Proceedings of CONF-MPCS 2025 Symposium: Mastering Optimization: Strategies for Maximum Efficiency
DOI: 10.54254/2753-8818/2024.CH21606

Convergence Analysis and Improvement of Iterative Solutions for Linear Equation Systems

Yang Liu

Xiangtan University, Xiangtan, China
13330689940@163.com

Abstract: Many physics and mathematics problems generate linear equation systems, and solving linear equation systems has become an important proposition. Therefore, combined with the characteristics of modern computers, various methods for solutions need to be sought. This article introduces applications of linear equation systems in different fields, as well as representative figures and works with outstanding achievements. It focuses on providing formulas and corresponding Matlab codes for Gauss elimination, Jacobi iteration, and G-S iteration to solve equation systems, and rigorously proves the sufficient and necessary condition for convergence of iterative formula and also proves the convergence of different iteration methods under different types of coefficient matrices. Based on these solving methods, two examples are practiced in Matlab, the running time and iteration times of different methods are comprehensively compared. Thus, the superiority of the G-S iterative method is obtained. Finally, when there are zero elements in the diagonal elements of the coefficient matrix, the article proposes an improved method to solve this problem.

Keywords: system of linear equations, iterative methods, convergence

1. Introduction

The study of linear equation systems can be traced back to ancient China. The earliest mathematical work in China, the Book of Arithmetic, included some early problems with linear equation systems. The Nine Chapters on Arithmetic, discussed the solutions to linear equation systems in the “Equations” chapter.

In the late summer of 1949, Harvard University professor Wassily Leontief decomposed the American economy into 500 sectors, such as the coal industry, automobile industry, transportation system, and so on. For each department, he wrote a linear equation describing how the output of that department is allocated to other economic sectors. Leontief simplified the problem into a system of 42 equations containing 42 unknowns. To solve Leontief’s 42 equations, the Mark II computer took 56 hours of computation to obtain the final answer. Leontief was awarded the Nobel Prize in Economics in 1973 and opened the door to a new era of economic mathematical modeling. The work at Harvard in 1949 marked the beginning applying computers to analyse large-scale mathematical models, and since then, many researchers in other fields have applied computers to analyze mathematical models. Due to the large amount of data involved, these models are usually linear, that is, they are described by a system of linear equations.

In addition, linear equation systems have wide applications in different fields and aspects, and many mathematical and physical models ultimately boil down to solving linear equation systems.

DOI: 10.54254/2753-8818/2024.CH21606

Consider n-order linear differential equation problems

$$x^{(n)} + a_1(t)x^{(n-1)} + \cdots + a_{n-1}(t)\dot{x} + a_n(t)x = f(t) \tag{1}$$

where the coefficient polynomials and right-hand polynomial are continuous functions.

Set

$$x_1 = x, x_2 = \dot{x}, x_3 = \ddot{x}, \cdots, x_n = x^{(n-1)} \tag{2}$$

there is

$$\dot{x_1} = x_2, \dot{x_2} = x_3, \cdots, \dot{x_{n-1}} = x_n \tag{3}$$

$$\dot{x_n} = -a_n(t)x_1 - a_{n-1}(t)x_2 - \cdots a_1(t)x_n + f(t) \tag{4}$$

set up

$$\tilde{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \dot{\tilde{x}} = \begin{pmatrix} \dot{x_1} \\ \dot{x_2} \\ \vdots \\ \dot{x_n} \end{pmatrix} \tag{5}$$

then

$$\dot{\tilde{x}} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n(t) & -a_{n-1}(t) & -a_{n-2}(t) & \cdots & -a_1(t) \end{pmatrix} \tilde{x} + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ f(t) \end{pmatrix} \tag{6}$$

This becomes a problem of linear differential equation systems, then we can explore the general theory of linear differential equation systems.

When scientists and engineers study the flow in some networks, they will derive a system of linear equations. The problem of network analysis is to determine the traffic of each branch when local information (such as the input and output of the network) is known. The basic assumption of network flow is that the total inflow of the network is equal to the total outflow. Because traffic is conserved in each node, we have similarly, the traffic of each node can be described by an equation, and multiple nodes can be represented by a system of linear equations.

2. Three common methods

Studying how to solve linear equation systems is an important issue, and we have the following three common methods for solving the system of linear equations.

2.1. Gauss elimination method

Set the initial linear equation system

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n = b_n \end{cases} \Leftrightarrow \begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{21}^{(1)}x_1 + a_{22}^{(1)}x_2 + \cdots + a_{2n}^{(1)}x_n = b_2^{(1)} \\ \vdots \\ a_{n1}^{(1)}x_1 + a_{n2}^{(1)}x_2 + \cdots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases} \tag{7}$$

the coefficient matrix is A. This article only discusses the reversible case of A.

DOI: 10.54254/2753-8818/2024.CH21606

If $a_{ll}^{(l)} \neq 0$,let $l_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, i = 2,3 \cdots n$,subtract the i-th equation from the first equation and multiply it by $l_{i1}(i = 2,3 \cdots n)$ to obtain the following system of equations

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \vdots \\ a_{n2}^{(2)}x_2 + \cdots + a_{nn}^{(2)}x_n = b_n^{(2)} \end{cases} \tag{8}$$

where

$$a_{ij}^{(2)} = a_{ij}^{(1)} - l_{i1}a_{1j}^{(1)}, b_2^{(2)} = b_i^{(1)} - l_{i1}b_1^{(1)} (i, j = 2,3 \cdots n) \tag{9}$$

If $a_{22}^{(2)} \neq 0$,let $l_{i2} = \frac{a_{i1}^{(2)}}{a_{22}^{(2)}}, i = 3,4 \cdots n$,Subtract the i-th equation from the second equation and multiply it by $l_{i2}, i = 3,4 \cdots n$.Generally, at step k, there is a system of equations.

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \vdots \\ a_{k2}^{(k)}x_k + \cdots + a_{kn}^{(k)}x_n = b_k^{(k)} \\ \vdots \\ a_{nk}^{(k)}x_k + \cdots + a_{nn}^{(k)}x_n = b_n^{(k)} \end{cases} \tag{10}$$

where

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - l_{i,k-1}a_{k-1,j}^{(k-1)}, b_i^{(k)} = b_i^{(k-1)} - l_{i,k-1}b_{k-1}^{(k-1)}, l_{i,k-1} = \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} (i, j = k, k+1 \cdots n) \tag{11}$$

If the value of each step $a_{ii}^{(i)} \neq 0, i = 1,2 \cdots n$,the system of equations can ultimately be

$$\begin{cases} a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \cdots + a_{1n}^{(1)}x_n = b_1^{(1)} \\ a_{22}^{(2)}x_2 + \cdots + a_{2n}^{(2)}x_n = b_2^{(2)} \\ \vdots \\ a_{nn}^{(n)}x_n = b_n^{(n)} \end{cases} \tag{12}$$

We can use back substitution

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}, x_k = (b_k^{(k)} - \sum_{j=k+1}^{n} a_{kj}^{(k)} x_j)/a_{kk}^{(k)}, k = 1,2 \cdots n-2, n-1. \tag{13}$$

As discussed above, this method requires $a_{ii}^{(i)} \neq 0$ at each step.It can be proven that the necessary and sufficient condition for $a_{ii}^{(i)} \neq 0$ is that each order principal minor determinant of the coefficient matrix A is all non-zero.

The Gauss elimination method in MATLAB algorithm is simple as follows:

```
for k=1:(n-1)
   m=A(k+1:n,k)/A(k,k);
   A(k+1:n,k+1:n)=A(k+1:n,k+1:n)-m*A(k,k+1:n);
   b(k+1:n)=b(k+1:n)-m*b(k);
```

DOI: 10.54254/2753-8818/2024.CH21606

```
    A(k+1:n,k)=zeros(n-k,1);
end
x=zeros(n,1);
x(n)=b(n)/A(n,n);
for k=n-1:-1:1
    x(k)=(b(k)-A(k,k+1:n)*x(k+1:n))/A(k,k);
end
```

In addition, there are also the following iterative methods. Consider a system of linear equations $Ax = b$,where $A = D - L - U$ and D is a matrix composed of the main diagonal elements of A,-L is the strict lower triangular matrix of A (with diagonal elements of 0),-U is the strict upper triangular matrix of A.The following only considers the case where D is an invertible matrix.

2.2. Jacobi iterative method

The following only considers the case where D is an invertible matrix.

$$Ax = b \Leftrightarrow (D - L - U)x = b \Leftrightarrow Dx = (L + U)x + b \Leftrightarrow x = D^{-1}(L + U)x + D^{-1}b \quad (14)$$

thus the Jacobi iteration method is

$$x^{(k+1)} = D^{-1}(L + U)x^{(k)} + D^{-1}b \quad (15)$$

This shows the reason why D needs to be reversible.

The Jacobi iterative method and Matlab algorithm are as follows:

```
ep=1e-6;
D=diag(diag(A));
L=-tril(A,-1);
U=-triu(A,1);
x=D\(L+U)*x0+D\b;
while norm(x-x0)>=ep
    x0=x;
    x=D\(L+U)*x0+D\b;
end
```

2.3. Gauss-Seidel iteration method (G-S iteration)

Similar to the Jacobi iteration method, except that

$$x^{(k+1)} = D^{-1}(L + U)x^{(k)} + D^{-1}b \quad (16)$$

this step is transformed to

$$x^{(k+1)} = D^{-1}(Lx^{(k+1)} + Ux^{(k)}) + D^{-1}b \quad (17)$$

and simply it to

$$x^{(k+1)} = (D - L)^{-1}Ux^{(k)} + (D - L)^{-1}b \quad (18)$$

Here appears $(D - L)^{-1}$,a lower triangular in common with the main diagonal elements of D. It can be inferred that this only requires D to be reversible to ensure reversibility.

The Gauss Seidel iterative method and Matlab algorithm are as follows:

```
ep=1e-6;
```

DOI: 10.54254/2753-8818/2024.CH21606

```
D=diag(diag(A));
L=-tril(A,-1);
U=-triu(A,1);
x=(D-L)\U*x0+(D-L)\b;
while norm(x-x0)>=ep
   x0=x;
   x=(D-L)\U*x0+(D-L)\b;
end
```

3. Convergence

In 2.2 and 2.3, it is natural to ask when the iterative vector sequence converges to the exact solution of the system of equations, thus leading to the following convergence discussion.

3.1. Convergence of iterative method

Consider transforming $Ax = b$into$x = Gx + d$(simply split A and perform a simple transformation), which means the iterative formula is $x^{(k+1)} = Gx^{(k)} + d$,where G is called the iterative matrix,and the spectral radius is set to $\rho(G)$.

Theorem 3.1.1: *The sufficient and necessary condition for iterative formula* $x^{(k+1)} = Gx^{(k)} + d$ *converge is that* $\rho(G)<1$.

Proof: Let

$$\varepsilon^{(k)} = x^{(k)} - x^{*} \tag{19}$$

x^{*} be the solution of

$$Ax = b \tag{20}$$

so as to

$$x^{(k+1)} = Gx^{(k)} + d \text{ converge} \Leftrightarrow \varepsilon^{(k)} \to 0, k \to \infty \tag{21}$$

We can know that

$$\varepsilon^{(k+1)} = G\varepsilon^{(k)} = G^{k}\varepsilon^{(1)} \tag{22}$$

thus

$$\varepsilon^{(k)} \to 0, k \to \infty \Leftrightarrow G^{k} \to 0, k \to \infty \tag{23}$$

It can be seen that the proof of theorem 3.1.1 turns into the proof of

$$G^{k} \to 0, k \to \infty \Leftrightarrow \rho<1 \tag{24}$$

This requires knowledge of the subordinate norm of matrix, the subordinate norm of matrix A is

$$||A|| = \max_{x \neq 0} \frac{||Ax||}{||x||} = \max_{||x||=1} ||Ax|| \tag{25}$$

and the norm of vector x can be p-norm $||x||_p$(p=1,2,$\cdots \infty$)or any other norm.We provide following properties without proof:

(i) $||A|| \geq 0, ||A|| = 0$ if and only if A=0;

(ii) $|\alpha A| = |\alpha||A||$,for $\forall \alpha \in R$;

DOI: 10.54254/2753-8818/2024.CH21606

(iii) $||A+B|| \leq ||A|| + ||B||,\ \forall A, B \in R^{n\times n}$;
(iv) $||Ax|| \leq ||A|| \cdot ||x||, \forall x \in R^{n}$;
(v) $||AB|| \leq ||A|| \cdot ||B||, \forall A, B \in R^{n\times n}$;
(vi) $||G||_{\infty} = \max_{i} \sum_{j=1}^{n} |a_{ij}|$;

On the one hand, consider "⇒".Assume a certain eigenvalue λ of G and $|\lambda| \geq 1$ and

$$Gx = \lambda x (x \neq 0) \tag{26}$$

then

$$G^k x = \lambda^k x \tag{27}$$

$$||G^k x|| = ||\lambda^k x|| \geq ||x|| \tag{28}$$

By definition of the subordinate of G,

$$||G^k|| \geq \frac{||G^k x||}{||x||} \geq 1 \tag{29}$$

This is contradictory to

$$G^k \to 0,\ k \to \infty \tag{30}$$

On the other hand, another derivation direction is more difficult. We can easily deduce that $\rho(G) \leq ||G||$.

In fact, for $\forall\lambda$, λ is the eigenvalue of G, x is a real eigenvector belonging to λ .Combine property(iv)

$$||Gx|| \leq ||G|| \cdot ||x|| \tag{31}$$

with

$$||Gx|| = ||\lambda x|| = |\lambda| ||x|| \tag{32}$$then

$$|\lambda| \leq ||G|| \tag{33}$$

Therefore

$$\rho(G) \leq ||G|| \tag{34}$$

What if it adds a little more?

When $\rho(G)$ has a slight perturbation, it becomes to $\rho(G)+\varepsilon$,then we have following theorem.

Theorem 3.1.2:*For* $\forall \varepsilon > 0$, *there is a subordinate norm of* G *that* $||G|| \leq \rho(G)+\varepsilon$.

Proof:Let

$$G = PJP^{-1} \tag{35}$$

J is the standard type of A,

$$J = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_S \end{bmatrix}, J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}_{n_i \times n_i}, \sum_{i=1}^{s} n_i = n \tag{36}$$

Let

DOI: 10.54254/2753-8818/2024.CH21606

$$D = \begin{bmatrix} 1 & & & \\ & \varepsilon & & \\ & & \ddots & \\ & & & \varepsilon^{n-1} \end{bmatrix}_{n\times n} \tag{37}$$

Then

$$\tilde{J} = D^{-1}JD \tag{38}$$

has the following form:

$$\tilde{J} = \begin{bmatrix} \tilde{J}_1 & & & \\ & \tilde{J}_2 & & \\ & & \ddots & \\ & & & \tilde{J}_S \end{bmatrix}, \tilde{J}_i = \begin{bmatrix} \lambda_i & \varepsilon & & \\ & \lambda_i & \ddots & \\ & & \ddots & \varepsilon \\ & & & \lambda_i \end{bmatrix}_{n_i \times n_i} \tag{39}$$

Thus

$$||\tilde{J}||_\infty \le \rho(G) + \varepsilon \text{ (in fact,the equal sign holds)} \tag{40}$$

Consider the relationship between G and $\tilde{J}$,Let

$$Q = PD \tag{41}$$

Then

$$G = Q\tilde{J}Q^{-1}, \tilde{J} = Q^{-1}GQ \tag{42}$$

$$||\tilde{J}||_\infty = \max_{||y||_\infty=1} ||\tilde{J}y||_\infty = \max_{||y||_\infty=1} ||Q^{-1}GQy||_\infty \tag{43}$$

Let

$$Qy = x, \tag{44}$$

Then

$$||\tilde{J}||_\infty = \max_{||Q^{-1}x||_\infty=1} ||Q^{-1}Gx||_\infty \tag{45}$$

Through observation, we can let

$$||x|| = ||Q^{-1}x||_\infty \tag{46}$$

$$||\tilde{J}||_\infty = \max_{||Q^{-1}x||_\infty=1} ||Q^{-1}Gx||_\infty = \max_{||x||=1} ||Gx|| = ||G|| \tag{47}$$

Then

$$||G|| \le \rho(G) + \varepsilon \tag{48}$$

Now we can prove "⇐" of theorem 3.1.1. We can always find a ε that

$$\rho(G) + \varepsilon < 1 \tag{49}$$

holds. By property(v),

$$||G^k|| \le ||G||^k \le (\rho(G) + \varepsilon)^k \to 0 \tag{50}$$

By property(i),

$$||G^k|| \to 0 \Leftrightarrow G^k \to 0 \tag{51}$$

DOI: 10.54254/2753-8818/2024.CH21606

proof completed.

When the coefficient matrix is some special matrix, there will be some special results.

3.2. Special coefficient matrix

3.2.1. The coefficient matrix is a strictly diagonally dominant matrix

Theorem 3.2.1:If A is a strictly diagonally dominant matrix, then the G-S iteration method and Jacobi iteration method for solving the linear equation system Ax=b both converge.

Proof: Firstly, prove the convergence of the G-S iterative method.

In the discussion of theorem 3.1.1 and the G-S iterative method mentioned above, it is necessary to prove that the spectral radius of $G = (D - L)^{-1}U$ is less than 1.Proof by contradiction, assuming a certain eigenvalue of G $|\lambda| \geq 1$.

$$|\lambda I - (D - L)^{-1}U| = |(D - L)^{-1}||\lambda(D - L) - U| = |(D - L)^{-1}||\lambda D - \lambda L - U| \quad (52)$$

because

$$|(D - L)^{-1}| \neq 0(\text{A is a strictly diagonally dominant matrix}) \quad (53)$$

$$|\lambda D - \lambda L - U| \text{ only needs to be considered,}\lambda D - \lambda L - U \text{ is denoted as B} \quad (54)$$

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} |b_{ij}| = |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \leq |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + |\lambda| \sum_{j=i+1}^{n} |a_{ij}| \leq |\lambda|(\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|) < |\lambda||a_{ii}| = |b_{ii}| \quad (55)$$

This indicates that B is also a strictly diagonally matrix, thus B is a non-singular matrix,

$$|B| \neq 0 \quad (56)$$

This λ contradicts the eigenvalue of G, so the spectral radius of G is less than 1.In accordance with the theorem 3.1.1,proof completed.

Further, prove the convergence of Jacobi iteration method.

At this point,$G = D^{-1}(L + U)$, it is actually similar to the convergence proof of the G-S iteration method, just note that

$$|\lambda I - D^{-1}(L + U)| = |D^{-1}||\lambda D - L - U| \quad (57)$$

$$B = \lambda D - L - U \quad (58)$$

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} |b_{ij}| = \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}| \leq |\lambda|(\sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|) < |\lambda||a_{ii}| = |b_{ii}| \quad (59)$$

B is a strictly diagonally matrix, therefore B is a non-singular matrix,

$$|B| \neq 0 \quad (60)$$

so this λ contradicts the eigenvalues of G.

3.2.2. The coefficient matrix is a positive definite real symmetric matrix

Theorem 3.2.2:If A is a real symmetric positive definite matrix, then when 2D-A is positive, the Jacobi iteration method converges.

Proof: The iterative matrix

$$G = D^{-1}(L + U) = D^{-1}(D - A) = I - D^{-1}A, x \neq 0, x \in R^n \quad (61)$$

$$||Gx||_A^2 = (AGx, Gx) = (A(I - D^{-1}A)x, (I - D^{-1}A)x) = ||x||_A^2 - ((2D - A)D^{-1}Ax, D^{-1}Ax) \quad (62)$$

DOI: 10.54254/2753-8818/2024.CH21606

we easily know $D^{-1}A$ positive definite, and $2D-A$ is a positive definite.

Therefore,

$$((2D-A)D^{-1}Ax, D^{-1}Ax) > 0 \tag{63}$$

then

$$||Gx||_A < ||x||_A \text{ holds} \tag{64}$$

Thus

$$||G||_A = \max_{x\neq 0} \frac{||Gx||_A}{||x||_A} < 1, \rho(G) \leq ||Gx||_A < 1 \tag{65}$$

then the Jacobi iterative method converges.

4. Example

We will give two examples of system of linear equations Ax=b, run them in MATLAB, solve them using the three methods mentioned above, and compare the results of their runs. For the iterative methods, an initial value x0 needs to be given.

$$A=\begin{pmatrix}4 & 2 & 1\\ 3 & 7 & 2\\ 1 & -1 & 3\end{pmatrix}, b=\begin{pmatrix}10\\ 3\\ 5\end{pmatrix}, x0=\begin{pmatrix}0\\ 0\\ 0\end{pmatrix} \tag{66}$$

Operation results	Run solution	Duration time/second	Number of iterations
Gauss elimination method	[2.8529,-0.9118,0.4118]'	0.001969	\
Gauss-Seidel iterative method	[2.8529,-0.9118,0.4118]'	0.000453	11
Jacobi iterative method	[2.8529,-0.9118,0.4118]'	0.001062	22

$$A=\begin{pmatrix}10 & 2 & 3 & 4\\ 4 & 13 & 5 & 3\\ 7 & 1 & 9 & 0\\ 9 & 8 & 3 & 21\end{pmatrix}, b=\begin{pmatrix}2\\ 7\\ 5\\ 3\end{pmatrix}, x0=\begin{pmatrix}0\\ 0\\ 0\\ 0\end{pmatrix} \tag{67}$$

Operation results	Run solution/10^-1	Duration time/second	Number of iterations
Gauss elimination method	[0.009,3.553,5.153,-0.665]'	0.001742	\
Gauss-Seidel iterative method	[0.009,3.553,5.153,-0.665]'	0.000374	14
Jacobi iterative method	[0.009,3.553,5.153,-0.665]'	0.000728	149

From the table, it can be seen that the Gauss-Seidel iteration method has the shortest duration, the fastest solution, and the least number of iterations. Therefore, if Gauss-Seidel iteration can be used among the three methods, this method should be adopted.

5. Improvement of the iterative method

The Jacobi iterative and G-S iterative methods require the condition that D is reversible, so what improvement should be taken when D is irreversible. At this point, we need to make some changes

DOI: 10.54254/2753-8818/2024.CH21606

to A. Obviously, doing so will result in the solution and we are supposed to take the impact of disturbance of A on the solution into consideration.

Set Δ be a matrix with very small absolute values of its component elements, the exact solution of $Ax = b$ is x^*.When $Ax = b$ transforms into $(A + \Delta)x = b$,the solution of the equations become $x^* + \delta$,thus $(A + \Delta)(x^* + \delta) = b$.Expand and simplify it,we can obtain

$$\delta = -A^{-1} \cdot \Delta \cdot x^* - A^{-1} \cdot \Delta \cdot \delta. \tag{68}$$

Therefore,

$$||\delta|| \leq ||A^{-1}|| \cdot ||\Delta|| \cdot ||x^*|| + ||A^{-1}|| \cdot ||\Delta|| \cdot ||\delta|| \tag{69}$$

or

$$(1 - ||A^{-1}|| \cdot ||\Delta||)||\delta|| \leq ||A^{-1}|| \cdot ||\Delta|| \cdot ||x^*||. \tag{70}$$

Due to the character of Δ, $||A^{-1}|| \cdot ||\Delta|| < 1$ can always hold.

Thereby

$$\frac{||\delta||}{||x^*||} \leq \frac{||A^{-1}|| \cdot ||A||}{1 - ||A^{-1}|| \cdot ||A||} = \frac{||A^{-1}|| \cdot ||A|| \cdot (||\Delta|| / ||A||)}{1 - ||A^{-1}|| \cdot ||A|| \cdot (||\Delta|| / ||A||)} \tag{71}$$

Using the condition number of A,$\text{cond}(A) = ||A|| \cdot ||A^{-1}||$,further obtain

$$\frac{||\delta||}{||x^*||} \leq \frac{\text{cond}(A) \cdot (||\Delta|| / ||A||)}{1 - \text{cond}(A) \cdot (||\Delta|| / ||A||)}. \tag{72}$$

Through the character of the formula, when $\text{cond}(A)$ is very small, the slight perturbation of the coefficient matrix A makes the relative error of solution very small.So adding a very small non-zero number to the zero diagonal element of A makes D reversible,thus Jacobi and G-S methods can be used.

6. Conclusion

This article summarizes the formulas and corresponding Matlab codes for Gauss elimination, Jacobi iteration, and G-S iteration in solving system of equations. The convergence of different iteration methods under different types of coefficient matrices is studied, and it is found that when the coefficient matrix is strictly diagonal, both Jacobi iteration and Gauss-Seidel iteration converge. The coefficient matrix is a positive definite real symmetric matrix, and when 2D-A is positive, Jacobi iteration method converges Meanwhile, through simulation programming, two tables were obtained, from which the running results show that the Gauss-Seidel iteration method has the shortest duration, the fastest solution, and the least number of iterations. Therefore, when the diagonal elements of the coefficient matrix are non-zero, the Gauss-Seidel iteration method has the best effect. When there are zero elements in the diagonal elements and the condition number of A is small, the iterative methods can be used by perturbing at the zero elements.

References

[1] *David C.Lay Steven R.Lay Judi J.McDonald.Linear Algebra and Its Applications.*
[2] *Wang G.X.Zhou Z.M.Zhu S.M.Ordinary Differential Equation.*
[3] *Huang Y.Q.Shu S.Yang Y..Numerical Method.*
[4] *Shams M, Kausar N, Agarwal P, et al.Triangular intuitionistic fuzzy linear system of equations with applications: an analytical approach[J].Applied Mathematics in Science and Engineering, 2024, 32(1):*
[5] *Darvishi T M , Khani F , Godarzi M A , et al.Symmetric modified AOR method to solve systems of linear equations[J]. Journal of Applied Mathematics and Computing, 2011, 36(1-2):41-59.*

DOI: 10.54254/2753-8818/2024.CH21606

[6] *Zhong-Zhi B.On convergence of the matrix splitting iteration paradigm for solving systems of linear equations[J]. Applied Mathematics Letters, 2024, 150108969-.*

[7] *EisenstatC S, ElmanC H, SchultzH M.Variational Iterative Methods for Nonsymmetric Systems of Linear Equations[J].SIAM Journal on Numerical Analysis, 2006, 20(2):345-357.*