Chapter 9

Iterative Methods for Solving Linear Systems

9.1 Convergence of Sequences of Vectors and Matrices

In Chapter 7 we 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 in $\mathrm{M}_n(\mathbb{C})$), find another matrix $B \in \mathrm{M}_n(\mathbb{C})$ and a vector $c \in \mathbb{C}^n$, 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 $Ax = b$.

Consequently, it is important to find conditions that ensure the convergence of the above sequences and to have tools to compare the "rate" of convergence of these sequences. Thus, we begin with some general results about the convergence of sequences of vectors and matrices.

Let $(E, \| \ \|)$ be a normed vector space. Recall from Section 8.7 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 8.5 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 9.1. *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 $\| \, \|$.*

Proof. Assume (1) and let $\| \, \|$ be a vector norm on E and $\| \, \|$ be the corresponding matrix norm. For every vector $v \in E$, because $\| \, \|$ is a matrix norm, we have
$$\|B^k v\| \leq \|B^k\| \|v\|,$$
and since $\lim_{k \mapsto \infty} B^k = 0$ means that $\lim_{k \mapsto \infty} \|B^k\| = 0$, we conclude that $\lim_{k \mapsto \infty} \|B^k v\| = 0$, that is, $\lim_{k \mapsto \infty} B^k v = 0$. This proves that (1) implies (2).

Assume (2). If we had $\rho(B) \geq 1$, then there would be some eigenvector $u \, (\neq 0)$ and some eigenvalue λ such that
$$Bu = \lambda u, \quad |\lambda| = \rho(B) \geq 1,$$
but then the sequence $(B^k u)$ would not converge to 0, because $B^k u = \lambda^k u$ and $|\lambda^k| = |\lambda|^k \geq 1$. It follows that (2) implies (3).

Assume that (3) holds, that is, $\rho(B) < 1$. By Proposition 8.12, we can find $\epsilon > 0$ small enough that $\rho(B) + \epsilon < 1$, and a subordinate matrix norm $\| \, \|$ such that
$$\|B\| \leq \rho(B) + \epsilon,$$
which is (4).

Finally, assume (4). Because $\| \, \|$ is a matrix norm,
$$\|B^k\| \leq \|B\|^k,$$
and since $\|B\| < 1$, we deduce that (1) holds. □

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

Proposition 9.2. *For every square matrix $B \in \mathrm{M}_n(\mathbb{C})$ and every matrix norm $\| \, \|$, we have*

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

Proof. We know from Proposition 8.6 that $\rho(B) \leq \|B\|$, and since $\rho(B) = (\rho(B^k))^{1/k}$, we deduce that

$$\rho(B) \leq \|B^k\|^{1/k} \quad \text{for all } k \geq 1,$$

and so

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

Now let us prove that for every $\epsilon > 0$, there is some integer $N(\epsilon)$ such that

$$\|B^k\|^{1/k} \leq \rho(B) + \epsilon \quad \text{for all } k \geq N(\epsilon),$$

which proves that

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

and our proposition.

For any given $\epsilon > 0$, let B_ϵ be the matrix

$$B_\epsilon = \frac{B}{\rho(B) + \epsilon}.$$

Since $\|B_\epsilon\| < 1$, Theorem 9.1 implies that $\lim_{k \mapsto \infty} B_\epsilon^k = 0$. Consequently, there is some integer $N(\epsilon)$ such that for all $k \geq N(\epsilon)$, we have

$$\|B^k\| = \frac{\|B^k\|}{(\rho(B) + \epsilon)^k} \leq 1,$$

which implies that

$$\|B^k\|^{1/k} \leq \rho(B) + \epsilon,$$

as claimed. □

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

9.2 Convergence of Iterative Methods

Recall that iterative methods for solving a linear system $Ax = b$ (with $A \in \mathrm{M}_n(\mathbb{C})$ 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 9.3. *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 $\| \, \|$.

Proof. Define the vector e_k (*error vector*) by

$$e_k = u_k - \widetilde{u},$$

where $\widetilde{u}$ is the unique solution of the system $u = Bu + c$. Clearly, the iterative method is convergent iff

$$\lim_{k \mapsto \infty} e_k = 0.$$

We claim that

$$e_k = B^k e_0, \quad k \geq 0,$$

where $e_0 = u_0 - \widetilde{u}$.

This is proven by induction on k. The base case $k = 0$ is trivial. By the induction hypothesis, $e_k = B^k e_0$, and since $u_{k+1} = Bu_k + c$, we get

$$u_{k+1} - \widetilde{u} = Bu_k + c - \widetilde{u},$$

and because $\widetilde{u} = B\widetilde{u} + c$ and $e_k = B^k e_0$ (by the induction hypothesis), we obtain

$$u_{k+1} - \widetilde{u} = Bu_k - B\widetilde{u} = B(u_k - \widetilde{u}) = Be_k = BB^k e_0 = B^{k+1} e_0,$$

proving the induction step. Thus, the iterative method converges iff

$$\lim_{k \mapsto \infty} B^k e_0 = 0.$$

Consequently, our theorem follows by Theorem 9.1. □

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 9.4. *Let $\| \ \|$ be any vector norm, let $B \in \mathrm{M}_n(\mathbb{C})$ 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 $u = B_1 u + c_1$ and $u = B_2 u + c_2$ have the same unique solution $\widetilde{u}$, and consider any two sequences (u_k) and (v_k) defined inductively by

$$\begin{aligned} u_{k+1} &= B_1 u_k + c_1 \\ v_{k+1} &= B_2 v_k + c_2, \end{aligned}$$

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

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

Proof. Let $\| \ \|$ be the subordinate matrix norm. Recall that

$$u_k - \widetilde{u} = B^k e_0,$$

with $e_0 = u_0 - \widetilde{u}$. For every $k \in \mathbb{N}$, we have

$$(\rho(B_1))^k = \rho(B_1^k) \leq \|B_1^k\| = \sup_{\|e_0\|=1} \|B_1^k e_0\|,$$

which implies

$$\rho(B_1) = \sup_{\|e_0\|=1} \|B_1^k e_0\|^{1/k} = \|B_1^k\|^{1/k},$$

and Statement (1) follows from Proposition 9.2.

Because $u_0 = v_0$, we have

$$\begin{aligned} u_k - \widetilde{u} &= B_1^k e_0 \\ v_k - \widetilde{u} &= B_2^k e_0, \end{aligned}$$

with $e_0 = u_0 - \widetilde{u} = v_0 - \widetilde{u}$. Again, by Proposition 9.2, for every $\epsilon > 0$, there is some natural number $N(\epsilon)$ such that if $k \geq N(\epsilon)$, then

$$\sup_{\|e_0\|=1} \|B_1^k e_0\|^{1/k} \leq \rho(B_1) + \epsilon.$$

Furthermore, for all $k \geq N(\epsilon)$, there exists a vector $e_0 = e_0(k)$ such that

$$\|e_0\| = 1 \quad \text{and} \quad \|B_2^k e_0\|^{1/k} = \|B_2^k\|^{1/k} \geq \rho(B_2),$$

which implies Statement (2). □

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 8.11, this implies that $I - B$ is invertible (since $\| - B\| = \|B\|$, Proposition 8.11 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.

9.3 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

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

which shows that $I - B = M^{-1}A$ is invertible. The iterative method associated with the matrix $B = M^{-1}N$ is given by

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

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 by $(*)$,

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

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:

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

In `Matlab` one step of Jacobi iteration is achieved by the following function:

```
function v = Jacobi2(A,b,u)
n = size(A,1);
v = zeros(n,1);
   for i = 1:n
      v(i,1)  = u(i,1) + (-A(i,:)*u + b(i))/A(i,i);
   end
end
```

In order to run m iteration steps, run the following function:

```
function u = jacobi(A,b,u0,m)
  u = u0;
  for j = 1:m
    u = Jacobi2(A,b,u);
  end
end
```

Example 9.1. Consider the linear system

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 25 \\ -24 \\ 21 \\ -15 \end{pmatrix}.$$

We check immediately that the solution is

$$x_1 = 11,\ x_2 = -3,\ x_3 = 7,\ x_4 = -4.$$

It is easy to see that the Jacobi matrix is

$$J = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

After 10 Jacobi iterations, we find the approximate solution

$$x_1 = 10.2588,\ x_2 = -2.5244,\ x_3 = 5.8008,\ x_4 = -3.7061.$$

After 20 iterations, we find the approximate solution

$$x_1 = 10.9110,\ x_2 = -2.9429,\ x_3 = 6.8560,\ x_4 = -3.9647.$$

After 50 iterations, we find the approximate solution

$$x_1 = 10.9998,\ x_2 = -2.9999,\ x_3 = 6.9998,\ x_4 = -3.9999,$$

and After 60 iterations, we find the approximate solution

$$x_1 = 11.0000,\ x_2 = -3.0000,\ x_3 = 7.0000,\ x_4 = -4.0000,$$

correct up to at least four decimals.

It can be shown (see Problem 9.6) that the eigenvalues of J are

$$\cos\left(\frac{\pi}{5}\right),\ \cos\left(\frac{2\pi}{5}\right),\ \cos\left(\frac{3\pi}{5}\right),\ \cos\left(\frac{4\pi}{5}\right),$$

so the spectral radius of $J = B$ is

$$\rho(J) = \cos\left(\frac{\pi}{5}\right) = 0.8090 < 1.$$

By Theorem 9.3, Jacobi's method converges for the matrix of this example.

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

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

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).

In Matlab one step of Gauss–Seidel iteration is achieved by the following function:

```
function u = GaussSeidel3(A,b,u)
n = size(A,1);
for i = 1:n
   u(i,1)  = u(i,1) + (-A(i,:)*u + b(i))/A(i,i);
end
end
```

It is remarkable that the only difference with Jacobi2 is that the same variable u is used on both sides of the assignment. In order to run m iteration steps, run the following function:

```
function u = GaussSeidel1(A,b,u0,m)
  u = u0;
  for j = 1:m
    u = GaussSeidel3(A,b,u);
  end
end
```

Example 9.2. Consider the same linear system

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 25 \\ -24 \\ 21 \\ -15 \end{pmatrix}$$

as in Example 9.1, whose solution is

$$x_1 = 11, \; x_2 = -3, \; x_3 = 7, \; x_4 = -4.$$

After 10 Gauss–Seidel iterations, we find the approximate solution

$$x_1 = 10.9966, \; x_2 = -3.0044, \; x_3 = 6.9964, \; x_4 = -4.0018.$$

After 20 iterations, we find the approximate solution

$$x_1 = 11.0000,\ x_2 = -3.0001,\ x_3 = 6.9999,\ x_4 = -4.0000.$$

After 25 iterations, we find the approximate solution

$$x_1 = 11.0000,\ x_2 = -3.0000,\ x_3 = 7.0000,\ x_4 = -4.0000,$$

correct up to at least four decimals. We observe that for this example, Gauss–Seidel's method converges about twice as fast as Jacobi's method. It will be shown in Proposition 9.8 that for a tridiagonal matrix, the spectral radius of the Gauss–Seidel matrix $\mathcal{L}_1$ is given by

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

so our observation is consistent with the theory.

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

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

where $\omega \neq 0$ is a real parameter to be suitably chosen. Actually, we show in Section 9.4 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

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

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

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

In Matlab one step of relaxation iteration is achieved by the following function:

```
function u = relax3(A,b,u,omega)
n = size(A,1);
for i = 1:n
   u(i,1)  = u(i,1) + omega*(-A(i,:)*u + b(i))/A(i,i);
end
end
```

Observe that function relax3 is obtained from the function GaussSeidel3 by simply inserting ω in front of the expression $(-A(i,:)*u+b(i))/A(i,i)$. In order to run m iteration steps, run the following function:

```
function u = relax(A,b,u0,omega,m)
  u = u0;
  for j = 1:m
    u = relax3(A,b,u,omega);
  end
end
```

Example 9.3. Consider the same linear system as in Examples 9.1 and 9.2, whose solution is

$$x_1 = 11,\ x_2 = -3,\ x_3 = 7,\ x_4 = -4.$$

After 10 relaxation iterations with $\omega = 1.1$, we find the approximate solution

$$x_1 = 11.0026,\ x_2 = -2.9968,\ x_3 = 7.0024,\ x_4 = -3.9989.$$

After 10 iterations with $\omega = 1.2$, we find the approximate solution

$$x_1 = 11.0014,\ x_2 = -2.9985,\ x_3 = 7.0010,\ x_4 = -3.9996.$$

After 10 iterations with $\omega = 1.3$, we find the approximate solution

$$x_1 = 10.9996,\ x_2 = -3.0001,\ x_3 = 6.9999,\ x_4 = -4.0000.$$

After 10 iterations with $\omega = 1.27$, we find the approximate solution

$$x_1 = 11.0000,\ x_2 = -3.0000,\ x_3 = 7.0000,\ x_4 = -4.0000,$$

correct up to at least four decimals. We observe that for this example the method of relaxation with $\omega = 1.27$ converges faster than the method of Gauss–Seidel. This observation will be confirmed by Proposition 9.10.

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 D is an invertible block-diagonal matrix. See Figure 9.1.

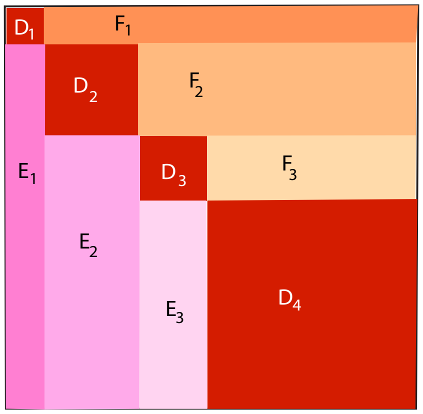

Figure 9.1: A schematic representation of a block decomposition $A = D - E - F$, where $D = \cup_{i=1}^4 D_i$, $E = \cup_{i=1}^3 E_i$, and $F = \cup_{i=1}^3 F_i$.

9.4 Convergence of the Methods of 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 9.5. *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.

Proof. Since $M = A + N$ and A is Hermitian, $A^* = A$, so we get

$$M^* + N = A^* + N^* + N = A + N + N^* = M + N^* = (M^* + N)^*,$$

which shows that $M^* + N$ is indeed Hermitian.

Because A is Hermitian positive definite, the function

$$v \mapsto (v^*Av)^{1/2}$$

from $\mathbb{C}^n$ to $\mathbb{R}$ is a vector norm $\| \ \|$, and let $\| \ \|$ also denote its subordinate matrix norm. We prove that

$$\|M^{-1}N\| < 1,$$

which by Theorem 9.1 proves that $\rho(M^{-1}N) < 1$. By definition

$$\|M^{-1}N\| = \|I - M^{-1}A\| = \sup_{\|v\|=1} \|v - M^{-1}Av\|,$$

which leads us to evaluate $\|v - M^{-1}Av\|$ when $\|v\| = 1$. If we write $w = M^{-1}Av$, using the facts that $\|v\| = 1$, $v = A^{-1}Mw$, $A^* = A$, and $A = M - N$, we have

$$\begin{aligned}
\|v - w\|^2 &= (v - w)^*A(v - w) \\
&= \|v\|^2 - v^*Aw - w^*Av + w^*Aw \\
&= 1 - w^*M^*w - w^*Mw + w^*Aw \\
&= 1 - w^*(M^* + N)w.
\end{aligned}$$

Now since we assumed that $M^* + N$ is positive definite, if $w \neq 0$, then $w^*(M^* + N)w > 0$, and we conclude that

$$\text{if} \quad \|v\| = 1, \quad \text{then} \quad \|v - M^{-1}Av\| < 1.$$

Finally, the function

$$v \mapsto \|v - M^{-1}Av\|$$

is continuous as a composition of continuous functions, therefore it achieves its maximum on the compact subset $\{v \in \mathbb{C}^n \mid \|v\| = 1\}$, which proves that

$$\sup_{\|v\|=1} \|v - M^{-1}Av\| < 1,$$

and completes the proof. □

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 9.6. *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.*

Proof. Recall that for the relaxation method, $A = M - N$ with

$$M = \frac{D}{\omega} - E$$

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

and because $D^* = D$, $E^* = F$ (since A is Hermitian) and $\omega \neq 0$ is real, we have

$$M^* + N = \frac{D^*}{\omega} - E^* + \frac{1-\omega}{\omega}D + F = \frac{2-\omega}{\omega}D.$$

If D consists of the diagonal entries of A, then we know from Section 7.8 that these entries are all positive, and since $\omega \in (0, 2)$, we see that the matrix $((2-\omega)/\omega)D$ is positive definite. If D consists of diagonal blocks of A, because A is positive, definite, by choosing vectors z obtained by picking a nonzero vector for each block of D and padding with zeros, we see that each block of D is positive definite, and thus D itself is positive definite. Therefore, in all cases, $M^* + N$ is positive definite, and we conclude by using Proposition 9.5. □

Remark: What if we allow the parameter ω to be a nonzero complex number $\omega \in \mathbb{C}$? In this case, we get

$$M^* + N = \frac{D^*}{\overline{\omega}} - E^* + \frac{1-\omega}{\omega}D + F = \left(\frac{1}{\omega} + \frac{1}{\overline{\omega}} - 1\right)D.$$

But,

$$\frac{1}{\omega} + \frac{1}{\overline{\omega}} - 1 = \frac{\omega + \overline{\omega} - \omega\overline{\omega}}{\omega\overline{\omega}} = \frac{1 - (\omega - 1)(\overline{\omega} - 1)}{|\omega|^2} = \frac{1 - |\omega - 1|^2}{|\omega|^2},$$

so 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 9.6 does not apply to Jacobi's method, but in special cases, Proposition 9.5 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 9.7. *Given any matrix $A = D - E - F$, with A and D invertible, for any $\omega \neq 0$, we have*

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

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

Proof. Observe that the product $\lambda_1 \cdots \lambda_n$ of the eigenvalues of $\mathcal{L}_\omega$, which is equal to $\det(\mathcal{L}_\omega)$, is given by

$$\lambda_1 \cdots \lambda_n = \det(\mathcal{L}_\omega) = \frac{\det\left(\dfrac{1-\omega}{\omega}D + F\right)}{\det\left(\dfrac{D}{\omega} - E\right)} = (1 - \omega)^n.$$

It follows that

$$\rho(\mathcal{L}_\omega) \geq |\lambda_1 \cdots \lambda_n|^{1/n} = |\omega - 1|.$$

The proof is the same if $\omega \in \mathbb{C}$. □

9.5 Convergence of the Methods of Jacobi, Gauss–Seidel, and Relaxation for Tridiagonal Matrices

We now consider the case where A is a *tridiagonal matrix*, possibly by blocks. In this case, we obtain precise results about the spectral radius of J and $\mathcal{L}_\omega$, and as a consequence, about the convergence of these methods. We also obtain some information about the rate of convergence of these methods. 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 9.8. *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.

Proof. We begin with a preliminary result. Let $A(\mu)$ with a tridiagonal matrix by block of the form

$$A(\mu) = \begin{pmatrix} A_1 & \mu^{-1}C_1 & 0 & 0 & \cdots & 0 \\ \mu B_1 & A_2 & \mu^{-1}C_2 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \cdots & \vdots \\ \vdots & \cdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mu B_{p-2} & A_{p-1} & \mu^{-1}C_{p-1} \\ 0 & \cdots & \cdots & 0 & \mu B_{p-1} & A_p \end{pmatrix},$$

then

$$\det(A(\mu)) = \det(A(1)), \quad \mu \neq 0.$$

To prove this fact, form the block diagonal matrix

$$P(\mu) = \mathrm{diag}(\mu I_1, \mu^2 I_2, \ldots, \mu^p I_p),$$

where I_j is the identity matrix of the same dimension as the block A_j. Then it is easy to see that

$$A(\mu) = P(\mu)A(1)P(\mu)^{-1},$$

and thus,

$$\det(A(\mu)) = \det(P(\mu)A(1)P(\mu)^{-1}) = \det(A(1)).$$

Since the Jacobi matrix is $J = D^{-1}(E + F)$, the eigenvalues of J are the zeros of the characteristic polynomial

$$p_J(\lambda) = \det(\lambda I - D^{-1}(E + F)),$$

and thus, they are also the zeros of the polynomial

$$q_J(\lambda) = \det(\lambda D - E - F) = \det(D)p_J(\lambda).$$

Similarly, since the Gauss–Seidel matrix is $\mathcal{L}_1 = (D - E)^{-1}F$, the zeros of the characteristic polynomial

$$p_{\mathcal{L}_1}(\lambda) = \det(\lambda I - (D - E)^{-1}F)$$

are also the zeros of the polynomial

$$q_{\mathcal{L}_1}(\lambda) = \det(\lambda D - \lambda E - F) = \det(D - E) p_{\mathcal{L}_1}(\lambda).$$

Since $A = D - E - F$ is tridiagonal (or tridiagonal by blocks), $\lambda^2 D - \lambda^2 E - F$ is also tridiagonal (or tridiagonal by blocks), and by using our preliminary result with $\mu = \lambda \neq 0$, we get

$$q_{\mathcal{L}_1}(\lambda^2) = \det(\lambda^2 D - \lambda^2 E - F) = \det(\lambda^2 D - \lambda E - \lambda F) = \lambda^n q_J(\lambda).$$

By continuity, the above equation also holds for $\lambda = 0$. But then we deduce that:

1. For any $\beta \neq 0$, if β is an eigenvalue of $\mathcal{L}_1$, then $\beta^{1/2}$ and $-\beta^{1/2}$ are both eigenvalues of J, where $\beta^{1/2}$ is one of the complex square roots of β.

2. For any $\alpha \neq 0$, if α and $-\alpha$ are both eigenvalues of J, then α^2 is an eigenvalue of $\mathcal{L}_1$.

The above immediately implies that $\rho(\mathcal{L}_1) = (\rho(J))^2$. □

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

Proposition 9.9. *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.

Proof. The proof is very technical and can be found in Serre [57] and Ciarlet [14]. As in the proof of the previous proposition, we begin by showing that the eigenvalues of the matrix $\mathcal{L}_\omega$ are the zeros of the polynomial

$$q_{\mathcal{L}_\omega}(\lambda) = \det\left(\frac{\lambda + \omega - 1}{\omega} D - \lambda E - F\right) = \det\left(\frac{D}{\omega} - E\right) p_{\mathcal{L}_\omega}(\lambda),$$

where $p_{\mathcal{L}_\omega}(\lambda)$ is the characteristic polynomial of $\mathcal{L}_\omega$. Then using the preliminary fact from Proposition 9.8, it is easy to show that

$$q_{\mathcal{L}_\omega}(\lambda^2) = \lambda^n q_J\left(\frac{\lambda^2 + \omega - 1}{\lambda\omega}\right),$$

for all $\lambda \in \mathbb{C}$, with $\lambda \neq 0$. This time we cannot extend the above equation to $\lambda = 0$. This leads us to consider the equation

$$\frac{\lambda^2 + \omega - 1}{\lambda\omega} = \alpha,$$

which is equivalent to

$$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0,$$

for all $\lambda \neq 0$. Since $\lambda \neq 0$, the above equivalence does not hold for $\omega = 1$, but this is not a problem since the case $\omega = 1$ has already been considered in the previous proposition. Then we can show the following:

1. For any $\beta \neq 0$, if β is an eigenvalue of $\mathcal{L}_\omega$, then

   $$\frac{\beta + \omega - 1}{\beta^{1/2}\omega}, \quad -\frac{\beta + \omega - 1}{\beta^{1/2}\omega}$$

   are eigenvalues of J.

2. For every $\alpha \neq 0$, if α and $-\alpha$ are eigenvalues of J, then $\mu_+(\alpha, \omega)$ and $\mu_-(\alpha, \omega)$ are eigenvalues of $\mathcal{L}_\omega$, where $\mu_+(\alpha, \omega)$ and $\mu_-(\alpha, \omega)$ are the squares of the roots of the equation

   $$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0.$$

It follows that

$$\rho(\mathcal{L}_\omega) = \max_{\lambda \mid p_J(\lambda) = 0} \{\max(|\mu_+(\alpha, \omega)|, |\mu_-(\alpha, \omega)|)\},$$

and since we are assuming that J has real roots, we are led to study the function

$$M(\alpha, \omega) = \max\{|\mu_+(\alpha, \omega)|, |\mu_-(\alpha, \omega)|\},$$

where $\alpha \in \mathbb{R}$ and $\omega \in (0, 2)$. Actually, because $M(-\alpha, \omega) = M(\alpha, \omega)$, it is only necessary to consider the case where $\alpha \geq 0$.

Note that for $\alpha \neq 0$, the roots of the equation

$$\lambda^2 - \alpha\omega\lambda + \omega - 1 = 0.$$

are

$$\frac{\alpha\omega \pm \sqrt{\alpha^2\omega^2 - 4\omega + 4}}{2}.$$

In turn, this leads to consider the roots of the equation

$$\omega^2\alpha^2 - 4\omega + 4 = 0,$$

which are

$$\frac{2(1 \pm \sqrt{1 - \alpha^2})}{\alpha^2},$$

for $\alpha \neq 0$. Since we have

$$\frac{2(1 + \sqrt{1 - \alpha^2})}{\alpha^2} = \frac{2(1 + \sqrt{1 - \alpha^2})(1 - \sqrt{1 - \alpha^2})}{\alpha^2(1 - \sqrt{1 - \alpha^2})} = \frac{2}{1 - \sqrt{1 - \alpha^2}}$$

and

$$\frac{2(1-\sqrt{1-\alpha^2})}{\alpha^2} = \frac{2(1+\sqrt{1-\alpha^2})(1-\sqrt{1-\alpha^2})}{\alpha^2(1+\sqrt{1-\alpha^2})} = \frac{2}{1+\sqrt{1-\alpha^2}},$$

these roots are

$$\omega_0(\alpha) = \frac{2}{1+\sqrt{1-\alpha^2}}, \quad \omega_1(\alpha) = \frac{2}{1-\sqrt{1-\alpha^2}}.$$

Observe that the expression for $\omega_0(\alpha)$ is exactly the expression in the statement of our proposition! The rest of the proof consists in analyzing the variations of the function $M(\alpha, \omega)$ by considering various cases for α. In the end, we find that the minimum of $\rho(\mathcal{L}_\omega)$ is obtained for $\omega_0(\rho(J))$. The details are tedious and we omit them. The reader will find complete proofs in Serre [57] and Ciarlet [14]. □

Combining the results of Theorem 9.6 and Proposition 9.9, 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 9.10. *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 $\rho(\mathcal{L}_1) = \rho(J) = 0$.

Proof. In order to apply Proposition 9.9, we have to check that $J = D^{-1}(E + F)$ has real eigenvalues. However, if α is any eigenvalue of J and if u is any corresponding eigenvector, then

$$D^{-1}(E+F)u = \alpha u$$

implies that

$$(E+F)u = \alpha D u,$$

and since $A = D - E - F$, the above shows that $(D - A)u = \alpha D u$, that is,

$$Au = (1-\alpha)Du.$$

Consequently,

$$u^*Au = (1-\alpha)u^*Du,$$

and since A and D are Hermitian positive definite, we have $u^*Au > 0$ and $u^*Du > 0$ if $u \neq 0$, which proves that $\alpha \in \mathbb{R}$. The rest follows from Theorem 9.6 and Proposition 9.9. □

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

9.6 Summary

The main concepts and results of this chapter are listed below:

- Iterative methods. Splitting A as $A = M - N$.
- *Convergence of a sequence of vectors or matrices*.
- A criterion for the convergence of the sequence (B^k) of powers of a matrix B to zero in terms of the spectral radius $\rho(B)$.
- A characterization of the spectral radius $\rho(B)$ as the limit of the sequence $(\|B^k\|^{1/k})$.
- A criterion of the convergence of iterative methods.
- Asymptotic behavior of iterative methods.
- Splitting A as $A = D-E-F$, and the methods of *Jacobi*, *Gauss–Seidel*, and *relaxation* (and *SOR*).
- The *Jacobi matrix*, $J = D^{-1}(E + F)$.
- The *Gauss–Seidel matrix*, $\mathcal{L}_1 = (D - E)^{-1}F$.
- The *matrix of relaxation*, $\mathcal{L}_\omega = (D - \omega E)^{-1}((1 - \omega)D + \omega F)$.
- Convergence of iterative methods: a general result when $A = M - N$ is Hermitian positive definite.
- A sufficient condition for the convergence of the methods of Jacobi, Gauss–Seidel, and relaxation. The *Ostrowski-Reich theorem*: A is Hermitian positive definite and $\omega \in (0, 2)$.
- A necessary condition for the convergence of the methods of Jacobi , Gauss–Seidel, and relaxation: $\omega \in (0, 2)$.
- The case of tridiagonal matrices (possibly by blocks). Simultaneous convergence or divergence of Jacobi's method and Gauss–Seidel's method, and comparison of the spectral radii of $\rho(J)$ and $\rho(\mathcal{L}_1)$: $\rho(\mathcal{L}_1) = (\rho(J))^2$.
- The case of tridiagonal Hermitian positive definite matrices (possibly by blocks). The methods of Jacobi, Gauss–Seidel, and relaxation, all converge.
- In the above case, there is a unique optimal relaxation parameter for which $\rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_1) = (\rho(J))^2 < \rho(J)$ (if $\rho(J) \neq 0$).

9.7 Problems

Problem 9.1. Consider the matrix

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}.$$

Prove that $\rho(J) = 0$ and $\rho(\mathcal{L}_1) = 2$, so

$$\rho(J) < 1 < \rho(\mathcal{L}_1),$$

where J is Jacobi's matrix and $\mathcal{L}_1$ is the matrix of Gauss–Seidel.

Problem 9.2. Consider the matrix

$$A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}.$$

Prove that $\rho(J) = \sqrt{5}/2$ and $\rho(\mathcal{L}_1) = 1/2$, so

$$\rho(\mathcal{L}_1) < \rho(J),$$

where where J is Jacobi's matrix and $\mathcal{L}_1$ is the matrix of Gauss–Seidel.

Problem 9.3. Consider the following linear system:

$$\begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 19 \\ 19 \\ -3 \\ -12 \end{pmatrix}.$$

(1) Solve the above system by Gaussian elimination.

(2) Compute the sequences of vectors $u_k = (u_1^k, u_2^k, u_3^k, u_4^k)$ for $k = 1, \ldots, 10$, using the methods of Jacobi, Gauss–Seidel, and relaxation for the following values of ω: $\omega = 1.1, 1.2, \ldots, 1.9$. In all cases, the initial vector is $u_0 = (0, 0, 0, 0)$.

Problem 9.4. Recall that a complex or real $n \times n$ matrix A is *strictly row diagonally dominant* if $|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|$ for $i = 1, \ldots, n$.

(1) Prove that if A is strictly row diagonally dominant, then Jacobi's method converges.

(2) Prove that if A is strictly row diagonally dominant, then Gauss–Seidel's method converges.

Problem 9.5. Prove that the converse of Proposition 9.5 holds. That is, if A is a Hermitian positive definite matrix writen as $A = M - N$ with M invertible, if the Hermitan matrix $M^* + N$ is positive definite, and if $\rho(M^{-1}N) < 1$, then A is positive definite.

Problem 9.6. Consider the following tridiagonal $n \times n$ matrix:

$$A = \frac{1}{(n+1)^2}\begin{pmatrix} 2 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & 0 & -1 & 2 \end{pmatrix}.$$

(1) Prove that the eigenvalues of the Jacobi matrix J are given by

$$\lambda_k = \cos\left(\frac{k\pi}{n+1}\right), \quad k = 1, \ldots, n.$$

Hint. First show that the Jacobi matrix is

$$J = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & & \\ 1 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & 0 & 1 \\ & & 0 & 1 & 0 \end{pmatrix}.$$

Then the eigenvalues and the eigenvectors of J are solutions of the system of equations

$$\begin{aligned} y_0 &= 0 \\ y_{k+1} + y_{k-1} &= 2\lambda y_k, \quad k = 1, \ldots, n \\ y_{n+1} &= 0. \end{aligned}$$

It is well known that the general solution to the above recurrence is given by

$$y_k = \alpha z_1^k + \beta z_2^k, \quad k = 0, \ldots, n+1,$$

(with $\alpha, \beta \neq 0$) where z_1 and z_2 are the zeros of the equation

$$z^2 - 2\lambda z + 1 = 0.$$

It follows that $z_2 = z_1^{-1}$ and $z_1 + z_2 = 2\lambda$. The boundary condition $y_0 = 0$ yields $\alpha + \beta = 0$, so $y_k = \alpha(z_1^k - z_1^{-k})$, and the boundary condition $y_{n+1} = 0$ yields

$$z_1^{2(n+1)} = 1.$$

Deduce that we may assume that the n possible values $(z_1)_k$ for z_1 are given by

$$(z_1)_k = e^{\frac{k\pi i}{n+1}}, \quad k = 1, \ldots, n,$$

and find

$$2\lambda_k = (z_1)_k + (z_1)_k^{-1}.$$

Show that an eigenvector $(y_1^{(k)}, \ldots, y_n^{(k)})$ associated wih the eigenvalue λ_k is given by

$$y_j^{(k)} = \sin\left(\frac{kj\pi}{n+1}\right), \quad j = 1, \ldots, n.$$

(2) Find the spectral radius $\rho(J)$, $\rho(\mathcal{L}_1)$, and $\rho(\mathcal{L}_{\omega_0})$, as functions of $h = 1/(n+1)$.