 12 \\ -6x_1 + 7x_2 + 6.5x_3 - 6x_4 &= -6.5 \\ x_1 + 7.5_2 + 6.25x_3 + 5.5x_4 &= 16 \\ -12x_1 + 22x_2 + 15.5x_3 - x_4 &= 17 \end{aligned}$$

1.3 LU Decomposition Method

Consider the system of equations

$$Ax = b.$$

The LU decomposition consists of transforming the coefficient matrix A into the product of two matrices, L and U, where L is a lower triangular matrix and U is an upper triangular matrix

having 1's on its diagonal.

LU Decomposition Algorithm:

Given a real nonsingular matrix A, apply LU decomposition first:

$$A = LU.$$

Given also a right-hand-side vector b:

1. Forward substitution: solve
$$Ly = b.$$
2. Backward substitution: solve
$$Ux = y.$$

Two types of factorizations will now be presented, the first one uses Crout's and Cholesky's methods and the second one uses the Gaussian elimination method.

1.3.1 Crout and Doolittle's decomposition method

We shall illustrate the method of finding L and U in the case of a 4-by-4 matrix: We wish to find L, having nonzero diagonal entries, and U such that

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

Multiplying the rows of L by the first column of U, one gets

$$l_{i1} = a_{i1}, i = 1, 2, 3, 4.$$

Hence, the first column of L is given by the first column of A. Next, multiply the columns of U by the first row of L to get

$$l_{11}u_{1i} = a_{1i}, i = 2, 3, 4.$$

Thus,

$$u_{1i} = \frac{a_{1i}}{l_{11}}, i = 2, 3, 4,$$

which give the first row of U. We continue in this way by getting alternatively a column of L and a row of U. The result is

$$\begin{aligned} l_{i2} &= a_{i2} - l_{i1}u_{12}, i = 2, 3, 4. \\ u_{2i} &= \frac{a_{2i} - l_{21}u_{1i}}{l_{22}}, i = 3, 4. \\ l_{i3} &= a_{i3} - l_{i1}u_{13} - l_{i2}u_{23}, i = 3, 4. \\ u_{34} &= \frac{a_{34} - l_{31}u_{14} - l_{32}u_{24}}{l_{33}}, \\ l_{44} &= a_{44} - l_{41}u_{14} - l_{42}u_{24} - l_{43}u_{34}. \end{aligned}$$

In algorithmic form, the factorization may be presented as follows for an $n \times n$ matrix:

$$l_{ij} = a_{ij} - \sum_{k=1}^{j-1} l_{ik}u_{kj}, j \leq i, i = 1, 2, \cdots, n. \tag{1.6}$$

$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} l_{ik}u_{kj}}{l_{ii}}, i \leq j, j = 2, 3, \cdots, n. \tag{1.7}$$

Note that this algorithm can be applied if the diagonal elements l_{ii}, for each $i = 1, \cdots, n$, of L, are nonzero.

The LU factorization that we have just described, requiring the diagonal elements of U to be one, is known as Crout's method. If instead the diagonal of L is required to be one, the factorization is called **Doolittle's method**.

- $l_{ii} = 1, (i = 1, 2, 3, \cdots, n)$ the method is called Doolittle's method.
- $U_{ii} = 1, (i = 1, 2, 3, \cdots, n)$ the method is called Crout's method.

Example 1.9

Use Crout's method to solve the system

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 \\ -1 & 1 & -5 & 3 \\ 3 & 1 & 7 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 10 \\ 31 \\ -2 \\ 18 \end{bmatrix}.$$

Solution: If A has a direct factorization LU, then

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 \\ -1 & 1 & -5 & 3 \\ 3 & 1 & 7 & -2 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

By multiplying L with U and comparing the elements of the product matrix with those of A, we obtain:

1. Multiplication of the first row of L with the columns of U gives

$$\begin{aligned} &l_{11} = 1, \\ &l_{11}u_{12} = 1 \implies u_{12} = 1, \\ &l_{11}u_{13} = 1 \implies u_{13} = 1, \\ &l_{11}u_{14} = 1 \implies u_{14} = 1. \end{aligned}$$

2. Multiplication of the second row of L with the columns of U gives

$$\begin{aligned} &l_{21} = 2, \\ &l_{21}u_{12} + l_{22} = 3 \implies l_{22} = 3 - l_{21}u_{12} = 1, \\ &l_{21}u_{13} + l22u_{23} = 1 \implies u_{23} = (1 - l_{21}u_{13})/l_{22} = -1, \\ &l_{21}u_{14} + l_{22}u_{24} = 5 \implies u_{24} = (5 - l_{21}u_{14})/l_{22} = 3. \end{aligned}$$

3. Multiplication of the third row of$L$ with the columns of U gives

$$
\begin{aligned}
&l_{31} = -1,\\
&l_{31}u_{12} + l_{32} = 1 \implies l_{32} = 1 - l_{31}u_{12} = 2,\\
&l_{31}u_{13} + l_{32}u_{23} + l_{33} = -5 \implies l_{33} = -5 - l_{31}u_{13}l_{32}u_{23} = -2,\\
&l_{31}u_{14} + l_{32}u_{24} + l_{33}u_{34} = 3 \implies u_{34} = (3 - l_{31}u_{14} - l_{32}u_{24})/l_{33} = 1.
\end{aligned}
$$

4. Multiplication of the fourth row of L with the columns of U gives

$$
\begin{aligned}
&l_{41} = 3,\\
&l_{41}u_{12} + l_{42} = 1 \implies l_{42} = 1 - l_{41}u_{12} = -2,\\
&l_{41}u_{13} + l_{42}u_{23} + l_{43} = 7 \implies l_{43} = 7 - l_{41}u_{13} - l_{42}u_{23} = 2,\\
&l_{41}u_{14} + l_{42}u_{24} + l_{43}u_{34} + l_{44} = -2 \implies l_{44} = -2 - l_{41}u_{14} - l_{42}u_{24} - l_{43}u_{34} = -1.
\end{aligned}
$$

Hence,

$$
L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 2 & -2 & 0 \\ 3 & -2 & 2 & -1 \end{bmatrix} \text{ and } U = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.
$$

By applying the forward substitution to the lower triangular system $Ly = b$, we get

$$
\begin{aligned}
&y_1 = 10\\
&y_2 = 31 - 2(10) = 11\\
&y_3 = [-2 + 10 - 2(11)]/(-2) = 7\\
&y_4 = -[18 - 3(10) + 2(11) - 2(7)] = 4.
\end{aligned}
$$

Finally, by applying the back substitution to the upper triangular system $Ux = y$, we get

$$
\begin{aligned}
&x_1 = 10 - 4 - 3 - 2 = 1\\
&x_2 = -[11 - 4 - 3(3)] = 2\\
&x_3 = 7 - 4 = 3\\
&x_4 = 4.
\end{aligned}
$$

Computed results with MATLAB output

```
A=[1 1 1 1;2 3 1 5;-1 1 -5 3;3 1 7 -2];
 b=[10 31 -2 18];
[L, U]=LUdecompCrout(A)
y=ForwardSub(L,b)
x=BackwardSub(U,y)
L =
     1     0     0     0
     2     1     0     0
    -1     2    -2     0
     3    -2     2    -1
U =
     1     1     1     1
     0     1    -1     3
     0     0     1     1
```


```
     0     0     0     1
The forward substitution gives
y =

    10
    11
     7
     4
The Backward substitution gives
x =
     1
     2
     3
     4
```

Exercise 1.5

Find the solution of system of linear equation using Crout's and Doolittle's method?

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ 4x_1 + 3x_2 - x_3 &= 6 \\ 3x_1 + 5x_2 + 3x_3 &= 4 \end{aligned}$$

1.3.2 Cholesky Decomposition

Definition 1.2

Symmetric matrix *is a square matrix that is equal to its transpose. Formally, matrix A is symmetric if*

$$A = A^{\mathrm{T}}.$$

Definition 1.3

A symmetric $n \times n$ real matrix A is said to be ***positive definite*** *if the scalar $X^{\mathrm{T}}AX$ is positive for every non-zero column vector z of n real numbers. Here X^{T} denotes the transpose of X.*

This factorization is known as Cholesky's method, and A can be factored in the form

$$A = LL^T$$

where L is a lower triangular matrix. The construction of L is similar to the one used for Crout's method.

If we write out the equation

$$\begin{aligned} \mathbf{A} = \mathbf{L}\mathbf{L}^T &= \begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{pmatrix} \begin{pmatrix} L_{11} & L_{21} & L_{31} \\ 0 & L_{22} & L_{32} \\ 0 & 0 & L_{33} \end{pmatrix} \\ &= \begin{pmatrix} L_{11}^2 & & (\text{symmetric}) \\ L_{21}L_{11} & L_{21}^2 + L_{22}^2 & \\ L_{31}L_{11} & L_{31}L_{21} + L_{32}L_{22} & L_{31}^2 + L_{32}^2 + L_{33}^2 \end{pmatrix}, \end{aligned}$$

$$\mathbf{L} = \begin{pmatrix} \sqrt{A_{11}} & 0 & 0 \\ A_{21}/L_{11} & \sqrt{A_{22} - L_{21}^2} & 0 \\ A_{31}/L_{11} & \left(A_{32} - L_{31}L_{21}\right)/L_{22} & \sqrt{A_{33} - L_{31}^2 - L_{32}^2} \end{pmatrix}$$

and therefore the following formulae for the entries of L:

$$L_{j,j} = \sqrt{A_{j,j} - \sum_{k=1}^{j-1} L_{j,k}^2},$$
$$L_{i,j} = \frac{1}{L_{j,j}} \left(A_{i,j} - \sum_{k=1}^{j-1} L_{i,k} L_{j,k} \right) \quad \text{for } i > j.$$

The expression under the square root is always positive if A is real and positive-definite.

Example 1.10

Here is the Cholesky decomposition of a symmetric real matrix:

$$\begin{pmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ -8 & 5 & 3 \end{pmatrix} \begin{pmatrix} 2 & 6 & -8 \\ 0 & 1 & 5 \\ 0 & 0 & 3 \end{pmatrix}.$$

And here is its LDL^T decomposition:

$$\begin{pmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -4 & 5 & 1 \end{pmatrix} \begin{pmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 9 \end{pmatrix} \begin{pmatrix} 1 & 3 & -4 \\ 0 & 1 & 5 \\ 0 & 0 & 1 \end{pmatrix}.$$

Example 1.11

Solve the system of equations

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 22 \\ 3 & 22 & 82 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \\ -10 \end{bmatrix}$$

Using the cholesky method. We write

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 8 & 22 \\ 3 & 22 & 82 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix} = \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}l_{31} \\ l_{21}l_{11} & l_{21}^2 + l_{22}^2 & l_{21}l_{31} + l_{22}l_{32} \\ l_{31}l_{11} & l_{31}l_{21} + l_{32}l_{22} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{bmatrix}$$

Comparing the corresponding elements on both sides, we get

$$\begin{aligned} l_{11}^2 = 1, \text{ or } l_{11} &= 1 \\ l_{11}^2 = 1, \text{ or } l_{11} &= 1 \\ l_{11}l_{31} = 3, \quad \textit{or} \quad l_{31} &= 3 \\ l_{21}^2 + l_{22}^2 = 8, \quad \textit{or} \quad l_{22} &= 2 \\ l_{31}l_{21} + l_{32}l_{22} = 22 \quad \textit{or} \quad l_{32} &= 8 \\ l_{31}^2 + l_{32}^2 + l_{33}^2 = 82 \quad \textit{or} \quad l_{33} &= 3 \end{aligned}$$

Example

Hence we get $A = LL^T$ *Where* $L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 8 & 3 \end{bmatrix}$

We write the given system of equations as

$$LL^T x = b$$

$$Ly = b \quad and \quad L^T x = y.$$

From $Ly = b$*, we obtain*

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 8 & 3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 5 \\ 6 \\ -10 \end{bmatrix} \implies \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 5 \\ -2 \\ -3 \end{bmatrix}$$

From $L^T x = y$*, we obtained*

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 2 & 8 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ -2 \\ -3 \end{bmatrix} \implies \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \\ -1 \end{bmatrix}$$

Exercise 1.6

Determine if the following matrix is hermitian positive definite. Also find its Cholessky factorization if possible

$$A = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & 3 \\ 1 & 3 & 2 \end{bmatrix} \ \& \ B = \begin{bmatrix} 1 & 2 & 2 \\ 2 & 8 & 0 \\ 2 & 0 & 24 \end{bmatrix}$$

1.4 Indirect Iteration Method

1.4.1 Introduction

There are occasions when direct methods (like Gaussian Elimination or the use of an LU decomposition) are not the best way to solve a system of equations. An alternative approach is to use an iterative method.

Because of round-off errors, direct methods become less efficient than iterative methods when they are applied to large systems, sometimes with as many as 100,000 variables. Examples of these large systems arise in the solution of partial differential equations. In these cases, an iterative method is preferable. In addition to round-off errors, the amount of storage space required for iterative solutions on a computer is far less than the one required for direct methods when the coefficient matrix of the system is sparse, that is, matrices that contain a high proportion of zeros. Thus, for sparse matrices, iterative methods are more attractive than direct methods.

An iterative scheme for linear systems consists of converting the system (1.2) to the form

$$x = b' - Bx.$$

After an initial guess, $x^{(0)}$ is selected, the sequence of approximation solution vectors is generated by computing $x^{(k)} = b' - Bx^{(k-1)}$ for each $k = 1, 2, 3, \cdots$.

1.4.2 Jacobi Method

Suppose that $x^{(0)} = \begin{bmatrix} x_1^{(0)} \\ x_2^{(0)} \\ \vdots \\ x_n^{(0)} \end{bmatrix}$ is an initial approximation to the solution x of the following system of n equations in n unknowns:

$$\begin{array}{lccccccc} E(1): & a_{11}x_1 & + & a_{12}x_2 & + \cdots + & a_{1n}x_n & = & b_1 \\ E(2): & a_{21}x_1 & + & a_{22}x_2 & + \cdots + & a_{2n}x_n & = & b_2 \\ \vdots & \vdots & & \vdots & \ddots & \vdots & & \vdots \\ E(n): & a_{n1}x_1 & + & a_{n2}x_2 & + \cdots + & a_{nn}x_n & = & b_n \end{array} \tag{1.8}$$

Solving of the system of equations, we assume that the quantities a_{ii} in the system are pivot elements. The the system equation may be written as:

$$\begin{cases} a_{11}x_1 &= b_1 - (a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n) \\ a_{22}x_2 &= b_2 - (a_{21}x_1 + a_{23}x_3 + \cdots + a_{2n}x_n) \\ a_{33}x_3 &= b_3 - (a_{31}x_1 + a_{32}x_2 + \cdots + a_{3n}x_n) \\ &\vdots \\ a_{nn}x_n &= b_n - (a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn-1}x_{n-1}) \end{cases} \tag{1.9}$$

For the Jacobi Iteration method, from the i^{th} equation in the system $Ax = b$ we isolate for the each variable x_i. Provided that $a_{ii} \neq 0$ for $i = 1, 2, \cdots, n$ we get that:

$$\begin{aligned} x_1 &= \frac{b_1 - [a_{12}x_2 + a_{13}x_3 + ... + a_{1n}x_n]}{a_{11}} \\ x_2 &= \frac{b_2 - [a_{21}x_1 + a_{23}x_3 + ... + a_{2n}x_n]}{a_{22}} \\ &\vdots \\ x_n &= \frac{b_n - [a_{n1}x_1 + a_{n2}x_2 + ... + a_{n,n-1}x_{n-1}]}{a_{nn}} \end{aligned}$$

in sigma notation, for each $i = 1, 2, \cdots, n$, we have that:

$$x_i = \frac{b_i - \sum_{j=1, j\neq i}^{n} a_{ij}x_j}{a_{ii}} \tag{1.10}$$

To obtain our first approximation $x^{(1)}$ of the solution x using the Jacobi Iteration Method, we take the isolated equations above and input the values of our initial approximation $x^{(0)}$ to get:

$$\begin{aligned} x_1^{(1)} &= \frac{b_1 - \left[a_{12}x_2^{(0)} + a_{13}x_3^{(0)} + ... + a_{1n}x_n^{(0)}\right]}{a_{11}} \\ x_2^{(1)} &= \frac{b_2 - \left[a_{21}x_1^{(0)} + a_{23}x_3^{(0)} + ... + a_{2n}x_n^{(0)}\right]}{a_{22}} \\ &\vdots \\ x_n^{(1)} &= \frac{b_n - \left[a_{n1}x_1^{(0)} + a_{n2}x_2^{(0)} + ... + a_{n,n-1}x_{n-1}^{(0)}\right]}{a_{nn}} \end{aligned} \tag{1.11}$$

Or in sigma notation, for $i = 1, 2, \cdots, n$:

$$x_i^{(1)} = \frac{b_i - \sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(0)}}{a_{ii}} \tag{1.12}$$

We can then use our approximation $x^{(1)}$ in a similar manner to obtain another approximation, $x^{(2)}$, and so forth in the hopes that these successive approximations converge to the actual solution x of $Ax = b$. For each $k \geq 1$, the $(k+1)^{th}$ iteration of the Jacobi Iteration Method yields:

$$\begin{cases} x_1^{k+1} &= \frac{1}{a_{11}}(b_1 - a_{12}x_2^k + a_{13}x_3^k + \cdots + a_{1n}x_n^k) \\ x_2^{k+1} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^k + a_{23}x_3^k + \cdots + a_{2n}x_n^k) \\ x_3^{k+1} &= \frac{1}{a_{33}}(b_3 - a_{31}x_1^k + a_{32}x_2^k + \cdots + a_{3n}x_n^k) \\ &\vdots \\ x_n^{k+1} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^k + a_{n2}x_2^k + \cdots + a_{nn-1}x_{n-1}^k) \end{cases} \tag{1.13}$$

Matrix form

Let

$$A\mathbf{x} = \mathbf{b}$$

be a square system of n linear equations, where:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Then A can be decomposed into a diagonal component D, and the remainder R:

$$A = D + R \qquad \text{where} \qquad D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix} \text{ and } R = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix}.$$

The solution is then obtained iteratively via

$$\mathbf{x}^{(k+1)} = D^{-1}(\mathbf{b} - R\mathbf{x}^{(k)}),$$

where $\mathbf{x}^{(k)}$ is the kth approximation or iteration of $\mathbf{x}$ and $\mathbf{x}^{(k+1)}$ is the next or $k+1$ iteration of $\mathbf{x}$. The element-based formula is thus:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j \neq i} a_{ij} x_j^{(k)} \right), \quad i = 1, 2, \ldots, n.$$

The computation of $x_i^{(k+1)}$ requires each element in $x^{(k)}$ except itself.

Convergence

The standard convergence condition (for any iterative method) is when the spectral radius of the iteration matrix is less than 1:

$$\rho(D^{-1}R) < 1.$$

A sufficient (but not necessary) condition for the method to converge is that the matrix A is strictly or irreducibly diagonally dominant. Strict row diagonal dominance means that for each row, the absolute value of the diagonal term is greater than the sum of absolute values of other terms:

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

The Jacobi method sometimes converges even if these conditions are not satisfied.

The iterative process is terminated when a convergence criterion is satisfied. One commonly used stopping criterion, known as the relative change criteria, is to iterate until

$$\frac{|x^{(k)} - x^{(k-1)}|}{|x^{(k)}|}, \; x^{(k)} = (x_1^{(k)}, \cdots, x_n^{(k)})^T$$

is less than a prescribed tolerance $\epsilon > 0$. Contrary to Newton?s method for finding the roots of an equation, the convergence or divergence of the iterative process in the Jacobi method does not depend on the initial guess, but depends only on the character of the matrices themselves. However, a good first guess in case of convergence will make for a relatively small number of iterations.

Example 1.12

Suppose we are given the following linear system:

$$\begin{aligned} 10x_1 - x_2 + 2x_3 &= 6, \\ -x_1 + 11x_2 - x_3 + 3x_4 &= 25, \\ 2x_1 - x_2 + 10x_3 - x_4 &= -11, \\ 3x_2 - x_3 + 8x_4 &= 15. \end{aligned}$$

If we choose $(0, 0, 0, 0)$ *as the initial approximation, then the first approximate solution is given by*

$$\begin{aligned} x_1 &= (6 + 0 - 0)/10 = 0.6, \\ x_2 &= (25 - 0 - 0)/11 = 25/11 = 2.2727, \\ x_3 &= (-11 - 0 - 0)/10 = -1.1, \\ x_4 &= (15 - 0 - 0)/8 = 1.875. \end{aligned}$$

Using the approximations obtained, the iterative procedure is repeated until the desired accuracy has been reached. The following are the approximated solutions after five iterations.

x_1	x_2	x_3	x_4
0.6	*2.27272*	*-1.1*	*1.875*
1.04727	*1.7159*	*-0.80522*	*0.88522*
0.93263	*2.05330*	*-1.0493*	*1.13088*
1.01519	*1.95369*	*-0.9681*	*0.97384*
0.98899	*2.0114*	*-1.0102*	*1.02135*

The exact solution of the system is $(1, 2, -1, 1)$.

Example 1.13

A linear system of the form $Ax = b$ with initial estimate $x^{(0)}$ is given by

$$A = \begin{bmatrix} 2 & 1 \\ 5 & 7 \end{bmatrix}, \; b = \begin{bmatrix} 11 \\ 13 \end{bmatrix} \quad \text{and} \quad x^{(0)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

We use the equation $x^{(k+1)} = D^{-1}(b - Rx^{(k)})$, described above, to estimate x. First, we rewrite the equation in a more convenient form $D^{-1}(b - Rx^{(k)}) = Tx^{(k)} + C$, where $T = -D^{-1}R$ and $C = D^{-1}b$. Note that $R = L + U$ where L and U are the strictly lower and upper parts of A. From the known values

$$D^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/7 \end{bmatrix}, \; L = \begin{bmatrix} 0 & 0 \\ 5 & 0 \end{bmatrix} \quad \text{and} \quad U = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

we determine $T = -D^{-1}(L + U)$ as

$$T = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/7 \end{bmatrix} \left\{ \begin{bmatrix} 0 & 0 \\ -5 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix} \right\} = \begin{bmatrix} 0 & -1/2 \\ -5/7 & 0 \end{bmatrix}.$$

Further, C is found as

$$C = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/7 \end{bmatrix} \begin{bmatrix} 11 \\ 13 \end{bmatrix} = \begin{bmatrix} 11/2 \\ 13/7 \end{bmatrix}.$$

With T and C calculated, we estimate x as $x^{(1)} = Tx^{(0)} + C$:

$$x^{(1)} = \begin{bmatrix} 0 & -1/2 \\ -5/7 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 11/2 \\ 13/7 \end{bmatrix} = \begin{bmatrix} 5.0 \\ 8/7 \end{bmatrix} \approx \begin{bmatrix} 5 \\ 1.143 \end{bmatrix}.$$

The next iteration yields

$$x^{(2)} = \begin{bmatrix} 0 & -1/2 \\ -5/7 & 0 \end{bmatrix} \begin{bmatrix} 5.0 \\ 8/7 \end{bmatrix} + \begin{bmatrix} 11/2 \\ 13/7 \end{bmatrix} = \begin{bmatrix} 69/14 \\ -12/7 \end{bmatrix} \approx \begin{bmatrix} 4.929 \\ -1.714 \end{bmatrix}.$$

This process is repeated until convergence (i.e., until $\|Ax^{(n)} - b\|$ is small). The solution after 25 iterations is

$$x = \begin{bmatrix} 7.111 \\ -3.222 \end{bmatrix}.$$

1.4.3 Gauss-Seidel Method

We recently saw The Jacobi Iteration Method for solving a system of linear equations $Ax = b$ where A is an $n \times n$ matrix. We will now look at another method known as the Gauss-Seidel Iteration Method that is somewhat of an improvement of the Jacobi Iteration Method.

We will now obtain a first approximation to the solution $x^{(1)}$ of the actual solution x by using the Gauss-Seidel Iteration Method. We compute $x_1^{(1)}$ by plugging in the values of our initial solution approximation $x^{(0)}$. We then obtain an approximation to the entry $x_1^{(1)}$ of $x_1^{(1)}$. We use this entry and the remaining entries from $x^{(0)}$ to obtain an approximation of the entry $x_2^{(1)}$. We then use both $x_1^{(1)}$ and $x_2^{(1)}$ as well as the remaining entries from $x^{(0)}$ to obtain an

approximation of the entry $x_3^{(1)}$ and so forth, and thus:

$$
\begin{aligned}
x_1^{(1)} &= \frac{b_1 - \left[a_{12}x_2^{(0)} + a_{13}x_3^{(0)} + a_{14}x_4^{(0)} + \ldots + a_{1n}x_n^{(0)}\right]}{a_{11}} \\
x_2^{(1)} &= \frac{b_2 - \left[a_{21}x_1^{(1)} + a_{23}x_3^{(0)} + a_{24}x_4^{(0)} + \ldots + a_{2n}x_n^{(0)}\right]}{a_{22}} \\
x_3^{(1)} &= \frac{b_3 - \left[a_{31}x_1^{(1)} + a_{32}x_2^{(1)} + a_{34}x_4^{(0)} + \ldots + a_{3n}x_n^{(0)}\right]}{a_{33}} \\
&\vdots \\
x_n^{(1)} &= \frac{b_n - \left[a_{n1}x_1^{(1)} + a_{n2}x_2^{(1)} + a_{n3}x_3^{(1)} + \ldots + a_{n,n-1}x_{n-1}^{(1)}\right]}{a_{nn}}
\end{aligned}
$$

In sigma notation we have that each component $x_i^{(1)}$ for $i = 1, 2, \cdots, n$ of $x^{(1)}$ is given by:

$$
x_i^{(1)} = \frac{b_i - \left[\sum_{j=1}^{i-1} a_{ij}x_j^{(1)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(0)}\right]}{a_{ii}}
$$

To obtain the second approximation $x^{(2)}$ of x using the Gauss-Seidel method, we would have that:

$$
\begin{aligned}
x_1^{(2)} &= \frac{b_1 - \left[a_{12}x_2^{(1)} + a_{13}x_3^{(1)} + a_{14}x_4^{(1)} + \ldots + a_{1n}x_n^{(1)}\right]}{a_{11}} \\
x_2^{(2)} &= \frac{b_2 - \left[a_{21}x_1^{(2)} + a_{23}x_3^{(1)} + a_{24}x_4^{(1)} + \ldots + a_{2n}x_n^{(1)}\right]}{a_{22}} \\
x_3^{(2)} &= \frac{b_3 - \left[a_{31}x_1^{(2)} + a_{32}x_2^{(2)} + a_{34}x_4^{(1)} + \ldots + a_{3n}x_n^{(1)}\right]}{a_{33}} \\
&\vdots \\
x_n^{(2)} &= \frac{b_n - \left[a_{n1}x_1^{(2)} + a_{n2}x_2^{(2)} + a_{n3}x_3^{(2)} + \ldots + a_{n,n-1}x_{n-1}^{(2)}\right]}{a_{nn}}
\end{aligned}
$$

In sigma notation we have that each component $x_i^{(2)}$ for $i = 1, 2, \cdots, n$ of $x^{(2)}$ is given by:

$$
x_i^{(2)} = \frac{b_i - \left[\sum_{j=1}^{i-1} a_{ij}x_j^{(2)} + \sum_{j=i+1}^{n} a_{ij}x_j^{(1)}\right]}{a_{ii}}
$$

We can continue approximating x with these solutions in the hopes that the sequence of approximations with the Gauss-Seidel method converges to the true solution. Thus for $k \geq 1$ and for $i = 1, 2, \cdots, n$, the $(k+1)^{th}$ iteration of the Gauss-Seidel method can be defined as:

$$
\begin{cases}
x_1^{k+1} &= \frac{1}{a_{11}}(b_1 - a_{12}x_2^k + a_{13}x_3^k + \cdots + a_{1n}x_n^k) \\
x_2^{k+1} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{k+1} + a_{23}x_3^k + \cdots + a_{2n}x_n^k) \\
x_3^{k+1} &= \frac{1}{a_{33}}(b_3 - a_{31}x_1^{k+1} + a_{32}x_2^{k+1} + \cdots + a_{3n}x_n^k) \\
&\vdots \\
x_n^{k+1} &= \frac{1}{a_{nn}}(b_n - a_{n1}x_1^{k+1} + a_{n2}x_2^{k+1} + \cdots + a_{nn-1}x_{n-1}^{k+1})
\end{cases}
\tag{1.14}
$$

The Gauss-Seidel method is an iterative technique for solving a square system of n linear equations with unknown x:

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

It is defined by the iteration

$$L_*\mathbf{x}^{(k+1)} = \mathbf{b} - U\mathbf{x}^{(k)},$$

where $\mathbf{x}^{(k)}$ is the kth approximation or iteration of $\mathbf{x}$, $\mathbf{x}^{(k+1)}$ is the next or $k+1$ iteration of $\mathbf{x}$, and the matrix A is decomposed into a lower triangular component L_*, and a strictly upper triangular component U: $A = L_* + U$
In more detail, write out A, x and b in their components:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Then the decomposition of A into its lower triangular component and its strictly upper triangular component is given by:

$$A = L_* + U \qquad \text{where} \qquad L_* = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad U = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

The system of linear equations may be rewritten as:

$$L_*\mathbf{x} = \mathbf{b} - U\mathbf{x}$$

The Gauss-Seidel method now solves the left hand side of this expression for x, using previous value for x on the right hand side. Analytically, this may be written as:

$$\mathbf{x}^{(k+1)} = L_*^{-1}(\mathbf{b} - U\mathbf{x}^{(k)}).$$

However, by taking advantage of the triangular form of L_*, the elements of $x^{(k+1)}$ can be computed sequentially using forward substitution:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right), \quad i = 1, 2, \ldots, n.$$

The procedure is generally continued until the changes made by an iteration are below some tolerance, such as a sufficiently small residual.

The element-wise formula for the Gauss-Seidel method is extremely similar to that of the Jacobi method. The computation of $x_i^{(k+1)}$ uses only the elements of $x^{(k+1)}$ that have already been computed, and only the elements of $x^{(k)}$ that have not yet to be advanced to iteration $k+1$. This means that, unlike the Jacobi method, only one storage vector is required as elements can be overwritten as they are computed, which can be advantageous for very large problems.
However, unlike the Jacobi method, the computations for each element cannot be done in parallel. Furthermore, the values at each iteration are dependent on the order of the original equations.

Convergence

The convergence properties of the Gauss-Seidel method are dependent on the matrix A. Namely, the procedure is known to converge if either:

- A is symmetric positive-definite, or
- A is strictly or irreducibly diagonally dominant.

The Gauss-Seidel method sometimes converges even if these conditions are not satisfied.

Example 1.14

A linear system shown as $A\mathbf{x} = \mathbf{b}$ *is given by:* $A = \begin{bmatrix} 16 & 3 \\ 7 & -11 \end{bmatrix}$ *and* $b = \begin{bmatrix} 11 \\ 13 \end{bmatrix}$. *We want to use the equation* $\mathbf{x}^{(k+1)} = L_*^{-1}(\mathbf{b} - U\mathbf{x}^{(k)})$ *in the form* $\mathbf{x}^{(k+1)} = T\mathbf{x}^{(k)} + C$ *where:* $T = -L_*^{-1}U$ *and* $C = L_*^{-1}\mathbf{b}$. *We must decompose* A *into the sum of a lower triangular component* L_* *and a strict upper triangular component* U*:* $L_* = \begin{bmatrix} 16 & 0 \\ 7 & -11 \end{bmatrix}$ *and* $U = \begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix}$. *The inverse of* L_* *is:*

$$L_*^{-1} = \begin{bmatrix} 16 & 0 \\ 7 & -11 \end{bmatrix}^{-1} = \begin{bmatrix} 0.0625 & 0.0000 \\ 0.0398 & -0.0909 \end{bmatrix}$$

Now we can find:

$$T = -\begin{bmatrix} 0.0625 & 0.0000 \\ 0.0398 & -0.0909 \end{bmatrix} \times \begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.000 & -0.1875 \\ 0.000 & -0.1193 \end{bmatrix},$$

$$C = \begin{bmatrix} 0.0625 & 0.0000 \\ 0.0398 & -0.0909 \end{bmatrix} \times \begin{bmatrix} 11 \\ 13 \end{bmatrix} = \begin{bmatrix} 0.6875 \\ -0.7443 \end{bmatrix}.$$

Now we have T *and* C *and we can use them to obtain the vectors* $\mathbf{x}$ *iteratively. First of all, we have to choose* $\mathbf{x}^{(0)}$*: we can only guess. The better the guess, the quicker the algorithm will perform. We suppose:* $x^{(0)} = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}$. *We can then calculate:*

$$x^{(1)} = \begin{bmatrix} 0.000 & -0.1875 \\ 0.000 & -0.1193 \end{bmatrix} \times \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix} + \begin{bmatrix} 0.6875 \\ -0.7443 \end{bmatrix} = \begin{bmatrix} 0.5000 \\ -0.8636 \end{bmatrix}.$$

$$x^{(2)} = \begin{bmatrix} 0.000 & -0.1875 \\ 0.000 & -0.1193 \end{bmatrix} \times \begin{bmatrix} 0.5000 \\ -0.8636 \end{bmatrix} + \begin{bmatrix} 0.6875 \\ -0.7443 \end{bmatrix} = \begin{bmatrix} 0.8494 \\ -0.6413 \end{bmatrix}.$$

$$x^{(3)} = \begin{bmatrix} 0.000 & -0.1875 \\ 0.000 & -0.1193 \end{bmatrix} \times \begin{bmatrix} 0.8494 \\ -0.6413 \end{bmatrix} + \begin{bmatrix} 0.6875 \\ -0.7443 \end{bmatrix} = \begin{bmatrix} 0.8077 \\ -0.6678 \end{bmatrix}.$$

$$x^{(4)} = \begin{bmatrix} 0.000 & -0.1875 \\ 0.000 & -0.1193 \end{bmatrix} \times \begin{bmatrix} 0.8077 \\ -0.6678 \end{bmatrix} + \begin{bmatrix} 0.6875 \\ -0.7443 \end{bmatrix} = \begin{bmatrix} 0.8127 \\ -0.6646 \end{bmatrix}.$$

$$x^{(5)} = \begin{bmatrix} 0.000 & -0.1875 \\ 0.000 & -0.1193 \end{bmatrix} \times \begin{bmatrix} 0.8127 \\ -0.6646 \end{bmatrix} + \begin{bmatrix} 0.6875 \\ -0.7443 \end{bmatrix} = \begin{bmatrix} 0.8121 \\ -0.6650 \end{bmatrix}.$$

$$x^{(6)} = \begin{bmatrix} 0.000 & -0.1875 \\ 0.000 & -0.1193 \end{bmatrix} \times \begin{bmatrix} 0.8121 \\ -0.6650 \end{bmatrix} + \begin{bmatrix} 0.6875 \\ -0.7443 \end{bmatrix} = \begin{bmatrix} 0.8122 \\ -0.6650 \end{bmatrix}.$$

As expected, the algorithm converges to the exact solution: $\mathbf{x} = A^{-1}\mathbf{b} \approx \begin{bmatrix} 0.8122 \\ -0.6650 \end{bmatrix}$. *In fact, the matrix* A *is strictly diagonally dominant (but not positive definite).*

Example 1.15

Another linear system shown as $A\mathbf{x} = \mathbf{b}$ *is given by:* $A = \begin{bmatrix} 2 & 3 \\ 5 & 7 \end{bmatrix}$ *and* $b = \begin{bmatrix} 11 \\ 13 \end{bmatrix}$. *We want to use the equation*

$$\mathbf{x}^{(k+1)} = L_*^{-1}(\mathbf{b} - U\mathbf{x}^{(k)})$$

in the form

$$\mathbf{x}^{(k+1)} = T\mathbf{x}^{(k)} + C$$

where: $T = -L_*^{-1}U$ *and* $C = L_*^{-1}\mathbf{b}$. *We must decompose* A *into the sum of a lower triangular component* L_* *and a strict upper triangular component* U*:* $L_* = \begin{bmatrix} 2 & 0 \\ 5 & 7 \end{bmatrix}$ *and* $U = \begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix}$. *is:*

$$L_*^{-1} = \begin{bmatrix} 2 & 0 \\ 5 & 7 \end{bmatrix}^{-1} = \begin{bmatrix} 0.500 & 0.000 \\ -0.357 & 0.143 \end{bmatrix}$$

Now we can find:

$$T = -\begin{bmatrix} 0.500 & 0.000 \\ -0.357 & 0.143 \end{bmatrix} \times \begin{bmatrix} 0 & 3 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0.000 & -1.500 \\ 0.000 & 1.071 \end{bmatrix},$$

$$C = \begin{bmatrix} 0.500 & 0.000 \\ -0.357 & 0.143 \end{bmatrix} \times \begin{bmatrix} 11 \\ 13 \end{bmatrix} = \begin{bmatrix} 5.500 \\ -2.071 \end{bmatrix}.$$

Now we have T *and* C *and we can use them to obtain the vectors* $\mathbf{x}$ *iteratively.*
First of all, we have to choose $\mathbf{x}^{(0)}$*: we can only guess. The better the guess, the quicker will perform the algorithm.*
We suppose:

$$x^{(0)} = \begin{bmatrix} 1.1 \\ 2.3 \end{bmatrix}.$$

We can then calculate:

$$x^{(1)} = \begin{bmatrix} 0 & -1.500 \\ 0 & 1.071 \end{bmatrix} \times \begin{bmatrix} 1.1 \\ 2.3 \end{bmatrix} + \begin{bmatrix} 5.500 \\ -2.071 \end{bmatrix} = \begin{bmatrix} 2.050 \\ 0.393 \end{bmatrix}.$$

$$x^{(2)} = \begin{bmatrix} 0 & -1.500 \\ 0 & 1.071 \end{bmatrix} \times \begin{bmatrix} 2.050 \\ 0.393 \end{bmatrix} + \begin{bmatrix} 5.500 \\ -2.071 \end{bmatrix} = \begin{bmatrix} 4.911 \\ -1.651 \end{bmatrix}.$$

$$x^{(3)} = \cdots .$$

If we test for convergence we'll find that the algorithm diverges. In fact, the matrix A *is neither diagonally dominant nor positive definite. Then, convergence to the exact solution*

$$\mathbf{x} = A^{-1}\mathbf{b} = \begin{bmatrix} -38 \\ 29 \end{bmatrix}$$

is not guaranteed and, in this case, will not occur.

Example 1.16

Suppose given k equations where x_n are vectors of these equations and starting point x_0. From the first equation solve for x_1 in terms of $x_{n+1}, x_{n+2}, \ldots, x_n$.. For the next equations substitute the previous values of xs.
To make it clear let's consider an example.

$$\begin{array}{rrrrl} 10x_1 & -x_2 & +2x_3 & & = 6, \\ -x_1 & +11x_2 & -x_3 & +3x_4 & = 25, \\ 2x_1 & -x_2 & +10x_3 & -x_4 & = -11, \\ & 3x_2 & -x_3 & +8x_4 & = 15. \end{array}$$

Solving for x_1, x_2, x_3 and x_4 gives:

$$\begin{aligned} x_1 &= x_2/10 - x_3/5 + 3/5, \\ x_2 &= x_1/11 + x_3/11 - 3x_4/11 + 25/11, \\ x_3 &= -x_1/5 + x_2/10 + x_4/10 - 11/10, \\ x_4 &= -3x_2/8 + x_3/8 + 15/8. \end{aligned}$$

Suppose we choose $(0, 0, 0, 0)$ as the initial approximation, then the first approximate solution is given by

$$\begin{aligned} x_1 &= 3/5 = 0.6, \\ x_2 &= (3/5)/11 + 25/11 = 3/55 + 25/11 = 2.3272, \\ x_3 &= -(3/5)/5 + (2.3272)/10 - 11/10 = -3/25 + 0.23272 - 1.1 = -0.9873, \\ x_4 &= -3(2.3272)/8 + (-0.9873)/8 + 15/8 = 0.8789. \end{aligned}$$

Using the approximations obtained, the iterative procedure is repeated until the desired accuracy has been reached. The following are the approximated solutions after four iterations.

x_1	x_2	x_3	x_4
0.6	2.32727	−0.987273	0.878864
1.03018	2.03694	−1.01446	0.984341
1.00659	2.00356	−1.00253	0.998351
1.00086	2.0003	−1.00031	0.99985

The exact solution of the system is $(1, 2, -1, 1)$.

1.5 Eigenvalue Problem

The calculation of eigenvalues and eigenvectors is a problem that plays an important part in a large number of applications, both theoretical and practical. They touch most areas in science, engineering, and economics. Some examples are the solution of the Schrödinger equation in quantum mechanics, the various eigenvalues representing the energy levels of the resulting orbital, the solution of ordinary equations, space dynamics, elasticity, fluid mechanics, and many others.

1.5.1 Basic Introduction

Le A be a real square, $n \times n$ matrix and let x be a vector of dimension n. We want to find scalars λ for which there exists a nonzero vector x such that

$$Ax = \lambda x. \tag{1.15}$$

When this occurs, we call λ an eigenvalue and x an eigenvector that corresponds to λ. Together they form an eigenpair (λ, x) of A. Note that Eqn. (1.15) will have a nontrivial solution only if

$$p(\lambda) = det(A - \lambda I) = 0. \tag{1.16}$$

The function $p(\lambda)$ is a polynomial of degree n and is known as the characteristic polynomial. The determinant in Eqn. (1.16) can be written in the form

$$\begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} \cdots & a_{nn} - \lambda & \end{vmatrix} = 0$$

It is known that p is an nth degree polynomial with real coefficients and has at most n distinct zeros not necessarily real. Each root λ can be substituted into Eqn. (1.15) to obtain a system of equations that has a nontrivial solution vector x. We now state the following definitions and theorems necessary for the study of eigenvalues.

Definition 1.4

The spectral radius $\rho(A)$ of an $n \times n$ matrix A is defined by

$$\rho(A) = max_{1 \leq i \leq n} |\lambda_i|$$

where λ_i are the eigenvalues of A.

Theorem 1.3

1. *The eigenvalues of a symmetric matrix are all real numbers.*
2. *For distinct eigenvalues λ there exists at least one eigenvector v corresponding to λ.*
3. *If the eigenvalues of an $n \times n$ matrix A are all distinct, then there exists n eigenvectors v_j , for $j = 1, 2, \cdots, n$.*

Example 1.17

Find the eigenpairs for the matrix

$$\begin{bmatrix} 2 & -3 & 6 \\ 0 & 3 & -4 \\ 0 & 2 & -3 \end{bmatrix}.$$

Solution: *The characteristic equation* $\det(A - \lambda I) = 0$*is*

$$-\lambda^3 + 2\lambda^2 + \lambda - 2 = 0.$$

The roots of the equation are the three eigenvalues $\lambda_1 = 1, \lambda_2 = 2$, *and* $\lambda_3 = -1$. *To find the eigenvector* x_1 *corresponding to* λ_1, *we substitute* $\lambda_1 = 1$ *to Eqn. (1.15) to get the system of equations*

$$\begin{aligned} x_1 - 3x_2 + 6x_3 &= 0 \\ 2x_2 - 4x_3 &= 0 \\ 2x_2 - 4x_3 &= 0. \end{aligned}$$

Since the last two equations are identical, the system is reduced to two equations in three unknowns. Set $x_3 = \alpha$, *where* α *is an arbitrary constant, to get* $x_2 = 2\alpha$ *and* $x_1 = 0$. *Hence, by setting* $\alpha = 1$, *the first eigenpair is* $\lambda_1 = 1$ *and* $x_1 = (0, 2, 1)^T$. *To find* x_2, *substitute* $\lambda_2 = 2$ *to Eqn. (1.15) to get the system of equations*

$$\begin{aligned} -3x_2 + 6x_3 &= 0 \\ x_2 - 4x_3 &= 0 \\ 2x_2 - 5x_3 &= 0. \end{aligned}$$

The solution of this system is $x_1 = \alpha, x_2 = x_3 = 0$. *Hence, by setting* $\alpha = 1$, *the second eigenpair is* $\lambda_2 = 2$ *and* $x_2 = (1, 0, 0)^T$. *Finally, to find* x_3 *substitute* $\lambda_3 = -1$ *to Eqn. (1.15) to get the system of equations*

$$\begin{aligned} 3x_1 - 3x_2 + 6x_3 &= 0 \\ 4x_2 - 4x_3 &= 0 \\ 2x_2 - 2x_3 &= 0. \end{aligned}$$

The solution of this system is $x_1 = -\alpha, x_2 = x_3 = \alpha$. *Hence, by setting* $\alpha = 1$ *the third eigenpair is* $\lambda_3 = -1$ *and* $x_3 = (-1, 1, 1)^T$.

Example 1.18

Consider the matrix $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$. *Taking the determinant to find characteristic polynomial of* A,

$$|A - \lambda I| = \left| \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix},$$
$$= 3 - 4\lambda + \lambda^2.$$

Setting the characteristic polynomial equal to zero, it has roots at $\lambda = 1$ *and* $\lambda = 3$, *which are the two eigenvalues of* A.
For $\lambda = 1$***, the cxc equation becomes,***

$$(A - I)v_{\lambda=1} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Any non-zero vector with $v_1 = -v_2$ *solves this equation. Therefore,*

$$v_{\lambda=1} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

is an eigenvector of A *corresponding to* $\lambda = 1$, *as is any scalar multiple of this vector.*
For $\lambda = 3$***, CXC Equation becomes***

$$(A - 3I)v_{\lambda=3} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Any non-zero vector with $v_1 = v_2$ *solves this equation. Therefore,*

$$v_{\lambda=3} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

is an eigenvector of A *corresponding to* $\lambda = 3$, *as is any scalar multiple of this vector. Thus, the vectors* $v_{\lambda=1}$ *and* $v_{\lambda=3}$ *are eigenvectors of* A *associated with the eigenvalues* $\lambda = 1$ *and* $\lambda = 3$, *respectively.*

As mentioned above, the eigenvalues and eigenvectors of an $n \times n$ matrix where $n \geq 4$ must be found numerically instead of by hand. The essence of all these methods is captured in the Power method, which we now introduce.

Definition 1.5

Let $\lambda_1, \lambda_2, \lambda_3, \cdots, \lambda_n$ *be the eigenvalues of an matrix* A. λ_1 *is called the dominant eigenvalue of* A *if* $|\lambda_1| > |\lambda_i|$ *for* $i = 2,\ 3,\ 4,\ \cdots,\ n$. *The eigenvectors corresponding to* λ_1 *are called dominant eigenvectors of* A.

1.5.2 Power Method

The power method is a classical method of use mainly to determine the largest eigenvalue in magnitude, called the dominant eigenvalue, and the corresponding eigenvector of the system

$$Ax = \lambda x.$$

Theorem 1.4

If A is an diagonalizable matrix with a dominant eigenvalue, then there exists a nonzero vector x_0 such that the sequence of vectors given by

$$Ax_0,\ A^2x_0,\ A^3x_0,\ A^4x_0,\ A^5x_0,\ A^6x_0,\ \cdots,\ A^nx_0,\ \cdots$$

approaches a multiple of the dominant eigenvector of A.

Proof Because A is diagonalizable, you know that it has n linearly independent eigenvectors $x_1,\ x_2,\ ,x_3,\ \cdots,\ x_n$ with corresponding eigenvalues of $\lambda_1,\ \lambda_2,\ \lambda_3,\ \cdots,\ \lambda_n$. Assume that these eigenvalues are ordered so that λ_1 is the dominant eigenvalue (with a corresponding eigenvector of x_1). Because the n eigenvectors $x_1,\ x_2,\ ,x_3,\ \cdots,\ x_n$ are linearly independent, they must form a basis for R^n. For the initial approximation x_0 choose a nonzero vector such that the linear combination

$$x_0 = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

has nonzero leading coefficients. (If $c_1 = 0$ the power method may not converge, and a different x_0 must be used as the initial approximation.) Now, multiplying both sides of this equation by A produces

$$\begin{aligned} Ax_0 &= A(c_1x_1 + c_2x_2 + \cdots + c_nx_n) \\ Ax_0 &= c_1(Ax_1) + c_2(Ax_2) + \cdots + c_n(Ax_n) \\ Ax_0 &= c_1(\lambda_1x_1) + c_2(\lambda_2x_2) + \cdots + c_n(\lambda_nx_n) \end{aligned}$$

Repeated multiplication of both sides of this equation by A produces

$$A^kx_0 = c_1(\lambda_1^kx_1) + c_2(\lambda_2^kx_2) + \cdots + c_n(\lambda_n^kx_n)$$

which implies that

$$A^kx_0 = \lambda_1^k[c_1x_1 + c_2(\frac{\lambda_2}{\lambda_1})^kx_2 + \cdots + c_n(\frac{\lambda_n}{\lambda_1})^kx_n]$$

Now, from the original assumption that λ_1 is larger in absolute value than the other eigenvalues it follows that each of the fractions

$$\frac{\lambda_2}{\lambda_1},\ \frac{\lambda_3}{\lambda_1},\ \cdots,\ \frac{\lambda_n}{\lambda_1}$$

is less than 1 in absolute value. So each of the factors

$$(\frac{\lambda_2}{\lambda_1})^k,\ (\frac{\lambda_3}{\lambda_1})^k,\ \cdots,\ (\frac{\lambda_n}{\lambda_1})^k$$

must approach 0 as k approaches infinity. This implies that the approximation

$$A^kx_0 \approx c_1\lambda_1x_1$$

improves as k increases. Because x_1 is a dominant eigenvector, it follows that any scalar multiple of x_1 is also a dominant eigenvector, so showing that A^kx_0 approaches a multiple of the dominant eigenvector of A.

Note The power method will converge quickly if $\frac{\lambda_i}{\lambda_1},\ \ i = 2,\ 3,\ \cdots,\ n$ is small, and slowly if $\frac{\lambda_i}{\lambda_1},\ \ i = 2,\ 3,\ \cdots,\ n$ is close to 1.

Hence first assume that the matrix A has a dominant eigenvalue with corresponding dominant eigenvectors. Then choose an initial approximation x_0 of one of the dominant eigenvectors of A. This initial approximation must be a nonzero vector in R^n Finally, form the sequence given by

$$\begin{aligned} x_1 &= Ax_0 \\ x_2 &= Ax_1 = A(Ax_0) = A^2x_0 \\ x_3 &= Ax_2 = A(A^2x_0) = A^3x_0 \\ &\vdots \\ x_k &= Ax_{k-1} = A(A^{k-1}x_0) = A^kx_0 \end{aligned}$$

For large powers of k, and by properly scaling this sequence, you will see that you obtain a good approximation of the dominant eigenvector of A. This procedure is illustrated in the following Example.

Example 1.19

Approximating a Dominant Eigenvector by the Power Method Complete six iterations of the power method to approximate a dominant eigenvector of

$$\begin{bmatrix} 4 & 2 & -2 \\ -2 & 8 & 1 \\ 2 & 4 & -4 \end{bmatrix}$$

by the Power Method

Solution: *Begin with an initial nonzero approximation of*

$$x_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Then obtain the following approximations.

$$x_1 = Ax_0 = \begin{bmatrix} 4 & 2 & -2 \\ -2 & 8 & 1 \\ 2 & 4 & -4 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \\ 2 \end{bmatrix} \implies 7 \begin{bmatrix} 0.5714 \\ 1 \\ 0.2857 \end{bmatrix}$$

$$x_2 = Ax_1 = \begin{bmatrix} 4 & 2 & -2 \\ -2 & 8 & 1 \\ 2 & 4 & -4 \end{bmatrix} \begin{bmatrix} 0.5714 \\ 1 \\ 0.2857 \end{bmatrix} = \begin{bmatrix} 3.7143 \\ 7.1429 \\ 4 \end{bmatrix} \implies 7.1429 \begin{bmatrix} 0.52 \\ 1.00 \\ 0.56 \end{bmatrix}$$

$$x_3 = Ax_2 = \begin{bmatrix} 4 & 2 & -2 \\ -2 & 8 & 1 \\ 2 & 4 & -4 \end{bmatrix} \begin{bmatrix} 0.52 \\ 1.00 \\ 0.56 \end{bmatrix} = \begin{bmatrix} 2.96 \\ 7.52 \\ 2.8 \end{bmatrix} \implies 7.52 \begin{bmatrix} 0.3936 \\ 1.000 \\ 0.3723 \end{bmatrix}$$

$$x_4 = Ax_3 = \begin{bmatrix} 4 & 2 & -2 \\ -2 & 8 & 1 \\ 2 & 4 & -4 \end{bmatrix} \begin{bmatrix} 0.3936 \\ 1.000 \\ 0.3723 \end{bmatrix} = \begin{bmatrix} 2.8298 \\ 7.5851 \\ 2.2979 \end{bmatrix} \implies 7.5851 \begin{bmatrix} 0.3731 \\ 1.00 \\ 0.4348 \end{bmatrix}$$

$$x_5 = Ax_4 = \begin{bmatrix} 4 & 2 & -2 \\ -2 & 8 & 1 \\ 2 & 4 & -4 \end{bmatrix} \begin{bmatrix} 0.3731 \\ 1.00 \\ 0.4348 \end{bmatrix} = \begin{bmatrix} 2.6227 \\ 7.6886 \\ 3.0070 \end{bmatrix} \implies 7.6886 \begin{bmatrix} 0.3411 \\ 1.00 \\ 0.3911 \end{bmatrix}$$

Example

$$x_6 = Ax_5 = \begin{bmatrix} 4 & 2 & -2 \\ -2 & 8 & 1 \\ 2 & 4 & -4 \end{bmatrix} \begin{bmatrix} 0.3411 \\ 1.00 \\ 0.3911 \end{bmatrix} = \begin{bmatrix} 2.5197 \\ 7.7401 \\ 3.0760 \end{bmatrix} \implies 7.7401 \begin{bmatrix} 0.3255 \\ 1.00 \\ 0.3974 \end{bmatrix}$$

The results show that the differences between the vector $[x_i]$ and the normalized vector $[x_{i+i}]$ are getting smaller. The value of the multiplicative factor (7.7401) is an estimate of the largest eigenvalue.

Theorem 1.5

Determining an Eigenvalue from an Eigenvector If x is an eigenvector of a matrix A, then its corresponding eigenvalue is given by

$$\lambda = \frac{Ax^t * x}{x^t * x}$$

This quotient is called the Rayleigh quotient

Proof Because x is an eigenvector of A, you know that $Ax = \lambda x$ and can write

$$\frac{(Ax) * x^t}{x * x^t} = \frac{(\lambda x) * x^t}{x * x^t} = \lambda \frac{x * x^t}{x * x^t} = \lambda$$

In cases for which the power method generates a good approximation of a dominant eigenvector, the Rayleigh quotient provides a correspondingly good approximation of the dominant eigenvalue

Example 1.20

Consider the eigenvalue problem

$$\begin{bmatrix} -9 & 14 & 4 \\ -7 & 12 & 4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

where the eigenvalues of A are $\lambda_1 = 5, \lambda_2 = 1$, and $\lambda_3 = -2$.
Solution: *As a first guess, we choose $x_0 = (1, 1, 1)^T$. Now compute*

$$\begin{bmatrix} -9 & 14 & 4 \\ -7 & 12 & 4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 9 \\ 9 \\ 1 \end{bmatrix} = 9 \begin{bmatrix} 1 \\ 1 \\ 1/9 \end{bmatrix} = \lambda_1 x_1.$$

We have normalized the vector by dividing through by its largest element. The next iteration yields

$$\begin{bmatrix} -9 & 14 & 4 \\ -7 & 12 & 4 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1/9 \end{bmatrix} == 49/9 \begin{bmatrix} 1 \\ 1 \\ 1/49 \end{bmatrix} = \lambda_2 x_2.$$

After 10 iterations, the sequence of vectors converges to

$$x = [1, 1, 1.02 \times 10^{-8}]^T,$$

and the sequence λ_k of constants converges to $\lambda = 5$.

Example 1.21

The Power Method with Scaling Calculate seven iterations of the power method with scaling to approximate a dominant eigenvector of the matrix

$$\begin{bmatrix} 1 & 2 & 0 \\ -2 & 1 & 2 \\ 1 & 3 & 1 \end{bmatrix}$$

Use $x_0 = (1, 1, 1)^t$ *as the initial approximation.*
***Solution**:*
One iteration of the power method produces

$$Ax_0 = \begin{bmatrix} 1 & 2 & 0 \\ -2 & 1 & 2 \\ 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix} = 5 \begin{bmatrix} 0.60 \\ 0.20 \\ 1.00 \end{bmatrix}$$

and by scaling you obtain the approximation

$$x_1 = \frac{1}{5} \begin{bmatrix} 3 \\ 1 \\ 5 \end{bmatrix} = \begin{bmatrix} 0.60 \\ 0.20 \\ 1.00 \end{bmatrix}$$

A second iteration yields

$$Ax_1 = \begin{bmatrix} 1 & 2 & 0 \\ -2 & 1 & 2 \\ 1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 0.60 \\ 0.20 \\ 1.00 \end{bmatrix} = \begin{bmatrix} 1.00 \\ 1.00 \\ 2.20 \end{bmatrix} = 2.2 \begin{bmatrix} 0.45 \\ 0.45 \\ 1.00 \end{bmatrix}$$

and

$$x_2 = \frac{1}{2.2} \begin{bmatrix} 0.45 \\ 0.45 \\ 1.00 \end{bmatrix} = \begin{bmatrix} 0.45 \\ 0.45 \\ 1.00 \end{bmatrix}$$

Continuing this process, you obtain the sequence of approximations shown in the following Table

x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7
$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.6 \\ 0.2 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.45 \\ 0.45 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.48 \\ 0.55 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.51 \\ 0.51 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.50 \\ 0.49 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.50 \\ 0.50 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 0.50 \\ 0.50 \\ 1 \end{bmatrix}$

From the Table above you can approximate a dominant eigenvector of A *to be* $\begin{bmatrix} 0.50 \\ 0.50 \\ 1 \end{bmatrix}$
Using the Rayleigh quotient, you can approximate the dominant eigenvalue of A *to be* $\lambda = 3$ *(For this example you can check that the approximations of* x *and* λ *are exact.)*

x_1	x_2	x_3	x_4	x_5	x_6	x_7
$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$	$\Downarrow$
5.00	2.20	2.82	3.13	3.02	2.99	3.00

are approaching the dominant eigenvalue $\lambda = 3$

1.5.3 Inverse Power Method

Inverse power method can give approximation to any eigenvalue. However, it is used usually to find the smallest eigenvalue in magnitude and the corresponding eigenvector of a given matrix A. The eigenvectors are computed very accurately by this method. Further, the method is powerful to calculate accurately the eigenvectors, when the eigenvalues are not well separated. In this case, power method converges very slowly.

If λ is an eigenvalue of A, then $\frac{1}{\lambda}$ is an eigenvalue of A^{-1} corresponding to the same eigenvector. The smallest eigenvalue λ in magnitude of A is the largest eigenvalue $\frac{1}{\lambda}$ in magnitude of A^{-1}. Then choose an initial approximation x_0 of one of the dominant eigenvectors of A^{-1}. This initial approximation must be a nonzero vector in R^n Finally, Applying the power method on A^{-1}, we have

$$\begin{aligned}
x_1 &= A^{-1}x_0 \\
x_2 &= A^{-1}x_1 = A^{-1}(A^{-1}x_0) = (A^{-1})^2x_0 \\
x_3 &= A^{-1}x_2 = A^{-1}((A^{-1})^2x_0) = (A^{-1})^3x_0 \\
&\vdots \\
x_k &= A^{-1}x_{k-1} = A^{-1}((A^{-1})^{k-1}x_0) = (A^{-1})^kx_0
\end{aligned}$$

For large powers of k, and by properly scaling this sequence, you will see that you obtain a good approximation of the dominant eigenvector of A. This procedure is illustrated in the following. Then using Rayleigh quotient we can fined the dominant eigenvalue of A^{-1}

$$\frac{1}{\lambda} = \frac{A^{-1}x * x^t}{x * x^t}$$

Example 1.22

Find the smallest eigenvalue in magnitude of the matrix

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

use four iteration of the inverse power method.
***Solution**:*
The smallest eigenvalue in magnitude of A is the largest eigenvalue in magnitude of A^{-1}.
We have

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

Then use $x_0 = (1,\ 1,\ 1)^t$ and apply inverse power method with scaling.
First approximation

$$A^{-1}x_0 = A^{-1} = \frac{1}{4}\begin{bmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 2 \\ 1.5 \end{bmatrix} \implies x_1 = \begin{bmatrix} 1 \\ 1.333 \\ 1 \end{bmatrix}$$

Example: Solution

Second Approximation

$$A^{-1}x_1 = \frac{1}{4}\begin{bmatrix}3 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 3\end{bmatrix}\begin{bmatrix}1\\ 1.333\\ 1\end{bmatrix} = \begin{bmatrix}1.6667\\ 2.3333\\ 1.6667\end{bmatrix} \implies x_2 = \begin{bmatrix}1.0000\\ 1.4000\\ 1.0000\end{bmatrix}.$$

Third approximation

$$A^{-1}x_2 = \frac{1}{4}\begin{bmatrix}3 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 3\end{bmatrix}\begin{bmatrix}1.0000\\ 1.4000\\ 1.0000\end{bmatrix} = \begin{bmatrix}1.7000\\ 2.4000\\ 1.7000\end{bmatrix} \implies x_3 = \begin{bmatrix}1.0000\\ 1.4118\\ 1.0000\end{bmatrix}.$$

Fourth approximation

$$A^{-1}x_3 = \frac{1}{4}\begin{bmatrix}3 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 3\end{bmatrix}\begin{bmatrix}1.0000\\ 1.4118\\ 1.0000\end{bmatrix} = \begin{bmatrix}1.7059\\ 2.4118\\ 1.7059\end{bmatrix} \implies x_4 = \begin{bmatrix}1.0000\\ 1.4138\\ 1.0000\end{bmatrix}.$$

From the above we can approximate a dominant eigenvector of A^{-1} *to be* $\begin{bmatrix}1.0000\\ 1.4138\\ 1.0000\end{bmatrix}$. *After four iteration using the Rayleigh quotient, you can approximate the dominant eigenvalue of* A^{-1} *is*

$$\frac{1}{\lambda} = \frac{\frac{1}{4}\left(\begin{bmatrix}3 & 2 & 1\\ 2 & 4 & 2\\ 1 & 2 & 3\end{bmatrix}\begin{bmatrix}1.0000\\ 1.4138\\ 1.0000\end{bmatrix}\right)^t\begin{bmatrix}1.0000\\ 1.4138\\ 1.0000\end{bmatrix}}{\left(\begin{bmatrix}1.0000 & 1.4138 & 1.0000\end{bmatrix}\right)\begin{bmatrix}1.0000\\ 1.4138\\ 1.0000\end{bmatrix}} = 1.7071.$$

Therefore $\lambda = 0.5858$ *is required eigenvalue. The corresponding eigenvector is* $\begin{bmatrix}1.0000 & 1.4138 & 1.0000\end{bmatrix}^t$.
The smallest eigenvalue of A is $2 - \sqrt{2} = 0.5858$.

1.6 System of Non-linear Equations

Recall that at the end of Chap. 2 we presented an approach to solve two nonlinear equations with one unknowns. This approach can be extended to the general case of solving n simultaneous nonlinear equations.

$$\begin{aligned} f_1(x_1, x_2, \cdots, x_n) &= 0\\ f_2(x_1, x_2, \cdots, x_n) &= 0\\ f_3(x_1, x_2, \cdots, x_n) &= 0\\ &\vdots\\ f_n(x_1, x_2, \cdots, x_n) &= 0 \end{aligned}$$

The solution of this system consists of the set of x values that simultaneously result in all the equations equaling zero.

1.6.1 Newton Raphson method

Newton's Method for Solving Systems of Two Nonlinear Equations

Recall from the **Newton's Method for Approximating Roots** section that if we have a function $y = f(x)$ and α is a root of this function, then if we have an initial approximation x_0 of this root, then we can define the tangent line of the point $(x_0, f(x_0))$ as:

$$p_1(x) = f(x_0) + f'(x_0)(x - x_0) \tag{1.17}$$

We then take the root of this line which we denote as $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$. Under ideal circumstances, this value of x_1 will be a better approximation of the root x_0. We then repeat the process to obtain a sequence of approximations $\{x_0, x_1, \cdots, x_n, \cdots\}$ that once again, under ideal circumstances, will converge to the root α. The general formula for the x-intercept approximations is:

$$x_{n+1} = f(x_n) - \frac{f(x_n)}{f'(x_n)} \tag{1.18}$$

We will now look at a slightly modified form of Newton's Method in approximating the solutions to a system of two nonlinear equations with two unknowns. Consider the following system of two nonlinear equations of the two variables x and y:

$$\begin{cases} f(x, y) = 0 \\ g(x, y) = 0 \end{cases} \tag{1.19}$$

Now suppose that a solution (α, β) to this system exists and that (x_0, y_0) is an initial approximation to this solution. Now note that $z = f(x, y)$ and $z = g(x, y)$ will represent surfaces in $\mathbb{R}^3$. The best approximation of these surfaces at the point $(x_0, y_0, f(x_0, y_0))$ will be the tangent plane that passes through this point. The general equation for this tangent plane is given by:

$$p_1(x, y) = f(x_0, y_0) + (x - x_0)\frac{\partial}{\partial x} f(x_0, y_0) + (y - y_0)\frac{\partial}{\partial y} f(x_0, y_0) \tag{1.20}$$

Provided that $f(x_0, y_0)$ is close enough to 0 (such as to be close enough to satisfy f(x,y)=0) then the level curve,$p_1(x, y) = 0$ which actually represents a straight line in $\mathbb{R}^2$ can be used to approximate the level curve $f(x, y) = 0$ for points (x, y) that are near (x_0, y_0).
Furthermore, we can apply the same procedure to the surface $z = g(x, y)$. The best approximation to this surface at the point $(x_0, y_0, g(x_0, y_0))$ is the tangent plane that passes through this point and given by the following equation:

$$q_1(x, y) = g(x_0, y_0) + (x - x_0)\frac{\partial}{\partial x} g(x_0, y_0) + (y - y_0)\frac{\partial}{\partial y} g(x_0, y_0) \tag{1.21}$$

Provided that $g(x_0, y_0)$ is close enough to 0 then the level curve $q_1(x, y) = 0$ (which represents a straight line in $\mathbb{R}^2$) can be used to approximate the level curve $g(x, y) = 0$ for points (x, y) near (x_0, y_0).
Now we can approximate the solution (α, β) of interest between the curves $f(x, y) = 0$ and $g(x, y) = 0$ with the solution between the lines $p_1(x, y) = 0$ and $q_1(x, y) = 0$. Let (x_1, y_1) be the solution to the now linear system, $\begin{cases} p_1(x, y) = 0 \\ q_1(x, y) = 0 \end{cases}$. Then (x_1, y_1) will hopefully be a better approximation to the solution (α, β) of the nonlinear system from earlier.
Now to find the intersection of $p_1(x, y) = 0$ and $q_1(x, y) = 0$ is simple. We only need to solve the following system of equations:

$$f(x_0,y_0)+(x-x_0)\frac{\partial}{\partial x}f(x_0,y_0)+(y-y_0)\frac{\partial}{\partial y}f(x_0,y_0)=0$$
$$g(x_0,y_0)+(x-x_0)\frac{\partial}{\partial x}g(x_0,y_0)+(y-y_0)\frac{\partial}{\partial y}g(x_0,y_0)=0$$

This system can be more nicely compressed using matrices. If we let $x-x_0=\delta_x$ and $y-y_0=\delta_y$ then:

$$\begin{bmatrix}\frac{\partial}{\partial x}f(x_n,y_n) & \frac{\partial}{\partial y}f(x_n,y_n)\\ \frac{\partial}{\partial x}g(x_n,y_n) & \frac{\partial}{\partial x}g(x_n,y_n)\end{bmatrix}\begin{bmatrix}\delta_{x,n}\\ \delta_{y,n}\end{bmatrix}=-\begin{bmatrix}f(x_n,y_n)\\ g(x_n,y_n)\end{bmatrix}$$

The (hopefully) better approximation to the solution (α,β) will be (x_1,y_1) where $x_1=x_0+\delta_x$ and $y_1=y_0+\delta_y$. We can then repeat this process in hopes that the sequence of approximations $(x_0,y_0),(x_1,y_1),\cdots,(x_n,y_n),\cdots$ converges to the solution (α,β). More generally, each iteration of this algorithm for $n=0,1,2,\cdots$ can be computed in matrix form as:

$$\begin{bmatrix}\frac{\partial}{\partial x}f(x_n,y_n) & \frac{\partial}{\partial y}f(x_n,y_n)\\ \frac{\partial}{\partial x}g(x_n,y_n) & \frac{\partial}{\partial x}g(x_n,y_n)\end{bmatrix}\begin{bmatrix}\delta_{x,n}\\ \delta_{y,n}\end{bmatrix}=-\begin{bmatrix}f(x_n,y_n)\\ g(x_n,y_n)\end{bmatrix}$$

And each successive approximation is given by $x_{n+1}=x_n+\delta_{x,n}$ and $y_{n+1}=y_n+\delta_{y,n}$.

Example 1.23

Consider the following non-linear system of equations $\begin{cases}x^3+y=1\\ y^3-x=-1\end{cases}$. *There exists a solution* (α,β) *such that* $\alpha,\beta>0$. *Let* $(0.9,0.9)$ *be an initial approximation to this system. Use Newton's method with three iterations to approximate this solution.*
It's not hard to see that the solution of interest is $(\alpha,\beta)=(1,1)$ *which can be obtained by substituting one of the equations into the other. Regardless, we will still use Newton's method to demonstrate the algorithm.*
We first rewrite our system of equations as:

$$\begin{cases}f(x,y)=x^3+y-1=0\\ g(x,y)=y^3-x+1=0\end{cases}$$

We now compute the partial derivatives of f *and* g. *We have that:*

$$\begin{bmatrix}\frac{\partial f}{\partial x} & \frac{\partial f}{\partial y}\\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y}\end{bmatrix}=\begin{bmatrix}3x^2 & 1\\ -1 & 3y^2\end{bmatrix}$$

We will use the matrix above for each iteration. For the first iteration, we need to solve the following system of equations:

$$\begin{bmatrix}3(0.9)^2 & 1\\ -1 & 3(0.9)^2\end{bmatrix}\begin{bmatrix}\delta_{x,0}\\ \delta_{y,0}\end{bmatrix}=\begin{bmatrix}f(0.9,0.9)\\ g(0.9,0.9)\end{bmatrix}$$
$$\begin{bmatrix}2.43 & 1\\ -1 & 2.43\end{bmatrix}\begin{bmatrix}\delta_{x,0}\\ \delta_{y,0}\end{bmatrix}=\begin{bmatrix}0.629\\ 0.829\end{bmatrix}$$

In solving this system, we get that $\delta_{x,1}=0.101$ *and that* $\delta_{y,1}=0.383$. *Therefore we have that:*

$$x_1=x_0+\delta_{x,0}=0.9+0.101=1.001\quad y_1=y_0+\delta_{y,0}=0.9+0.383=1.283$$

Example

For the second iteration, we want to solve the following system of equations:

$$\begin{bmatrix} 3(1.001)^2 & 1 \\ -1 & 3(1.283)^2 \end{bmatrix} \begin{bmatrix} \delta_{x,1} \\ \delta_{y,1} \end{bmatrix} = \begin{bmatrix} f(1.001, 1.283) \\ g(1.001, 1.283) \end{bmatrix}$$
$$\begin{bmatrix} 3.006003 & 1 \\ -1 & 4.938267 \end{bmatrix} \begin{bmatrix} \delta_{x,1} \\ \delta_{y,1} \end{bmatrix} = \begin{bmatrix} 1.286003 \\ 2.110932 \end{bmatrix}$$

When we solve this system, we get that $\delta_{x,2} = 0.201143$ and $\delta_{y,2} = 0.400208$. Therefore we have that:

$$x_2 = x_1 + \delta_{x,1} = 1.001 + 0.201143 = 1.202173$$
$$y_2 = y_1 + \delta_{y,1} = 1.283 + 0.400208 = 1.683208$$

Newton's Method for Solving Systems of Many Nonlinear Equations

One approach to solving such systems is based on a multidimensional version of the Newton-Raphson method. Thus, a Taylor series expansion is written for each equation about the point $(x_1^k,\ x_2^k,\ \cdots,\ x_n^k)$ we get,

$$\begin{aligned}
&f_1(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) = f_1(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) \\
&\quad + [\triangle x_1\frac{\partial}{\partial x_1} + \triangle x_2\frac{\partial}{\partial x_2} + \cdots + \triangle x_n\frac{\partial}{\partial x_n}]f_1(x_1^k,\ x_2^k,\ \cdots,\ x_n^k) + \\
&\quad \frac{1}{2!}[\triangle x_1\frac{\partial}{\partial x_1} + \triangle x_2\frac{\partial}{\partial x_2} + \cdots + \triangle x_n\frac{\partial}{\partial x_n}]^2 f_1(x_1^k,\ x_2^k,\ \cdots,\ x_n^k) + \cdots = 0 \\
&f_2(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) = f_2(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) \\
&\quad + [\triangle x_1\frac{\partial}{\partial x_1} + \triangle x_2\frac{\partial}{\partial x_2} + \cdots + \triangle x_n\frac{\partial}{\partial x_n}]f_2(x_1^k,\ x_2^k,\ \cdots,\ x_n^k) + \\
&\quad \frac{1}{2!}[\triangle x_1\frac{\partial}{\partial x_1} + \triangle x_2\frac{\partial}{\partial x_2} + \cdots + \triangle x_n\frac{\partial}{\partial x_n}]^2 f_2(x_1^k,\ x_2^k,\ \cdots,\ x_n^k) + \cdots = 0 \\
&\qquad\qquad \cdots \\
&\qquad\qquad \cdots \\
&f_n(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) = f_n(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) \\
&\quad + [\triangle x_1\frac{\partial}{\partial x_1} + \triangle x_2\frac{\partial}{\partial x_2} + \cdots + \triangle x_n\frac{\partial}{\partial x_n}]f_n(x_1^k,\ x_2^k,\ \cdots,\ x_n^k) + \\
&\quad \frac{1}{2!}[\triangle x_1\frac{\partial}{\partial x_1} + \triangle x_2\frac{\partial}{\partial x_2} + \cdots + \triangle x_n\frac{\partial}{\partial x_n}]^2 f_n(x_1^k,\ x_2^k,\ \cdots,\ x_n^k) + \cdots = 0
\end{aligned} \tag{1.22}$$

Neglecting 2^{nd} and higher powers of $\triangle x_1,\ \triangle x_2,\ \cdots$ and $\triangle x_n$, we obtain

$$\begin{aligned}
&f_1(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) = f_1(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) \\
&\quad + [\triangle x_1\frac{\partial}{\partial x_1} + \triangle x_2\frac{\partial}{\partial x_2} + \cdots + \triangle x_n\frac{\partial}{\partial x_n}]f_1(x_1^k,\ x_2^k,\ \cdots,\ x_n^k) \\
&f_2(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) = f_2(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) \\
&\quad + [\triangle x_1\frac{\partial}{\partial x_1} + \triangle x_2\frac{\partial}{\partial x_2} + \cdots + \triangle x_n\frac{\partial}{\partial x_n}]f_2(x_1^k,\ x_2^k,\ \cdots,\ x_n^k) \\
&\qquad\qquad \cdots \\
&\qquad\qquad \cdots \\
&f_n(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) = f_n(x_1^k+\triangle x_1,\ x_2^k+\triangle x_2,\ \cdots,\ x_n^k+\triangle x_n) \\
&\quad + [\triangle x_1\frac{\partial}{\partial x_1} + \triangle x_2\frac{\partial}{\partial x_2} + \cdots + \triangle x_n\frac{\partial}{\partial x_n}]f_n(x_1^k,\ x_2^k,\ \cdots,\ x_n^k)
\end{aligned} \tag{1.23}$$

Since $x_1^{k+1} = x_1^k + \triangle x_1,\ x_2^{k+1} = x_2^k + \triangle x_2,\ \cdots$ and $x_n^{k+1} = x_n^k + \triangle x_n$ writing the equation in matrix form, we get

$$J_k \triangle X^k = -F(X^k) \tag{1.24}$$

where $J_k = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ & \cdots & & \\ & \cdots & & \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}$ $at\ \ (x_1^k,\ x_2^k,\ \cdots,\ x_n^k)$ $, \triangle X^k = \begin{bmatrix} x_1^{k+1} - x_1^k \\ x_2^{k+1} - x_2^k \\ \cdot \\ \cdot \\ x_n^{k+1} - x_n^k \end{bmatrix}$ and

$F(X^k) = \begin{bmatrix} f_1(x_1^k,\ x_2^k,\ x_3^k, \cdots,\ x_n^k) \\ f_2(x_1^k,\ x_2^k,\ x_3^k, \cdots,\ x_n^k) \\ \cdot \\ \cdot \\ f_n(x_1^k,\ x_2^k,\ x_3^k, \cdots,\ x_n^k) \end{bmatrix}$ Therefore equation can be written as $\triangle X^k = -J_k^{-1} F(X^k)$

$$\left[X^{k+1}\right] = \left[X^k\right] - J_k^{-1} F(X^k) \qquad , k = 0,\ 1,\ 2,\ 3,\ \cdots \tag{1.25}$$

The convergence of the method depends on the initial approximation X_0. A sufficient condition for convergence is that for each k

$$\|J_k^{-1}\| < 1.$$

whereas a necessary and sufficient condition for convergence is

$$\rho(J_k^{-1}) < 1$$

Where $\|.\|$ is suitable norm and $\rho(J_k^{-1})$ is the spectral radius (large eigenvalue in magnitude) of the matrix J_k^{-1}
if the method converges, then its rate of convergence is two. The iterations stopped when

$$\|X^{k+1} - X^k\| < \epsilon$$

Where ϵ is the given error tolerance.

Example 1.24

perform three iterations of the Newton-Raphson Method to solve the system of equations

$$x^2 + xy + y^2 = 7$$
$$x^3 + y^3 = 9$$

Take initial approximation as $x_0 = 1.5$ *and* $y_0 = 0.5$. *The exact solution is* $x = 2,\ y = 1$
Solution: *We have*

$$f(x) = x^2 + xy + y^2 - 7 = 0$$
$$g(x) = x^3 + y^3 - 9 = 0$$

$$J_k = \begin{bmatrix} f_x(x_k, y_k) & f_y(x_k, y_k) \\ g_x(x_k, y_k) & g_y(x_k, y_k) \end{bmatrix} = \begin{bmatrix} 2x_k + y_k & x_k + 2y_k \\ 3x_k^2 & 3y_k^2 \end{bmatrix}$$

$$J_k^{-1} = \frac{1}{D_k} \begin{bmatrix} 3y_k^2 & -(x_k + 2y_k) \\ -3x_k^2 & 2x_k + y_k \end{bmatrix}$$

Where $D_k = |J_k| = 3y_k^2(2x_k + y_k) - 3x_k^2(x_k + 2y_k)$. *Know we can write the method as*

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - \frac{1}{D_k} \begin{bmatrix} 3y_k^2 & -(x_k + 2y_k) \\ -3x_k^2 & 2x_k + y_k \end{bmatrix} \begin{bmatrix} x_k^2 + x_k y_k + y_k^2 - 7 \\ x_k^3 + y_k^3 - 9 \end{bmatrix} \qquad k = 0,\ 1,\ 2,\ 3,\ \cdots$$

Using $(x_0, y_0) = (1.5, 0.5)$, *we get*

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 0.5 \end{bmatrix} - \frac{1}{-14.25} \begin{bmatrix} 0.75 & -2.5 \\ -6.75 & 3.5 \end{bmatrix} \begin{bmatrix} -3.75 \\ -5.5 \end{bmatrix} = \begin{bmatrix} 2.2675 \\ 0.9254 \end{bmatrix}$$

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2.2675 \\ 0.9254 \end{bmatrix} - \frac{1}{-49.4951} \begin{bmatrix} 2.5691 & -4.1183 \\ -15.4247 & 5.4604 \end{bmatrix} \begin{bmatrix} 1.0963 \\ 3.4510 \end{bmatrix} = \begin{bmatrix} 2.0373 \\ 0.9645 \end{bmatrix}$$

$$\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} 2.0373 \\ 0.9645 \end{bmatrix} - \frac{1}{-35.3244} \begin{bmatrix} 2.7908 & -3.9663 \\ -12.4518 & 5.0391 \end{bmatrix} \begin{bmatrix} 0.0458 \\ 0.3532 \end{bmatrix} = \begin{bmatrix} 2.0013 \\ 0.9987 \end{bmatrix}$$

Exercise 1.7

Multivariate Newton Examples

$$x_1^2 + 2x_2^2 - x_2 - 2x_3 = 0$$
$$x_1^2 - 8x_2^2 + 10x_3 = 0$$
$$x_1^2 - 7x_2x_3 = 0$$

Exercise 1.8

Three intersecting radius-1 spheres: 

$$(x_1 - 1)^2 + (x_2 - 1)^2 + x_3 = 1$$
$$(x_1 - 1)^2 + x_2 + (x_3 - 1)^2 = 1$$
$$x_1 + (x_2 - 1)^2 + (x_3 - 1)^2 = 1$$