Chapter 8

Iterative Methods for Solving Linear Systems

8.1 Convergence of Sequences of Vectors and Matrices

In Chapter 4 we have discussed some of the main methods for solving systems of linear equations. These methods are *direct methods*, in the sense that they yield exact solutions (assuming infinite precision!).

Another class of methods for solving linear systems consists in approximating solutions using *iterative methods*.

The basic idea is this: Given a linear system $Ax = b$ (with A a square invertible matrix), find another matrix B and a vector c, such that

1. The matrix $I - B$ is invertible
2. The unique solution $\widetilde{x}$ of the system $Ax = b$ is identical to the unique solution $\widetilde{u}$ of the system

$$u = Bu + c,$$

and then, starting from any vector u_0, compute the sequence (u_k) given by

$$u_{k+1} = Bu_k + c, \quad k \in \mathbb{N}.$$

Under certain conditions (to be clarified soon), the sequence (u_k) converges to a limit $\widetilde{u}$ which is the unique solution of $u = Bu + c$, and thus of $Ax = b$.

Let $(E, \| \|)$ be a normed vector space. Recall that a sequence (u_k) of vectors $u_k \in E$ *converges to a limit* $u \in E$, if for every $\epsilon > 0$, there some natural number N such that

$$\|u_k - u\| \leq \epsilon, \quad \textit{for all } k \geq N.$$

We write

$$u = \lim_{k \mapsto \infty} u_k.$$

If E is a finite-dimensional vector space and $\dim(E) = n$, we know from Theorem 6.3 that any two norms are equivalent, and if we choose the norm $\| \|_\infty$, we see that the convergence of the sequence of vectors u_k is equivalent to the convergence of the n sequences of scalars formed by the components of these vectors (over any basis).

The same property applies to the finite-dimensional vector space $\mathrm{M}_{m,n}(K)$ of $m \times n$ matrices (with $K = \mathbb{R}$ or $K = \mathbb{C}$), which means that the convergence of a sequence of matrices $A_k = (a_{ij}^{(k)})$ is equivalent to the convergence of the $m \times n$ sequences of scalars $(a_{ij}^{(k)})$, with i, j fixed $(1 \leq i \leq m,\ 1 \leq j \leq n)$.

The first theorem below gives a necessary and sufficient condition for the sequence (B^k) of powers of a matrix B to converge to the zero matrix.

Recall that the spectral radius $\rho(B)$ of a matrix B is the maximum of the moduli $|\lambda_i|$ of the eigenvalues of B.

Theorem 8.1. *For any square matrix B, the following conditions are equivalent:*

(1) $\lim_{k\mapsto\infty} B^k = 0$,

(2) $\lim_{k\mapsto\infty} B^k v = 0$, *for all vectors* v,

(3) $\rho(B) < 1$,

(4) $\|B\| < 1$, *for some subordinate matrix norm* $\|\ \|$.

The following proposition is needed to study the rate of convergence of iterative methods.

Proposition 8.2. *For every square matrix B and every matrix norm $\|\ \|$, we have*

$$\lim_{k\mapsto\infty} \|B^k\|^{1/k} = \rho(B).$$

We now apply the above results to the convergence of iterative methods.

8.2 Convergence of Iterative Methods

Recall that iterative methods for solving a linear system $Ax = b$ (with A invertible) consists in finding some matrix B and some vector c, such that $I - B$ is invertible, and the unique solution $\widetilde{x}$ of $Ax = b$ is equal to the unique solution $\widetilde{u}$ of $u = Bu + c$.

Then, starting from any vector u_0, compute the sequence (u_k) given by

$$u_{k+1} = Bu_k + c, \quad k \in \mathbb{N},$$

and say that the iterative method is *convergent* iff

$$\lim_{k \mapsto \infty} u_k = \widetilde{u},$$

for *every* initial vector u_0.

Here is a fundamental criterion for the convergence of any iterative methods based on a matrix B, called the *matrix of the iterative method*.

Theorem 8.3. *Given a system $u = Bu + c$ as above, where $I - B$ is invertible, the following statements are equivalent:*

(1) The iterative method is convergent.

(2) $\rho(B) < 1$.

(3) $\|B\| < 1$, for some subordinate matrix norm $\| \|$.

The next proposition is needed to compare the rate of convergence of iterative methods.

It shows that *asymptotically, the error vector $e_k = B^k e_0$ behaves at worst like $(\rho(B))^k$*.

Proposition 8.4. *Let $\| \|$ be any vector norm, let B be a matrix such that $I - B$ is invertible, and let $\widetilde{u}$ be the unique solution of $u = Bu + c$.*

(1) If (u_k) is any sequence defined iteratively by

$$u_{k+1} = Bu_k + c, \quad k \in \mathbb{N},$$

then

$$\lim_{k \mapsto \infty} \left[\sup_{\|u_0 - \widetilde{u}\| = 1} \|u_k - \widetilde{u}\|^{1/k} \right] = \rho(B).$$

(2) Let B_1 and B_2 be two matrices such that $I - B_1$ and $I - B_2$ are invertibe, assume that both $u = B_1u + c_1$ and $u = B_2u + c_2$ have the same unique solution $\widetilde{u}$, and consider any two sequences (u_k) and (v_k) defined inductively by

$$u_{k+1} = B_1u_k + c_1$$
$$v_{k+1} = B_2v_k + c_2,$$

with $u_0 = v_0$. If $\rho(B_1) < \rho(B_2)$, then for any $\epsilon > 0$, there is some integer $N(\epsilon)$, such that for all $k \geq N(\epsilon)$, we have

$$\sup_{\|u_0 - \widetilde{u}\| = 1} \left[\frac{\|v_k - \widetilde{u}\|}{\|u_k - \widetilde{u}\|}\right]^{1/k} \geq \frac{\rho(B_2)}{\rho(B_1) + \epsilon}.$$

In light of the above, we see that when we investigate new iterative methods, we have to deal with the following two problems:

1. Given an iterative method with matrix B, determine whether the method is convergent. This involves determining whether $\rho(B) < 1$, or equivalently whether there is a subordinate matrix norm such that $\|B\| < 1$.

By Proposition 6.8, this implies that $I - B$ is invertible (since $\| - B\| = \|B\|$, Proposition 6.8 applies).

2. Given two convergent iterative methods, compare them. The iterative method which is faster is that whose matrix has the smaller spectral radius.

We now discuss three iterative methods for solving linear systems:

1. Jacobi's method
2. Gauss-Seidel's method
3. The relaxation method.

8.3 Description of the Methods of Jacobi, Gauss-Seidel, and Relaxation

The methods described in this section are instances of the following scheme: Given a linear system $Ax = b$, with A invertible, suppose we can write A in the form

$$A = M - N,$$

with M invertible, and "easy to invert," which means that M is close to being a diagonal or a triangular matrix (perhaps by blocks).

Then, $Au = b$ is equivalent to

$$Mu = Nu + b,$$

that is,

$$u = M^{-1}Nu + M^{-1}b.$$

Therefore, we are in the situation described in the previous sections with $B = M^{-1}N$ and $c = M^{-1}b$.

In fact, since $A = M - N$, we have

$$B = M^{-1}N = M^{-1}(M - A) = I - M^{-1}A,$$

which shows that $I - B = M^{-1}A$ is invertible.

The iterative method associated with the matrix $B = M^{-1}N$ is given by

$$u_{k+1} = M^{-1}Nu_k + M^{-1}b, \quad k \geq 0,$$

starting from any arbitrary vector u_0.

From a practical point of view, we do not invert M, and instead we solve iteratively the systems

$$Mu_{k+1} = Nu_k + b, \quad k \geq 0.$$

Various methods correspond to various ways of choosing M and N from A. The first two methods choose M and N as disjoint submatrices of A, but the relaxation method allows some overlapping of M and N.

To describe the various choices of M and N, it is convenient to write A in terms of three submatrices D, E, F, as

$$A = D - E - F,$$

where the only nonzero entries in D are the diagonal entries in A, the only nonzero entries in E are entries in A below the the diagonal, and the only nonzero entries in F are entries in A above the diagonal.

More explicitly, if

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n-1} & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n-1\,1} & a_{n-1\,2} & a_{n-1\,3} & \cdots & a_{n-1\,n-1} & a_{n-1\,n} \\ a_{n\,1} & a_{n\,2} & a_{n\,3} & \cdots & a_{n\,n-1} & a_{n\,n} \end{pmatrix},$$

then

$$D = \begin{pmatrix} a_{11} & 0 & 0 & \cdots & 0 & 0 \\ 0 & a_{22} & 0 & \cdots & 0 & 0 \\ 0 & 0 & a_{33} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-1\,n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & a_{n\,n} \end{pmatrix},$$

$$-E = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ a_{n-1\,1} & a_{n-1\,2} & a_{n-1\,3} & \ddots & 0 & 0 \\ a_{n\,1} & a_{n\,2} & a_{n\,3} & \cdots & a_{n\,n-1} & 0 \end{pmatrix},$$

$$-F = \begin{pmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n-1} & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n-1} & a_{2n} \\ 0 & 0 & 0 & \ddots & a_{3n-1} & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{n-1\,n} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{pmatrix}.$$

In *Jacobi's method*, we assume that all diagonal entries in A are nonzero, and we pick

$$\begin{aligned} M &= D \\ N &= E + F, \end{aligned}$$

so that

$$B = M^{-1}N = D^{-1}(E + F) = I - D^{-1}A.$$

As a matter of notation, we let

$$J = I - D^{-1}A = D^{-1}(E + F),$$

which is called *Jacobi's matrix*.

The corresponding method, *Jacobi's iterative method*, computes the sequence (u_k) using the recurrence

$$u_{k+1} = D^{-1}(E+F)u_k + D^{-1}b, \quad k \geq 0.$$

In practice, we iteratively solve the systems

$$Du_{k+1} = (E+F)u_k + b, \quad k \geq 0.$$

If we write $u_k = (u_1^k, \ldots, u_n^k)$, we solve iteratively the following system:

$$\begin{array}{ccccccccc}
a_{11}u_1^{k+1} & = & & -a_{12}u_2^k & -a_{13}u_3^k & \cdots & -a_{1n}u_n^k & + b_1 \\
a_{22}u_2^{k+1} & = & -a_{21}u_1^k & & -a_{23}u_3^k & \cdots & -a_{2n}u_n^k & + b_2 \\
\vdots & \vdots & \vdots & & & & & \\
a_{n-1\,n-1}u_{n-1}^{k+1} & = & -a_{n-1\,1}u_1^k & \cdots & -a_{n-1\,n-2}u_{n-2}^k & & -a_{n-1\,n}u_n^k & + b_{n-1} \\
a_{n\,n}u_n^{k+1} & = & -a_{n\,1}u_1^k & -a_{n\,2}u_2^k & \cdots & -a_{n\,n-1}u_{n-1}^k & & + b_n
\end{array}.$$

Observe that we can try to "speed up" the method by using the new value u_1^{k+1} instead of u_1^k in solving for u_2^{k+2} using the second equations, and more generally, use $u_1^{k+1}, \ldots, u_{i-1}^{k+1}$ instead of $u_1^k, \ldots, u_{i-1}^k$ in solving for u_i^{k+1} in the ith equation.

This observation leads to the system

$$\begin{array}{ccccccccc}
a_{11}u_1^{k+1} & = & & -a_{12}u_2^k & -a_{13}u_3^k & \cdots & -a_{1n}u_n^k & + b_1 \\
a_{22}u_2^{k+1} & = & -a_{21}u_1^{k+1} & & -a_{23}u_3^k & \cdots & -a_{2n}u_n^k & + b_2 \\
\vdots & \vdots & \vdots & & & & & \\
a_{n-1\,n-1}u_{n-1}^{k+1} & = & -a_{n-1\,1}u_1^{k+1} & \cdots & -a_{n-1\,n-2}u_{n-2}^{k+1} & & -a_{n-1\,n}u_n^k & + b_{n-1} \\
a_{n\,n}u_n^{k+1} & = & -a_{n\,1}u_1^{k+1} & -a_{n\,2}u_2^{k+1} & \cdots & -a_{n\,n-1}u_{n-1}^{k+1} & & + b_n,
\end{array}$$

which, in matrix form, is written

$$Du_{k+1} = Eu_{k+1} + Fu_k + b.$$

Because D is invertible and E is lower triangular, the matrix $D - E$ is invertible, so the above equation is equivalent to

$$u_{k+1} = (D - E)^{-1}Fu_k + (D - E)^{-1}b, \quad k \geq 0.$$

The above corresponds to choosing M and N to be

$$\begin{aligned} M &= D - E \\ N &= F, \end{aligned}$$

and the matrix B is given by

$$B = M^{-1}N = (D - E)^{-1}F.$$

Since $M = D - E$ is invertible, we know that $I - B = M^{-1}A$ is also invertible.

The method that we just described is the *iterative method of Gauss-Seidel*, and the matrix B is called the *matrix of Gauss-Seidel* and denoted by $\mathcal{L}_1$, with

$$\mathcal{L}_1 = (D - E)^{-1}F.$$

One of the advantages of the method of Gauss-Seidel is that is requires only half of the memory used by Jacobi's method, since we only need

$$u_1^{k+1}, \ldots, u_{i-1}^{k+1}, u_{i+1}^{k}, \ldots, u_n^{k}$$

to compute u_i^{k+1}.

We also show that in certain important cases (for example, if A is a tridiagonal matrix), the method of Gauss-Seidel converges faster than Jacobi's method (in this case, they both converge or diverge simultaneously).

The new ingredient in the *relaxation method* is to incorporate part of the matrix D into N: we define M and N by

$$M = \frac{D}{\omega} - E$$
$$N = \frac{1-\omega}{\omega} D + F,$$

where $\omega \neq 0$ is a real parameter to be suitably chosen.

Actually, we show in Section 8.4 that for the relaxation method to converge, we must have $\omega \in (0, 2)$.

Note that the case $\omega = 1$ corresponds to the method of Gauss-Seidel.

If we assume that all diagonal entries of D are nonzero, the matrix M is invertible.

The matrix B is denoted by $\mathcal{L}_\omega$ and called the *matrix of relaxation*, with

$$\begin{aligned}\mathcal{L}_\omega &= \left(\frac{D}{\omega} - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right)\\ &= (D - \omega E)^{-1}((1-\omega)D + \omega F).\end{aligned}$$

The number ω is called the *parameter of relaxation*. When $\omega > 1$, the relaxation method is known as *successive overrelaxation*, abbreviated as *SOR*.

At first glance, the relaxation matrix $\mathcal{L}_\omega$ seems at lot more complicated than the Gauss-Seidel matrix $\mathcal{L}_1$, but the iterative system associated with the relaxation method is very similar to the method of Gauss-Seidel, and is quite simple.

Indeed, the system associated with the relaxation method is given by

$$\left(\frac{D}{\omega} - E\right)u_{k+1} = \left(\frac{1-\omega}{\omega}D + F\right)u_k + b,$$

which is equivalent to

$$(D - \omega E)u_{k+1} = ((1-\omega)D + \omega F)u_k + \omega b,$$

and can be written

$$Du_{k+1} = Du_k - \omega(Du_k - Eu_{k+1} - Fu_k - b).$$

Explicitly, this is the system

$$\begin{aligned}
a_{11}u_1^{k+1} &= a_{11}u_1^k - \omega(a_{11}u_1^k + a_{12}u_2^k + a_{13}u_3^k + \cdots + a_{1n-2}u_{n-2}^k + a_{1n-1}u_{n-1}^k + a_{1n}u_n^k - b_1)\\
a_{22}u_2^{k+1} &= a_{22}u_2^k - \omega(a_{21}u_1^{k+1} + a_{22}u_2^k + a_{23}u_3^k + \cdots + a_{2n-2}u_{n-2}^k + a_{2n-1}u_{n-1}^k + a_{2n}u_n^k - b_2)\\
&\vdots\\
a_{n\,n}u_n^{k+1} &= a_{n\,n}u_n^k - \omega(a_{n\,1}u_1^{k+1} + a_{n\,2}u_2^{k+1} + \cdots + a_{n\,n-2}u_{n-2}^{k+1} + a_{n\,n-1}u_{n-1}^{k+1} + a_{n\,n}u_n^k - b_n).
\end{aligned}$$

What remains to be done is to find conditions that ensure the convergence of the relaxation method (and the Gauss-Seidel method), that is:

1. Find conditions on ω, namely some interval $I \subseteq \mathbb{R}$ so that $\omega \in I$ implies $\rho(\mathcal{L}_\omega) < 1$; we will prove that $\omega \in (0, 2)$ is a necessary condition.
2. Find if there exist some *optimal value* ω_0 of $\omega \in I$, so that
$$\rho(\mathcal{L}_{\omega_0}) = \inf_{\omega \in I} \rho(\mathcal{L}_\omega).$$

We will give partial answers to the above questions in the next section.

It is also possible to extend the methods of this section by using *block decompositions* of the form $A = D - E - F$, where D, E, and F consist of blocks, and with D an invertible block-diagonal matrix.

8.4 Convergence of the Methods of Jacobi, Gauss-Seidel, and Relaxation

We begin with a general criterion for the convergence of an iterative method associated with a (complex) Hermitian, positive, definite matrix, $A = M - N$. Next, we apply this result to the relaxation method.

Proposition 8.5. *Let A be any Hermitian, positive, definite matrix, written as*

$$A = M - N,$$

with M invertible. Then, $M^* + N$ is Hermitian, and if it is positive, definite, then

$$\rho(M^{-1}N) < 1,$$

so that the iterative method converges.

Now, as in the previous sections, we assume that A is written as $A = D - E - F$, with D invertible, possibly in block form.

The next theorem provides a sufficient condition (which turns out to be also necessary) for the relaxation method to converge (and thus, for the method of Gauss-Seidel to converge).

This theorem is known as the *Ostrowski-Reich theorem*.

Theorem 8.6. *If $A = D - E - F$ is Hermitian, positive, definite, and if $0 < \omega < 2$, then the relaxation method converges. This also holds for a block decomposition of A.*

Remark: What if we allow the parameter ω to be a nonzero complex number $\omega \in \mathbb{C}$? In this case, the relaxation method also converges for $\omega \in \mathbb{C}$, provided that

$$|\omega - 1| < 1.$$

This condition reduces to $0 < \omega < 2$ if ω is real.

Unfortunately, Theorem 8.6 does not apply to Jacobi's method, but in special cases, Proposition 8.5 can be used to prove its convergence.

On the positive side, if a matrix is strictly column (or row) diagonally dominant, then it can be shown that the method of Jacobi and the method of Gauss-Seidel both converge.

The relaxation method also converges if $\omega \in (0, 1]$, but this is not a very useful result because the speed-up of convergence usually occurs for $\omega > 1$.

We now prove that, without any assumption on $A = D - E - F$, other than the fact that A and D are invertible, in order for the relaxation method to converge, we must have $\omega \in (0, 2)$.

Proposition 8.7. *Given any matrix $A = D - E - F$, with A and D invertible, for any $\omega \neq 0$, we have*

$$\rho(\mathcal{L}_\omega) \geq |\omega - 1|.$$

Therefore, the relaxation method (possibly by blocks) does not converge unless $\omega \in (0, 2)$. If we allow ω to be complex, then we must have

$$|\omega - 1| < 1$$

for the relaxation method to converge.

We now consider the case where A is a *tridiagonal matrix*, possibly by blocks.

We begin with the case $\omega = 1$, which is technically easier to deal with.

The following proposition gives us the precise relationship between the spectral radii $\rho(J)$ and $\rho(\mathcal{L}_1)$ of the Jacobi matrix and the Gauss-Seidel matrix.

Proposition 8.8. *Let A be a tridiagonal matrix (possibly by blocks). If $\rho(J)$ is the spectral radius of the Jacobi matrix and $\rho(\mathcal{L}_1)$ is the spectral radius of the Gauss-Seidel matrix, then we have*

$$\rho(\mathcal{L}_1) = (\rho(J))^2.$$

Consequently, the method of Jacobi and the method of Gauss-Seidel both converge or both diverge simultaneously (even when A is tridiagonal by blocks);

when they converge, the method of Gauss-Seidel converges faster than Jacobi's method.

We now consider the more general situation where ω is any real in $(0, 2)$.

Proposition 8.9. *Let A be a tridiagonal matrix (possibly by blocks), and assume that the eigenvalues of the Jacobi matrix are all real. If $\omega \in (0, 2)$, then the method of Jacobi and the method of relaxation both converge or both diverge simultaneously (even when A is tridiagonal by blocks).*

When they converge, the function $\omega \mapsto \rho(\mathcal{L}_\omega)$ (for $\omega \in (0, 2)$) has a unique minimum equal to $\omega_0 - 1$ for

$$\omega_0 = \frac{2}{1 + \sqrt{1 - (\rho(J))^2}},$$

where $1 < \omega_0 < 2$ if $\rho(J) > 0$. We also have $\rho(\mathcal{L}_1) = (\rho(J))^2$, as before.

Combining the results of Theorem 8.6 and Proposition 8.9, we obtain the following result which gives precise information about the spectral radii of the matrices J, $\mathcal{L}_1$, and $\mathcal{L}_\omega$.

Proposition 8.10. *Let A be a tridiagonal matrix (possibly by blocks) which is Hermitian, positive, definite. Then, the methods of Jacobi, Gauss-Seidel, and relaxation, all converge for $\omega \in (0, 2)$. There is a unique optimal relaxation parameter*

$$\omega_0 = \frac{2}{1 + \sqrt{1 - (\rho(J))^2}},$$

such that

$$\rho(\mathcal{L}_{\omega_0}) = \inf_{0<\omega<2} \rho(\mathcal{L}_\omega) = \omega_0 - 1.$$

Furthermore, if $\rho(J) > 0$, *then*

$$\rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_1) = (\rho(J))^2 < \rho(J),$$

and if $\rho(J) = 0$, *then* $\omega_0 = 1$ *and* $\rho(\mathcal{L}_1) = \rho(J) = 0$.

Remark: It is preferable to overestimate rather than underestimate the relaxation parameter when the optimum relaxation parameter is not known exactly.