4

Iterative Methods for Solving Linear Systems

Iterative methods formally yield the solution $\mathbf{x}$ of a linear system after an infinite number of steps. At each step they require the computation of the residual of the system. In the case of a full matrix, their computational cost is therefore of the order of n^2 operations for each iteration, to be compared with an overall cost of the order of $\frac{2}{3}n^3$ operations needed by direct methods. Iterative methods can therefore become competitive with direct methods provided the number of iterations that are required to converge (within a prescribed tolerance) is either independent of n or scales sublinearly with respect to n.

In the case of large sparse matrices, as discussed in Section 3.9, direct methods may be unconvenient due to the dramatic fill-in, although extremely efficient direct solvers can be devised on sparse matrices featuring special structures like, for example, those encountered in the approximation of partial differential equations (see Chapters 12 and 13).

Finally, we notice that, when A is ill-conditioned, a combined use of direct and iterative methods is made possible by preconditioning techniques that will be addressed in Section 4.3.2.

4.1 On the Convergence of Iterative Methods

The basic idea of iterative methods is to construct a sequence of vectors $\mathbf{x}^{(k)}$ that enjoy the property of *convergence*

$$\mathbf{x} = \lim_{k \to \infty} \mathbf{x}^{(k)}, \tag{4.1}$$

where $\mathbf{x}$ is the solution to (3.2). In practice, the iterative process is stopped at the minimum value of n such that $\|\mathbf{x}^{(n)} - \mathbf{x}\| < \varepsilon$, where ε is a fixed tolerance and $\|\cdot\|$ is any convenient vector norm. However, since the exact solution is obviously not available, it is necessary to introduce suitable stopping criteria to monitor the convergence of the iteration (see Section 4.6).

To start with, we consider iterative methods of the form

$$\text{given } \mathbf{x}^{(0)},\ \mathbf{x}^{(k+1)} = \mathrm{B}\mathbf{x}^{(k)} + \mathbf{f}, \quad k \geq 0, \tag{4.2}$$

having denoted by B an $n \times n$ square matrix called the *iteration matrix* and by $\mathbf{f}$ a vector that is obtained from the right hand side $\mathbf{b}$.

Definition 4.1 An iterative method of the form (4.2) is said to be *consistent* with (3.2) if $\mathbf{f}$ and B are such that $\mathbf{x} = \mathrm{B}\mathbf{x} + \mathbf{f}$. Equivalently,

$$\mathbf{f} = (\mathrm{I} - \mathrm{B})\mathrm{A}^{-1}\mathbf{b}.$$

■

Having denoted by

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x} \tag{4.3}$$

the error at the k-th step of the iteration, the condition for convergence (4.1) amounts to requiring that $\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}$ for any choice of the initial datum $\mathbf{x}^{(0)}$ (often called the *initial guess*).

Consistency alone does not suffice to ensure the convergence of the iterative method (4.2), as shown in the following example.

Example 4.1 To solve the linear system $2\mathrm{I}\mathbf{x} = \mathbf{b}$, consider the iterative method

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

which is obviously consistent. This scheme is not convergent for any choice of the initial guess. If, for instance, $\mathbf{x}^{(0)} = \mathbf{0}$, the method generates the sequence $\mathbf{x}^{(2k)} = \mathbf{0}$, $\mathbf{x}^{(2k+1)} = \mathbf{b}$, $k = 0, 1, \ldots$.

On the other hand, if $\mathbf{x}^{(0)} = \frac{1}{2}\mathbf{b}$ the method is convergent. •

Theorem 4.1 *Let* (4.2) *be a consistent method. Then, the sequence of vectors* $\{\mathbf{x}^{(k)}\}$ *converges to the solution of* (3.2) *for any choice of* $\mathbf{x}^{(0)}$ *iff* $\rho(\mathrm{B}) < 1$.

Proof. From (4.3) and the consistency assumption, the recursive relation $\mathbf{e}^{(k+1)} = \mathrm{B}\mathbf{e}^{(k)}$ is obtained. Therefore,

$$\mathbf{e}^{(k)} = \mathrm{B}^k\mathbf{e}^{(0)}, \qquad \forall k = 0, 1, \ldots \tag{4.4}$$

Thus, thanks to Theorem 1.5, it follows that $\lim_{k\to\infty} \mathrm{B}^k\mathbf{e}^{(0)} = \mathbf{0}$ for any $\mathbf{e}^{(0)}$ iff $\rho(\mathrm{B}) < 1$.

Conversely, suppose that $\rho(\mathrm{B}) > 1$, then there exists at least one eigenvalue $\lambda(\mathrm{B})$ with module greater than 1. Let $\mathbf{e}^{(0)}$ be an eigenvector associated with λ; then $\mathrm{B}\mathbf{e}^{(0)} = \lambda\mathbf{e}^{(0)}$ and, therefore, $\mathbf{e}^{(k)} = \lambda^k\mathbf{e}^{(0)}$. As a consequence, $\mathbf{e}^{(k)}$ cannot tend to $\mathbf{0}$ as $k \to \infty$, since $|\lambda| > 1$. ◇

From (1.23) and Theorem 1.4 it follows that a sufficient condition for convergence to hold is that $\|\mathrm{B}\| < 1$, for any consistent matrix norm. It is reasonable

to expect that the convergence is faster when $\rho(\mathrm{B})$ is smaller so that an estimate of $\rho(\mathrm{B})$ might provide a sound indication of the convergence of the algorithm. Other remarkable quantities in convergence analysis are contained in the following definition.

Definition 4.2 Let B be the iteration matrix. We call:

1. $\|\mathrm{B}^m\|$ the *convergence factor* after m steps of the iteration;
2. $\|\mathrm{B}^m\|^{1/m}$ the *average convergence factor* after m steps;
3. $R_m(\mathrm{B}) = -\frac{1}{m}\log\|\mathrm{B}^m\|$ the *average convergence rate* after m steps.

■

These quantities are too expensive to compute since they require evaluating B^m. Therefore, it is usually preferred to estimate the *asymptotic convergence rate*, which is defined as

$$R(\mathrm{B}) = \lim_{k\to\infty} R_k(\mathrm{B}) = -\log\rho(\mathrm{B}), \tag{4.5}$$

where Property 1.14 has been accounted for. In particular, if B were symmetric, we would have

$$R_m(\mathrm{B}) = -\frac{1}{m}\log\|\mathrm{B}^m\|_2 = -\log\rho(\mathrm{B}).$$

In the case of nonsymmetric matrices, $\rho(\mathrm{B})$ sometimes provides an overoptimistic estimate of $\|\mathrm{B}^m\|^{1/m}$ (see [Axe94], Section 5.1). Indeed, although $\rho(\mathrm{B}) < 1$, the convergence to zero of the sequence $\|\mathrm{B}^m\|$ might be nonmonotone (see Exercise 1). We finally notice that, due to (4.5), $\rho(\mathrm{B})$ is the *asymptotic convergence factor*. Criteria for estimating the quantities defined so far will be addressed in Section 4.6.

Remark 4.1 The iterations introduced in (4.2) are a special instance of iterative methods of the form

$$\mathbf{x}^{(0)} = \mathbf{f}_0(\mathrm{A},\mathbf{b}),$$

$$\mathbf{x}^{(n+1)} = \mathbf{f}_{n+1}(\mathbf{x}^{(n)},\mathbf{x}^{(n-1)},\ldots,\mathbf{x}^{(n-m)},\mathrm{A},\mathbf{b}),\ \text{for } n \geq m,$$

where $\mathbf{f}_i$ and $\mathbf{x}^{(m)},\ldots,\mathbf{x}^{(1)}$ are given functions and vectors, respectively. The number of steps which the current iteration depends on is called the *order of the method*. If the functions $\mathbf{f}_i$ are independent of the step index i, the method is called *stationary*, otherwise it is *nonstationary*. Finally, if $\mathbf{f}_i$ depends linearly on $\mathbf{x}^{(0)},\ldots,\mathbf{x}^{(m)}$, the method is called *linear*, otherwise it is *nonlinear*.

In the light of these definitions, the methods considered so far are therefore *stationary linear iterative methods of first order*. In Section 4.3, examples of nonstationary linear methods will be provided. ■

4.2 Linear Iterative Methods

A general technique to devise consistent linear iterative methods is based on an additive *splitting* of the matrix A of the form A=P−N, where P and N are two suitable matrices and P is nonsingular. For reasons that will be clear in the later sections, P is called *preconditioning matrix* or *preconditioner*.

Precisely, given $\mathbf{x}^{(0)}$, one can compute $\mathbf{x}^{(k)}$ for $k \geq 1$, solving the systems

$$\mathrm{P}\mathbf{x}^{(k+1)} = \mathrm{N}\mathbf{x}^{(k)} + \mathbf{b}, \quad k \geq 0. \tag{4.6}$$

The iteration matrix of method (4.6) is $\mathrm{B} = \mathrm{P}^{-1}\mathrm{N}$, while $\mathbf{f} = \mathrm{P}^{-1}\mathbf{b}$. Alternatively, (4.6) can be written in the form

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathrm{P}^{-1}\mathbf{r}^{(k)}, \tag{4.7}$$

where

$$\mathbf{r}^{(k)} = \mathbf{b} - \mathrm{A}\mathbf{x}^{(k)} \tag{4.8}$$

denotes the *residual* vector at step k. Relation (4.7) outlines the fact that a linear system, with coefficient matrix P, must be solved to update the solution at step $k+1$. Thus P, besides being nonsingular, ought to be easily invertible, in order to keep the overall computational cost low. (Notice that, if P were equal to A and N=0, method (4.7) would converge in one iteration, but at the same cost of a direct method).

Let us mention two results that ensure convergence of the iteration (4.7), provided suitable conditions on the splitting of A are fulfilled (for their proof, we refer to [Hac94]).

Property 4.1 *Let* $\mathrm{A} = \mathrm{P} - \mathrm{N}$, *with* A *and* P *symmetric and positive definite. If the matrix* $2\mathrm{P} - \mathrm{A}$ *is positive definite, then the iterative method defined in* (4.7) *is convergent for any choice of the initial datum* $\mathbf{x}^{(0)}$ *and*

$$\rho(\mathrm{B}) = \|\mathrm{B}\|_{\mathrm{A}} = \|\mathrm{B}\|_{\mathrm{P}} < 1.$$

Moreover, the convergence of the iteration is monotone with respect to the norms $\|\cdot\|_{\mathrm{P}}$ *and* $\|\cdot\|_{\mathrm{A}}$ *(i.e.,* $\|\mathbf{e}^{(k+1)}\|_{\mathrm{P}} < \|\mathbf{e}^{(k)}\|_{\mathrm{P}}$ *and* $\|\mathbf{e}^{(k+1)}\|_{\mathrm{A}} < \|\mathbf{e}^{(k)}\|_{\mathrm{A}}$ $k = 0, 1, \ldots$*).*

Property 4.2 *Let* $\mathrm{A} = \mathrm{P} - \mathrm{N}$ *with* A *being symmetric and positive definite. If the matrix* $\mathrm{P} + \mathrm{P}^T - \mathrm{A}$ *is positive definite, then* P *is invertible, the iterative method defined in* (4.7) *is monotonically convergent with respect to norm* $\|\cdot\|_{\mathrm{A}}$ *and* $\rho(\mathrm{B}) \leq \|\mathrm{B}\|_{\mathrm{A}} < 1$.

4.2.1 Jacobi, Gauss-Seidel and Relaxation Methods

In this section we consider some classical linear iterative methods.

If the diagonal entries of A are nonzero, we can single out in each equation the corresponding unknown, obtaining the equivalent linear system

$$x_i = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j \right], \qquad i = 1, \ldots, n. \tag{4.9}$$

In the Jacobi method, once an arbitrarily initial guess $\mathbf{x}^{(0)}$ has been chosen, $\mathbf{x}^{(k+1)}$ is computed by the formulae

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right], \; i = 1, \ldots, n. \tag{4.10}$$

This amounts to performing the following splitting for A

$$\mathrm{P} = \mathrm{D}, \; \mathrm{N} = \mathrm{D} - \mathrm{A} = \mathrm{E} + \mathrm{F},$$

where D is the diagonal matrix of the diagonal entries of A, E is the lower triangular matrix of entries $e_{ij} = -a_{ij}$ if $i > j$, $e_{ij} = 0$ if $i \leq j$, and F is the upper triangular matrix of entries $f_{ij} = -a_{ij}$ if $j > i$, $f_{ij} = 0$ if $j \leq i$. As a consequence, $\mathrm{A} = \mathrm{D} - (\mathrm{E} + \mathrm{F})$.

The iteration matrix of the Jacobi method is thus given by

$$\mathrm{B}_J = \mathrm{D}^{-1}(\mathrm{E} + \mathrm{F}) = \mathrm{I} - \mathrm{D}^{-1}\mathrm{A}. \tag{4.11}$$

A generalization of the Jacobi method is the over-relaxation method (or JOR), in which, having introduced a relaxation parameter ω, (4.10) is replaced by

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

The corresponding iteration matrix is

$$\mathrm{B}_{J_\omega} = \omega \mathrm{B}_J + (1 - \omega)\mathrm{I}. \tag{4.12}$$

In the form (4.7), the JOR method corresponds to

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \omega \mathrm{D}^{-1} \mathbf{r}^{(k)}.$$

This method is consistent for any $\omega \neq 0$ and for $\omega = 1$ it coincides with the Jacobi method.

The Gauss-Seidel method differs from the Jacobi method in the fact that at the $k+1$-th step the available values of $x_i^{(k+1)}$ are being used to update the solution, so that, instead of (4.10), one has

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right], \; i = 1, \ldots, n. \quad (4.13)$$

This method amounts to performing the following splitting for A

$$\mathrm{P} = \mathrm{D} - \mathrm{E}, \; \mathrm{N} = \mathrm{F},$$

and the associated iteration matrix is

$$\mathrm{B}_{GS} = (\mathrm{D} - \mathrm{E})^{-1}\mathrm{F}. \quad (4.14)$$

Starting from Gauss-Seidel method, in analogy to what was done for Jacobi iterations, we introduce the successive over-relaxation method (or SOR method)

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right] + (1-\omega) x_i^{(k)}, \quad (4.15)$$

for $i = 1, \ldots, n$. The method (4.15) can be written in vector form as

$$(\mathrm{I} - \omega \mathrm{D}^{-1}\mathrm{E})\mathbf{x}^{(k+1)} = [(1-\omega)\mathrm{I} + \omega \mathrm{D}^{-1}\mathrm{F}]\mathbf{x}^{(k)} + \omega \mathrm{D}^{-1}\mathbf{b}, \quad (4.16)$$

from which the iteration matrix is

$$\mathrm{B}(\omega) = (\mathrm{I} - \omega \mathrm{D}^{-1}\mathrm{E})^{-1}[(1-\omega)\mathrm{I} + \omega \mathrm{D}^{-1}\mathrm{F}]. \quad (4.17)$$

Multiplying by D both sides of (4.16) and recalling that $\mathrm{A} = \mathrm{D} - (\mathrm{E} + \mathrm{F})$ yields the following form (4.7) of the SOR method

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \left(\frac{1}{\omega}\mathrm{D} - \mathrm{E} \right)^{-1} \mathbf{r}^{(k)}.$$

It is consistent for any $\omega \neq 0$ and for $\omega = 1$ it coincides with Gauss-Seidel method. In particular, if $\omega \in (0,1)$ the method is called under-relaxation, while if $\omega > 1$ it is called over-relaxation.

4.2.2 Convergence Results for Jacobi and Gauss-Seidel Methods

There exist special classes of matrices for which it is possible to state a priori some convergence results for the methods examined in the previous section. The first result in this direction is the following.

Theorem 4.2 *If* A *is a strictly diagonally dominant matrix by rows, the Jacobi and Gauss-Seidel methods are convergent.*

Proof. Let us prove the part of the theorem concerning the Jacobi method, while for the Gauss-Seidel method we refer to [Axe94]. Since A is strictly diagonally dominant by rows, $|a_{ii}| > \sum_{j=1}^{n} |a_{ij}|$ for $j \neq i$ and $i = 1, \ldots, n$. As a consequence, $\|\mathrm{B}_J\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1, j\neq i}^{n} |a_{ij}|/|a_{ii}| < 1$, so that the Jacobi method is convergent. ◇

Theorem 4.3 *If* A *and* $2\mathrm{D} - \mathrm{A}$ *are symmetric and positive definite matrices, then the Jacobi method is convergent and* $\rho(\mathrm{B}_J) = \|\mathrm{B}_J\|_\mathrm{A} = \|\mathrm{B}_J\|_\mathrm{D}$.

Proof. The theorem follows from Property 4.1 taking P=D. ◇

In the case of the JOR method, the assumption on $2\mathrm{D} - \mathrm{A}$ can be removed, yielding the following result.

Theorem 4.4 *If* A *is symmetric positive definite, then the JOR method is convergent if* $0 < \omega < 2/\rho(\mathrm{D}^{-1}\mathrm{A})$.

Proof. The result immediately follows from (4.12) and noting that A has real positive eigenvalues. ◇

Concerning the Gauss-Seidel method, the following result holds.

Theorem 4.5 *If* A *is symmetric positive definite, the Gauss-Seidel method is monotonically convergent with respect to the norm* $\|\cdot\|_\mathrm{A}$.

Proof. We can apply Property 4.2 to the matrix P=D−E, upon checking that $\mathrm{P} + \mathrm{P}^T - \mathrm{A}$ is positive definite. Indeed

$$\mathrm{P} + \mathrm{P}^T - \mathrm{A} = 2\mathrm{D} - \mathrm{E} - \mathrm{F} - \mathrm{A} = \mathrm{D},$$

having observed that $(\mathrm{D} - \mathrm{E})^T = \mathrm{D} - \mathrm{F}$. We conclude by noticing that D is positive definite, since it is the diagonal of A. ◇

Finally, if A is tridiagonal (or block tridiagonal), it can be shown that

$$\rho(\mathrm{B}_{GS}) = \rho^2(\mathrm{B}_J) \tag{4.18}$$

(see [You71] for the proof). From (4.18) we can conclude that both methods converge or fail to converge at the same time. In the former case, the Gauss-Seidel method is more rapidly convergent than the Jacobi method, and the asymptotic convergence rate of the Gauss-Seidel method is twice than that of the Jacobi method. In particular, if A is tridiagonal and symmetric positive definite, Theorem 4.5 implies convergence of the Gauss-Seidel method, and (4.18) ensures convergence also for the Jacobi method.

Relation (4.18) holds even if A enjoys the following *A-property*.

Definition 4.3 A *consistently ordered* matrix $\mathrm{M} \in \mathbb{R}^{n\times n}$ (that is, a matrix such that $\alpha\mathrm{D}^{-1}\mathrm{E} + \alpha^{-1}\mathrm{D}^{-1}\mathrm{F}$, for $\alpha \neq 0$, has eigenvalues that do not depend

on α, where $M = D - E - F$, $D = \text{diag}(m_{11}, \ldots, m_{nn})$, E and F are strictly lower and upper triangular matrices, respectively) enjoys the A-property if it can be partitioned in the 2×2 block form

$$M = \begin{bmatrix} \tilde{D}_1 & M_{12} \\ M_{21} & \tilde{D}_2 \end{bmatrix},$$

where $\tilde{D}_1$ and $\tilde{D}_2$ are diagonal matrices. ■

When dealing with general matrices, no a priori conclusions on the convergence properties of the Jacobi and Gauss-Seidel methods can be drawn, as shown in Example 4.2.

Example 4.2 Consider the 3×3 linear systems of the form $A_i\mathbf{x} = \mathbf{b}_i$, where $\mathbf{b}_i$ is always taken in such a way that the solution of the system is the unit vector, and the matrices A_i are

$$A_1 = \begin{bmatrix} 3 & 0 & 4 \\ 7 & 4 & 2 \\ -1 & 1 & 2 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} -3 & 3 & -6 \\ -4 & 7 & -8 \\ 5 & 7 & -9 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 4 & 1 & 1 \\ 2 & -9 & 0 \\ 0 & -8 & -6 \end{bmatrix}, \qquad A_4 = \begin{bmatrix} 7 & 6 & 9 \\ 4 & 5 & -4 \\ -7 & -3 & 8 \end{bmatrix}.$$

It can be checked that the Jacobi method does fail to converge for A_1 ($\rho(B_J) = 1.33$), while the Gauss-Seidel scheme is convergent. Conversely, in the case of A_2, the Jacobi method is convergent, while the Gauss-Seidel method fails to converge ($\rho(B_{GS}) = 1.\bar{1}$). In the remaining two cases, the Jacobi method is more slowly convergent than the Gauss-Seidel method for matrix A_3 ($\rho(B_J) = 0.44$ against $\rho(B_{GS}) = 0.018$), and the converse is true for A_4 ($\rho(B_J) = 0.64$ while $\rho(B_{GS}) = 0.77$). •

We conclude the section with the following result.

Theorem 4.6 *If the Jacobi method is convergent, then the JOR method converges if* $0 < \omega \leq 1$.

Proof. From (4.12) we obtain that the eigenvalues of B_{J_ω} are

$$\mu_k = \omega\lambda_k + 1 - \omega, \qquad k = 1, \ldots, n,$$

where λ_k are the eigenvalues of B_J. Then, recalling the Euler formula for the representation of a complex number, we let $\lambda_k = r_k e^{i\theta_k}$ and get

$$|\mu_k|^2 = \omega^2 r_k^2 + 2\omega r_k \cos(\theta_k)(1-\omega) + (1-\omega)^2 \leq (\omega r_k + 1 - \omega)^2,$$

which is less than 1 if $0 < \omega \leq 1$. ◇

4.2.3 Convergence Results for the Relaxation Method

The following result provides a necessary condition on ω in order the SOR method to be convergent.