Numerical Analysis Lecture Notes

Peter J. Olver

7. Iterative Methods for Linear Systems

Linear iteration coincides with multiplication by successive powers of a matrix; convergence of the iterates depends on the magnitude of its eigenvalues. We discuss in some detail a variety of convergence criteria based on the spectral radius, on matrix norms, and on eigenvalue estimates provided by the Gerschgorin Circle Theorem.

We will then turn our attention to the three most important iterative schemes used to accurately approximate the solutions to linear algebraic systems. The classical Jacobi method is the simplest, while an evident serialization leads to the popular Gauss–Seidel method. Completely general convergence criteria are hard to formulate, although convergence is assured for the important class of diagonally dominant matrices that arise in many applications. A simple modification of the Gauss–Seidel scheme, known as Successive Over-Relaxation (SOR), can dramatically speed up the convergence rate, and is the method of choice in many modern applications. Finally, we introduce the method of conjugate gradients, a powerful "semi-direct" iterative scheme that, in contrast to the classical iterative schemes, is guaranteed to eventually produce the exact solution.

7.1. Linear Iterative Systems.

We begin with the basic definition of an iterative system of linear equations.

Definition 7.1. A *linear iterative system* takes the form

$$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}, \qquad \mathbf{u}^{(0)} = \mathbf{a}. \tag{7.1}$$

The *coefficient matrix* T has size $n \times n$. We will consider both real and complex systems, and so the *iterates*[†] $\mathbf{u}^{(k)}$ are vectors either in $\mathbb{R}^n$ (which assumes that the coefficient matrix T is also real) or in $\mathbb{C}^n$. For $k = 1, 2, 3, \ldots$, the solution $\mathbf{u}^{(k)}$ is uniquely determined by the *initial conditions* $\mathbf{u}^{(0)} = \mathbf{a}$.

Powers of Matrices

The solution to the general linear iterative system (7.1) is, at least at first glance, immediate. Clearly,

$$\mathbf{u}^{(1)} = T\,\mathbf{u}^{(0)} = T\,\mathbf{a}, \qquad \mathbf{u}^{(2)} = T\,\mathbf{u}^{(1)} = T^2\mathbf{a}, \qquad \mathbf{u}^{(3)} = T\,\mathbf{u}^{(2)} = T^3\mathbf{a},$$

[†] *Warning*: The superscripts on $\mathbf{u}^{(k)}$ refer to the iterate number, and should not be mistaken for derivatives.

and, in general,

$$\mathbf{u}^{(k)} = T^k \mathbf{a}. \tag{7.2}$$

Thus, the iterates are simply determined by multiplying the initial vector $\mathbf{a}$ by the successive powers of the coefficient matrix T. And so, unlike differential equations, proving the existence and uniqueness of solutions to an iterative system is completely trivial.

However, unlike real or complex scalars, the general formulae and qualitative behavior of the powers of a square matrix are not nearly so immediately apparent. (Before continuing, the reader is urged to experiment with simple 2×2 matrices, trying to detect patterns.) To make progress, recall how we managed to solve linear systems of differential equations by suitably adapting the known exponential solution from the scalar version. In the iterative case, the scalar solution formula (2.8) is written in terms of powers, not exponentials. This motivates us to try the power ansatz

$$\mathbf{u}^{(k)} = \lambda^k \mathbf{v}, \tag{7.3}$$

in which λ is a scalar and $\mathbf{v}$ is a fixed vector, as a possible solution to the system. We find

$$\mathbf{u}^{(k+1)} = \lambda^{k+1} \mathbf{v}, \qquad \text{while} \qquad T\,\mathbf{u}^{(k)} = T(\lambda^k \mathbf{v}) = \lambda^k\, T\mathbf{v}.$$

These two expressions will be equal if and only if

$$T\mathbf{v} = \lambda \mathbf{v}.$$

Therefore, (7.3) is a nontrivial solution to (7.1) if and only if λ is an *eigenvalue* of the coefficient matrix T and $\mathbf{v} \neq \mathbf{0}$ an associated *eigenvector*.

Thus, to each eigenvector and eigenvalue of the coefficient matrix, we can construct a solution to the iterative system. We can then appeal to linear superposition to combine the basic power solutions to form more general solutions. In particular, if the coefficient matrix is complete, then this method will, as in the case of linear ordinary differential equations, produce the general solution.

Theorem 7.2. *If the coefficient matrix T is complete, then the general solution to the linear iterative system $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$ is given by*

$$\mathbf{u}^{(k)} = c_1 \lambda_1^k \mathbf{v}_1 + c_2 \lambda_2^k \mathbf{v}_2 + \cdots + c_n \lambda_n^k \mathbf{v}_n, \tag{7.4}$$

where $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are the linearly independent eigenvectors and $\lambda_1, \ldots, \lambda_n$ the corresponding eigenvalues of T. The coefficients $c_1, \ldots, c_n$ are arbitrary scalars and are uniquely prescribed by the initial conditions $\mathbf{u}^{(0)} = \mathbf{a}$.

Proof: Since we already know that (7.4) is a solution to the system for arbitrary $c_1, \ldots, c_n$, it suffices to show that we can match any prescribed initial conditions. To this end, we need to solve the linear system

$$\mathbf{u}^{(0)} = c_1 \mathbf{v}_1 + \cdots + c_n \mathbf{v}_n = \mathbf{a}. \tag{7.5}$$

Completeness of T implies that its eigenvectors form a basis of $\mathbb{C}^n$, and hence (7.5) always admits a solution. In matrix form, we can rewrite (7.5) as

$$S\,\mathbf{c} = \mathbf{a}, \qquad \text{so that} \qquad \mathbf{c} = S^{-1}\mathbf{a},$$

where $S = (\, \mathbf{v}_1 \; \mathbf{v}_2 \; \ldots \; \mathbf{v}_n \,)$ is the (nonsingular) matrix whose columns are the eigenvectors. *Q.E.D.*

Remark: Solutions in the incomplete cases rely on the Jordan canonical form. As with systems of differential equations, the formulas are more complicated, and will not be written out.

Example 7.3. Consider the iterative system

$$x^{(k+1)} = \tfrac{3}{5}\, x^{(k)} + \tfrac{1}{5}\, y^{(k)}, \qquad y^{(k+1)} = \tfrac{1}{5}\, x^{(k)} + \tfrac{3}{5}\, y^{(k)}, \tag{7.6}$$

with initial conditions

$$x^{(0)} = a, \qquad y^{(0)} = b. \tag{7.7}$$

The system can be rewritten in our matrix form (7.1), with

$$T = \begin{pmatrix} .6 & .2 \\ .2 & .6 \end{pmatrix}, \qquad \mathbf{u}^{(k)} = \begin{pmatrix} x^{(k)} \\ y^{(k)} \end{pmatrix}, \qquad \mathbf{a} = \begin{pmatrix} a \\ b \end{pmatrix}.$$

Solving the characteristic equation

$$\det(T - \lambda\, \mathrm{I}\,) = \lambda^2 - 1.2\, \lambda - .32 = 0$$

produces the eigenvalues $\lambda_1 = .8, \lambda_2 = .4$. We then solve the associated linear systems $(T - \lambda_j\, \mathrm{I}\,)\mathbf{v}_j = \mathbf{0}$ for the corresponding eigenvectors:

$$\lambda_1 = .8\,, \qquad \mathbf{v}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \lambda_2 = .4\,, \qquad \mathbf{v}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Therefore, the basic power solutions are

$$\mathbf{u}_1^{(k)} = (\,.8)^k \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad \mathbf{u}_2^{(k)} = (\,.4)^k \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Theorem 7.2 tells us that the general solution is given as a linear combination,

$$\mathbf{u}^{(k)} = c_1\, \mathbf{u}_1^{(k)} + c_2\, \mathbf{u}_2^{(k)} = c_1\,(\,.8)^k \begin{pmatrix} 1 \\ 1 \end{pmatrix} + c_2\,(\,.4)^k \begin{pmatrix} -1 \\ 1 \end{pmatrix} = \begin{pmatrix} c_1\,(\,.8)^k - c_2\,(\,.4)^k \\ c_1\,(\,.8)^k + c_2\,(\,.4)^k \end{pmatrix},$$

where c_1, c_2 are determined by the initial conditions:

$$\mathbf{u}^{(0)} = \begin{pmatrix} c_1 - c_2 \\ c_1 + c_2 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}, \qquad \text{and hence} \qquad c_1 = \frac{a+b}{2}\,, \qquad c_2 = \frac{b-a}{2}\,.$$

Therefore, the explicit formula for the solution to the initial value problem (7.6–7) is

$$x^{(k)} = (\,.8)^k\, \frac{a+b}{2} + (\,.4)^k\, \frac{a-b}{2}, \qquad y^{(k)} = (\,.8)^k\, \frac{a+b}{2} + (\,.4)^k\, \frac{b-a}{2}.$$

In particular, as $k \to \infty$, the iterates $\mathbf{u}^{(k)} \to \mathbf{0}$ converge to zero at a rate governed by the dominant eigenvalue $\lambda_1 = .8$. Thus, (7.6) defines a stable iterative system. Figure 7.1 illustrates the cumulative effect of the iteration. The initial conditions consist of a large number of points on the unit circle $x^2 + y^2 = 1$, which are successively mapped to points on progressively smaller and flatter ellipses, all converging towards the origin.

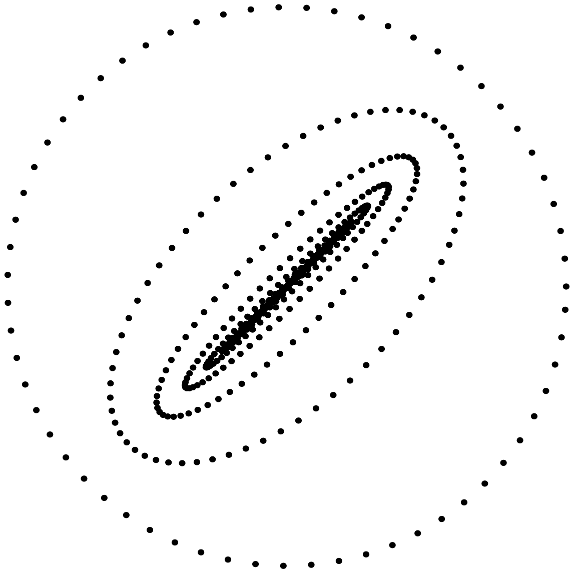

Figure 7.1. Stable Iterative System.

Example 7.4. The *Fibonacci numbers* are defined by the second order[†] iterative scheme

$$u^{(k+2)} = u^{(k+1)} + u^{(k)}, \tag{7.8}$$

with initial conditions

$$u^{(0)} = a, \qquad u^{(1)} = b. \tag{7.9}$$

In short, to obtain the next Fibonacci number, add the previous two. The classical *Fibonacci integers* start with $a = 0$, $b = 1$; the next few are

$$u^{(0)} = 0, \;\; u^{(1)} = 1, \;\; u^{(2)} = 1, \;\; u^{(3)} = 2, \;\; u^{(4)} = 3, \;\; u^{(5)} = 5, \;\; u^{(6)} = 8, \;\; u^{(7)} = 13, \;\; \ldots .$$

The Fibonacci integers occur in a surprising variety of natural objects, including leaves, flowers, and fruit, [**52**]. They were originally introduced by the Italian Renaissance mathematician Fibonacci (Leonardo of Pisa) as a crude model of the growth of a population of rabbits. In Fibonacci's model, the k^{th} Fibonacci number $u^{(k)}$ measures the total number of pairs of rabbits at year k. We start the process with a single juvenile pair[‡] at year 0. Once a year, each pair of rabbits produces a new pair of offspring, but it takes a full year for a rabbit pair to mature enough to produce offspring of their own.

Just as every higher order ordinary differential equation can be replaced by an equivalent first order system, so every higher order iterative equation can be replaced by a first

[†] In general, an iterative system $\mathbf{u}^{(k+j)} = T_1 \mathbf{u}^{(k+j-1)} + \cdots + T_j \mathbf{u}^{(k)}$ in which the new iterate depends upon the preceding j values is said to have *order j*.

[‡] We ignore important details like the sex of the offspring.

order iterative system. In this particular case, we define the vector

$$\mathbf{u}^{(k)} = \begin{pmatrix} u^{(k)} \\ u^{(k+1)} \end{pmatrix} \in \mathbb{R}^2,$$

and note that (7.8) is equivalent to the matrix system

$$\begin{pmatrix} u^{(k+1)} \\ u^{(k+2)} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u^{(k)} \\ u^{(k+1)} \end{pmatrix}, \qquad \text{or} \qquad \mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}, \qquad \text{where} \qquad T = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

To find the explicit formula for the Fibonacci numbers, we must determine the eigenvalues and eigenvectors of the coefficient matrix T. A straightforward computation produces

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \;=\; 1.618034\ldots\,, \qquad \lambda_2 = \frac{1-\sqrt{5}}{2} \;=\; -.618034\ldots\,,$$
$$\mathbf{v}_1 = \begin{pmatrix} \frac{-1+\sqrt{5}}{2} \\ 1 \end{pmatrix}, \qquad \mathbf{v}_2 = \begin{pmatrix} \frac{-1-\sqrt{5}}{2} \\ 1 \end{pmatrix}.$$

Therefore, according to (7.4), the general solution to the Fibonacci system is

$$\mathbf{u}^{(k)} = \begin{pmatrix} u^{(k)} \\ u^{(k+1)} \end{pmatrix} = c_1 \left(\frac{1+\sqrt{5}}{2}\right)^k \begin{pmatrix} \frac{-1+\sqrt{5}}{2} \\ 1 \end{pmatrix} + c_2 \left(\frac{1-\sqrt{5}}{2}\right)^k \begin{pmatrix} \frac{-1-\sqrt{5}}{2} \\ 1 \end{pmatrix}. \tag{7.10}$$

The initial data

$$\mathbf{u}^{(0)} = c_1 \begin{pmatrix} \frac{-1+\sqrt{5}}{2} \\ 1 \end{pmatrix} + c_2 \begin{pmatrix} \frac{-1-\sqrt{5}}{2} \\ 1 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

uniquely specifies the coefficients

$$c_1 = \frac{2\,a + (1+\sqrt{5})\,b}{2\sqrt{5}}\,, \qquad c_2 = -\,\frac{2\,a + (1-\sqrt{5})\,b}{2\sqrt{5}}\,.$$

The first entry of the solution vector (7.10) produces the explicit formula

$$u^{(k)} \;=\; \frac{(-1+\sqrt{5})\,a + 2\,b}{2\sqrt{5}} \left(\frac{1+\sqrt{5}}{2}\right)^k \;+\; \frac{(1+\sqrt{5})\,a - 2\,b}{2\sqrt{5}} \left(\frac{1-\sqrt{5}}{2}\right)^k \tag{7.11}$$

for the k^{th} Fibonacci number. For the particular initial conditions $a = 0$, $b = 1$, (7.11) reduces to the classical *Binet formula*

$$u^{(k)} = \frac{1}{\sqrt{5}} \left[\left(\frac{1+\sqrt{5}}{2}\right)^k - \left(\frac{1-\sqrt{5}}{2}\right)^k \right] \tag{7.12}$$

for the k^{th} Fibonacci integer. It is a remarkable fact that, for every value of k, all the $\sqrt{5}$'s cancel out, and the Binet formula does indeed produce the Fibonacci integers listed above. Another useful observation is that, since

$$0 < |\,\lambda_2\,| = \frac{\sqrt{5}-1}{2} < 1 < \lambda_1 = \frac{1+\sqrt{5}}{2}\,,$$

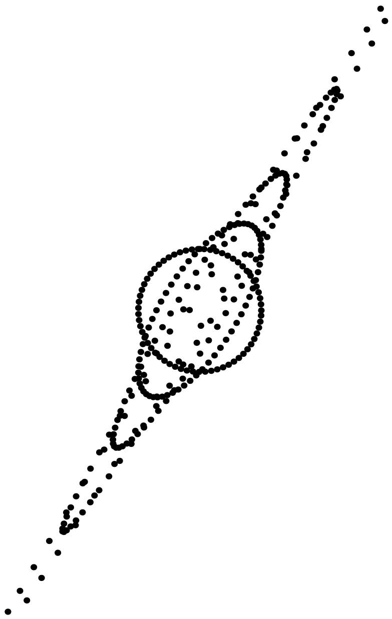

Figure 7.2. Fibonacci Iteration.

the terms involving λ_1^k go to ∞ (and so the zero solution to this iterative system is unstable) while the terms involving λ_2^k go to zero. Therefore, even for k moderately large, the first term in (7.11) is an excellent approximation (and one that gets more and more accurate with increasing k) to the k^{th} Fibonacci number. A plot of the first 4 iterates, starting with the initial data consisting of equally spaced points on the unit circle, can be seen in Figure 7.2. As in the previous example, the circle is mapped to a sequence of progressively more eccentric ellipses; however, their major semi-axes become more and more stretched out, and almost all points end up going off to ∞.

The dominant eigenvalue $\lambda_1 = \frac{1}{2}\left(1+\sqrt{5}\,\right) = 1.618034\ldots$ is known as the *golden ratio* and plays an important role in spiral growth in nature, as well as in art, architecture and design, [**52**]. It describes the overall growth rate of the Fibonacci integers, and, in fact, any sequence of Fibonacci numbers with initial conditions $b \neq \frac{1}{2}\left(1-\sqrt{5}\,\right) a$.

Example 7.5. Let $T = \begin{pmatrix} -3 & 1 & 6 \\ 1 & -1 & -2 \\ -1 & -1 & 0 \end{pmatrix}$ be the coefficient matrix for a three-dimensional iterative system $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$. Its eigenvalues and corresponding eigenvectors are

$$\begin{aligned} &\lambda_1 = -2, && \lambda_2 = -1 + \mathrm{i}\,, && \lambda_3 = -1 - \mathrm{i}\,, \\ &\mathbf{v}_1 = \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix}, && \mathbf{v}_2 = \begin{pmatrix} 2-\mathrm{i} \\ -1 \\ 1 \end{pmatrix}, && \mathbf{v}_3 = \begin{pmatrix} 2+\mathrm{i} \\ -1 \\ 1 \end{pmatrix}. \end{aligned}$$

Therefore, according to (7.4), the general complex solution to the iterative system is

$$\mathbf{u}^{(k)} = b_1\,(-2)^k \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + b_2\,(-1+\mathrm{i}\,)^k \begin{pmatrix} 2-\mathrm{i} \\ -1 \\ 1 \end{pmatrix} + b_3\,(-1-\mathrm{i}\,)^k \begin{pmatrix} 2+\mathrm{i} \\ -1 \\ 1 \end{pmatrix},$$

where b_1, b_2, b_3 are arbitrary complex scalars.

If we are only interested in real solutions, we can, as in the case of systems of differential equations, break up any complex solution into its real and imaginary parts, each of which constitutes a real solution. We begin by writing $\lambda_2 = -1 + \mathrm{i} = \sqrt{2}\, e^{3\pi\,\mathrm{i}\,/4}$, and hence

$$(-1+\mathrm{i}\,)^k = 2^{k/2}\, e^{3\,k\,\pi\,\mathrm{i}\,/4} = 2^{k/2}\left(\cos\tfrac{3}{4}\,k\,\pi + \mathrm{i}\,\sin\tfrac{3}{4}\,k\,\pi\,\right).$$

Therefore, the complex solution

$$(-1+\mathrm{i}\,)^k \begin{pmatrix} 2-\mathrm{i} \\ -1 \\ 1 \end{pmatrix} = 2^{k/2} \begin{pmatrix} 2\cos\frac{3}{4}\,k\,\pi + \sin\frac{3}{4}\,k\,\pi \\ -\cos\frac{3}{4}\,k\,\pi \\ \cos\frac{3}{4}\,k\,\pi \end{pmatrix} + \mathrm{i}\;2^{k/2} \begin{pmatrix} 2\sin\frac{3}{4}\,k\,\pi - \cos\frac{3}{4}\,k\,\pi \\ -\sin\frac{3}{4}\,k\,\pi \\ \sin\frac{3}{4}\,k\,\pi \end{pmatrix}$$

is a combination of two independent real solutions. The complex conjugate eigenvalue $\lambda_3 = -1 - \mathrm{i}$ leads, as before, to the complex conjugate solution — and the same two real solutions. The general real solution $\mathbf{u}^{(k)}$ to the system can be written as a linear combination of the three independent real solutions:

$$c_1\,(-2)^k \begin{pmatrix} 4 \\ -2 \\ 1 \end{pmatrix} + c_2\,2^{k/2} \begin{pmatrix} 2\cos\frac{3}{4}\,k\,\pi + \sin\frac{3}{4}\,k\,\pi \\ -\cos\frac{3}{4}\,k\,\pi \\ \cos\frac{3}{4}\,k\,\pi \end{pmatrix} + c_3\,2^{k/2} \begin{pmatrix} 2\sin\frac{3}{4}\,k\,\pi - \cos\frac{3}{4}\,k\,\pi \\ -\sin\frac{3}{4}\,k\,\pi \\ \sin\frac{3}{4}\,k\,\pi \end{pmatrix}, \tag{7.13}$$

where c_1, c_2, c_3 are arbitrary real scalars, uniquely prescribed by the initial conditions.

7.2. Stability.

With the solution formula (7.4) in hand, we are now in a position to understand the qualitative behavior of solutions to (complete) linear iterative systems. The most important case for applications is when all the iterates converge to $\mathbf{0}$.

Definition 7.6. The equilibrium solution $\mathbf{u}^\star = \mathbf{0}$ to a linear iterative system (7.1) is called *asymptotically stable* if and only if all solutions $\mathbf{u}^{(k)} \to \mathbf{0}$ as $k \to \infty$.

Asymptotic stability relies on the following property of the coefficient matrix.

Definition 7.7. A matrix T is called *convergent* if its powers converge to the zero matrix, $T^k \to \mathrm{O}$, meaning that the individual entries of T^k all go to 0 as $k \to \infty$.

The equivalence of the convergence condition and stability of the iterative system follows immediately from the solution formula (7.2).

Proposition 7.8. *The linear iterative system* $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)}$ *has asymptotically stable zero solution if and only if* T *is a convergent matrix.*

Proof: If $T^k \to \mathrm{O}$, and $\mathbf{u}^{(k)} = T^k \mathbf{a}$ is any solution, then clearly $\mathbf{u}^{(k)} \to \mathbf{0}$ as $k \to \infty$, proving stability. Conversely, the solution $\mathbf{u}_j^{(k)} = T^k \mathbf{e}_j$ is the same as the j^{th} column of T^k. If the origin is asymptotically stable, then $\mathbf{u}_j^{(k)} \to \mathbf{0}$. Thus, the individual columns of T^k all tend to $\mathbf{0}$, proving that $T^k \to \mathrm{O}$. *Q.E.D.*

To facilitate the analysis of convergence, we shall adopt a norm $\| \cdot \|$ on our underlying vector space, $\mathbb{R}^n$ or $\mathbb{C}^n$. The reader may be inclined to choose the Euclidean (or Hermitian) norm, but, in practice, the ∞ norm

$$\| \mathbf{u} \|_\infty = \max\{\, | u_1 |, \ \ldots \ , | u_n | \,\} \tag{7.14}$$

prescribed by the vector's maximal entry (in modulus) is usually much easier to work with. Convergence of the iterates is equivalent to convergence of their norms:

$$\mathbf{u}^{(k)} \to \mathbf{0} \qquad \text{if and only if} \qquad \| \mathbf{u}^{(k)} \| \to \mathbf{0} \qquad \text{as} \qquad k \to \infty.$$

The fundamental stability criterion for linear iterative systems relies on the size of the eigenvalues of the coefficient matrix.

Theorem 7.9. *A linear iterative system* (7.1) *is asymptotically stable if and only if all its* (*complex*) *eigenvalues have modulus strictly less than one*: $| \lambda_j | < 1$.

Proof: Let us prove this result assuming that the coefficient matrix T is complete. (The proof in the incomplete case relies on the Jordan canonical form, and is outlined in the exercises.) If λ_j is an eigenvalue such that $| \lambda_j | < 1$, then the corresponding basis solution $\mathbf{u}_j^{(k)} = \lambda_j^k \mathbf{v}_j$ tends to zero as $k \to \infty$; indeed,

$$\| \mathbf{u}_j^{(k)} \| = \| \lambda_j^k \mathbf{v}_j \| = | \lambda_j |^k \, \| \mathbf{v}_j \| \ \longrightarrow \ 0 \qquad \text{since} \qquad | \lambda_j | < 1.$$

Therefore, if all eigenvalues are less than 1 in modulus, all terms in the solution formula (7.4) tend to zero, which proves asymptotic stability: $\mathbf{u}^{(k)} \to \mathbf{0}$. Conversely, if any eigenvalue satisfies $| \lambda_j | \geq 1$, then the solution $\mathbf{u}^{(k)} = \lambda_j^k \mathbf{v}_j$ does not tend to $\mathbf{0}$ as $k \to \infty$, and hence $\mathbf{0}$ is not asymptotically stable. *Q.E.D.*

Consequently, the necessary and sufficient condition for asymptotic stability of a linear iterative system is that all the eigenvalues of the coefficient matrix lie strictly inside the unit circle in the complex plane: $| \lambda_j | < 1$.

Definition 7.10. The *spectral radius* of a matrix T is defined as the maximal modulus of all of its real and complex eigenvalues: $\rho(T) = \max \{\, | \lambda_1 |, \ldots, | \lambda_k | \,\}$.

We can then restate the Stability Theorem 7.9 as follows:

Theorem 7.11. *The matrix T is convergent if and only if its spectral radius is strictly less than one*: $\rho(T) < 1$.

If T is complete, then we can apply the triangle inequality to (7.4) to estimate

$$\begin{aligned} \| \mathbf{u}^{(k)} \| &= \| c_1 \lambda_1^k \mathbf{v}_1 + \cdots + c_n \lambda_n^k \mathbf{v}_n \| \\ &\le | \lambda_1 |^k \, \| c_1 \mathbf{v}_1 \| + \cdots + | \lambda_n |^k \, \| c_n \mathbf{v}_n \| \\ &\le \rho(T)^k \bigl(| c_1 | \, \| \mathbf{v}_1 \| + \cdots + | c_n | \, \| \mathbf{v}_n \| \bigr) = C \, \rho(T)^k, \end{aligned} \tag{7.15}$$

for some constant $C > 0$ that depends only upon the initial conditions. In particular, if $\rho(T) < 1$, then

$$\| \mathbf{u}^{(k)} \| \;\le\; C \, \rho(T)^k \;\longrightarrow\; 0 \qquad \text{as} \qquad k \to \infty, \tag{7.16}$$

in accordance with Theorem 7.11. Thus, the spectral radius prescribes the rate of convergence of the solutions to equilibrium. The smaller the spectral radius, the faster the solutions go to $\mathbf{0}$.

If T has only one largest (simple) eigenvalue, so $| \lambda_1 | > | \lambda_j |$ for all $j > 1$, then the first term in the solution formula (7.4) will eventually dominate all the others: $\| \lambda_1^k \mathbf{v}_1 \| \gg \| \lambda_j^k \mathbf{v}_j \|$ for $j > 1$ and $k \gg 0$. Therefore, provided that $c_1 \neq 0$, the solution (7.4) has the asymptotic formula

$$\mathbf{u}^{(k)} \;\approx\; c_1 \lambda_1^k \mathbf{v}_1, \tag{7.17}$$

and so most solutions end up parallel to $\mathbf{v}_1$. In particular, if $| \lambda_1 | = \rho(T) < 1$, such a solution approaches $\mathbf{0}$ along the direction of the dominant eigenvector $\mathbf{v}_1$ at a rate governed by the modulus of the dominant eigenvalue. The exceptional solutions, with $c_1 = 0$, tend to $\mathbf{0}$ at a faster rate, along one of the other eigendirections. In practical computations, one rarely observes the exceptional solutions. Indeed, even if the initial condition does not involve the dominant eigenvector, round-off error during the iteration will almost inevitably introduce a small component in the direction of $\mathbf{v}_1$, which will, if you wait long enough, eventually dominate the computation.

Warning: The inequality (7.15) only applies to complete matrices. In the general case, one can prove that the solution satisfies the slightly weaker inequality

$$\| \mathbf{u}^{(k)} \| \le C \, \sigma^k \qquad \text{for all} \qquad k \ge 0, \qquad \text{where} \qquad \sigma > \rho(T) \tag{7.18}$$

is any number larger than the spectral radius, while $C > 0$ is a positive constant (whose value may depend on how close σ is to ρ).

Example 7.12. According to Example 7.5, the matrix

$$T = \begin{pmatrix} -3 & 1 & 6 \\ 1 & -1 & -2 \\ -1 & -1 & 0 \end{pmatrix} \qquad \text{has eigenvalues} \qquad \begin{aligned} \lambda_1 &= -2, \\ \lambda_2 &= -1 + \mathrm{i}\,, \\ \lambda_3 &= -1 - \mathrm{i}\,. \end{aligned}$$

Since $| \lambda_1 | = 2 > | \lambda_2 | = | \lambda_3 | = \sqrt{2}$, the spectral radius is $\rho(T) = | \lambda_1 | = 2$. We conclude that T is not a convergent matrix. As the reader can check, either directly, or from the solution formula (7.13), the vectors $\mathbf{u}^{(k)} = T^k \mathbf{u}^{(0)}$ obtained by repeatedly multiplying any nonzero initial vector $\mathbf{u}^{(0)}$ by T rapidly go off to ∞, at a rate roughly equal to $\rho(T)^k = 2^k$.

On the other hand, the matrix

$$\widetilde{T} = -\tfrac{1}{3}\,T = \begin{pmatrix} 1 & -\frac{1}{3} & -2 \\ -\frac{1}{3} & \frac{1}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{1}{3} & 0 \end{pmatrix} \qquad \text{with eigenvalues} \qquad \begin{aligned} \lambda_1 &= \tfrac{2}{3}, \\ \lambda_2 &= \tfrac{1}{3} - \tfrac{1}{3}\,\mathrm{i}\,, \\ \lambda_3 &= \tfrac{1}{3} + \tfrac{1}{3}\,\mathrm{i}\,, \end{aligned}$$

has spectral radius $\rho(\widetilde{T}) = \frac{2}{3}$, and hence is a convergent matrix. According to (7.17), if we write the initial data $\mathbf{u}^{(0)} = c_1\,\mathbf{v}_1 + c_2\,\mathbf{v}_2 + c_3\,\mathbf{v}_3$ as a linear combination of the eigenvectors, then, provided $c_1 \neq 0$, the iterates have the asymptotic form $\mathbf{u}^{(k)} \approx c_1\,\left(-\frac{2}{3}\right)^k\,\mathbf{v}_1$, where $\mathbf{v}_1 = (\,4, -2, 1\,)^T$ is the eigenvector corresponding to the dominant eigenvalue $\lambda_1 = -\frac{2}{3}$. Thus, for most initial vectors, the iterates end up decreasing in length by a factor of almost exactly $\frac{2}{3}$, eventually becoming parallel to the dominant eigenvector $\mathbf{v}_1$. This is borne out by a sample computation: starting with $\mathbf{u}^{(0)} = (\,1, 1, 1\,)^T$, the first ten iterates are

$$\begin{pmatrix} -.0936 \\ .0462 \\ -.0231 \end{pmatrix}, \quad \begin{pmatrix} -.0627 \\ .0312 \\ -.0158 \end{pmatrix}, \quad \begin{pmatrix} -.0416 \\ .0208 \\ -.0105 \end{pmatrix}, \quad \begin{pmatrix} -.0275 \\ .0138 \\ -.0069 \end{pmatrix}, \quad \begin{pmatrix} -.0182 \\ .0091 \\ -.0046 \end{pmatrix},$$

$$\begin{pmatrix} -.0121 \\ .0061 \\ -.0030 \end{pmatrix}, \quad \begin{pmatrix} -.0081 \\ .0040 \\ -.0020 \end{pmatrix}, \quad \begin{pmatrix} -.0054 \\ .0027 \\ -.0013 \end{pmatrix}, \quad \begin{pmatrix} -.0036 \\ .0018 \\ -.0009 \end{pmatrix}, \quad \begin{pmatrix} -.0024 \\ .0012 \\ -.0006 \end{pmatrix}.$$

7.3. Matrix Norms.

The convergence of a linear iterative system is governed by the spectral radius or largest eigenvalue (in modulus) of the coefficient matrix. Unfortunately, finding accurate approximations to the eigenvalues of most matrices is a nontrivial computational task. Indeed, all practical numerical algorithms rely on some form of iteration. But using iteration to determine the spectral radius defeats the purpose, which is to predict the behavior of the iterative system in advance!

In this section, we present two alternative approaches for directly investigating convergence and stability issues. Matrix norms form a natural class of norms on the vector space of $n \times n$ matrices and can, in many instances, be used to establish convergence with a minimal effort.

Matrix Norms

We work exclusively with real $n \times n$ matrices in this section, although the results straightforwardly extend to complex matrices. We begin by fixing a norm $\|\cdot\|$ on $\mathbb{R}^n$. The norm may or may not come from an inner product — this is irrelevant as far as the construction goes. Each norm on $\mathbb{R}^n$ will naturally induce a norm on the vector space $\mathcal{M}_{n\times n}$ of all $n \times n$ matrices. Roughly speaking, the matrix norm tells us how much a linear transformation stretches vectors relative to the given norm.

Theorem 7.13. *If* $\| \cdot \|$ *is any norm on* $\mathbb{R}^n$, *then the quantity*

$$\| A \| = \max \{ \| A \mathbf{u} \| \mid \| \mathbf{u} \| = 1 \} \tag{7.19}$$

defines a norm on $\mathcal{M}_{n \times n}$, *known as the* natural matrix norm.

Proof: First note that $\| A \| < \infty$, since the maximum is taken on a closed and bounded subset, namely the unit sphere $S_1 = \{ \| \mathbf{u} \| = 1 \}$ for the given norm. To show that (7.19) defines a norm, we need to verify the three basic axioms of Definition 5.8.

Non-negativity, $\| A \| \geq 0$, is immediate. Suppose $\| A \| = 0$. This means that, for every unit vector, $\| A \mathbf{u} \| = 0$, and hence $A \mathbf{u} = \mathbf{0}$ whenever $\| \mathbf{u} \| = 1$. If $\mathbf{0} \neq \mathbf{v} \in \mathbb{R}^n$ is any nonzero vector, then $\mathbf{u} = \mathbf{v}/r$, where $r = \| \mathbf{v} \|$, is a unit vector, so

$$A \mathbf{v} = A(r \, \mathbf{u}) = r \, A \mathbf{u} = \mathbf{0}. \tag{7.20}$$

Therefore, $A \mathbf{v} = \mathbf{0}$ for every $\mathbf{v} \in \mathbb{R}^n$, which implies $A = \mathrm{O}$ is the zero matrix. This serves to prove the positivity property. As for homogeneity, if $c \in \mathbb{R}$ is any scalar,

$$\| c \, A \| = \max \{ \| c A \mathbf{u} \| \} = \max \{ | c | \, \| A \mathbf{u} \| \} = | c | \max \{ \| A \mathbf{u} \| \} = | c | \, \| A \|.$$

Finally, to prove the triangle inequality, we use the fact that the maximum of the sum of quantities is bounded by the sum of their individual maxima. Therefore, since the norm on $\mathbb{R}^n$ satisfies the triangle inequality,

$$\begin{aligned} \| A + B \| &= \max \{ \| A \mathbf{u} + B \mathbf{u} \| \} \leq \max \{ \| A \mathbf{u} \| + \| B \mathbf{u} \| \} \\ &\leq \max \{ \| A \mathbf{u} \| \} + \max \{ \| B \mathbf{u} \| \} = \| A \| + \| B \|. \end{aligned}$$

Q.E.D.

The property that distinguishes a matrix norm from a generic norm on the space of matrices is the fact that it also obeys a very useful *product inequality*.

Theorem 7.14. *A natural matrix norm satisfies*

$$\| A \mathbf{v} \| \leq \| A \| \, \| \mathbf{v} \|, \qquad \text{for all} \qquad A \in \mathcal{M}_{n \times n}, \quad \mathbf{v} \in \mathbb{R}^n. \tag{7.21}$$

Furthermore,

$$\| A B \| \leq \| A \| \, \| B \|, \qquad \text{for all} \qquad A, B \in \mathcal{M}_{n \times n}. \tag{7.22}$$

Proof: Note first that, by definition $\| A \mathbf{u} \| \leq \| A \|$ for all unit vectors $\| \mathbf{u} \| = 1$. Then, letting $\mathbf{v} = r \, \mathbf{u}$ where $\mathbf{u}$ is a unit vector and $r = \| \mathbf{v} \|$, we have

$$\| A \mathbf{v} \| = \| A (r \, \mathbf{u}) \| = r \, \| A \mathbf{u} \| \leq r \, \| A \| = \| \mathbf{v} \| \, \| A \|,$$

proving the first inequality. To prove the second, we apply the first to compute

$$\begin{aligned} \| A \, B \| &= \max \{ \| A \, B \mathbf{u} \| \} = \max \{ \| A \, (B \mathbf{u}) \| \} \\ &\leq \max \{ \| A \| \, \| B \mathbf{u} \| \} = \| A \| \max \{ \| B \mathbf{u} \| \} = \| A \| \, \| B \|. \end{aligned}$$

Q.E.D.

Remark: In general, a norm on the vector space of $n \times n$ matrices is called a *matrix norm* if it also satisfies the multiplicative inequality (7.22). Most, but not all, matrix norms used in applications come from norms on the underlying vector space.

The multiplicative inequality (7.22) implies, in particular, that $\| A^2 \| \leq \| A \|^2$; equality is not necessarily valid. More generally:

Proposition 7.15. *If A is a square matrix, then $\| A^k \| \leq \| A \|^k$. In particular, if $\| A \| < 1$, then $\| A^k \| \to 0$ as $k \to \infty$, and hence A is a convergent matrix*: $A^k \to \mathrm{O}$.

The converse is not quite true; a convergent matrix does not necessarily have matrix norm less than 1, or even ≤ 1 — see Example 7.20 below. An alternative proof of Proposition 7.15 can be based on the following useful estimate:

Theorem 7.16. *The spectral radius of a matrix is bounded by its matrix norm*:

$$\rho(A) \leq \| A \|. \tag{7.23}$$

Proof: If λ is a real eigenvalue, and $\mathbf{u}$ a corresponding unit eigenvector, so that $A\mathbf{u} = \lambda\,\mathbf{u}$ with $\| \mathbf{u} \| = 1$, then

$$\| A\mathbf{u} \| = \| \lambda\,\mathbf{u} \| = |\,\lambda\,|\;\| \mathbf{u} \| = |\,\lambda\,|. \tag{7.24}$$

Since $\| A \|$ is the maximum of $\| A\mathbf{u} \|$ over all possible unit vectors, this implies that

$$|\,\lambda\,| \leq \| A \|. \tag{7.25}$$

If all the eigenvalues of A are real, then the spectral radius is the maximum of their absolute values, and so it too is bounded by $\| A \|$, proving (7.23).

If A has complex eigenvalues, then we need to work a little harder to establish (7.25). (This is because the matrix norm is defined by the effect of A on *real* vectors, and so we cannot directly use the complex eigenvectors to establish the required bound.) Let $\lambda = r\,e^{\mathrm{i}\,\theta}$ be a complex eigenvalue with complex eigenvector $\mathbf{z} = \mathbf{x} + \mathrm{i}\,\mathbf{y}$. Define

$$m = \min \left\{ \| \operatorname{Re}\,(e^{\mathrm{i}\,\varphi}\,\mathbf{z}) \| = \| (\cos\varphi)\,\mathbf{x} - (\sin\varphi)\,\mathbf{y} \| \;\Big|\; 0 \leq \varphi \leq 2\pi \right\}. \tag{7.26}$$

Since the indicated subset is a closed curve (in fact, an ellipse) that does not go through the origin, $m > 0$. Let φ_0 denote the value of the angle that produces the minimum, so

$$m = \| (\cos\varphi_0)\,\mathbf{x} - (\sin\varphi_0)\,\mathbf{y} \| = \| \operatorname{Re}\left(e^{\mathrm{i}\,\varphi_0}\,\mathbf{z}\right) \|.$$

Define the real unit vector

$$\mathbf{u} = \frac{\operatorname{Re}\left(e^{\mathrm{i}\,\varphi_0}\,\mathbf{z}\right)}{m} = \frac{(\cos\varphi_0)\,\mathbf{x} - (\sin\varphi_0)\,\mathbf{y}}{m}, \qquad \text{so that} \qquad \| \mathbf{u} \| = 1.$$

Then

$$A\mathbf{u} = \frac{1}{m}\operatorname{Re}\left(e^{\mathrm{i}\,\varphi_0}\,A\,\mathbf{z}\right) = \frac{1}{m}\operatorname{Re}\left(e^{\mathrm{i}\,\varphi_0}\,r\,e^{\mathrm{i}\,\theta}\,\mathbf{z}\right) = \frac{r}{m}\operatorname{Re}\left(e^{\mathrm{i}\,(\varphi_0+\theta)}\,\mathbf{z}\right).$$

Therefore, keeping in mind that m is the minimal value in (7.26),

$$\| A \| \geq \| A\mathbf{u} \| = \frac{r}{m}\,\| \operatorname{Re}\left(e^{\mathrm{i}\,(\varphi_0+\theta)}\,\mathbf{z}\right) \| \geq r = |\,\lambda\,|, \tag{7.27}$$

and so (7.25) also holds for complex eigenvalues. *Q.E.D.*

Explicit Formulae

Let us now determine the explicit formulae for the matrix norms induced by our most important vector norms on $\mathbb{R}^n$. The simplest to handle is the ∞ norm

$$\| \mathbf{v} \|_\infty = \max \{ | v_1 |, \ \ldots \ , | v_n | \}.$$

Definition 7.17. The i^{th} *absolute row sum* of a matrix A is the sum of the absolute values of the entries in the i^{th} row:

$$s_i = | a_{i1} | + \ \cdots \ + | a_{in} | = \sum_{j=1}^{n} | a_{ij} |. \tag{7.28}$$

Proposition 7.18. *The* ∞ matrix norm *of a matrix A is equal to its maximal absolute row sum*:

$$\| A \|_\infty = \max\{ s_1, \ldots, s_n \} = \max \left\{ \sum_{j=1}^{n} | a_{ij} | \ \middle| \ 1 \leq i \leq n \right\}. \tag{7.29}$$

Proof: Let $s = \max\{ s_1, \ldots, s_n \}$ denote the right hand side of (7.29). Given any $\mathbf{v} \in \mathbb{R}^n$, we compute

$$\begin{aligned} \| A \mathbf{v} \|_\infty \ &= \ \max \left\{ \left| \sum_{j=1}^{n} a_{ij} v_j \right| \right\} \ \leq \ \max \left\{ \sum_{j=1}^{n} | a_{ij} v_j | \right\} \\ &\leq \ \max \left\{ \sum_{j=1}^{n} | a_{ij} | \right\} \ \max \{ | v_j | \} = s \, \| \mathbf{v} \|_\infty . \end{aligned}$$

In particular, by specializing to $\| \mathbf{v} \|_\infty = 1$, we deduce that $\| A \|_\infty \leq s$.

On the other hand, suppose the maximal absolute row sum occurs at row i, so

$$s_i = \sum_{j=1}^{n} | a_{ij} | = s. \tag{7.30}$$

Let $\mathbf{u} \in \mathbb{R}^n$ be the specific vector that has the following entries: $u_j = +1$ if $a_{ij} > 0$, while $u_j = -1$ if $a_{ij} < 0$. Then $\| \mathbf{u} \|_\infty = 1$. Moreover, since $a_{ij} u_j = | a_{ij} |$, the i^{th} entry of $A \mathbf{u}$ is equal to the i^{th} absolute row sum (7.30). This implies that

$$\| A \|_\infty \geq \| A \mathbf{u} \|_\infty \geq s. \qquad \textit{Q.E.D.}$$

Combining Propositions 7.15 and 7.18, we have established the following convergence criterion.

Corollary 7.19. *If all the absolute row sums of A are strictly less than* 1, *then* $\| A \|_\infty < 1$ *and hence A is a convergent matrix.*

Example 7.20. Consider the symmetric matrix $A = \begin{pmatrix} \frac{1}{2} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{1}{4} \end{pmatrix}$. Its two absolute row sums are $\left| \frac{1}{2} \right| + \left| -\frac{1}{3} \right| = \frac{5}{6}$, $\left| -\frac{1}{3} \right| + \left| \frac{1}{4} \right| = \frac{7}{12}$, so

$$\| A \|_\infty = \max \left\{ \tfrac{5}{6}, \tfrac{7}{12} \right\} = \tfrac{5}{6} \approx .83333\ldots .$$

Since the norm is less than 1, A is a convergent matrix. Indeed, its eigenvalues are

$$\lambda_1 = \frac{9 + \sqrt{73}}{24} \approx .7310\ldots , \qquad \lambda_2 = \frac{9 - \sqrt{73}}{24} \approx .0190\ldots ,$$

and hence the spectral radius is

$$\rho(A) = \frac{9 + \sqrt{73}}{24} \approx .7310\ldots ,$$

which is slightly smaller than its ∞ norm.

The row sum test for convergence is not always conclusive. For example, the matrix

$$A = \begin{pmatrix} \frac{1}{2} & -\frac{3}{5} \\ -\frac{3}{5} & \frac{1}{4} \end{pmatrix} \qquad \text{has matrix norm} \qquad \| A \|_\infty = \tfrac{11}{10} > 1. \tag{7.31}$$

On the other hand, its eigenvalues are $(15 \pm \sqrt{601}\,)/40$, and hence its spectral radius is

$$\rho(A) = \frac{15 + \sqrt{601}}{40} \approx .98788\ldots ,$$

which implies that A is (just barely) convergent, even though its maximal row sum is larger than 1.

The matrix norm associated with the Euclidean norm $\| \mathbf{v} \|_2 = \sqrt{v_1^2 + \cdots + v_n^2}$ is given by largest singular value.

Proposition 7.21. *The matrix norm corresponding to the Euclidean norm equals the maximal singular value*:

$$\| A \|_2 = \sigma_1 = \max \{ \sigma_1, \ldots, \sigma_r \}, \qquad r = \operatorname{rank} A > 0, \qquad \text{while} \qquad \| \mathrm{O} \|_2 = 0. \tag{7.32}$$

Unfortunately, as we discovered in Example 7.20, matrix norms are not a foolproof test of convergence. There exist convergent matrices such that $\rho(A) < 1$ and yet have matrix norm $\| A \| \geq 1$. In such cases, the matrix norm is not able to predict convergence of the iterative system, although one should expect the convergence to be quite slow. Although such pathology might show up in the chosen matrix norm, it turns out that one can always rig up some matrix norm for which $\| A \| < 1$. This follows from a more general result, whose proof can be found in [**44**].

Theorem 7.22. *Let A have spectral radius $\rho(A)$. If $\varepsilon > 0$ is any positive number, then there exists a matrix norm $\| \cdot \|$ such that*

$$\rho(A) \leq \| A \| < \rho(A) + \varepsilon. \tag{7.33}$$

Corollary 7.23. *If A is a convergent matrix, then there exists a matrix norm such that $\| A \| < 1$.*

Proof: By definition, A is convergent if and only if $\rho(A) < 1$. Choose $\varepsilon > 0$ such that $\rho(A) + \varepsilon < 1$. Any norm that satisfies (7.33) has the desired property. *Q.E.D.*

Remark: Based on the accumulated evidence, one might be tempted to speculate that the spectral radius itself defines a matrix norm. Unfortunately, this is not the case. For example, the nonzero matrix $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ has zero spectral radius, $\rho(A) = 0$, in violation of a basic norm axiom.

7.4. Iterative Solution of Linear Algebraic Systems.

In this section, we return to the most basic problem in linear algebra: solving the linear algebraic system

$$A\mathbf{u} = \mathbf{b}, \tag{7.34}$$

consisting of n equations in n unknowns. We assume that the coefficient matrix A is nonsingular, and so the solution $\mathbf{u} = A^{-1}\mathbf{b}$ is unique.

We will introduce several popular iterative methods that can be used to approximate the solution for certain classes of coefficient matrices. The resulting algorithms will provide an attractive alternative to Gaussian Elimination, particularly when dealing with the large, sparse systems that arise in the numerical solution to differential equations. One major advantage of an iterative technique is that it (typically) produces progressively more and more accurate approximations to the solution, and hence, by prolonging the iterations, can, at least in principle, compute the solution to any desired order of accuracy. Moreover, even performing just a few iterations may produce a reasonable approximation to the true solution — in stark contrast to Gaussian Elimination, where one must continue the algorithm through to the bitter end before any useful information can be extracted. A partially completed Gaussian Elimination is of scant use! A significant weakness is that iterative schemes are not universally applicable, and their design relies upon the detailed structure of the coefficient matrix.

We shall be attempting to solve the linear system (7.34) by replacing it with an iterative system of the form

$$\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)} + \mathbf{c}, \qquad \mathbf{u}^{(0)} = \mathbf{u}_0, \tag{7.35}$$

in which T is an $n \times n$ matrix and $\mathbf{c}$ a vector. This represents a slight generalization of our earlier iterative system (7.1), in that the right hand side is now an affine function of $\mathbf{u}^{(k)}$. Suppose that the solutions to the affine iterative system converge: $\mathbf{u}^{(k)} \to \mathbf{u}^\star$ as $k \to \infty$. Then, by taking the limit of both sides of (7.35), we discover that the limit point $\mathbf{u}^\star$ solves the *fixed-point equation*

$$\mathbf{u}^\star = T\mathbf{u}^\star + \mathbf{c}. \tag{7.36}$$

Thus, we need to design our iterative system so that

(*a*) te solution to the fixed-point system $\mathbf{u} = T\mathbf{u} + \mathbf{c}$ coincides with the solution to the original system $A\mathbf{u} = \mathbf{b}$, and

(*b*) the iterates defined by (7.35) are known to converge to the fixed point.

Before exploring these issues in depth, let us look at a simple example.

Example 7.24. Consider the linear system

$$3x + y - z = 3, \qquad x - 4y + 2z = -1, \qquad -2x - y + 5z = 2, \tag{7.37}$$

which has the vectorial form $A\mathbf{u} = \mathbf{b}$, with

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}, \qquad \mathbf{u} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix}.$$

One easy way to convert a linear system into a fixed-point form is to rewrite it as

$$\mathbf{u} = \mathrm{I}\,\mathbf{u} - A\mathbf{u} + A\mathbf{u} = (\mathrm{I} - A)\mathbf{u} + \mathbf{b} = T\mathbf{u} + \mathbf{c}, \qquad \text{where} \qquad T = \mathrm{I} - A, \qquad \mathbf{c} = \mathbf{b}.$$

In the present case,

$$T = \mathrm{I} - A = \begin{pmatrix} -2 & -1 & 1 \\ -1 & 5 & -2 \\ 2 & 1 & -4 \end{pmatrix}, \qquad \mathbf{c} = \mathbf{b} = \begin{pmatrix} 3 \\ -1 \\ 2 \end{pmatrix}.$$

The resulting iterative system $\mathbf{u}^{(k+1)} = T\mathbf{u}^{(k)} + \mathbf{c}$ has the explicit form

$$\begin{aligned} x^{(k+1)} &= -2x^{(k)} - y^{(k)} + z^{(k)} + 3, \\ y^{(k+1)} &= \;\; - x^{(k)} + 5y^{(k)} - 2z^{(k)} - 1, \\ z^{(k+1)} &= \;\;\; 2x^{(k)} + y^{(k)} - 4z^{(k)} + 2. \end{aligned} \tag{7.38}$$

Another possibility is to solve the first equation in (7.37) for x, the second for y, and the third for z, so that

$$x = -\tfrac{1}{3}\,y + \tfrac{1}{3}\,z + 1, \qquad y = \tfrac{1}{4}\,x + \tfrac{1}{2}\,z + \tfrac{1}{4}, \qquad z = \tfrac{2}{5}\,x + \tfrac{1}{5}\,y + \tfrac{2}{5}.$$

The resulting equations have the form of a fixed-point system

$$\mathbf{u} = \widehat{T}\mathbf{u} + \widehat{\mathbf{c}}, \qquad \text{in which} \qquad \widehat{T} = \begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{2}{5} & \frac{1}{5} & 0 \end{pmatrix}, \qquad \widehat{\mathbf{c}} = \begin{pmatrix} 1 \\ \frac{1}{4} \\ \frac{2}{5} \end{pmatrix}.$$

The corresponding iteration $\mathbf{u}^{(k+1)} = \widehat{T}\mathbf{u}^{(k)} + \widehat{\mathbf{c}}$ takes the explicit form

$$\begin{aligned} x^{(k+1)} &= -\tfrac{1}{3}\,y^{(k)} + \tfrac{1}{3}\,z^{(k)} + 1, \\ y^{(k+1)} &= \;\;\tfrac{1}{4}\,x^{(k)} + \tfrac{1}{2}\,z^{(k)} + \tfrac{1}{4}, \\ z^{(k+1)} &= \;\;\tfrac{2}{5}\,x^{(k)} + \tfrac{1}{5}\,y^{(k)} + \tfrac{2}{5}. \end{aligned} \tag{7.39}$$

Do the resulting iterative schemes converge to the solution $x = y = z = 1$? The results, starting with initial guess $\mathbf{u}^{(0)} = (0,0,0)$, are tabulated as follows.

k	$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)} + \mathbf{b}$			$\mathbf{u}^{(k+1)} = \widehat{T}\,\mathbf{u}^{(k)} + \widehat{\mathbf{c}}$		
0	0	0	0	0	0	0
1	3	-1	2	1	.25	.4
2	0	-13	-1	1.05	.7	.85
3	15	-64	-7	1.05	.9375	.96
4	30	-322	-4	1.0075	.9925	1.0075
5	261	-1633	-244	1.005	1.00562	1.0015
6	870	-7939	-133	.9986	1.002	1.0031
7	6069	-40300	-5665	1.0004	1.0012	.9999
8	22500	-196240	-5500	.9995	1.0000	1.0004
9	145743	-992701	-129238	1.0001	1.0001	.9998
10	571980	-4850773	-184261	.9999	.9999	1.0001
11	3522555	-24457324	-2969767	1.0000	1.0000	1.0000

For the first scheme, the answer is clearly no — the iterates become wilder and wilder. Indeed, this occurs no matter how close the initial guess $\mathbf{u}^{(0)}$ is to the actual solution — unless $\mathbf{u}^{(0)} = \mathbf{u}^\star$ happens to be exactly equal. In the second case, the iterates do converge to the solution, and it does not take too long, even starting from a poor initial guess, to obtain a reasonably accurate approximation. Of course, in such a simple example, it would be silly to use iteration, when Gaussian Elimination can be done by hand and produces the solution almost immediately. However, we use the small examples for illustrative purposes, bringing the full power of iterative schemes to bear on the large linear systems arising in applications.

The convergence of solutions to (7.35) to the fixed point $\mathbf{u}^\star$ is based on the behavior of the *error vectors*

$$\mathbf{e}^{(k)} = \mathbf{u}^{(k)} - \mathbf{u}^\star, \tag{7.40}$$

which measure how close the iterates are to the true solution. Let us find out how the successive error vectors are related. We compute

$$\mathbf{e}^{(k+1)} = \mathbf{u}^{(k+1)} - \mathbf{u}^\star = (T\,\mathbf{u}^{(k)} + \mathbf{a}) - (T\,\mathbf{u}^\star + \mathbf{a}) = T(\mathbf{u}^{(k)} - \mathbf{u}^\star) = T\,\mathbf{e}^{(k)},$$

showing that the error vectors satisfy a *linear* iterative system

$$\mathbf{e}^{(k+1)} = T\,\mathbf{e}^{(k)}, \tag{7.41}$$

with the *same* coefficient matrix T. Therefore, they are given by the explicit formula

$$\mathbf{e}^{(k)} = T^k\,\mathbf{e}^{(0)}.$$

Now, the solutions to (7.35) converge to the fixed point, $\mathbf{u}^{(k)} \to \mathbf{u}^\star$, if and only if the error vectors converge to zero: $\mathbf{e}^{(k)} \to \mathbf{0}$ as $k \to \infty$. Our analysis of linear iterative systems, as summarized in Proposition 7.8, establishes the following basic convergence result.

Proposition 7.25. *The affine iterative system* (7.35) *will converge to the solution to the fixed point equation* (7.36) *if and only if* T *is a convergent matrix:* $\rho(T) < 1$.

The spectral radius $\rho(T)$ of the coefficient matrix will govern the speed of convergence. Therefore, our main goal is to construct an iterative scheme whose coefficient matrix has as small a spectral radius as possible. At the very least, the spectral radius must be less than 1. For the two iterative schemes presented in Example 7.24, the spectral radii of the coefficient matrices are found to be

$$\rho(T) \approx 4.9675, \qquad\qquad \rho(\widehat{T}\,) = .5.$$

Therefore, T is not a convergent matrix, which explains the wild behavior of its iterates, whereas $\widehat{T}$ is convergent, and one expects the error to roughly decrease by a factor of $\frac{1}{2}$ at each step.

The Jacobi Method

The first general iterative scheme for solving linear systems is based on the same simple idea used in our illustrative Example 7.24. Namely, we solve the $i^{\rm th}$ equation in the system $A\mathbf{u} = \mathbf{b}$, which is

$$\sum_{j=1}^{n} a_{ij}\,u_j = b_i,$$

for the $i^{\rm th}$ variable u_i. To do this, we need to assume that all the diagonal entries of A are nonzero: $a_{ii} \neq 0$. The result is

$$u_i \;=\; -\,\frac{1}{a_{ii}} \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}\,u_j + \frac{b_i}{a_{ii}} \;=\; \sum_{j=1}^{n} t_{ij}\,u_j + c_i, \tag{7.42}$$

where

$$t_{ij} = \begin{cases} -\,\dfrac{a_{ij}}{a_{ii}}\,, & i \neq j, \\ 0, & i = j, \end{cases} \qquad\quad \text{and} \qquad\quad c_i = \frac{b_i}{a_{ii}}\,. \tag{7.43}$$

The result has the form of a fixed-point system $\mathbf{u} = T\,\mathbf{u} + \mathbf{c}$, and forms the basis of the *Jacobi method*

$$\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)} + \mathbf{c}, \qquad\qquad \mathbf{u}^{(0)} = \mathbf{u}_0, \tag{7.44}$$

named after the influential nineteenth century German analyst Carl Jacobi. The explicit form of the Jacobi iterative scheme is

$$u_i^{(k+1)} \;=\; -\,\frac{1}{a_{ii}} \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}\,u_j^{(k)} + \frac{b_i}{a_{ii}}\,. \tag{7.45}$$

It is instructive to rederive the Jacobi method in a direct matrix form. We begin by decomposing the coefficient matrix

$$A = L + D + U \tag{7.46}$$

into the sum of a strictly lower triangular matrix L, a diagonal matrix D, and a strictly upper triangular matrix U, each of which is uniquely specified. For example, when

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}, \tag{7.47}$$

the decomposition (7.46) yields

$$L = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -2 & -1 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 3 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 5 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Warning: The L, D, U in the elementary additive decomposition (7.46) have nothing to do with the L, D, U appearing in factorizations arising from Gaussian Elimination. The latter play no role in the iterative solution methods considered here.

We then rewrite the system

$$A\,\mathbf{u} = (L + D + U)\,\mathbf{u} = \mathbf{b} \qquad \text{in the alternative form} \qquad D\,\mathbf{u} = -\,(L + U)\,\mathbf{u} + \mathbf{b}.$$

The Jacobi fixed point equations (7.42) amounts to solving for

$$\mathbf{u} = T\,\mathbf{u} + \mathbf{c}, \qquad \text{where} \qquad T = -\,D^{-1}(L + U), \qquad \mathbf{c} = D^{-1}\mathbf{b}. \tag{7.48}$$

For the example (7.47), we recover the Jacobi iteration matrix as

$$T = -\,D^{-1}(L + U) = \begin{pmatrix} 0 & -\frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & 0 & \frac{1}{2} \\ \frac{2}{5} & \frac{1}{5} & 0 \end{pmatrix}.$$

Deciding in advance whether or not the Jacobi method will converge is not easy. However, it can be shown that Jacobi iteration *is* guaranteed to converge when the original coefficient matrix has large diagonal entries, in accordance with Definition 6.25.

Theorem 7.26. *If A is strictly diagonally dominant, then the associated Jacobi iteration scheme converges.*

Proof: We shall prove that $\|\,T\,\|_\infty < 1$, and so Corollary 7.19 implies that T is a convergent matrix. The absolute row sums of the Jacobi matrix $T = -\,D^{-1}(L + U)$ are, according to (7.43),

$$s_i = \sum_{j=1}^{n} |\, t_{ij} \,| = \frac{1}{|\, a_{ii} \,|} \sum_{\substack{j=1 \\ j \neq i}}^{n} |\, a_{ij} \,| < 1, \tag{7.49}$$

because A is strictly diagonally dominant. Thus, $\|\,T\,\|_\infty = \max\{s_1, \ldots, s_n\} < 1$, and the result follows. *Q.E.D.*

Example 7.27. Consider the linear system

$$\begin{aligned} 4x + y + w &= 1, \\ x + 4y + z + v &= 2, \\ y + 4z + w &= -1, \\ x + z + 4w + v &= 2, \\ y + w + 4v &= 1. \end{aligned}$$

The Jacobi method solves the respective equations for x, y, z, w, v, leading to the iterative scheme

$$\begin{aligned} x^{(k+1)} &= -\tfrac{1}{4}\,y^{(k)} - \tfrac{1}{4}\,w^{(k)} + \tfrac{1}{4}\,, \\ y^{(k+1)} &= -\tfrac{1}{4}\,x^{(k)} - \tfrac{1}{4}\,z^{(k)} - \tfrac{1}{4}\,v^{(k)} + \tfrac{1}{2}\,, \\ z^{(k+1)} &= -\tfrac{1}{4}\,y^{(k)} - \tfrac{1}{4}\,w^{(k)} - \tfrac{1}{4}\,, \\ w^{(k+1)} &= -\tfrac{1}{4}\,x^{(k)} - \tfrac{1}{4}\,z^{(k)} - \tfrac{1}{4}\,v^{(k)} + \tfrac{1}{2}\,, \\ v^{(k+1)} &= -\tfrac{1}{4}\,y^{(k)} - \tfrac{1}{4}\,w^{(k)} + \tfrac{1}{4}\,. \end{aligned}$$

The coefficient matrix of the original system,

$$A = \begin{pmatrix} 4 & 1 & 0 & 1 & 0 \\ 1 & 4 & 1 & 0 & 1 \\ 0 & 1 & 4 & 1 & 0 \\ 1 & 0 & 1 & 4 & 1 \\ 0 & 1 & 0 & 1 & 4 \end{pmatrix},$$

is diagonally dominant, and so we are guaranteed that the Jacobi iterations will eventually converge to the solution. Indeed, the Jacobi scheme takes the iterative form (7.48), with

$$T = \begin{pmatrix} 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} & 0 & -\frac{1}{4} \\ 0 & -\frac{1}{4} & 0 & -\frac{1}{4} & 0 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \\ -\frac{1}{4} \\ \frac{1}{2} \\ \frac{1}{4} \end{pmatrix}.$$

Note that $\| T \|_\infty = \frac{3}{4} < 1$, validating convergence of the scheme. Thus, to obtain, say, four decimal place accuracy in the solution, we estimate that it would take less than $\log(.5 \times 10^{-4}) / \log .75 \approx 34$ iterates, assuming a moderate initial error. But the matrix norm always underestimates the true rate of convergence, as prescribed by the spectral radius $\rho(T) = .6124$, which would imply about $\log(.5 \times 10^{-4}) / \log .6124 \approx 20$ iterations to obtain the desired accuracy. Indeed, starting with the initial guess $x^{(0)} = y^{(0)} = z^{(0)} = w^{(0)} = v^{(0)} = 0$, the Jacobi iterates converge to the exact solution

$$x = -.1, \qquad y = .7, \qquad z = -.6, \qquad w = .7, \qquad v = -.1,$$

to within four decimal places in exactly 20 iterations.

The Gauss–Seidel Method

The Gauss–Seidel method relies on a slightly more refined implementation of the Jacobi process. To understand how it works, it will help to write out the Jacobi iteration scheme (7.44) in full detail:

$$\begin{array}{rllllllll}
u_1^{(k+1)} &= & & t_{12}\,u_2^{(k)} & + \; t_{13}\,u_3^{(k)} & + \;\cdots & + \; t_{1,n-1}\,u_{n-1}^{(k)} & + \; t_{1n}\,u_n^{(k)} & + \; c_1, \\
u_2^{(k+1)} &= & t_{21}\,u_1^{(k)} & & + \; t_{23}\,u_3^{(k)} & + \;\cdots & + \; t_{2,n-1}\,u_{n-1}^{(k)} & + \; t_{2n}\,u_n^{(k)} & + \; c_2, \\
u_3^{(k+1)} &= & t_{31}\,u_1^{(k)} & + \; t_{32}\,u_2^{(k)} & & \;\cdots & + \; t_{3,n-1}\,u_{n-1}^{(k)} & + \; t_{3n}\,u_n^{(k)} & + \; c_3, \\
\vdots & & \vdots & \vdots & & \ddots & & \ddots & \vdots \\
u_n^{(k+1)} &= & t_{n1}\,u_1^{(k)} & + \; t_{n2}\,u_2^{(k)} & + \; t_{n3}\,u_3^{(k)} & + \;\cdots & + \; t_{n,n-1}\,u_{n-1}^{(k)} & & + \; c_n,
\end{array} \tag{7.50}$$

where we are explicitly noting the fact that the diagonal entries of T vanish. Observe that we are using the entries of $\mathbf{u}^{(k)}$ to compute *all* of the updated values of $\mathbf{u}^{(k+1)}$. Presumably, if the iterates $\mathbf{u}^{(k)}$ are converging to the solution $\mathbf{u}^{\star}$, then their individual entries are also converging, and so each $u_j^{(k+1)}$ should be a better approximation to $u_j^{\star}$ than $u_j^{(k)}$ is. Therefore, if we begin the k^{th} Jacobi iteration by computing $u_1^{(k+1)}$ using the first equation, then we are tempted to use this new and improved value to replace $u_1^{(k)}$ in each of the subsequent equations. In particular, we employ the modified equation

$$u_2^{(k+1)} = t_{21}\,u_1^{(k+1)} + t_{23}\,u_3^{(k)} + \;\cdots\; + t_{1n}\,u_n^{(k)} + c_2$$

to update the second component of our iterate. This more accurate value should then be used to update $u_3^{(k+1)}$, and so on.

The upshot of these considerations is the *Gauss–Seidel method*

$$u_i^{(k+1)} = t_{i1}\,u_1^{(k+1)} + \;\cdots\; + t_{i,i-1}\,u_{i-1}^{(k+1)} + t_{i,i+1}\,u_{i+1}^{(k)} + \;\cdots\; + t_{in}\,u_n^{(k)} + c_i, \quad i = 1, \ldots, n, \tag{7.51}$$

named after Gauss (as usual!) and the German astronomer/mathematician Philipp von Seidel. At the k^{th} stage of the iteration, we use (7.51) to compute the revised entries $u_1^{(k+1)}$, $u_2^{(k+1)}, \ldots, u_n^{(k+1)}$ in their numerical order. Once an entry has been updated, the new value is immediately used in all subsequent computations.

Example 7.28. For the linear system

$$3x + y - z = 3, \qquad x - 4y + 2z = -1, \qquad -2x - y + 5z = 2,$$

the Jacobi iteration method was given in (7.39). To construct the corresponding Gauss–Seidel scheme we use updated values of x, y and z as they become available. Explicitly,

$$\begin{aligned}
x^{(k+1)} &= -\tfrac{1}{3}\,y^{(k)} + \tfrac{1}{3}\,z^{(k)} + 1, \\
y^{(k+1)} &= \tfrac{1}{4}\,x^{(k+1)} + \tfrac{1}{2}\,z^{(k)} + \tfrac{1}{4}, \\
z^{(k+1)} &= \tfrac{2}{5}\,x^{(k+1)} + \tfrac{1}{5}\,y^{(k+1)} + \tfrac{2}{5}.
\end{aligned} \tag{7.52}$$

The resulting iterates starting with $\mathbf{u}^{(0)} = \mathbf{0}$ are

$$\mathbf{u}^{(1)} = \begin{pmatrix} 1.0000 \\ .5000 \\ .9000 \end{pmatrix}, \qquad \mathbf{u}^{(2)} = \begin{pmatrix} 1.1333 \\ .9833 \\ 1.0500 \end{pmatrix}, \qquad \mathbf{u}^{(3)} = \begin{pmatrix} 1.0222 \\ 1.0306 \\ 1.0150 \end{pmatrix}, \qquad \mathbf{u}^{(4)} = \begin{pmatrix} .9948 \\ 1.0062 \\ .9992 \end{pmatrix},$$

$$\mathbf{u}^{(5)} = \begin{pmatrix} .9977 \\ .9990 \\ .9989 \end{pmatrix}, \qquad \mathbf{u}^{(6)} = \begin{pmatrix} 1.0000 \\ .9994 \\ .9999 \end{pmatrix}, \qquad \mathbf{u}^{(7)} = \begin{pmatrix} 1.0001 \\ 1.0000 \\ 1.0001 \end{pmatrix}, \qquad \mathbf{u}^{(8)} = \begin{pmatrix} 1.0000 \\ 1.0000 \\ 1.0000 \end{pmatrix},$$

and have converged to the solution, to 4 decimal place accuracy, after only 8 iterations — as opposed to the 11 iterations required by the Jacobi method.

The Gauss–Seidel iteration scheme is particularly suited to implementation on a serial computer, since one can immediately replace each component $u_i^{(k)}$ by its updated value $u_i^{(k+1)}$, thereby also saving on storage in the computer's memory. In contrast, the Jacobi scheme requires us to retain all the old values $\mathbf{u}^{(k)}$ until the new approximation $\mathbf{u}^{(k+1)}$ has been computed. Moreover, Gauss–Seidel typically (although not always) converges faster than Jacobi, making it the iterative algorithm of choice for serial processors. On the other hand, with the advent of parallel processing machines, variants of the parallelizable Jacobi scheme have recently been making a comeback.

What is Gauss–Seidel really up to? Let us rewrite the basic iterative equation (7.51) by multiplying by a_{ii} and moving the terms involving $\mathbf{u}^{(k+1)}$ to the left hand side. In view of the formula (7.43) for the entries of T, the resulting equation is

$$a_{i1}\, u_1^{(k+1)} + \;\cdots\; + a_{i,i-1}\, u_{i-1}^{(k+1)} + a_{ii}\, u_i^{(k+1)} = -\, a_{i,i+1}\, u_{i+1}^{(k)} - \;\cdots\; - a_{in}\, u_n^{(k)} + b_i.$$

In matrix form, taking (7.46) into account, this reads

$$(L + D)\mathbf{u}^{(k+1)} = -\, U\, \mathbf{u}^{(k)} + \mathbf{b}, \tag{7.53}$$

and so can be viewed as a linear system of equations for $\mathbf{u}^{(k+1)}$ with lower triangular coefficient matrix $L + D$. Note that the fixed point of (7.53), namely the solution to

$$(L + D)\, \mathbf{u} = -\, U\, \mathbf{u} + \mathbf{b},$$

coincides with the solution to the original system

$$A\, \mathbf{u} = (L + D + U)\, \mathbf{u} = \mathbf{b}.$$

In other words, the Gauss–Seidel procedure is merely implementing Forward Substitution to solve the lower triangular system (7.53) for the next iterate:

$$\mathbf{u}^{(k+1)} = -\, (L + D)^{-1} U\, \mathbf{u}^{(k)} + (L + D)^{-1}\, \mathbf{b}.$$

The latter is in our more usual iterative form

$$\mathbf{u}^{(k+1)} = \widetilde{T}\, \mathbf{u}^{(k)} + \widetilde{\mathbf{c}}, \qquad \text{where} \qquad \widetilde{T} = -\, (L + D)^{-1} U, \qquad \widetilde{\mathbf{c}} = (L + D)^{-1}\, \mathbf{b}. \tag{7.54}$$

Consequently, the convergence of the Gauss–Seidel iterates is governed by the spectral radius of the coefficient matrix $\widetilde{T}$.

Returning to Example 7.28, we have

$$A = \begin{pmatrix} 3 & 1 & -1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}, \qquad L + D = \begin{pmatrix} 3 & 0 & 0 \\ 1 & -4 & 0 \\ -2 & -1 & 5 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{pmatrix}.$$

Therefore, the Gauss–Seidel matrix is

$$\widetilde{T} = -(L+D)^{-1}U = \begin{pmatrix} 0 & -.3333 & .3333 \\ 0 & -.0833 & .5833 \\ 0 & -.1500 & .2500 \end{pmatrix}.$$

Its eigenvalues are 0 and $.0833 \pm .2444\,\mathrm{i}$, and hence its spectral radius is $\rho(\widetilde{T}\,) \approx .2582$. This is roughly the square of the Jacobi spectral radius of .5, which tell us that the Gauss–Seidel iterations will converge about twice as fast to the solution. This can be verified by more extensive computations. Although examples can be constructed where the Jacobi method converges faster, in many practical situations Gauss–Seidel tends to converge roughly twice as fast as Jacobi.

Completely general conditions guaranteeing convergence of the Gauss–Seidel method are hard to establish. But, like the Jacobi scheme, it is guaranteed to converge when the original coefficient matrix is strictly diagonally dominant.

Theorem 7.29. *If A is strictly diagonally dominant, then the Gauss–Seidel iteration scheme for solving $A\mathbf{u} = \mathbf{b}$ converges.*

Proof: Let $\mathbf{e}^{(k)} = \mathbf{u}^{(k)} - \mathbf{u}^\star$ denote the k^{th} Gauss–Seidel error vector. As in (7.41), the error vectors satisfy the linear iterative system $\mathbf{e}^{(k+1)} = \widetilde{T}\,\mathbf{e}^{(k)}$, but a direct estimate of $\|\,\widetilde{T}\,\|_\infty$ is not so easy. Instead, let us write out the linear iterative system in components:

$$e_i^{(k+1)} = t_{i1}\, e_1^{(k+1)} + \;\cdots\; + t_{i,i-1}\, e_{i-1}^{(k+1)} + t_{i,i+1}\, e_{i+1}^{(k)} + \;\cdots\; + t_{in}\, e_n^{(k)}. \tag{7.55}$$

Let

$$m^{(k)} = \|\,\mathbf{e}^{(k)}\,\|_\infty = \max\bigl\{\,|\,e_1^{(k)}\,|,\;\ldots\;,|\,e_n^{(k)}\,|\,\bigr\} \tag{7.56}$$

denote the ∞ norm of the k^{th} error vector. To prove convergence, $\mathbf{e}^{(k)} \to \mathbf{0}$, it suffices to show that $m^{(k)} \to 0$ as $k \to \infty$. We claim that diagonal dominance of A implies that

$$m^{(k+1)} \le s\, m^{(k)}, \qquad \text{where} \qquad s = \|\,T\,\|_\infty < 1 \tag{7.57}$$

denotes the ∞ matrix norm of the *Jacobi* matrix (not the Gauss–Seidel matrix), which, by (7.49), is less than 1. We infer that $m^{(k)} \le s^k\, m^{(0)} \to 0$ as $k \to \infty$, demonstrating the theorem.

To prove (7.57), we use induction on $i = 1, \ldots, n$. Our induction hypothesis is

$$|\,e_j^{(k+1)}\,| \le s\, m^{(k)} < m^{(k)} \qquad \text{for} \qquad j = 1, \ldots, i-1.$$

(When $i = 1$, there is no assumption.) Moreover, by (7.56),

$$|\,e_j^{(k)}\,| \le m^{(k)} \qquad \text{for all} \qquad j = 1, \ldots, n.$$

We use these two inequalities to estimate $|\, e_i^{(k+1)} \,|$ from (7.55):

$$\begin{aligned} |\, e_i^{(k+1)} \,| &\le |\, t_{i1} \,|\, |\, e_1^{(k+1)} \,| + \cdots + |\, t_{i,i-1} \,|\, |\, e_{i-1}^{(k+1)} \,| + |\, t_{i,i+1} \,|\, |\, e_{i+1}^{(k)} \,| + \cdots + |\, t_{in} \,|\, |\, e_n^{(k)} \,| \\ &\le \big(\, |\, t_{i1} \,| + \cdots + |\, t_{in} \,| \,\big)\, m^{(k)} \le s\, m^{(k)}, \end{aligned}$$

which completes the induction step. As a result, the maximum

$$m^{(k+1)} = \max\{\, |\, e_1^{(k+1)} \,|, \ \ldots\ , |\, e_n^{(k+1)} \,| \,\} \le s\, m^{(k)}$$

also satisfies the same bound, and hence (7.57) follows. *Q.E.D.*

Example 7.30. For the linear system considered in Example 7.27, the Gauss–Seidel iterations take the form

$$\begin{aligned} x^{(k+1)} &= -\tfrac{1}{4}\, y^{(k)} - \tfrac{1}{4}\, w^{(k)} + \tfrac{1}{4}\,, \\ y^{(k+1)} &= -\tfrac{1}{4}\, x^{(k+1)} - \tfrac{1}{4}\, z^{(k)} - \tfrac{1}{4}\, v^{(k)} + \tfrac{1}{2}\,, \\ z^{(k+1)} &= -\tfrac{1}{4}\, y^{(k+1)} - \tfrac{1}{4}\, w^{(k)} - \tfrac{1}{4}\,, \\ w^{(k+1)} &= -\tfrac{1}{4}\, x^{(k+1)} - \tfrac{1}{4}\, z^{(k+1)} - \tfrac{1}{4}\, v^{(k)} + \tfrac{1}{2}\,, \\ v^{(k+1)} &= -\tfrac{1}{4}\, y^{(k+1)} - \tfrac{1}{4}\, w^{(k+1)} + \tfrac{1}{4}\,. \end{aligned}$$

Starting with $x^{(0)} = y^{(0)} = z^{(0)} = w^{(0)} = v^{(0)} = 0$, the Gauss–Seidel iterates converge to the solution $x = -.1, y = .7, z = -.6, w = .7, v = -.1$, to four decimal places in 11 iterations, again roughly twice as fast as the Jacobi scheme. Indeed, the convergence rate is governed by the corresponding Gauss–Seidel matrix $\widetilde{T}$, which is

$$\begin{pmatrix} 4 & 0 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 & 0 \\ 0 & 1 & 4 & 0 & 0 \\ 1 & 0 & 1 & 4 & 0 \\ 0 & 1 & 0 & 1 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -.2500 & 0 & -.2500 & 0 \\ 0 & .0625 & -.2500 & .0625 & -.2500 \\ 0 & -.0156 & .0625 & -.2656 & .0625 \\ 0 & .0664 & -.0156 & .1289 & -.2656 \\ 0 & -.0322 & .0664 & -.0479 & .1289 \end{pmatrix}.$$

Its spectral radius is $\rho(\widetilde{T}) = .3936$, which is, as in the previous example, approximately the square of the spectral radius of the Jacobi coefficient matrix, which explains the speed up in convergence.

Successive Over–Relaxation (SOR)

As we know, the smaller the spectral radius (or matrix norm) of the coefficient matrix, the faster the convergence of the iterative scheme. One of the goals of researchers in numerical linear algebra is to design new methods for accelerating the convergence. In his 1950 thesis, the American mathematician David Young discovered a simple modification of the Jacobi and Gauss–Seidel methods that can, in favorable situations, lead to a dramatic speed up in the rate of convergence. The method, known as *successive over-relaxation*, and often abbreviated as SOR, has become the iterative method of choice in many modern applications, [**13**, **53**]. In this subsection, we provide a brief overview.

In practice, finding the optimal iterative algorithm to solve a given linear system is as hard as solving the system itself. Therefore, researchers have relied on a few tried and true techniques for designing iterative schemes that can be used in the more common applications. Consider a linear algebraic system

$$A\mathbf{u} = \mathbf{b}.$$

Every decomposition

$$A = M - N \tag{7.58}$$

of the coefficient matrix into the difference of two matrices leads to an equivalent system of the form

$$M\,\mathbf{u} = N\,\mathbf{u} + \mathbf{b}. \tag{7.59}$$

Provided that M is nonsingular, we can rewrite the system in the fixed point form

$$\mathbf{u} = M^{-1}N\,\mathbf{u} + M^{-1}\mathbf{b} = T\,\mathbf{u} + \mathbf{c}, \qquad \text{where} \qquad T = M^{-1}N, \quad \mathbf{c} = M^{-1}\mathbf{b}.$$

Now, we are free to choose any such M, which then specifies $N = A - M$ uniquely. However, for the resulting iterative scheme $\mathbf{u}^{(k+1)} = T\,\mathbf{u}^{(k)} + \mathbf{c}$ to be practical we must arrange that

(*a*) $T = M^{-1}N$ is a convergent matrix, and

(*b*) M can be easily inverted.

The second requirement ensures that the iterative equations

$$M\,\mathbf{u}^{(k+1)} = N\,\mathbf{u}^{(k)} + \mathbf{b} \tag{7.60}$$

can be solved for $\mathbf{u}^{(k+1)}$ with minimal computational effort. Typically, this requires that M be either a diagonal matrix, in which case the inversion is immediate, or upper or lower triangular, in which case one employs Back or Forward Substitution to solve for $\mathbf{u}^{(k+1)}$.

With this in mind, we now introduce the SOR method. It relies on a slight generalization of the Gauss–Seidel decomposition (7.53) of the matrix into lower plus diagonal and upper triangular parts. The starting point is to write

$$A = L + D + U = \bigl[\,L + \alpha\,D\,\bigr] - \bigl[\,(\alpha - 1)\,D - U\,\bigr], \tag{7.61}$$

where $0 \neq \alpha$ is an adjustable scalar parameter. We decompose the system $A\mathbf{u} = \mathbf{b}$ as

$$(L + \alpha\,D)\mathbf{u} = \bigl[\,(\alpha - 1)\,D - U\,\bigr]\mathbf{u} + \mathbf{b}. \tag{7.62}$$

It turns out to be slightly more convenient to divide (7.62) through by α and write the resulting iterative system in the form

$$(\omega\,L + D)\mathbf{u}^{(k+1)} = \bigl[\,(1 - \omega)\,D - \omega\,U\,\bigr]\mathbf{u}^{(k)} + \omega\,\mathbf{b}, \tag{7.63}$$

where $\omega = 1/\alpha$ is called the *relaxation parameter*. Assuming, as usual, that all diagonal entries of A are nonzero, the matrix $\omega\,L + D$ is an invertible lower triangular matrix, and so we can use Forward Substitution to solve the iterative system (7.63) to recover $\mathbf{u}^{(k+1)}$. The explicit formula for its i^{th} entry is

$$\begin{aligned} u_i^{(k+1)} = \omega\,t_{i1}\,u_1^{(k+1)} + \;\cdots\; + \omega\,t_{i,i-1}\,u_{i-1}^{(k+1)} + (1-\omega)\,u_i^{(k)} + \\ + \omega\,t_{i,i+1}\,u_{i+1}^{(k)} + \;\cdots\; + \omega\,t_{in}\,u_n^{(k)} + \omega\,c_i, \end{aligned} \tag{7.64}$$

where t_{ij} and c_i denote the original Jacobi values (7.43). As in the Gauss–Seidel approach, we update the entries $u_i^{(k+1)}$ in numerical order $i = 1, \ldots, n$. Thus, to obtain the SOR scheme (7.64), we merely multiply the right hand side of the Gauss–Seidel scheme (7.51) by the adjustable relaxation parameter ω and append the diagonal term $(1-\omega)\,u_i^{(k)}$. In particular, if we set $\omega = 1$, then the SOR method reduces to the Gauss–Seidel method. Choosing $\omega < 1$ leads to an *under-relaxed* method, while $\omega > 1$, known as *over-relaxation*, is the choice that works in most practical instances.

To analyze the SOR scheme in detail, we rewrite (7.63) in the fixed point form

$$\mathbf{u}^{(k+1)} = T_\omega\, \mathbf{u}^{(k)} + \mathbf{c}_\omega, \tag{7.65}$$

where

$$T_\omega = (\omega\, L + D)^{-1} \big[\, (1-\omega)\, D - \omega\, U \,\big], \qquad \mathbf{c}_\omega = (\omega\, L + D)^{-1}\, \omega\, \mathbf{b}. \tag{7.66}$$

The rate of convergence is governed by the spectral radius of the matrix T_ω. The goal is to choose the relaxation parameter ω so as to make the spectral radius of T_ω as small as possible. As we will see, a clever choice of ω can result in a dramatic speed up in the convergence rate. Let us look at an elementary example.

Example 7.31. Consider the matrix $A = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$, which we decompose as $A = L + D + U$, where

$$L = \begin{pmatrix} 0 & 0 \\ -1 & 0 \end{pmatrix}, \qquad D = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}, \qquad U = \begin{pmatrix} 0 & -1 \\ 0 & 0 \end{pmatrix}.$$

Jacobi iteration is based on the coefficient matrix $T = -D^{-1}(L+U) = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}$. Its spectral radius is $\rho(T) = .5$, and hence the Jacobi scheme takes, on average, roughly $3.3 \approx -1/\log_{10} .5$ iterations to produce each new decimal place in the solution.

The SOR scheme (7.63) takes the explicit form

$$\begin{pmatrix} 2 & 0 \\ -\omega & 2 \end{pmatrix} \mathbf{u}^{(k+1)} = \begin{pmatrix} 2(1-\omega) & \omega \\ 0 & 2\,(1-\omega) \end{pmatrix} \mathbf{u}^{(k)} + \omega\, \mathbf{b},$$

where Gauss–Seidel is the particular case $\omega = 1$. The SOR coefficient matrix is

$$T_\omega = \begin{pmatrix} 2 & 0 \\ -\omega & 2 \end{pmatrix}^{-1} \begin{pmatrix} 2\,(1-\omega) & \omega \\ 0 & 2\,(1-\omega) \end{pmatrix} = \begin{pmatrix} 1-\omega & \frac{1}{2}\,\omega \\ \frac{1}{2}\,\omega(1-\omega) & \frac{1}{4}\,(2-\omega)^2 \end{pmatrix}.$$

To compute the eigenvalues of T_ω, we form its characteristic equation

$$0 = \det(T_\omega - \lambda\, \mathrm{I}\,) = \lambda^2 - \big(\, 2 - 2\omega + \tfrac{1}{4}\,\omega^2 \,\big)\lambda + (1-\omega)^2 = (\lambda + \omega - 1)^2 - \tfrac{1}{4}\,\lambda\, \omega^2. \tag{7.67}$$

Our goal is to choose ω so that

(*a*) both eigenvalues are less than 1 in modulus, so $|\,\lambda_1\,|, |\,\lambda_2\,| < 1$. This is the minimal requirement for convergence of the method.

(*b*) the largest eigenvalue (in modulus) is as small as possible. This will give the smallest spectral radius for T_ω and hence the fastest convergence rate.

The product of the two eigenvalues is the determinant,

$$\lambda_1 \lambda_2 = \det T_\omega = (1-\omega)^2.$$

If $\omega \leq 0$ or $\omega \geq 2$, then $\det T_\omega \geq 1$, and hence at least one of the eigenvalues would have modulus larger than 1. Thus, in order to ensure convergence, we must require $0 < \omega < 2$. For Gauss–Seidel, at $\omega = 1$, the eigenvalues are $\lambda_1 = \frac{1}{4}$, $\lambda_2 = 0$, and the spectral radius is $\rho(T_1) = .25$. This is exactly the square of the Jacobi spectral radius, and hence the Gauss–Seidel iterates converge twice as fast; so it only takes, on average, about $-1/\log_{10} .25 = 1.66$ Gauss–Seidel iterations to produce each new decimal place of accuracy. It can be shown that as ω increases above 1, the two eigenvalues move along the real axis towards each other. They coincide when

$$\omega = \omega_\star = 8 - 4\sqrt{3} \;\approx\; 1.07, \qquad \text{at which point} \qquad \lambda_1 = \lambda_2 = \omega_\star - 1 = .07 = \rho(T_\omega),$$

which is the convergence rate of the optimal SOR scheme. Each iteration produces slightly more than one new decimal place in the solution, which represents a significant improvement over the Gauss–Seidel convergence rate. It takes about twice as many Gauss–Seidel iterations (and four times as many Jacobi iterations) to produce the same accuracy as this optimal SOR method.

Of course, in such a simple 2×2 example, it is not so surprising that we can construct the best value for the relaxation parameter by hand. Young was able to find the optimal value of the relaxation parameter for a broad class of matrices that includes most of those arising in the finite difference and finite element numerical solutions to ordinary and partial differential equations. For the matrices in Young's class, the Jacobi eigenvalues occur in signed pairs. If $\pm\mu$ are a pair of eigenvalues for the Jacobi method, then the corresponding eigenvalues of the SOR iteration matrix satisfy the quadratic equation

$$(\lambda + \omega - 1)^2 = \lambda\, \omega^2 \mu^2. \tag{7.68}$$

If $\omega = 1$, so we have standard Gauss–Seidel, then $\lambda^2 = \lambda \mu^2$, and so the eigenvalues are $\lambda = 0$, $\lambda = \mu^2$. The Gauss–Seidel spectral radius is therefore the square of the Jacobi spectral radius, and so (at least for matrices in the Young class) its iterates converge twice as fast. The quadratic equation (7.68) has the same properties as in the 2×2 version (7.67) (which corresponds to the case $\mu = \frac{1}{2}$), and hence the optimal value of ω will be the one at which the two roots are equal:

$$\lambda_1 = \lambda_2 = \omega - 1, \qquad \text{which occurs when} \qquad \omega = \frac{2 - 2\sqrt{1-\mu^2}}{\mu^2} = \frac{2}{1 + \sqrt{1-\mu^2}}\,.$$

Therefore, if $\rho_J = \max |\,\mu\,|$ denotes the spectral radius of the Jacobi method, then the Gauss–Seidel has spectral radius $\rho_{GS} = \rho_J^2$, while the SOR method with optimal relaxation parameter

$$\omega_\star = \frac{2}{1 + \sqrt{1 - \rho_J^2}}, \qquad \text{has spectral radius} \qquad \rho_\star = \omega_\star - 1. \tag{7.69}$$

For example, if $\rho_J = .99$, which is rather slow convergence (but common for iterative numerical solution schemes for partial differential equations), then $\rho_{GS} = .9801$, which is twice as fast, but still quite slow, while SOR with $\omega_\star = 1.7527$ has $\rho_\star = .7527$, which is dramatically faster[†]. Indeed, since $\rho_\star \approx (\rho_{GS})^{14} \approx (\rho_J)^{28}$, it takes about 14 Gauss–Seidel (and 28 Jacobi) iterations to produce the same accuracy as one SOR step. It is amazing that such a simple idea can have such a dramatic effect.

[†] More precisely, since the SOR matrix is not diagonalizable, the overall convergence rate is slightly slower than the spectral radius. However, this technical detail does not affect the overall conclusion.