CHAPTER 6

Linear Systems: Iterative Methods

In Section 4.1.1 and Section 4.1.4 we have discussed methods that obtain exact solution of a linear system $A\boldsymbol{x} = \boldsymbol{b}$ in the absence of floating point errors (*i.e.*, when the infinite precision arithmetic is used). Such methods are called the ***direct methods***. The solution of a linear system can also be obtained using iterative procedures. Such methods are called ***iterative methods***. There are many iterative procedures out of which *Jacobi* and *Gauss-Seidel* methods are the simplest ones. In this chapter we introduce these two basic iterative methods and discuss a sufficient condition under which these methods converge. We introduce the notion of spectral radius of a matrix and prove a necessary and sufficient condition on the spectral radius of the *iterative matrix* for the convergence of a general iterative method to solve system of linear equations.

6.1 Jacobi Method

In Section 4.1.4, we have seen that when a linear system $A\boldsymbol{x} = \boldsymbol{b}$ is such that the coefficient matrix A is a diagonal matrix, then this system can be solved very easily. We explore this idea to build a new method based on iterative procedure. For this, we first rewrite the matrix A as

$$A = D - C,$$

where $D, C \in M_n(\mathbb{R})$ are such that

$$D = \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix},$$

where a_{ii}, $i = 1, 2, \cdots, n$ are the diagonal elements of the matrix A. Then the given system of linear equations can be rewritten as

$$D\boldsymbol{x} = C\boldsymbol{x} + \boldsymbol{b}. \tag{6.1}$$

If we assume that the right hand side is fully known to us, then the above system can be solved very easily as D is a diagonal matrix. But the right hand side involves the unknown vector $\boldsymbol{x}$ and therefore is unknown to us. Rather, if we choose (arbitrarily) some specific value for $\boldsymbol{x}$, say $\boldsymbol{x} = \boldsymbol{x}^{(0)}$, on the right hand side then the resulting system

$$D\boldsymbol{x} = C\boldsymbol{x}^{(0)} + \boldsymbol{b}$$

can be readily solved for $\boldsymbol{x}$ on the left hand side. Let us call the solution of this system as $\boldsymbol{x}^{(1)}$. That is,

$$D\boldsymbol{x}^{(1)} = C\boldsymbol{x}^{(0)} + \boldsymbol{b}.$$

Now taking $\boldsymbol{x} = \boldsymbol{x}^{(1)}$ on the right hand side of (6.1) we can obtain the value of $\boldsymbol{x}$ on the left hand side, which we denote as $\boldsymbol{x}^{(2)}$. Repeat this procedure to get a general iterative procedure as

$$D\boldsymbol{x}^{(k+1)} = C\boldsymbol{x}^{(k)} + \boldsymbol{b}, \quad k = 0, 1, 2, \cdots .$$

If D is invertible, then the above iterative procedure can be written as

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \quad k = 0, 1, 2, \cdots , \tag{6.2}$$

where $B = D^{-1}C$ and $\boldsymbol{c} = D^{-1}\boldsymbol{b}$. The iterative procedure (6.2) is called the ***Jacobi method*** and the matrix B is called the *Jacobi iterative matrix*.

Example 6.1.1.

Let us illustrate the Jacobi method in the case of 3×3 system

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\
a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\
a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3,
\end{aligned}$$

where $a_{11} \neq 0$, $a_{22} \neq 0$, and $a_{33} \neq 0$. We can rewrite the above system of linear equations as

$$\begin{aligned}
x_1 &= \frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3) \\
x_2 &= \frac{1}{a_{22}}(b_2 - a_{21}x_1 - a_{23}x_3) \\
x_3 &= \frac{1}{a_{33}}(b_3 - a_{31}x_1 - a_{32}x_2)
\end{aligned}$$

Let $\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^T$ be an initial guess to the true solution $\boldsymbol{x}$, which is chosen arbitrarily. Define a sequence of iterates (for $k = 0, 1, 2, \cdots$) by

$$\left.\begin{aligned} x_1^{(k+1)} &= \frac{1}{a_{11}}(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}) \\ x_2^{(k+1)} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{(k)} - a_{23}x_3^{(k)}) \\ x_3^{(k+1)} &= \frac{1}{a_{33}}(b_3 - a_{31}x_1^{(k)} - a_{32}x_2^{(k)}). \end{aligned}\right\} \tag{JS}$$

which is the ***Jacobi iterative sequence*** given by (6.2) in the case of 3×3 system.

Now, the question is that will the sequence of vectors $\{\boldsymbol{x}^{(k+1)}\}$ generated by the iterative procedure (6.2) always converge to the exact solution $\boldsymbol{x}$ of the given linear system?

The following example gives a system for which the Jacobi iterative sequence converges to the exact solution.

Example 6.1.2.

The Jacobi iterative sequence for the system

$$\begin{aligned} 6x_1 + x_2 + 2x_3 &= -2, \\ x_1 + 4x_2 + 0.5x_3 &= 1, \\ -x_1 + 0.5x_2 - 4x_3 &= 0. \end{aligned}$$

is given by

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{6}(-2 - x_2^{(k)} - 2x_3^{(k)}), \\ x_2^{(k+1)} &= \frac{1}{4}(1 - x_1^{(k)} - 0.5x_3^{(k)}), \\ x_3^{(k+1)} &= -\frac{1}{4}(0 + x_1^{(k)} - 0.5x_2^{(k)}). \end{aligned}$$

The exact solution (upto 6-digit rounding) of this system is

$$\boldsymbol{x} \approx (-0.441176, 0.341176, 0.152941).$$

Choosing the initial guess $\boldsymbol{x}^{(0)} = (0, 0, 0)$, we get

$$\begin{aligned} \boldsymbol{x}^{(1)} &\approx (-0.333333, 0.250000, 0.000000), \\ \boldsymbol{x}^{(2)} &\approx (-0.375000, 0.333333, 0.114583), \\ \boldsymbol{x}^{(3)} &\approx (-0.427083, 0.329427, 0.135417), \\ \boldsymbol{x}^{(4)} &\approx (-0.433377, 0.339844, 0.147949), \\ \boldsymbol{x}^{(5)} &\approx (-0.439290, 0.339851, 0.150825), \end{aligned}$$

and so on. We observe from the above computed results that the sequence $\{\boldsymbol{x}^{(k)}\}$ seems to be approaching the exact solution.

In the following example we discuss a system for which the Jacobi iterative sequence does not converge to the exact solution.

Example 6.1.3.

Consider the system

$$\begin{aligned} x_1 + 4x_2 + 0.5x_3 &= 1, \\ 6x_1 + x_2 + 2x_3 &= -2, \\ -x_1 + 0.5x_2 - 4x_3 &= 0. \end{aligned}$$

which is exactly the same as the system discussed in Example 6.1.2 but the only difference is the interchange of first and second equation. Hence, the exact solution is same as given in (**??**). The Jacobi iterative sequence for this system is given by

$$\begin{aligned} x_1^{(k+1)} &= (1 - 4x_2^{(k)} - 0.5x_3^{(k)}), \\ x_2^{(k+1)} &= (-2 - 6x_1^{(k)} - 2x_3^{(k)}), \\ x_3^{(k+1)} &= -\frac{1}{4}(0 + x_1^{(k)} - 0.5x_2^{(k)}). \end{aligned}$$

Choosing the initial guess $\boldsymbol{x}^{(0)} = (0, 0, 0)$, we get

$$\begin{aligned} \boldsymbol{x}^{(1)} &\approx (1, -2, 0), \\ \boldsymbol{x}^{(2)} &\approx (9, -8, -0.5), \\ \boldsymbol{x}^{(3)} &\approx (33.25, -55, -3.25), \\ \boldsymbol{x}^{(4)} &\approx (222.625, -195, -15.1875), \\ \boldsymbol{x}^{(5)} &\approx (788.59375, -1307.375, -80.03125), \end{aligned}$$

and so on. Here, we observe a diverging trend in the sequence $\{\boldsymbol{x}^{(k)}\}$.

The above two examples shows that the Jacobi sequence need not converge always and so we need to look for a condition on the system for which the Jacobi iterative sequence converges to the exact solution.

Define the error in the k^{th} iterate $\boldsymbol{x}^{(k)}$ compared to the exact solution by

$$\boldsymbol{e}^{(k)} = \boldsymbol{x} - \boldsymbol{x}^{(k)}.$$

It follows easily that $\boldsymbol{e}^{(k)}$ satisfies the system

$$\boldsymbol{e}^{(k+1)} = B\boldsymbol{e}^{(k)},$$

where B is as defined in (6.2). Using any vector norm and the matrix norm subordinate to it in the above equation, we get

$$\begin{aligned}\|\boldsymbol{e}^{(k+1)}\| &= \|B\boldsymbol{e}^{(k)}\| \\ &\leq \|B\|\|\boldsymbol{e}^{(k)}\| \\ &\leq \cdots \\ &\leq \cdots \\ &\leq \|B\|^{k+1}\|\boldsymbol{e}^{(0)}\|.\end{aligned}$$

Thus, when $\|B\| < 1$, the iteration method (6.2) always converges for any initial guess $\boldsymbol{x}^{(0)}$.

Again the question is

“what are all the matrices A for which the corresponding matrices B in (6.2) have the property $\|B\| < 1$, for some matrix norm subordinate to some vector norm?”

We would like an answer that is ‘’easily verifiable” using the entries of the matrix A. One such class of matrices are the *diagonally dominant* matrices, which we define now.

Definition 6.1.4 [Diagonally Dominant Matrices].

A matrix A is said to be ***diagonally dominant*** if it satisfies the inequality

$$\sum_{j=1, j\neq i}^{n} |a_{ij}| < |a_{ii}|, \quad i = 1, 2, \cdots, n.$$

We now prove the sufficient condition for the convergence of the Jacobi method. This theorem asserts that if A is a diagonally dominant matrix, then B in (6.2) of the Jacobi method is such that $\|B\|_\infty < 1$.

Theorem 6.1.5 [Convergence theorem for Jacobi method].

If the coefficient matrix A is diagonally dominant, then the Jacobi method (6.2) converges to the exact solution of $A\boldsymbol{x} = \boldsymbol{b}$.

Proof.

Since A is diagonally dominant, the diagonal entries are all non-zero and hence the Jacobi iterating sequence $\boldsymbol{x}^{(k)}$ given by

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(k)} \right), \quad i = 1, 2, \cdots, n. \tag{6.3}$$

is well-defined. Each component of the error satisfies

$$e_i^{(k+1)} = -\sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}} e_j^{(k)}, \quad i = 1, 2, \cdots, n. \tag{6.4}$$

which gives

$$|e_i^{(k+1)}| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right| \|\mathbf{e}^{(k)}\|_\infty. \tag{6.5}$$

Define

$$\mu = \max_{1\leq i\leq n} \sum_{\substack{j=1 \\ j\neq i}}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|. \tag{6.6}$$

Then

$$|e_i^{(k+1)}| \leq \mu \|\boldsymbol{e}^{(k)}\|_\infty, \tag{6.7}$$

which is true for all $i = 1, 2, \cdots, n$. Therefore, we have

$$\|\boldsymbol{e}^{(k+1)}\|_\infty \leq \mu \|\boldsymbol{e}^{(k)}\|_\infty. \tag{6.8}$$

The matrix A is diagonally dominant if and only if $\mu < 1$. Then iterating the last inequality we get

$$\|\boldsymbol{e}^{(k+1)}\|_\infty \leq \mu^{k+1} \|\boldsymbol{e}^{(0)}\|_\infty. \tag{6.9}$$

Therefore, if A is diagonally dominant, the Jacobi method converges.

Remark 6.1.6.

Observe that the system given in Example 6.1.2 is diagonally dominant, whereas the system in Example 6.1.3 is not so.

6.2 Gauss-Seidel Method

Gauss-Seidel method is a modified version of the Jacobi method discussed in Section 6.1. We demonstrate the method in the case of a 3×3 system and the method for a general $n \times n$ system can be obtained in a similar way.

Example 6.2.1.

Consider the 3×3 system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3. \end{aligned}$$

When the diagonal elements of this system are non-zero, we can rewrite the above system as

$$\begin{aligned} x_1 &= \frac{1}{a_{11}}(b_1 - a_{12}x_2 - a_{13}x_3), \\ x_2 &= \frac{1}{a_{22}}(b_2 - a_{21}x_1 - a_{23}x_3), \\ x_3 &= \frac{1}{a_{33}}(b_3 - a_{31}x_1 - a_{32}x_2). \end{aligned}$$

Let

$$\boldsymbol{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^T$$

be an initial guess to the true solution $\boldsymbol{x}$. Define a sequence of iterates (for $k = 0, 1, 2, \cdots$) by

$$\left.\begin{aligned} x_1^{(k+1)} &= \frac{1}{a_{11}}(b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)}), \\ x_2^{(k+1)} &= \frac{1}{a_{22}}(b_2 - a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)}), \\ x_3^{(k+1)} &= \frac{1}{a_{33}}(b_3 - a_{31}x_1^{(k+1)} - a_{32}x_2^{(k+1)}). \end{aligned}\right\} \tag{GSS}$$

This sequence of iterates is called the ***Gauss-Seidel iterative sequence*** and the method is called ***Gauss-Seidel Iteration method***.

Remark 6.2.2.

Compare (JS) and (GSS).

Theorem 6.2.3 [Convergence theorem for Gauss-Seidel method].

If the coefficient matrix is diagonally dominant, then the Gauss-Seidel method converges to the exact solution of the system $A\boldsymbol{x} = \boldsymbol{b}$.

Proof.

Since A is diagonally dominant, all the diagonal elements of A are non-zero, and hence the Gauss-Seidel iterative sequence given by

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left\{ b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right\}, \quad i = 1, 2, \cdots, n. \tag{6.10}$$

is well-defined. The error in each component is given by

$$e_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{(k+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} e_j^{(k)}, \quad i = 1, 2, \cdots, n. \tag{6.11}$$

For $i = 1, 2, \cdots, n,$, define

$$\alpha_i = \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right|, \quad \beta_i = \sum_{j=i+1}^{n} \left| \frac{a_{ij}}{a_{ii}} \right|,$$

with the convention that $\alpha_1 = \beta_n = 0$. Note that μ given in (6.6) can be written as

$$\mu = \max_{1 \le i \le n} (\alpha_i + \beta_i)$$

Since A is a diagonally dominant matrix, we have $\mu < 1$. Now

$$|e_i^{(k+1)}| \le \alpha_i \|\boldsymbol{e}^{(k+1)}\|_\infty + \beta_i \|\boldsymbol{e}^{(k)}\|_\infty, \quad i = 1, 2, \cdots, n. \tag{6.12}$$

Let l be such that

$$\|\boldsymbol{e}^{(k+1)}\|_\infty = |e_l^{(k+1)}|.$$

Then with $i = l$ in (6.12),

$$\|\boldsymbol{e}^{(k+1)}\|_\infty \le \alpha_l \|\boldsymbol{e}^{(k+1)}\|_\infty + \beta_l \|\boldsymbol{e}^{(k)}\|_\infty. \tag{6.13}$$

Since $\mu < 1$, we have $\alpha_l < 1$ and therefore the above inequality gives

$$\|\boldsymbol{e}^{(k+1)}\|_\infty \leq \frac{\beta_l}{1-\alpha_l}\|\boldsymbol{e}^{(k)}\|_\infty. \tag{6.14}$$

Define

$$\eta = \max_{1\leq i\leq n}\frac{\beta_i}{1-\alpha_i}. \tag{6.15}$$

Then the above inequality yields

$$\|\boldsymbol{e}^{(k+1)}\|_\infty \leq \eta\|\boldsymbol{e}^{(k)}\|_\infty. \tag{6.16}$$

Since for each i,

$$(\alpha_i + \beta_i) - \frac{\beta_i}{1-\alpha_i} = \frac{\alpha_i[1-(\alpha_i+\beta_i)]}{1-\alpha_i} \geq \frac{\alpha_i}{1-\alpha_i}[1-\mu] \geq 0, \tag{6.17}$$

we have

$$\eta \leq \mu < 1. \tag{6.18}$$

Thus, when the coefficient matrix A is diagonally dominant, Gauss-Seidel method converges.

Remark 6.2.4.

A careful observation of the proof of the above theorem reveals that the Gauss-Seidel method converges faster than the Jacobi method by comparing (6.18) and (6.9).

6.3 Stopping Criteria

Let $\boldsymbol{x}^*$ denote the computed solution using some method. The mathematical error in the approximate solution when compared to the exact solution of a linear system $A\boldsymbol{x} = \boldsymbol{b}$ is given by

$$\boldsymbol{e} = \boldsymbol{x} - \boldsymbol{x}^*. \tag{6.19}$$

Recall from Chapter 3 that the mathematical error is due to the approximation made in the numerical method where the computation is done without any floating-point approximation (ie., without rounding or chopping). Observe that to get the mathematical error, we need to know the exact solution. But an astonishing feature of linear systems (which is not there in nonlinear equations) is that this error can be obtained exactly without the knowledge of the exact solution. To do this, we first define the ***residual vector***

$$\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x}^* \tag{6.20}$$

in the approximation of $\boldsymbol{b}$ by $A\boldsymbol{x}^*$. This vector is also referred to as ***residual error***. Since $\boldsymbol{b} = A\boldsymbol{x}$, we have

$$\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x}^* = A\boldsymbol{x} - A\boldsymbol{x}^* = A(\boldsymbol{x} - \boldsymbol{x}^*).$$

The above identity can be written as

$$A\boldsymbol{e} = \boldsymbol{r}. \tag{6.21}$$

This shows that the error $\boldsymbol{e}$ satisfies a linear system with the same coefficient matrix A as in the original system $A\boldsymbol{x} = \boldsymbol{b}$, but a different right hand side vector. Thus, by having the approximate solution $\boldsymbol{x}^*$ in hand, we can obtain the error $\boldsymbol{e}$ without knowing the exact solution $\boldsymbol{x}$ of the system.

In the iterative methods discussed above, we have a sequence $\{\boldsymbol{x}^{(k)}\}$ that is expected to converge to the exact solution of the given linear system. In practical situation, we cannot go on computing the $\boldsymbol{x}^{(k)}$ indefinitely and we need to terminate our computation once the value of $\boldsymbol{x}^{(k)}$ reaches a desired accuracy for a sufficiently large k. That is, when the error

$$\|\boldsymbol{e}^{(k)}\| = \|\boldsymbol{x} - \boldsymbol{x}^{(k)}\|$$

in the k^{th} iteration in some norm is sufficiently small. Since, we do not know the exact solution $\boldsymbol{x}$, the error given above cannot be computed easily and needs another linear system (6.21) to be solved. Therefore, the question is how to decide where we have to stop our computation (without solving this linear system)? In other words, how do we know whether the computed vector $\boldsymbol{x}^{(k)}$ at the k^{th} iteration is sufficiently close to the exact solution or not. This can be decided by looking at the residual error vector of the k^{th} iteration defined as

$$\boldsymbol{r}^{(k)} = \boldsymbol{b} - A\boldsymbol{x}^{(k)}. \tag{6.22}$$

Thus, for a given sufficiently small positive number ϵ, we stop the iteration if

$$\|\boldsymbol{r}^{(k)}\| < \epsilon,$$

for some vector norm $\|\cdot\|$.

6.4 Exercises

Iterative Methods

1. Let A be a diagonally dominant matrix such that $a_{ij} = 0$ for every $i, j \in \{1, 2, \cdots, n\}$ such that $i > j + 1$. Does naive Gaussian elimination method preserve the diagonal dominance? Justify your answer.

2. Let A be a diagonally dominant matrix. Show that all the diagonal elements of A are non-zero (*i.e.*, $a_{ii} \neq 0$ for $i = 1, 2, \cdots, n$.). As a consequence, the iterating sequences of Jacobi and Gauss-Seidel methods are well-defined if the coefficient matrix A in the linear system $A\boldsymbol{x} = \boldsymbol{b}$ is a diagonally dominant matrix.

3. Write the formula for the Jacobi iterative sequence of the system

$$\begin{aligned} 7x_1 - 15x_2 - 21x_3 &= 2, \\ 7x_1 - x_2 - 5x_3 &= -3, \\ 7x_1 + 5x_2 + x_3 &= 1. \end{aligned}$$

   Without performing the iterations, show that the sequence does not converge to the exact solution of this system. Can you make a suitable interchange of rows so that the resulting system is diagonally dominants?

4. Find the $n \times n$ matrix B and the n-dimensional vector $\boldsymbol{c}$ such that the Gauss-Seidel method can be written in the form

$$\boldsymbol{x}^{(k+1)} = B\boldsymbol{x}^{(k)} + \boldsymbol{c}, \qquad k = 0, 1, 2, \cdots$$

   Here B is called the *iterative matrix* for the Gauss-Seidel method.