Chapter 8

Iterative Methods

1 Basic Iterative Methods

In this section, we study iterative methods for solving a linear system $A\mathbf{x} = \mathbf{b}$. Iterative methods start out with an initial approximation $\mathbf{x}^{(0)}$ to the solution and go through a fixed procedure to obtain a better approximation, $\mathbf{x}^{(1)}$. The same procedure is then repeated on $\mathbf{x}^{(1)}$ to obtain an improved approximation, $\mathbf{x}^{(2)}$; and so on. The iterations terminate when a desired accuracy has been achieved.

Iterative methods are most useful in solving large sparse systems. Such systems occur, for example, in the solution of boundary value problems for partial differential equations. The number of flops necessary to solve an $n \times n$ using iterative methods is proportional to n^2, whereas, the amount necessary using Gaussian elimination is proportional to n^3. Thus for large values of n iterative methods provide the only practical way of solving the system. Furthermore, the amount of memory required for a sparse coefficient matrix A is proportional to n, whereas, Gaussian elimination and the other direct methods studied in earlier chapters usually tend to fill in the zeros of A and hence require an amount of storage proportional to n^2. This can present quite a problem when n is very large, say $n \geq 20000$.

The iterative methods we will describe only require that in each iteration we can multiply A times a vector in R^n. If A is sparse, this can usually be accomplished in a systematic way so that only a small proportion of the entry of A need be accessed. The one disadvantage of iterative methods is that after solving $A\mathbf{x} = \mathbf{b_1}$, one must start all over again from the beginning in order to solve $A\mathbf{x} = \mathbf{b_2}$.

Matrix Splittings

Given a system $A\mathbf{x} = \mathbf{b}$, we write the coefficient matrix A in the form $A = C - M$, where C is a nonsingular matrix which is in some form that is easily invertible (e.g., diagonal or triangular). The representation $A = C - M$ is referred to as a *matrix splitting*. The system can then be rewritten in the form

$$C\mathbf{x} = M\mathbf{x} + \mathbf{b}$$

$$\mathbf{x} = C^{-1}M\mathbf{x} + C^{-1}\mathbf{b}$$

If we set

$$B = C^{-1}M = I - C^{-1}A \qquad \text{and} \qquad \mathbf{c} = C^{-1}\mathbf{b}$$

then

$$\mathbf{x} = B\mathbf{x} + \mathbf{c} \tag{1}$$

To solve the system, we start out with an initial guess $\mathbf{x}^{(0)}$, which may be any vector in R^n. We then set

$$\begin{aligned} \mathbf{x}^{(1)} &= B\mathbf{x}^{(0)} + \mathbf{c} \\ \mathbf{x}^{(2)} &= B\mathbf{x}^{(1)} + \mathbf{c} \end{aligned}$$

and in general

$$\mathbf{x}^{(k+1)} = B\mathbf{x}^{(k)} + \mathbf{c}$$

Let $\mathbf{x}$ be a solution of the linear system. If $\|\cdot\|$ denotes some vector norm on R^n and the corresponding matrix norm of B is less than 1, we claim that $\|\mathbf{x}^{(k)} - \mathbf{x}\| \to 0$ as $k \to \infty$. Indeed,

$$\begin{aligned} \mathbf{x}^{(1)} - \mathbf{x} &= (B\mathbf{x}^{(0)} + \mathbf{c}) - (B\mathbf{x} + \mathbf{c}) = B(\mathbf{x}^{(0)} - \mathbf{x}) \\ \mathbf{x}^{(2)} - \mathbf{x} &= (B\mathbf{x}^{(1)} + \mathbf{c}) - (B\mathbf{x} + \mathbf{c}) = B(\mathbf{x}^{(1)} - \mathbf{x}) = B^2(\mathbf{x}^{(0)} - \mathbf{x}) \end{aligned}$$

and so on. In general,

$$\mathbf{x}^{(k)} - \mathbf{x} = B^k(\mathbf{x}^{(0)} - \mathbf{x}) \tag{2}$$

and hence

$$\begin{aligned} \|\mathbf{x}^{(k)} - \mathbf{x}\| &= \|B^k(\mathbf{x}^{(0)} - \mathbf{x})\| \\ &\leq \|B^k\|\,\|\mathbf{x}^{(0)} - \mathbf{x}\| \\ &\leq \|B\|^k\|\mathbf{x}^{(0)} - \mathbf{x}\| \end{aligned}$$

Thus, if $\|B\| < 1$, then $\|\mathbf{x}^{(k)} - \mathbf{x}\| \to 0$ as $k \to \infty$.

The foregoing result holds for any norm on R^n, although in practice it is simplest to use the $\|\cdot\|_\infty$ or the $\|\cdot\|_1$. Essentially, then, we require that the matrix C be easily invertible and that C^{-1} be a good enough approximation to A^{-1} so that

$$\|I - C^{-1}A\| = \|B\| < 1$$

This last condition implies that all the eigenvalues of B are less than 1 in modulus.

Definition. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of B and let $\rho(B) = \max\limits_{1 \leq i \leq n} |\lambda_i|$. The constant $\rho(B)$ is called the **spectral radius** of B.

Theorem 8.1.1. *Let $\mathbf{x}^{(0)}$ be an arbitrary vector in R^n and define $\mathbf{x}^{(i+1)} = B\mathbf{x}^{(i)} + \mathbf{c}$ for $i = 0, 1, \ldots$. If $\mathbf{x}$ is the solution to (1), then a necessary and sufficient condition for $\mathbf{x}^{(k)} \to \mathbf{x}$ is that $\rho(B) < 1$.*

Proof. We will prove the theorem only in the case where B has n linearly independent eigenvectors. The case where B is not diagonalizable is beyond the scope of this book. If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are n linearly independent eigenvectors of B, we can write

$$\mathbf{x}^{(0)} - \mathbf{x} = \alpha_1\mathbf{x}_1 + \cdots + \alpha_n\mathbf{x}_n$$

and it follows from (2) that

$$\begin{aligned} \mathbf{x}^{(k)} - \mathbf{x} &= B^k(\alpha_1\mathbf{x}_1 + \cdots + \alpha_n\mathbf{x}_n) \\ &= \alpha_1\lambda_1^k\mathbf{x}_1 + \cdots + \alpha_n\lambda_n^k\mathbf{x}_n \end{aligned}$$

Thus

$$\mathbf{x}^{(k)} - \mathbf{x} \to \mathbf{0}$$

if and only if $|\lambda_i| < 1$ for $i = 1, \ldots, n$. Thus $\mathbf{x}^{(k)} \to \mathbf{x}$ if and only if $\rho(B) < 1$. □

The simplest choice of C is to let C be a diagonal matrix whose diagonal elements are the diagonal elements of A. The iteration scheme with this choice of C is called *Jacobi iteration*.

Jacobi Iteration

Let

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

and

$$M = -\begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ a_{21} & 0 & & a_{2n} \\ \vdots & & \ddots & \\ a_{n1} & a_{n2} & & 0 \end{pmatrix}$$

and set $B = C^{-1}M$ and $\mathbf{c} = C^{-1}\mathbf{b}$. Thus

$$B = \begin{pmatrix} 0 & \dfrac{-a_{12}}{a_{11}} & \cdots & \dfrac{-a_{1n}}{a_{11}} \\ \dfrac{-a_{21}}{a_{22}} & 0 & \cdots & \dfrac{-a_{2n}}{a_{22}} \\ \vdots & & & \\ \dfrac{-a_{n1}}{a_{nn}} & \dfrac{-a_{n2}}{a_{nn}} & \cdots & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{c} = \begin{pmatrix} \dfrac{b_1}{a_{11}} \\ \dfrac{b_2}{a_{22}} \\ \vdots \\ \dfrac{b_n}{a_{nn}} \end{pmatrix}$$

At the $(i+1)$st iteration, the vector $\mathbf{x}^{(i+1)}$ is calculated by

$$x_j^{(i+1)} = \frac{1}{a_{jj}} \left(- \sum_{\substack{k=1 \\ k \neq j}}^{n} a_{jk} x_k^{(i)} + b_j \right) \qquad j = 1, \ldots, n \tag{3}$$

The vector $\mathbf{x}^{(i)}$ is used in calculating $\mathbf{x}^{(i+1)}$. Consequently, these two vectors must be stored separately.

If the diagonal elements of A are much larger than the off-diagonal elements, the entries of B should all be small and the Jacobi iteration should converge. We say that A is *diagonally dominant* if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| \qquad \text{for} \qquad i = 1, \ldots, n$$

If A is diagonally dominant, the matrix B of the Jacobi iteration will have the property

$$\sum_{j=1}^{n} |b_{ij}| = \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} < 1 \qquad \text{for} \qquad i = 1, \ldots, n$$

Thus

$$\|B\|_\infty = \max_{1 \leq i \leq n} \left(\sum_{j=1}^{n} |b_{ij}| \right) < 1$$

It follows, then, that if A is diagonally dominant, the Jacobi iteration will converge to the solution of $A\mathbf{x} = \mathbf{b}$.

An alternative to the Jacobi iteration is to take C to be the lower triangular part of A (i.e., $c_{ij} = a_{ij}$ if $i \geq j$ and $c_{ij} = 0$ if $i < j$). Since C is a better approximation to A than the diagonal matrix in the Jacobi iteration, we would expect that C^{-1} is a better approximation to A^{-1}, and hopefully B will have a smaller norm. The iteration scheme with this choice of C is called *Gauss–Seidel iteration*. It usually converges faster than Jacobi iteration.

Gauss–Seidel Iteration

Let

$$L = - \begin{pmatrix} 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & \cdots & 0 & 0 \\ \vdots & & & & \\ a_{n-1,1} & a_{n-1,2} & & 0 & 0 \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & 0 \end{pmatrix}$$

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

and

$$U = - \begin{pmatrix} 0 & a_{12} & \cdots & a_{1,n-1} & a_{1n} \\ 0 & 0 & \cdots & a_{2,n-1} & a_{2n} \\ \vdots & & & & \\ 0 & 0 & & 0 & a_{n-1,n} \\ 0 & 0 & & 0 & 0 \end{pmatrix}$$

Set $C = D - L$ and $M = U$. Let $\mathbf{x}^{(0)}$ be an arbitrary nonzero vector in R^n. We have

$$\begin{aligned} C\mathbf{x}^{(i+1)} &= M\mathbf{x}^{(i)} + \mathbf{b} \\ (D-L)\mathbf{x}^{(i+1)} &= U\mathbf{x}^{(i)} + \mathbf{b} \\ D\mathbf{x}^{(i+1)} &= L\mathbf{x}^{(i+1)} + U\mathbf{x}^{(i)} + \mathbf{b} \end{aligned}$$

We can solve this last equation for $\mathbf{x}^{(i+1)}$ one coordinate at a time. The first coordinate of $\mathbf{x}^{(i+1)}$ is given by

$$x_1^{(i+1)} = \frac{1}{a_{11}} \left(-\sum_{k=2}^{n} a_{1k} x_k^{(i)} + b_1 \right)$$

The second coordinate of $\mathbf{x}^{(i+1)}$ can be solved for in terms of the first coordinate and the last $n-2$ coordinates of $x^{(i)}$.

$$x_2^{(i+1)} = \frac{1}{a_{22}} \left(-a_{21} x_1^{(i+1)} - \sum_{k=3}^{n} a_{2k} x_k^{(i)} + b_2 \right)$$

In general,

$$x_j^{(i+1)} = \frac{1}{a_{jj}} \left(-\sum_{k=1}^{j-1} a_{jk} x_k^{(i+1)} - \sum_{k=j+1}^{n} a_{jk} x_k^{(i)} + b_j \right) \tag{4}$$

It is interesting to compare (3) and (4). The difference between the Jacobi and Gauss–Seidel iterations is that in the latter case, one is using the coordinates of $\mathbf{x}^{(i+1)}$ as soon as they are calculated rather than in the next iteration. The program for the Gauss–Seidel iteration is actually simpler than the program for the Jacobi iteration. The vectors $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(i+1)}$ are both stored in the same vector, $\mathbf{x}$. As a coordinate of $\mathbf{x}^{(i+1)}$ is calculated, it replaces the corresponding coordinate of $\mathbf{x}^{(i)}$.

Theorem 8.1.2. *If A is diagonally dominant, then the Gauss–Seidel iteration converges to a solution of $A\mathbf{x} = \mathbf{b}$.*

Proof. For $j = 1, \ldots, n$, let

$$\alpha_j = \sum_{i=1}^{j-1} |a_{ji}|, \qquad \beta_j = \sum_{i=j+1}^{n} |a_{ji}|, \qquad \text{and} \qquad M_j = \frac{\beta_j}{(|a_{jj}| - \alpha_j)}$$

Since A is diagonally dominant, it follows that

$$|a_{jj}| > \alpha_j + \beta_j$$

and consequently $M_j < 1$ for $j = 1, \ldots, n$. Thus

$$M = \max_{1 \le j \le n} M_j < 1$$

We will show that

$$\|B\|_\infty = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|B\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \le M < 1$$

Let $\mathbf{x}$ be a nonzero vector in R^n and let $\mathbf{y} = B\mathbf{x}$. Choose k so that

$$\|\mathbf{y}\|_\infty = \max_{1 \le i \le n} |y_i| = |y_k|$$

It follows from the definition of B that

$$\mathbf{y} = B\mathbf{x} = (D - L)^{-1} U\mathbf{x}$$

and hence

$$\mathbf{y} = D^{-1}(L\mathbf{y} + U\mathbf{x})$$

Comparing the kth coordinates of each side, we see that

$$y_k = \frac{1}{a_{kk}} \left(-\sum_{i=1}^{k-1} a_{ki} y_i - \sum_{i=k+1}^{n} a_{ki} x_i \right)$$

and hence

$$\|\mathbf{y}\|_\infty = |y_k| \le \frac{1}{|a_{kk}|} (\alpha_k \|\mathbf{y}\|_\infty + \beta_k \|\mathbf{x}\|_\infty) \tag{5}$$

It follows from (5) that

$$\frac{\|B\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{\|\mathbf{y}\|_\infty}{\|\mathbf{x}\|_\infty} \le M_k \le M$$

Thus

$$\|B\|_\infty = \max_{\mathbf{x} \neq \mathbf{0}} \frac{\|B\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \le M < 1$$

and hence the iteration will converge to the solution of $A\mathbf{x} = \mathbf{b}$. □

Exercises

1. Let

$$A = \begin{pmatrix} 10 & 1 \\ 2 & 10 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 11 \\ 12 \end{pmatrix}, \qquad \text{and} \qquad \mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

   Use Jacobi iteration to compute $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$. [The exact solution is $\mathbf{x} = (1,1)^T$.]

2. Let

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \qquad \text{and} \qquad \mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

   Use Jacobi iteration to compute $\mathbf{x}^{(1)}$, $\mathbf{x}^{(2)}$, $\mathbf{x}^{(3)}$, and $\mathbf{x}^{(4)}$.

3. Repeat Exercise 1 using Gauss–Seidel iteration.

4. Let

$$A = \begin{pmatrix} 10 & 1 & 1 \\ 1 & 10 & 1 \\ 1 & 1 & 10 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 12 \\ 12 \\ 12 \end{pmatrix}, \qquad \text{and} \qquad \mathbf{x}^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

   (a) Calculate $\mathbf{x}^{(1)}$ using Jacobi iteration.

   (b) Calculate $\mathbf{x}^{(1)}$ using Gauss–Seidel iteration.

   (c) Compare your answers to (a) and (b) with the correct solution $\mathbf{x} = (1,1,1)^T$. Which is closer?

5. For which of the following matrices, will the iteration scheme

$$\mathbf{x}^{(k+1)} = B\mathbf{x}^{(k)} + \mathbf{c}$$

   converge to a solution of $\mathbf{x} = B\mathbf{x} + \mathbf{c}$? Explain.

   (a) $B = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$

   (b) $B = \begin{pmatrix} 0.9 & 1 & 1 \\ 0 & 0.9 & 1 \\ 0 & 0 & 0.9 \end{pmatrix}$

(c) $B = \begin{pmatrix} \frac{1}{2} & 10 & 100 \\ 0 & \frac{1}{2} & 10 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$

(d) $B = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{8} \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} \end{pmatrix}$

(e) $B = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{6} \\ 0 & \frac{1}{6} & \frac{1}{3} \end{pmatrix}$

6. Let $\mathbf{x}$ be the solution of $\mathbf{x} = B\mathbf{x} + \mathbf{c}$. Let $\mathbf{x}^{(0)}$ be an arbitrary vector in R^n and define
$$\mathbf{x}^{(k+1)} = B\mathbf{x}^{(k)} + \mathbf{c}$$
for $k = 0, 1, \ldots$. Prove that if B^m is the zero matrix, then $\mathbf{x}^{(m)} = \mathbf{x}$.

7. Let A be a nonsingular upper triangular matrix. Show that the Jacobi iteration will give the exact solution (assuming no roundoff erros) to $A\mathbf{x} = \mathbf{b}$ after n iterations.

8. For an iterative method based on the splitting $A = C - M$, C nonsingular, show that
$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + C^{-1}\mathbf{r}^{(k)}$$
where $\mathbf{r}^{(k)}$ denotes the residual $\mathbf{b} - A\mathbf{x}^{(k)}$.

9. Let $A = D - L - U$, where D, L, and U are defined as in Gauss–Seidel iteration and let ω be a nonzero scalar. The system $\omega A\mathbf{x} = \omega\mathbf{b}$ can be solved iteratively by splitting ωA into $C - M$, where $C = D - \omega L$. Determine the B and $\mathbf{c}$ corresponding to this splitting. (The constant ω is called a *relaxation parameter*. The case $\omega = 1$ corresponds to Gauss–Seidel iteration.)

10. Let $\mathbf{x}$ be the solution to $\mathbf{x} = B\mathbf{x} + \mathbf{c}$. Let $\mathbf{x}^{(0)}$ be an arbitrary vector in R^n and define
$$\mathbf{x}^{(i+1)} = B\mathbf{x}^{(i)} + \mathbf{c}$$
for $i = 0, 1, \ldots$. If $\|B\| = \alpha < 1$, show that
$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \leq \frac{\alpha}{1-\alpha}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|$$