Chapter 6

Iterative Methods for Solving Linear Systems

Consider a system of linear algebraic equations:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}, \quad \boldsymbol{f} \in \mathbb{L}, \quad \boldsymbol{x} \in \mathbb{L}, \tag{6.1}$$

where $\mathbb{L}$ is a vector space and $\boldsymbol{A} : \mathbb{L} \longmapsto \mathbb{L}$ is a linear operator. From the standpoint of applications, having a capability to compute its *exact* solution may be "nice," but it is typically not necessary. Quite the opposite, one can normally use an approximate solution, provided that it would guarantee the accuracy sufficient for a particular application. On the other hand, one usually cannot obtain the exact solution anyway. The reason is that the input data of the problem (the right-hand side $\boldsymbol{f}$, as well as the operator $\boldsymbol{A}$ itself) are always specified with some degree of uncertainty. This necessarily leads to a certain unavoidable error in the result. Besides, as all the numbers inside the machine are only specified with a finite precision, the round-off errors are also inevitable in the course of computations.

Therefore, instead of solving system (6.1) by a direct method, e.g., by Gaussian elimination, in many cases it may be advantageous to use an iterative method of solution. This is particularly true when the dimension n of system (6.1) is very large, and unless a special fast algorithm such as FFT (see Section 5.7.3) can be employed, the $\mathcal{O}(n^3)$ cost of a direct method (see Sections 5.4 and 5.6) would be unbearable.

A typical iterative method (or an iteration scheme) consists of building a sequence of vectors $\{\boldsymbol{x}^{(p)}\} \subset \mathbb{L}$, $p = 0, 1, 2, \ldots$, that are supposed to provide successively more accurate approximations of the exact solution $\boldsymbol{x}$. The initial guess $\boldsymbol{x}^{(0)} \in \mathbb{L}$ for an iteration scheme is normally taken arbitrarily. The notion of successively more accurate approximations can, of course, be quantified. It means that the sequence $\boldsymbol{x}^{(p)}$ has to converge to the exact solution $\boldsymbol{x}$ as the number p increases: $\boldsymbol{x}^{(p)} \longrightarrow \boldsymbol{x}$, when $p \longrightarrow \infty$. This means that for any $\varepsilon > 0$ we can always find $p = p(\varepsilon)$ such that the following inequality:

$$\|\boldsymbol{x} - \boldsymbol{x}^{(p)}\| \leq \varepsilon$$

will hold for all $p \geq p(\varepsilon)$. Accordingly, by specifying a sufficiently small $\varepsilon > 0$ we can terminate the iteration process after a finite number $p = p(\varepsilon)$ of steps, and subsequently use the iteration $\boldsymbol{x}^{(p)}$ in the capacity of an approximate solution that would meet the accuracy requirements for a given problem.

In this chapter, we will describe some popular iterative methods, and outline the conditions under which it may be advisable to use an iterative method rather than a direct method, or to prefer one particular iterative method over another.

6.1 Richardson Iterations and the Like

We will build a family of iterative methods by first recasting system (6.1) as follows:

$$\boldsymbol{x} = (\boldsymbol{I} - \tau \boldsymbol{A})\boldsymbol{x} + \tau \boldsymbol{f}. \tag{6.2}$$

In doing so, the new system (6.2) will be equivalent to the original one for any value of the parameter τ, $\tau > 0$. In general, there are many different ways of replacing the system $\boldsymbol{Ax} = \boldsymbol{f}$ by its equivalent of the type:

$$\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{\varphi}, \quad \boldsymbol{x} \in \mathbb{L}, \quad \boldsymbol{\varphi} \in \mathbb{L}. \tag{6.3}$$

System (6.2) is a particular case of (6.3) with $\boldsymbol{B} = (\boldsymbol{I} - \tau \boldsymbol{A})$ and $\boldsymbol{\varphi} = \tau \boldsymbol{f}$.

6.1.1 General Iteration Scheme

The general scheme of what is known as the first order linear stationary iteration process consists of successively computing the terms of the sequence:

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{B}\boldsymbol{x}^{(p)} + \boldsymbol{\varphi}, \quad p = 0, 1, 2, \ldots, \tag{6.4}$$

where the initial guess $\boldsymbol{x}^{(0)}$ is specified arbitrarily. The matrix $\boldsymbol{B}$ is known as the iteration matrix. Clearly, if the sequence $\boldsymbol{x}^{(p)}$ converges, i.e., if there is a limit: $\lim_{p\to\infty} \boldsymbol{x}^{(p)} = \boldsymbol{x}$, then $\boldsymbol{x}$ is the solution of system (6.1). Later we will identify the conditions that would guarantee convergence of the sequence (6.4).

Iterative method (6.4) is first order because the next iterate $\boldsymbol{x}^{(p+1)}$ depends only on one previous iterate, $\boldsymbol{x}^{(p)}$; it is linear because the latter dependence is linear; and finally it is stationary because if we formally rewrite (6.4) as $\boldsymbol{x}^{(p+1)} = F(\boldsymbol{x}^{(p)}, \boldsymbol{A}, \boldsymbol{f})$, then the function F does not depend on p.

A particular form of the iteration scheme (6.4) based on system (6.2) is known as the stationary Richardson method:

$$\boldsymbol{x}^{(p+1)} = (\boldsymbol{I} - \tau \boldsymbol{A})\boldsymbol{x}^{(p)} + \tau \boldsymbol{f}, \quad p = 0, 1, 2, \ldots. \tag{6.5}$$

Formula (6.5) can obviously be rewritten as:

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{x}^{(p)} - \tau \boldsymbol{r}^{(p)}, \tag{6.6}$$

where $\boldsymbol{r}^{(p)} = \boldsymbol{A}\boldsymbol{x}^{(p)} - \boldsymbol{f}$ is the residual of the iterate $\boldsymbol{x}^{(p)}$. Instead of keeping the parameter τ constant we can allow it to depend on p. Then, departing from formula (6.6), we arrive at the so-called non-stationary Richardson method:

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{x}^{(p)} - \tau_p \boldsymbol{r}^{(p)}. \tag{6.7}$$

Note that in order to actually compute the iterations according to formula (6.5) we only need to be able to obtain the vector $\boldsymbol{A}\boldsymbol{x}^{(p)}$ once $\boldsymbol{x}^{(p)} \in \mathbb{L}$ is given. This does not

necessarily entail the explicit knowledge of the matrix $\boldsymbol{A}$. In other words, an iterative method of solving $\boldsymbol{Ax} = \boldsymbol{f}$ can also be realized when the system is specified in an operator form. Building the iteration sequence does not require choosing a particular basis in $\mathbb{L}$ and reducing the system to its canonical form:

$$\sum_{j=1}^{n} a_{ij}x_j = f_i, \quad i = 1,2,\ldots,n. \tag{6.8}$$

Moreover, when computing the terms of the sequence $\boldsymbol{x}^{(p)}$ with the help of formula (6.5), we do not necessarily need to store all n^2 entries of the matrix $\boldsymbol{A}$ in the computer memory. Instead, to implement the iteration scheme (6.5) we may only store the current vector $\boldsymbol{x}^{(p)} \in \mathbb{L}$ that has n components. In addition to the memory savings and flexibility in specifying $\boldsymbol{A}$, we will see that for certain classes of linear systems the computational cost of obtaining a sufficiently accurate solution of $\boldsymbol{Ax} = \boldsymbol{f}$ with the help of an iterative method may be considerably lower than $\mathcal{O}(n^3)$ operations, which is characteristic of direct methods.

THEOREM 6.1

Let $\mathbb{L}$ be an n-dimensional normed vector space (say, $\mathbb{R}^n$ or $\mathbb{C}^n$), and assume that the induced operator norm of the iteration matrix $\boldsymbol{B}$ of (6.4) satisfies:

$$\|\boldsymbol{B}\| = q < 1. \tag{6.9}$$

Then, system (6.3) or equivalently, system (6.1), has a unique solution $\boldsymbol{x} \in \mathbb{L}$. Moreover, the iteration sequence (6.4) converges to this solution $\boldsymbol{x}$ for an arbitrary initial guess $\boldsymbol{x}^{(0)}$, and the error of the iterate number p:

$$\boldsymbol{\varepsilon}^{(p)} \stackrel{\text{def}}{=} \boldsymbol{x} - \boldsymbol{x}^{(p)}$$

satisfies the estimate:

$$\|\boldsymbol{\varepsilon}^{(p)}\| = \|\boldsymbol{x} - \boldsymbol{x}^{(p)}\| \leq q^p \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\| = q^p \|\boldsymbol{\varepsilon}^{(0)}\|. \tag{6.10}$$

In other words, the norm of the error $\|\boldsymbol{\varepsilon}^{(p)}\|$ vanishes when $p \longrightarrow \infty$ at least as fast as the geometric sequence q^p.

PROOF If $\boldsymbol{\varphi} = \boldsymbol{0}$, then system (6.3) only has a trivial solution $\boldsymbol{x} = \boldsymbol{0}$. Indeed, otherwise for a solution $\boldsymbol{x} \neq \boldsymbol{0}$, $\boldsymbol{\varphi} = \boldsymbol{0}$, we could write:

$$\|\boldsymbol{x}\| = \|\boldsymbol{Bx}\| \leq \|\boldsymbol{B}\|\|\boldsymbol{x}\| = q\|\boldsymbol{x}\| < \|\boldsymbol{x}\|,$$

i.e., $\|\boldsymbol{x}\| < \|\boldsymbol{x}\|$, which may not hold. The contradiction proves that system (6.3) is uniquely solvable for any $\boldsymbol{\varphi}$, and as such, so is system (6.1) for any $\boldsymbol{f}$.

Next, let $\boldsymbol{x}$ be the solution of system (6.3). Take an arbitrary $\boldsymbol{x}^{(0)} \in \mathbb{L}$ and subtract equality (6.3) from equality (6.4), which yields:

$$\boldsymbol{\varepsilon}^{(p+1)} = \boldsymbol{B\varepsilon}^{(p)}, \quad p = 0,1,2,\ldots.$$

Consequently, the following error estimate holds:

$$\|\boldsymbol{x}-\boldsymbol{x}^{(p)}\| = \|\boldsymbol{\varepsilon}^{(p)}\| = \|\boldsymbol{B}\boldsymbol{\varepsilon}^{(p-1)}\| \le q\|\boldsymbol{\varepsilon}^{(p-1)}\|$$
$$\le q^2\|\boldsymbol{\varepsilon}^{(p-2)}\| \le \ldots \le q^p\|\boldsymbol{\varepsilon}^{(0)}\| = q^p\|\boldsymbol{x}-\boldsymbol{x}^{(0)}\|,$$

which is equivalent to (6.10). ∎

REMARK 6.1 Condition (6.9) can be violated for a different choice of norm on the space $\mathbb{L}$: $\|\boldsymbol{x}\|'$ instead of $\|\boldsymbol{x}\|$. This, however, will not disrupt the convergence: $\boldsymbol{x}^{(p)} \longrightarrow \boldsymbol{x}$ as $p \longrightarrow \infty$. Moreover, the error estimate (6.10) will be replaced by

$$\|\boldsymbol{\varepsilon}^{(p)}\|' \le cq^p\|\boldsymbol{\varepsilon}^{(0)}\|', \tag{6.11}$$

where c is a constant that will, generally speaking, depend on the new norm $\|\cdot\|'$, whereas the value of q will remain the same.

To justify this remark one needs to employ the equivalence of any two norms on a vector (i.e., finite-dimensional) space, see [HJ85, Section 5.4]. This important result says that if $\|\cdot\|$ and $\|\cdot\|'$ are two norms on $\mathbb{L}$, then we can always find two constants $c_1 > 0$ and $c_2 > 0$, such that $\forall \boldsymbol{x} \in \mathbb{L}:\ c_1\|\boldsymbol{x}\|' \le \|\boldsymbol{x}\| \le c_2\|\boldsymbol{x}\|'$, where c_1 and c_2 do not depend on $\boldsymbol{x}$. Therefore, inequality (6.10) implies (6.11) with $c = c_2/c_1$. ∎

Example 1: The Jacobi Method

Let the matrix $\boldsymbol{A} = \{a_{ij}\}$ of system (6.1) be diagonally dominant:

$$|a_{ii}| > \sum_{j\ne i} |a_{ij}| + \delta, \quad i = 1,2,\ldots,n, \quad \delta > 0. \tag{6.12}$$

In the equation number i of system (6.1), we move all terms $a_{ij}x_j$, $j \ne i$, to the right-hand side and then divide this equation by a_{ii}. In doing so, we obtain a system of type (6.3) with the matrix:

$$\boldsymbol{B} = \begin{bmatrix} 0 & b_{12} & b_{13} & \ldots & b_{1,n-1} & b_{1n} \\ b_{21} & 0 & b_{23} & \ldots & b_{2,n-1} & b_{2n} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ b_{n1} & b_{n2} & b_{n3} & \ldots & b_{n,n-1} & 0 \end{bmatrix}.$$

Alternatively, if we define the diagonal $n \times n$ matrix $\boldsymbol{D} = \operatorname{diag}\{a_{ii}\}$, then in the resulting system (6.3) we have $\boldsymbol{B} = -\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{D})$ and $\boldsymbol{\varphi} = \boldsymbol{D}^{-1}\boldsymbol{f}$.

Due to the diagonal dominance of $\boldsymbol{A}$, one can find such a number $0 < q < 1$ that

$$\sum_{j=1}^{n} |b_{ij}| = \sum_{\substack{j=1,\\ j\ne i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \le q < 1.$$

Consequently, the maximum norm of the iteration matrix $\boldsymbol{B}$: $\|\boldsymbol{B}\|_\infty = \max_i \sum_j |b_{ij}|$, satisfies estimate (6.9). Then, according to Theorem 6.1, the Jacobi iterations:

$$\boldsymbol{x}^{(p+1)} = -\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{D})\boldsymbol{x}^{(p)} + \boldsymbol{D}^{-1}\boldsymbol{f} \tag{6.13}$$

converge to the solution $\boldsymbol{x}$ of system (6.1). In the component form, the Jacobi iterative method (6.13) is written as:

$$x_i^{(p+1)} = -\sum_{\substack{j=1,\\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(p)} + \frac{f_i}{a_{ii}}, \quad i = 1,2,\ldots,n. \tag{6.14}$$

Let us specify $\boldsymbol{x}^{(0)}$ arbitrarily, but so that the initial error $\|\boldsymbol{\varepsilon}^{(0)}\|_\infty = \|\boldsymbol{x}-\boldsymbol{x}^{(0)}\|_\infty$ will not exceed some prescribed quantity, e.g., will not exceed one. Let us also assume that the accuracy of the approximate solution will be satisfactory if in the course of the iteration the initial error drops by three orders of magnitude, i.e., if the error becomes no greater than 10^{-3}. Then it is sufficient to choose p so that to guarantee the inequality $q^p \leq 10^{-3}$. If, for example, $q = 1/2$, then one can take $p = 10$ regardless of the value of n. The overall computational cost will then be $\mathscr{O}(10n^2) = \mathscr{O}(n^2)$ arithmetic operations, as opposed to the cubic cost $\mathscr{O}(n^3)$ of Gaussian elimination. Indeed, every matrix-vector multiplication $\boldsymbol{B}\boldsymbol{x}^{(p)}$ requires $\mathscr{O}(n^2)$ operations. Of course, the key consideration here is the actual rate of convergence determined by the value of q. We will later provide accurate estimates for the convergence rates of several iteration schemes: the Richardson method (Section 6.1.3), the Chebyshev method (Section 6.2.1), and the method of conjugate gradients (Section 6.2.2).

Example 2: The Gauss-Seidel Method

The Gauss-Seidel method is similar to the Jacobi method, except that when computing the component $x_i^{(p+1)}$, the previously updated components $x_1^{(p+1)}, x_2^{(p+1)}, \ldots, x_{i-1}^{(p+1)}$ are immediately put to use, which yields the following iteration scheme [cf. formula (6.14)]:

$$x_i^{(p+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(p+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(p)} + \frac{f_i}{a_{ii}}, \quad i = 1,2,\ldots,n. \tag{6.15}$$

Using matrix notations, we can write the Gauss-Seidel method (6.15) as follows:

$$\boldsymbol{x}^{(p+1)} = -\boldsymbol{D}^{-1}\hat{\boldsymbol{L}}\boldsymbol{x}^{(p+1)} - \boldsymbol{D}^{-1}\hat{\boldsymbol{U}}\boldsymbol{x}^{(p)} + \boldsymbol{D}^{-1}\boldsymbol{f}, \tag{6.16}$$

where $\hat{\boldsymbol{L}} = \{\mathfrak{l}_{ij}\}$ is a lower triangular matrix with the entries: $\mathfrak{l}_{ij} = \begin{cases} a_{ij}, & j < i, \\ 0 & j \geq i, \end{cases}$ and $\hat{\boldsymbol{U}} = \{\mathfrak{u}_{ij}\}$ is an upper triangular matrix with the entries: $\mathfrak{u}_{ij} = \begin{cases} 0, & j \leq i, \\ a_{ij} & j > i, \end{cases}$ so that altogether $\boldsymbol{A} = \hat{\boldsymbol{L}} + \boldsymbol{D} + \hat{\boldsymbol{U}}$. By noticing that $(\boldsymbol{I} + \boldsymbol{D}^{-1}\hat{\boldsymbol{L}}) = \boldsymbol{D}^{-1}(\boldsymbol{D} + \hat{\boldsymbol{L}}) = \boldsymbol{D}^{-1}(\boldsymbol{A} - \hat{\boldsymbol{U}})$, we convert expression (6.16) to the form (6.4):

$$\boldsymbol{x}^{(p+1)} = -(\boldsymbol{A} - \hat{\boldsymbol{U}})^{-1}\hat{\boldsymbol{U}}\boldsymbol{x}^{(p)} + (\boldsymbol{A} - \hat{\boldsymbol{U}})^{-1}\boldsymbol{f} \tag{6.17}$$

and see that the iteration matrix $\boldsymbol{B}$ for the Gauss-Seidel method is given by: $\boldsymbol{B} = -(\boldsymbol{A} - \hat{\boldsymbol{U}})^{-1}\hat{\boldsymbol{U}}$. If the matrix $\boldsymbol{A}$ is diagonally dominant, see (6.12), then the Gauss-Seidel iterations (6.17) are known to converge to the solution $\boldsymbol{x}$ of system (6.1) for an arbitrary initial guess $\boldsymbol{x}^{(0)}$. We refer the reader to [Axe94] for the proof.

Example 3: The Over-Relaxation Methods

Note that the condition of diagonal dominance (6.12) of the matrix $\boldsymbol{A}$ that guarantees convergence of both the Jacobi method (6.14) and the Gauss-Seidel method (6.15) is only a sufficient and not a necessary condition of convergence. These methods may, in fact, converge for other types of matrices as well. Often, the convergence of an iteration is judged experimentally rather than studied theoretically. In this case, it may be beneficial to consider a broader family of algorithms that would provide more room for tuning the parameters in order to achieve a better convergence. A widely used generalization of both the Jacobi iteration and the Gauss-Seidel iteration is obtained by introducing the relaxation parameter γ and using a weighted update between the old and the new value, which leads to the Jacobi over-relaxation method (JOR):

$$x_i^{(p+1)} = \gamma\left(-\sum_{\substack{j=1,\\ j\neq i}}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(p)} + \frac{f_i}{a_{ii}} \right) + (1-\gamma) x_i^{(p)}, \quad i = 1,2,\ldots,n, \tag{6.18}$$

and to the successive over-relaxation method (SOR):

$$x_i^{(p+1)} = \gamma\left(-\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(p+1)} - \sum_{j=i+1}^{n} \frac{a_{ij}}{a_{ii}} x_j^{(p)} + \frac{f_i}{a_{ii}} \right) + (1-\gamma) x_i^{(p)}, \quad i = 1,2,\ldots,n. \tag{6.19}$$

Theoretical convergence results for the iterations (6.18) or (6.19) are only obtained for some particular cases. Again, we refer the reader to [Axe94] for detail. In general, the successive over-relaxation method (6.19) will not converge if $\gamma < 0$ or $\gamma > 2$. Otherwise, adjusting the value of γ can be used to speed up the convergence. Conversely, if it is observed in a numerical experiment that the convergence for a given case is "iffy," one may try and gain a better robustness by trading off the convergence rate, i.e., by assigning more weight to the old value and less weight to the new value, which means taking a small positive γ.

6.1.2 A Necessary and Sufficient Condition for Convergence

THEOREM 6.2

Let $\mathbb{L}$ be a complex n-dimensional linear space, and let $\boldsymbol{B}$ be an operator mapping this space onto itself $\boldsymbol{B} : \mathbb{L} \longmapsto \mathbb{L}$. A first order linear stationary iterative method (6.4):

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{B}\boldsymbol{x}^{(p)} + \boldsymbol{\varphi}, \quad p = 0,1,2,\ldots, \tag{6.20}$$

converges to the solution $\boldsymbol{x}$ of problem (6.1) in any norm and for an arbitrary initial guess $\boldsymbol{x}^{(0)} \in \mathbb{L}$ if and only if the spectral radius of the operator $\boldsymbol{B}$ is strictly less than one (here λ_j, $j = 1,\ldots,n$, are the eigenvalues of $\boldsymbol{B}$):

$$\rho(\boldsymbol{B}) \stackrel{\text{def}}{=} \max_j |\lambda_j| < 1. \tag{6.21}$$

PROOF We will first prove the sufficiency, that is, if inequality (6.21) holds then the iterations (6.20) converge. Inequality (6.21) obviously implies that the number $\lambda = 1$ is not an eigenvalue of the operator $\boldsymbol{B}$. Consequently, the linear system $\boldsymbol{Bx} = 1 \cdot \boldsymbol{x}$ only has a trivial solution $\boldsymbol{x} = 0$ and as such, the system $\boldsymbol{x} = \boldsymbol{Bx} + \boldsymbol{\varphi}$ has a unique solution for any $\boldsymbol{\varphi} \in \mathbb{L}$.

As before, we define the error of the iterate $\boldsymbol{x}^{(p)}$ as follows: $\boldsymbol{\varepsilon}^{(p)} = \boldsymbol{x} - \boldsymbol{x}^{(p)}$. Let us first note that convergence of the sequence $\boldsymbol{x}^{(p)}$, $p = 0, 1, 2, \ldots$, for an arbitrary $\boldsymbol{x}^{(0)}$ is equivalent to convergence of the sequence $\boldsymbol{\varepsilon}^{(p)}$ to zero for an arbitrary $\boldsymbol{\varepsilon}^{(0)}$: $\boldsymbol{\varepsilon}^{(p)} \longrightarrow \boldsymbol{0}$ as $p \longrightarrow \infty$. Indeed, let us first assume that the sequence $\boldsymbol{x}^{(p)}$ converges. Then it can only converge to the solution $\boldsymbol{x}$, because by taking both sides of equality (6.20) to the limit $p \longrightarrow \infty$ we obtain that the limit $\lim\limits_{p\to\infty} \boldsymbol{x}^{(p)}$ furnishes a solution to the system $\boldsymbol{x} = \boldsymbol{Bx} + \boldsymbol{\varphi}$. Because of the uniqueness, $\lim\limits_{p\to\infty} \boldsymbol{x}^{(p)} = \boldsymbol{x}$, and consequently, $\lim\limits_{p\to\infty} \boldsymbol{\varepsilon}^{(p)} = \boldsymbol{0}$. Conversely, if $\boldsymbol{\varepsilon}^{(p)} \longrightarrow \boldsymbol{0}$ as $p \longrightarrow \infty$, then clearly $\boldsymbol{x}^{(p)} \longrightarrow \boldsymbol{x}$ as $p \longrightarrow \infty$.

Let us fix some $\boldsymbol{\varepsilon}^{(0)}$. For the error $\boldsymbol{\varepsilon}^{(p)}$ we can write:

$$\boldsymbol{\varepsilon}^{(p+1)} = \boldsymbol{B}\boldsymbol{\varepsilon}^{(p)}, \quad p = 0, 1, 2, \ldots,$$

which yields: $\|\boldsymbol{\varepsilon}^{(p)}\| \le \|\boldsymbol{B}\|^p \|\boldsymbol{\varepsilon}^{(0)}\|$. Denote by $\boldsymbol{w}(\lambda)$, $\lambda \in \mathbb{C}$, the sum of the series of vector quantities:

$$\boldsymbol{w}(\lambda) = \sum_{p=0}^{\infty} \frac{\boldsymbol{\varepsilon}^{(p)}}{\lambda^p}. \tag{6.22}$$

This series converges uniformly (and absolutely) outside of any disk of radius $\|\boldsymbol{B}\| + \eta$, $\eta > 0$, centered at the origin on the complex plane of the variable λ. Indeed, $\forall \lambda \in \mathbb{C}$, $|\lambda| > \|\boldsymbol{B}\| + \eta$, series (6.22) is majorized (componentwise) by a convergent geometric series: $\|\boldsymbol{\varepsilon}^{(0)}\| \sum_{p=0}^{\infty} \frac{\|\boldsymbol{B}\|^p}{(\|\boldsymbol{B}\|+\eta)^p}$. According to the Weierstrass theorem proven in the courses of complex analysis, see, e.g., [Mar77, Chapter 3], the sum of a uniformly converging series of holomorphic functions is holomorphic. Therefore, in the region of convergence the function $\boldsymbol{w}(\lambda)$ is a holomorphic vector function of its argument λ, and the series (6.22) is its Laurent series. It is also easy to see that $\lambda \boldsymbol{w}(\lambda) - \lambda \boldsymbol{\varepsilon}^{(0)} = \boldsymbol{B}\boldsymbol{w}(\lambda)$, which, in turn, means: $\boldsymbol{w}(\lambda) = -\lambda(\boldsymbol{B} - \lambda \boldsymbol{I})^{-1}\boldsymbol{\varepsilon}^{(0)}$. Moreover, by multiplying the series (6.22) by λ^{p-1} and then integrating (counterclockwise) along the circle $|\lambda| = r$ on the complex plane, where the number r is to be chosen so that the contour of integration lie within the area of convergence, i.e., $r \ge \|\boldsymbol{B}\| + \eta$, we obtain:

$$\boldsymbol{\varepsilon}^{(p)} = \frac{1}{2\pi i} \int\limits_{|\lambda|=r} \lambda^{p-1} \boldsymbol{w}(\lambda) d\lambda = -\frac{1}{2\pi i} \int\limits_{|\lambda|=r} \lambda^p (\boldsymbol{B} - \lambda \boldsymbol{I})^{-1} \boldsymbol{\varepsilon}^{(0)} d\lambda. \tag{6.23}$$

Indeed, integrating the individual powers of λ on the complex plane, we have:

$$\int\limits_{|\lambda|=r} \lambda^k d\lambda = \begin{cases} 2\pi i, & k = -1, \\ 0, & k \ne -1. \end{cases}$$

In other words, formula (6.23) implies that $-\boldsymbol{\varepsilon}^{(p)}$ is the residue of the vector function $\lambda^{p-1}\boldsymbol{w}(\lambda)$ at infinity.

Next, according to inequality (6.21), all the eigenvalues of the operator $\boldsymbol{B}$ belong to the disk of radius $\rho < 1$ centered at the origin on the complex plane: $|\lambda_j| \le \rho < 1$, $j = 1,2,\ldots,n$. Then the integrand in the second integral of formula (6.23) is an analytic vector function of λ outside of this disk, i.e., for $|\lambda| > \rho$, because the operator $(\boldsymbol{B} - \lambda \boldsymbol{I})^{-1}$ exists (i.e., is bounded) for all $\lambda : \ |\lambda| > \rho$. This function is the analytic continuation of the function $\lambda^{p-1}\boldsymbol{w}(\lambda)$, where $\boldsymbol{w}(\lambda)$ is originally defined by the series (6.22) that can only be proven to converge outside of a larger disk $|\lambda| \le \|\boldsymbol{B}\| + \eta$. Consequently, the contour of integration in (6.23) can be altered, and instead of $r \ge \|\boldsymbol{B}\| + \eta$ one can take $r = \rho + \zeta$, where $\zeta > 0$ is arbitrary, without changing the value of the integral. Therefore, the error can be estimated as follows:

$$\begin{aligned} \|\boldsymbol{\varepsilon}^{(p)}\| &= \frac{1}{2\pi}\left\| \int\limits_{|\lambda|=\rho+\zeta} \lambda^p (\boldsymbol{B} - \lambda \boldsymbol{I})^{-1} \boldsymbol{\varepsilon}^{(0)} d\lambda \right\| \\ &\le (\rho+\zeta)^p \max_{|\lambda|=\rho+\zeta} \|(\boldsymbol{B} - \lambda \boldsymbol{I})^{-1}\| \|\boldsymbol{\varepsilon}^{(0)}\|. \end{aligned} \tag{6.24}$$

In formula (6.24), let us take $\zeta > 0$ sufficiently small so that $\rho + \zeta < 1$. Then, the right-hand side of inequality (6.24) vanishes as p increases, which implies the convergence: $\|\boldsymbol{\varepsilon}^{(p)}\| \longrightarrow 0$ when $p \longrightarrow \infty$. This completes the proof of sufficiency.

To prove the necessity, suppose that inequality (6.21) does not hold, i.e., that for some λ_k we have $|\lambda_k| \ge 1$. At the same time, contrary to the conclusion of the theorem, let us assume that the convergence still takes place for any choice of $\boldsymbol{x}^{(0)}$: $\boldsymbol{x}^{(p)} \longrightarrow \boldsymbol{x}$ as $p \longrightarrow \infty$. Then we can choose $\boldsymbol{x}^{(0)}$ so that $\boldsymbol{\varepsilon}^{(0)} = \boldsymbol{x} - \boldsymbol{x}^{(0)} = \boldsymbol{e}_k$, where $\boldsymbol{e}_k$ is the eigenvector of the operator $\boldsymbol{B}$ that corresponds to the eigenvalue λ_k. In this case, $\boldsymbol{\varepsilon}^{(p)} = \boldsymbol{B}^p \boldsymbol{\varepsilon}^{(0)} = \boldsymbol{B}^p \boldsymbol{e}_k = \lambda_k^p \boldsymbol{e}_k$. As $|\lambda_k| \ge 1$, the sequence $\lambda_k^p \boldsymbol{e}_k$ does not converge to $\boldsymbol{0}$ when p increases. The contradiction proves the necessity. ∎

REMARK 6.2 Let us make an interesting and important observation of a situation that we encounter here for the first time. The problem of computing the limit $\boldsymbol{x} = \lim\limits_{p\to\infty} \boldsymbol{x}^{(p)}$ is ultimately well conditioned, because the result $\boldsymbol{x}$ does not depend on the initial data at all, i.e., it does not depend on the initial guess $\boldsymbol{x}^{(0)}$. Yet the algorithm for computing the sequence $\boldsymbol{x}^{(p)}$ that converges according to Theorem 6.2 may still appear computationally unstable. The instability may take place if along with the inequality $\max_j |\lambda_j\| = \rho < 1$ we have $\|\boldsymbol{B}\| > 1$. This situation is typical for non-self-adjoint (or non-normal) matrices $\boldsymbol{B}$ (opposite of Theorem 5.2).

Indeed, if $\|\boldsymbol{B}\| < 1$, then the norm of the error $\|\boldsymbol{\varepsilon}^{(p)}\| = \|\boldsymbol{B}^p \boldsymbol{\varepsilon}^{(0)}\|$ decreases monotonically, this is the result of Theorem 6.1. Otherwise, if $\|\boldsymbol{B}\| > 1$, then for some $\boldsymbol{\varepsilon}^{(0)}$ the norm $\|\boldsymbol{\varepsilon}^{(p)}\|$ will initially grow, and only then decrease. The

behavior will be qualitatively similar to that shown in Figure 10.13 (see page 394) that pertains to the study of stability for finite-difference initial boundary value problems. In doing so, the height of the intermediate "hump" on the curve showing the dependence of $\|\boldsymbol{\varepsilon}^{(p)}\|$ on p may be arbitrarily high. A small relative error committed near the value of p that corresponds to the maximum of the "hump" will subsequently increase (i.e., its norm will increase) — this error will also evolve and undergo a maximum, etc. The resulting instability may appear so strong that the computation will become practically impossible already for moderate dimensions n and for the norms $\|\boldsymbol{B}\|$ that are only slightly larger than one. □

A rigorous definition of stability for the first order linear stationary iterative methods, as well as the classification of possible instabilities, the pertinent theorems, and examples can be found in work [Rya70].

6.1.3 The Richardson Method for $\boldsymbol{A} = \boldsymbol{A}^* > 0$

Consider equation (6.3): $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{\varphi}$, $\boldsymbol{x} \in \mathbb{L}$, assuming that $\mathbb{L}$ is an n-dimensional Euclidean space with the inner product $(\boldsymbol{x}, \boldsymbol{y})$ and the norm $\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x}, \boldsymbol{x})}$ (e.g., $\mathbb{L} = \mathbb{R}^n$). Also assume that $\boldsymbol{B} : \mathbb{L} \longmapsto \mathbb{L}$ is a self-adjoint operator: $\boldsymbol{B} = \boldsymbol{B}^*$, with respect to the chosen inner product. Let ν_j, $j = 1, 2, \ldots, n$, be the eigenvalues of $\boldsymbol{B}$ and let

$$\rho = \rho(\boldsymbol{B}) = \max_j |\nu_j|$$

be its spectral radius. Specify an arbitrary $\boldsymbol{x}^{(0)} \in \mathbb{L}$ and build a sequence of iterations:

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{B}\boldsymbol{x}^{(p)} + \boldsymbol{\varphi}, \quad p = 0, 1, 2, \ldots. \tag{6.25}$$

LEMMA 6.1

1. *If $\rho < 1$ then the system $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{\varphi}$ has a unique solution $\boldsymbol{x} \in \mathbb{L}$; the iterates $\boldsymbol{x}^{(p)}$ of (6.25) converge to $\boldsymbol{x}$; and the Euclidean norm of the error $\|\boldsymbol{x} - \boldsymbol{x}^{(p)}\|$ satisfies the estimate:*

$$\|\boldsymbol{x} - \boldsymbol{x}^{(p)}\| \le \rho^p \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\|, \quad p = 0, 1, 2, \ldots. \tag{6.26}$$

   Moreover, there is a particular $\boldsymbol{x}^{(0)} \in \mathbb{L}$ for which inequality (6.26) transforms into a precise equality.

2. *Let the system $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{\varphi}$ have a solution $\boldsymbol{x} \in \mathbb{L}$ for a given $\boldsymbol{\varphi} \in \mathbb{L}$, and let $\rho \ge 1$. Then there is an initial guess $\boldsymbol{x}^{(0)} \in \mathbb{L}$ such that the corresponding sequence of iterations (6.25) does not converge to $\boldsymbol{x}$.*

PROOF According to Theorem 5.2, the Euclidean norm of a self-adjoint operator $\boldsymbol{B} = \boldsymbol{B}^*$ coincides with its spectral radius ρ. Therefore, the first conclusion of the lemma except its last statement holds by virtue of Theorem 6.1.

To find the initial guess that would turn (6.26) into an equality, we first introduce our standard notation $\boldsymbol{\varepsilon}^{(p)} = \boldsymbol{x} - \boldsymbol{x}^{(p)}$ for the error of the iterate $\boldsymbol{x}^{(p)}$, and subtract equation (6.25) from $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{\varphi}$, which yields: $\boldsymbol{\varepsilon}^{(p+1)} = \boldsymbol{B}\boldsymbol{\varepsilon}^{(p)}$, $p = 0, 1, 2, \ldots$. Next, suppose that $|\nu_k| = \max_j |\nu_j| = \rho$ and take $\boldsymbol{\varepsilon}^{(0)} = \boldsymbol{x} - \boldsymbol{x}^{(0)} = \boldsymbol{e}_k$, where $\boldsymbol{e}_k$ is the eigenvector of $\boldsymbol{B}$ that corresponds to the eigenvalue with maximum magnitude. Then we obtain: $\|\boldsymbol{\varepsilon}^{(p)}\| = |\nu_k|^p \|\boldsymbol{\varepsilon}^{(0)}\| = \rho^p \|\boldsymbol{\varepsilon}^{(0)}\|$.

To prove the second conclusion of the lemma, we take the particular eigenvalue ν_k that delivers the maximum: $|\nu_k| = \max_j |\nu_j| = \rho \geq 1$, and again select $\boldsymbol{\varepsilon}^{(0)} = \boldsymbol{x} - \boldsymbol{x}^{(0)} = \boldsymbol{e}_k$, where $\boldsymbol{e}_k$ is the corresponding eigenvector. In this case the error obviously does not vanish as $p \longrightarrow \infty$, because:

$$\boldsymbol{\varepsilon}^{(p)} = \boldsymbol{B}\boldsymbol{\varepsilon}^{(p-1)} = \ldots = \boldsymbol{B}^p \boldsymbol{\varepsilon}^{(0)} = \nu_k^p \boldsymbol{e}_k,$$

and consequently, $\|\boldsymbol{\varepsilon}^{(p)}\| = \rho^p \|\boldsymbol{e}_k\|$, where ρ^p will either stay bounded but will not vanish, or will increase when $p \longrightarrow \infty$. □

Lemma 6.1 analyzes a special case $\boldsymbol{B} = \boldsymbol{B}^*$ and provides a simple illustration for the general conclusion of Theorem 6.2 that for the convergence of a first order linear stationary iteration it is necessary and sufficient that the spectral radius of the iteration matrix be strictly less than one. With the help of this lemma, we will now analyze convergence of the stationary Richardson iteration (6.5) for the case $\boldsymbol{A} = \boldsymbol{A}^* > 0$.

THEOREM 6.3
Consider a system of linear algebraic equations:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}, \quad \boldsymbol{A} = \boldsymbol{A}^* > 0, \tag{6.27}$$

where $\boldsymbol{x}, \boldsymbol{f} \in \mathbb{L}$, and $\mathbb{L}$ is an n-dimensional Euclidean space (e.g., $\mathbb{L} = \mathbb{R}^n$). Let $\lambda_{\min}$ and $\lambda_{\max}$ be the smallest and the largest eigenvalues of the operator $\boldsymbol{A}$, respectively. Specify some $\tau \neq 0$ and recast system (6.27) in an equivalent form:

$$\boldsymbol{x} = (\boldsymbol{I} - \tau\boldsymbol{A})\boldsymbol{x} + \tau\boldsymbol{f}. \tag{6.28}$$

Given an arbitrary initial guess $\boldsymbol{x}^{(0)} \in \mathbb{L}$, consider the sequence of Richardson iterations:

$$\boldsymbol{x}^{(p+1)} = (\boldsymbol{I} - \tau\boldsymbol{A})\boldsymbol{x}^{(p)} + \tau\boldsymbol{f}, \quad p = 0, 1, 2, \ldots. \tag{6.29}$$

1. *If the parameter τ satisfies the inequalities:*

$$0 < \tau < \frac{2}{\lambda_{\max}}, \tag{6.30}$$

then the sequence $\boldsymbol{x}^{(p)}$ of (6.29) converges to the solution $\boldsymbol{x}$ of system (6.27). Moreover, the norm of the error $\|\boldsymbol{x} - \boldsymbol{x}^{(p)}\|$ is guaranteed to decrease when p increases with the rate given by the following estimate:

$$\|\boldsymbol{x} - \boldsymbol{x}^{(p)}\| \leq \rho^p \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\|, \quad p = 0, 1, 2, \ldots. \tag{6.31}$$

The quantity ρ in formula (6.31) is defined as

$$\rho = \rho(\tau) = \max\{|1 - \tau\lambda_{\min}|, |1 - \tau\lambda_{\max}|\}. \tag{6.32}$$

This quantity is less than one, $\rho < 1$, as it is the maximum of two numbers, $|1 - \tau\lambda_{\min}|$ and $|1 - \tau\lambda_{\max}|$, neither of which may exceed one provided that inequalities (6.30) hold.

2. *Let the number τ satisfy (6.30). Then there is a special initial guess $\boldsymbol{x}^{(0)}$ for which estimate (6.31) cannot be improved, because for this $\boldsymbol{x}^{(0)}$ inequality (6.31) transforms into a precise equality.*

3. *If condition (6.30) is violated, so that either $\tau \geq 2/\lambda_{\max}$ or $\tau \leq 0$, then there is an initial guess $\boldsymbol{x}^{(0)}$ for which the sequence $\boldsymbol{x}^{(p)}$ of (6.29) does not converge to the solution $\boldsymbol{x}$ of system (6.27).*

4. *The number $\rho = \rho(\tau)$ given by formula (6.32) assumes its minimal (i.e., optimal) value $\rho_{\text{opt}} = \rho(\tau_{\text{opt}})$ when $\tau = \tau_{\text{opt}} = 2/(\lambda_{\min} + \lambda_{\max})$. In this case,*

$$\rho = \rho_{\text{opt}} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\mu(\boldsymbol{A}) - 1}{\mu(\boldsymbol{A}) + 1}, \tag{6.33}$$

where $\mu(\boldsymbol{A}) = \lambda_{\max}/\lambda_{\min}$ is the condition number of the operator $\boldsymbol{A}$ (see Theorem 5.3).

PROOF To prove Theorem 6.3, we will use Lemma 6.1. In this lemma, let us set $\boldsymbol{B} = \boldsymbol{I} - \tau\boldsymbol{A}$. Note that if $\boldsymbol{A} = \boldsymbol{A}^*$ then the operator $\boldsymbol{B} = \boldsymbol{I} - \tau\boldsymbol{A}$ is also self-adjoint, i.e., $\boldsymbol{B} = \boldsymbol{B}^*$:

$$\begin{aligned}(\boldsymbol{B}\boldsymbol{x}, \boldsymbol{y}) &= ((\boldsymbol{I} - \tau\boldsymbol{A})\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}, \boldsymbol{y}) - \tau(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{y}) \\ &= (\boldsymbol{x}, \boldsymbol{y}) - \tau(\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}) = (\boldsymbol{x}, (\boldsymbol{I} - \tau\boldsymbol{A})\boldsymbol{y}) = (\boldsymbol{x}, \boldsymbol{B}\boldsymbol{y}).\end{aligned}$$

Suppose that λ_j, $j = 1, 2, \ldots, n$, are the eigenvalues of the operator $\boldsymbol{A}$ arranged in the ascending order:

$$0 < \lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n = \lambda_{\max}, \tag{6.34}$$

and $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ are the corresponding eigenvectors: $\boldsymbol{A}\boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j$, $j = 1, 2, \ldots, n$, that form an orthonormal basis in the space $\mathbb{L}$. Then clearly, the same vectors $\boldsymbol{e}_j$, $j = 1, 2, \ldots, n$, are also eigenvectors of the operator $\boldsymbol{B}$, whereas the respective eigenvalues are given by:

$$\nu_j = \nu_j(\tau) = 1 - \tau\lambda_j, \quad j = 1, 2, \ldots, n. \tag{6.35}$$

Indeed,

$$\boldsymbol{B}\boldsymbol{e}_j = (\boldsymbol{I} - \tau\boldsymbol{A})\boldsymbol{e}_j = \boldsymbol{e}_j - \tau\lambda_j\boldsymbol{e}_j = (1 - \tau\lambda_j)\boldsymbol{e}_j = \nu_j\boldsymbol{e}_j, \quad j = 1, 2, \ldots, n.$$

According to (6.34), if $\tau > 0$ then the eigenvalues ν_j given by formula (6.35) are arranged in the descending order, see Figure 6.1:

$$1 > \nu_1 \geq \nu_2 \geq \ldots \geq \nu_n.$$

From Figure 6.1 it is also easy to see that the largest among the absolute values $|\nu_j|$, $j = 1,2,\ldots,n$, may be either $|\nu_1| = |1-\tau\lambda_1| \equiv |1-\tau\lambda_{\min}|$ or $|\nu_n| = |1-\tau\lambda_n| \equiv |1-\tau\lambda_{\max}|$; the case $|\nu_n| = \max_j |\nu_j|$ is realized when $\nu_n = 1-\tau\lambda_{\max} < 0$ and $|1-\tau\lambda_{\max}| > |1-\tau\lambda_{\min}|$. Consequently, the condition:

$$\rho = \max_j |\nu_j| < 1 \tag{6.36}$$

(see Lemma 6.1) coincides with the condition [see formula (6.32)]:

$$\rho = \max\{|1-\tau\lambda_{\min}|, |1-\tau\lambda_{\max}|\} < 1.$$

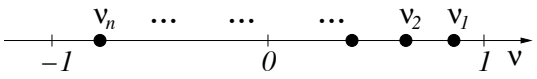

FIGURE 6.1: Eigenvalues of the matrix $\boldsymbol{B} = \boldsymbol{I} - \tau\boldsymbol{A}$.

Clearly, if $\tau > 0$ we can only guarantee $\rho < 1$ provided that the point ν_n on Figure 6.1 is located to the right of the point -1, i.e., if $\nu_n = 1 - \tau\lambda_{\max} > -1$. This means that along with $\tau > 0$ the second inequality of (6.30) also holds. Otherwise, if $\tau \geq 2/\lambda_{\max}$, then $\rho > 1$. If $\tau < 0$, then $\nu_j = 1-\tau\lambda_j = 1+|\tau|\lambda_j > 1$ for all $j = 1,2,\ldots,n$, and we will always have $\rho = \max_j |\nu_j| > 1$. Hence, condition (6.30) is equivalent to the requirement (6.36) for $\boldsymbol{B} = \boldsymbol{I} - \tau\boldsymbol{A}$ (or to requirement (6.21) of Theorem 6.2). Therefore, by virtue of Lemma 6.1, have proven the first three implications of Theorem 6.3.

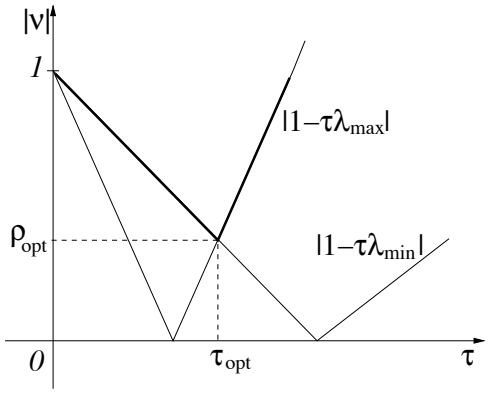

FIGURE 6.2: $|\nu_1|$ and $|\nu_n|$ as functions of τ.

To prove the remaining fourth implication, we need to analyze the behavior of the quantities $|\nu_1| = |1-\tau\lambda_{\min}|$ and $|\nu_n| = |1-\tau\lambda_{\max}|$ as functions of τ. We schematically show this behavior in Figure 6.2. From this figure, we determine that for smaller values of τ the quantity $|\nu_1|$ dominates, i.e., $|1-\tau\lambda_{\min}| > |1-\tau\lambda_{\max}|$, whereas for larger values of τ the quantity $|\nu_n|$ dominates, i.e., $|1-\tau\lambda_{\max}| > |1-\tau\lambda_{\min}|$. The value of $\rho(\tau) = \max\{|1-\tau\lambda_{\min}|, |1-\tau\lambda_{\max}|\}$ is shown by a bold

polygonal line in Figure 6.2; it coincides with $|1-\tau\lambda_{\min}|$ before the intersection point, and after this point it coincides with $|1-\tau\lambda_{\max}|$. Consequently, the minimum value of $\rho=\rho_{\text{opt}}$ is achieved precisely at the intersection, i.e., at the value of $\tau=\tau_{\text{opt}}$ obtained from the following condition: $\nu_1(\tau)=|\nu_n(\tau)|=-\nu_n(\tau)$. This condition reads:

$$1-\tau\lambda_{\min}=\tau\lambda_{\max}-1,$$

which yields:

$$\tau_{\text{opt}}=\frac{2}{\lambda_{\min}+\lambda_{\max}}.$$

Consequently,

$$\rho_{\text{opt}}=\rho(\tau_{\text{opt}})=1-\tau_{\text{opt}}\lambda_{\min}=\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}=\frac{\mu(\boldsymbol{A})-1}{\mu(\boldsymbol{A})+1}.$$

This expression is identical to (6.33), which completes the proof. □

Let us emphasize the following important consideration. Previously, we saw that the condition number of a matrix determines how sensitive the solution of the corresponding linear system will be to the perturbations of the input data (Section 5.3.2). The result of Theorem 6.3 provides the first evidence that the condition number also determines the rate of convergence of an iterative method. Indeed, from formula (6.33) it is clear that the closer the value of $\mu(\boldsymbol{A})$ to one, the closer the value of ρ_{opt} to zero, and consequently, the faster is the decay of the error according to estimate (6.31). When the condition number $\mu(\boldsymbol{A})$ increases, so does the quantity ρ_{opt} (while still remaining less than one) and the convergence slows down.

According to formulae (6.31) and (6.33), the optimal choice of the iteration parameter $\tau=\tau_{\text{opt}}$ enables the following error estimate:

$$\|\boldsymbol{\varepsilon}^{(p)}\|\le\left(\frac{1-\xi}{1+\xi}\right)^p\|\boldsymbol{\varepsilon}^{(0)}\|,\quad\text{where}\quad\xi=\frac{\lambda_{\min}}{\lambda_{\max}}=\frac{1}{\mu(\boldsymbol{A})}.$$

Moreover, Lemma 6.1 implies that this estimate cannot be improved, i.e., that there is a particular initial guess $\boldsymbol{x}^{(0)}$ (and hence $\boldsymbol{\varepsilon}^{(0)}$), for which the inequality transforms into a precise equality. Therefore, in order to guarantee that the initial error drops by a prescribed factor in the course of the iteration, i.e., in order to guarantee the estimate:

$$\|\boldsymbol{\varepsilon}^{(p)}\|\le\sigma\|\boldsymbol{\varepsilon}^{(0)}\|,\tag{6.37}$$

where $\sigma>0$ is given, it is necessary and sufficient to select p that would satisfy:

$$\left(\frac{1-\xi}{1+\xi}\right)^p\le\sigma,\quad\text{i.e.,}\quad p\ge-\frac{\ln\sigma}{\ln(1+\xi)-\ln(1-\xi)}.$$

A more practical estimate for the number p can also be obtained. Note that

$$\ln(1+\xi)-\ln(1-\xi)=2\xi\sum_{k=0}^{\infty}\frac{\xi^{2k}}{2k+1},$$

where

$$1 \le \sum_{k=0}^{\infty} \frac{\xi^{2k}}{2k+1} \le \frac{1}{1-\xi^2}.$$

Therefore, for the estimate (6.37) to hold it is sufficient that p satisfy:

$$p \ge -\frac{1}{2} \ln \sigma \cdot \mu(\boldsymbol{A}), \qquad \mu(\boldsymbol{A}) = \frac{1}{\xi}, \tag{6.38a}$$

and it is necessary that

$$p \ge -\frac{1}{2} \ln \sigma \cdot (1-\xi^2)\mu(\boldsymbol{A}). \tag{6.38b}$$

Altogether, the number of Richardson iterations required for reducing the initial error by a predetermined factor *is proportional to the condition number of the matrix*.

REMARK 6.3 In many cases, for example when approximating elliptic boundary value problems using finite differences (see, e.g., Section 5.1.3), the operator $\boldsymbol{A} : \mathbb{L} \longmapsto \mathbb{L}$ of the resulting linear system (typically, $\mathbb{L} = \mathbb{R}^n$) appears self-adjoint and positive definite ($\boldsymbol{A} = \boldsymbol{A}^* > 0$) in the sense of some natural inner product. However, most often one cannot find the precise minimum and maximum eigenvalues for such operators. Instead, only the estimates a and b for the boundaries of the spectrum may be available:

$$0 < a \le \lambda_{\min} \le \lambda_{\max} \le b. \tag{6.39}$$

In this case, the Richardson iteration (6.29) can still be used for solving the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}$. □

The key difference, though, between the more general case outlined in Remark 6.3 and the case of Theorem 6.3, for which the precise boundaries of the spectrum are known, is the way the iteration parameter τ is selected. If instead of $\lambda_{\min}$ and $\lambda_{\max}$ we only know a and b, see formula (6.39), then the best we can do is take $\tau' = 2/(a+b)$ instead of $\tau_{\text{opt}} = 2/(\lambda_{\min} + \lambda_{\max})$. Then, instead of ρ_{opt} given by formula (6.33):

$$\rho_{\text{opt}} = (\lambda_{\min} - \lambda_{\max})/(\lambda_{\min} + \lambda_{\max})$$

another quantity

$$\rho' = \max\{|1 - \tau'\lambda_{\min}|, |1 - \tau'\lambda_{\max}|\},$$

which is larger than ρ_{opt}, will appear in the guaranteed error estimate (6.31). As has been shown, for any value of τ within the limits (6.30), and for the respective value of $\rho = \rho(\tau)$ given by formula (6.32), there is always an initial guess $\boldsymbol{x}^{(0)}$ for which estimate (6.31) becomes a precise equality. Therefore, for $\tau = \tau' \ne \tau_{\text{opt}}$ we obtain an unimprovable estimate (6.31) with $\rho = \rho' > \rho_{\text{opt}}$. In doing so, the rougher the estimate for the boundaries of the spectrum, the slower the convergence.

Example

Let us apply the Richardson iterative method to solving the finite-difference Dirichlet problem for the Poisson equation: $-\Delta^{(h)}u^{(h)} = f^{(h)}$ that we introduced in Section 5.1.3. In this case, formula (6.29) becomes:

$$u^{(h,p+1)} = (\boldsymbol{I} + \tau\Delta^{(h)})u^{(h,p)} + \tau f^{(h)}, \quad p = 0,1,2,\ldots.$$

The eigenvalues of the operator $-\Delta^{(h)}$ are given by formula (5.109) of Section 5.7.2. In the same Section 5.7.2, we have shown that the operator $-\Delta^{(h)} : U^{(h)} \longmapsto U^{(h)}$ is self-adjoint with respect to the natural scalar product (5.20) on $U^{(h)}$:

$$(u^{(h)}, v^{(h)}) = h^2 \sum_{m_1,m_2=1}^{M-1} u_{m_1,m_2} v_{m_1,m_2}.$$

Finally, we have estimated its condition number: $\mu(-\Delta^{(h)}) = \mathcal{O}(h^{-2})$, see formula (5.115).

Therefore, according to formulae (6.37) and (6.38), when $\tau = \tau_{\text{opt}}$,[1] the number of iterations p required for reducing the initial error, say, by a factor of e is $p \approx \frac{1}{2}\mu(-\Delta^{(h)}) = \mathcal{O}(h^{-2})$. Every iteration requires $\mathcal{O}(h^{-2})$ arithmetic operations and consequently, the overall number of operations is $\mathcal{O}(h^{-4})$.

In Section 5.7.2, we represented the exact solution of the system $-\Delta^{(h)}u^{(h)} = f^{(h)}$ in the form of a finite Fourier series, see formula (5.110). However, in the case of a non-rectangular domain, or in the case of an equation with variable coefficients (as opposed to the Poisson equation):

$$\frac{\partial}{\partial x}\left(a\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(b\frac{\partial u}{\partial y}\right) = f, \;\; a = a(x,y) > 0, \;\; b = b(x,y) > 0, \tag{6.40}$$

we typically do not know the eigenvalues and eigenvectors of the problem and as such, cannot use the discrete Fourier series. At the same time, an iterative algorithm, such as the Richardson method, can still be implemented in quite the same way as it was done previously. In doing so, we only need to make sure that the discrete operator is self-adjoint, and also obtain reasonable estimates for the boundaries of its spectrum.

In Section 6.2, we will analyze other iterative methods for the system $\boldsymbol{Ax} = \boldsymbol{f}$, where $\boldsymbol{A} = \boldsymbol{A}^* > 0$, and will show that even for an ill conditioned operator $\boldsymbol{A}$, say, with the condition number $\mu(\boldsymbol{A}) = \mathcal{O}(h^{-2})$, it is possible to build the methods that will be far more efficient than the Richardson iteration. A better efficiency will be achieved by obtaining a more favorable dependence of the number of required iterations p on the condition number $\mu(\boldsymbol{A})$. We will have $p \gtrsim \sqrt{\mu(\boldsymbol{A})}$ as opposed to $p \gtrsim \mu(\boldsymbol{A})$, which is guaranteed by formulae (6.38).

[1] In this case, $\tau_{\text{opt}} = 2/(\lambda_{11} + \lambda_{M-1,M-1}) \approx h^2/[4(1+\sin^2\frac{\pi h}{2})]$, see formulae (5.112) and (5.114).

6.1.4 Preconditioning

As has been mentioned, some alternative methods that are less expensive computationally than the Richardson iteration will be described in Section 6.2. They will have a slower than linear rate of increase of p as a function of $\mu(\boldsymbol{A})$. A complementary strategy for reducing the number of iterations p consists of modifying the system $\boldsymbol{Ax} = \boldsymbol{f}$ itself, to keep the solution intact and at the same time make the condition number μ smaller.

Let $\boldsymbol{P}$ be a non-singular square matrix of the same dimension as that of $\boldsymbol{A}$. We can equivalently recast system (6.1) by multiplying it from the left by $\boldsymbol{P}^{-1}$:

$$\boldsymbol{P}^{-1}\boldsymbol{Ax} = \boldsymbol{P}^{-1}\boldsymbol{f}. \tag{6.41}$$

The matrix $\boldsymbol{P}$ is known as *a preconditioner.* The solution $\boldsymbol{x}$ of system (6.41) is obviously the same as that of system (6.1).

Accordingly, instead of the standard stationary Richardson iteration (6.5) or (6.6), we now obtain its preconditioned version:

$$\boldsymbol{x}^{(p+1)} = (\boldsymbol{I} - \tau\boldsymbol{P}^{-1}\boldsymbol{A})\boldsymbol{x}^{(p)} + \tau\boldsymbol{P}^{-1}\boldsymbol{f}, \quad p = 0, 1, 2, \ldots, \tag{6.42}$$

or equivalently,

$$\boldsymbol{x}^{(p+1)} = \boldsymbol{x}^{(p)} - \tau\boldsymbol{P}^{-1}\boldsymbol{r}^{(p)}, \quad \text{where} \quad \boldsymbol{r}^{(p)} = \boldsymbol{Ax}^{(p)} - \boldsymbol{f}^{(p)}. \tag{6.43}$$

As a matter of fact, we have already seen some examples of a preconditioned Richardson method. By comparing formulae (6.13) and (6.42), we conclude that the Jacobi method (Example 1 of Section 6.1.1) can be interpreted as a preconditioned Richardson iteration with $\tau = 1$ and $\boldsymbol{P} = \boldsymbol{D} = \operatorname{diag}\{a_{ii}\}$. Similarly, by comparing formulae (6.17) and (6.42) and by noticing that

$$-(\boldsymbol{A} - \hat{\boldsymbol{U}})^{-1}\hat{\boldsymbol{U}} = -(\boldsymbol{A} - \hat{\boldsymbol{U}})^{-1}(\boldsymbol{A} - (\boldsymbol{A} - \hat{\boldsymbol{U}})) = \boldsymbol{I} - (\boldsymbol{A} - \hat{\boldsymbol{U}})^{-1}\boldsymbol{A},$$

we conclude that the Gauss-Seidel method (Example 2 of Section 6.1.1) can be interpreted as a preconditioned Richardson iteration with $\tau = 1$ and $\boldsymbol{P} = \boldsymbol{A} - \hat{\boldsymbol{U}}$.

Of course, we need to remember that the purpose of preconditioning is not to analyze the equivalent system (6.41) "for the sake of it," but rather to reduce the condition number, so that $\mu(\boldsymbol{P}^{-1}\boldsymbol{A}) < \mu(\boldsymbol{A})$ or ideally, $\mu(\boldsymbol{P}^{-1}\boldsymbol{A}) \ll \mu(\boldsymbol{A})$. Unfortunately, relatively little systematic theory is available in the literature for the design of efficient preconditioners. Different types of problems may require special individual tools for analysis, and we refer the reader, e.g., to [Axe94] for detail.

For our subsequent considerations, let us assume that the operator $\boldsymbol{A}$ is self-adjoint and positive definite, as in Section 6.1.3, so that we need to solve the system:

$$\boldsymbol{Ax} = \boldsymbol{f}, \quad \boldsymbol{A} = \boldsymbol{A}^* > 0, \quad \boldsymbol{x} \in \mathbb{L}, \quad \boldsymbol{f} \in \mathbb{L}. \tag{6.44}$$

Introduce an operator $\boldsymbol{P} = \boldsymbol{P}^* > 0$, which can be taken arbitrarily in the meantime, and multiply both sides of system (6.44) by $\boldsymbol{P}^{-1}$, which yields an equivalent system:

$$\boldsymbol{Cx} = \boldsymbol{g}, \quad \boldsymbol{C} = \boldsymbol{P}^{-1}\boldsymbol{A}, \quad \boldsymbol{g} = \boldsymbol{P}^{-1}\boldsymbol{f}. \tag{6.45}$$

Note that the new operator $\boldsymbol{C}$ of (6.45) is, generally speaking, no longer self-adjoint.

Let us, however, introduce a new inner product on the space $\mathbb{L}$ by means of the operator $\boldsymbol{P}$: $[\boldsymbol{x},\boldsymbol{y}]_{\boldsymbol{P}} \stackrel{\text{def}}{=} (\boldsymbol{P}\boldsymbol{x},\boldsymbol{y})$. Then, the operator $\boldsymbol{C}$ of (6.45) appears self-adjoint and positive definite in the sense of this new inner product. Indeed, as the inverse of a self-adjoint operator is also self-adjoint, we can write:

$$\begin{aligned}[\boldsymbol{C}\boldsymbol{x},\boldsymbol{y}]_{\boldsymbol{P}} &= (\boldsymbol{P}\boldsymbol{C}\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{P}\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x},\boldsymbol{A}\boldsymbol{y}) \\ &= (\boldsymbol{P}^{-1}\boldsymbol{P}\boldsymbol{x},\boldsymbol{A}\boldsymbol{y}) = (\boldsymbol{P}\boldsymbol{x},\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{y}) = [\boldsymbol{x},\boldsymbol{C}\boldsymbol{y}]_{\boldsymbol{P}}, \\ [\boldsymbol{C}\boldsymbol{x},\boldsymbol{x}]_{\boldsymbol{P}} &= (\boldsymbol{P}\boldsymbol{C}\boldsymbol{x},\boldsymbol{x}) = (\boldsymbol{P}\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{x},\boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x},\boldsymbol{x}) > 0, \text{ if } \boldsymbol{x} \neq \boldsymbol{0}.\end{aligned}$$

As of yet, the choice of the preconditioner $\boldsymbol{P}$ for system (6.45) was arbitrary. For example, we can choose $\boldsymbol{P} = \boldsymbol{A}$ and obtain $\boldsymbol{C} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$, which immediately yields the solution $\boldsymbol{x}$. As such, $\boldsymbol{P} = \boldsymbol{A}$ can be interpreted as the ideal preconditioner; it provides an indication of what the ultimate goal should be. However, in real life setting $\boldsymbol{P} = \boldsymbol{A}$ is totally impractical. Indeed, the application of the operator $\boldsymbol{P}^{-1} = \boldsymbol{A}^{-1}$ is equivalent to solving the system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}$ directly, which is precisely what we are trying to avoid by employing an iterative scheme. Recall, an iterative method only requires computing $\boldsymbol{A}\boldsymbol{z}$ for a given $\boldsymbol{z} \in \mathbb{L}$, but does not require computing $\boldsymbol{A}^{-1}\boldsymbol{f}$. It therefore only makes sense to select the operator $\boldsymbol{P}$ among those for which the computation of $\boldsymbol{P}^{-1}\boldsymbol{z}$ for a given $\boldsymbol{z}$ is considerably easier than the computation of $\boldsymbol{A}^{-1}\boldsymbol{z}$. The other extreme, however, would be setting $\boldsymbol{P} = \boldsymbol{I}$, which does not require doing anything, but does not bring along any benefits either. In other words, the preconditioner $\boldsymbol{P}$ should be chosen so as to be easily invertible on one hand, and on the other hand, to "resemble" the operator $\boldsymbol{A}$. In this case we can expect that the operator $\boldsymbol{C} = \boldsymbol{P}^{-1}\boldsymbol{A}$ will "resemble" the unit operator $\boldsymbol{I}$, and the boundaries of its spectrum $\lambda_{\min}$ and $\lambda_{\max}$, as well as the condition number, will all be "closer" to one.

THEOREM 6.4

Let $\boldsymbol{P} = \boldsymbol{P}^* > 0$, let the two numbers $\gamma_1 > 0$ and $\gamma_2 > 0$ be fixed, and let the following inequalities hold:

$$\gamma_1(\boldsymbol{P}\boldsymbol{x},\boldsymbol{x}) \leq (\boldsymbol{A}\boldsymbol{x},\boldsymbol{x}) \leq \gamma_2(\boldsymbol{P}\boldsymbol{x},\boldsymbol{x}) \tag{6.46}$$

for all $\boldsymbol{x} \in \mathbb{L}$. Then the eigenvalues $\lambda_{\min}(\boldsymbol{C})$, $\lambda_{\max}(\boldsymbol{C})$ and the condition number $\mu_{\boldsymbol{P}}(\boldsymbol{C})$ of the operator $\boldsymbol{C} = \boldsymbol{P}^{-1}\boldsymbol{A}$ satisfy the inequalities:

$$\begin{gathered}\gamma_1 \leq \lambda_{\min}(\boldsymbol{C}) \leq \lambda_{\max}(\boldsymbol{C}) \leq \gamma_2, \\ \mu_{\boldsymbol{P}}(\boldsymbol{C}) \leq \gamma_2/\gamma_1.\end{gathered} \tag{6.47}$$

PROOF From the courses of linear algebra it is known that the eigenvalues of a self-adjoint operator can be obtained in the form of the Rayleigh-Ritz

quotients (see, e.g., [HJ85, Section 4.2]):

$$\lambda_{\min}(\boldsymbol{C}) = \min_{\boldsymbol{x}\in\mathbb{L},\ \boldsymbol{x}\neq\boldsymbol{0}} \frac{[\boldsymbol{C}\boldsymbol{x},\boldsymbol{x}]_{\boldsymbol{P}}}{[\boldsymbol{x},\boldsymbol{x}]_{\boldsymbol{P}}} = \min_{\boldsymbol{x}\in\mathbb{L},\ \boldsymbol{x}\neq\boldsymbol{0}} \frac{(\boldsymbol{A}\boldsymbol{x},\boldsymbol{x})}{(\boldsymbol{P}\boldsymbol{x},\boldsymbol{x})},$$
$$\lambda_{\max}(\boldsymbol{C}) = \max_{\boldsymbol{x}\in\mathbb{L},\ \boldsymbol{x}\neq\boldsymbol{0}} \frac{[\boldsymbol{C}\boldsymbol{x},\boldsymbol{x}]_{\boldsymbol{P}}}{[\boldsymbol{x},\boldsymbol{x}]_{\boldsymbol{P}}} = \max_{\boldsymbol{x}\in\mathbb{L},\ \boldsymbol{x}\neq\boldsymbol{0}} \frac{(\boldsymbol{A}\boldsymbol{x},\boldsymbol{x})}{(\boldsymbol{P}\boldsymbol{x},\boldsymbol{x})}.$$

By virtue of (6.46), these relations immediately yield (6.47). □

The operators $\boldsymbol{A}$ and $\boldsymbol{P}$ that satisfy inequalities (6.46) are called equivalent by spectrum or equivalent by energy, with the equivalence constants γ_1 and γ_2.

Let us emphasize that the transition from system (6.44) to system (6.45) is only justified if

$$\mu_{\boldsymbol{P}}(\boldsymbol{C}) \leq \frac{\gamma_2}{\gamma_1} \ll \mu(\boldsymbol{A}).$$

Indeed, since in this case the condition number of the transformed system becomes much smaller than that of the original system, the convergence of the iteration noticeably speeds up. In other words, the rate of decay of the error $\boldsymbol{\varepsilon}^{(p)} = \boldsymbol{x} - \boldsymbol{x}^{(p)}$ in the norm $\|\cdot\|_{\boldsymbol{P}}$ increases substantially compared to (6.31):

$$\|\boldsymbol{\varepsilon}^{(p)}\|_{\boldsymbol{P}} = \|\boldsymbol{x}-\boldsymbol{x}^{(p)}\|_{\boldsymbol{P}} \leq \rho_{\boldsymbol{P}}^p \|\boldsymbol{x}-\boldsymbol{x}^{(0)}\|_{\boldsymbol{P}} = \rho_{\boldsymbol{P}}^p \|\boldsymbol{\varepsilon}^{(0)}\|_{\boldsymbol{P}},$$
$$\rho_{\boldsymbol{P}} = \frac{\mu_{\boldsymbol{P}}(\boldsymbol{C})-1}{\mu_{\boldsymbol{P}}(\boldsymbol{C})+1}.$$

The mechanism of the increase is the drop $\rho_{\boldsymbol{P}} \ll \rho$ that takes place because $\rho = \dfrac{\mu(\boldsymbol{A})-1}{\mu(\boldsymbol{A})+1}$ according to formula (6.33), and $\mu_{\boldsymbol{P}}(\boldsymbol{C}) \ll \mu(\boldsymbol{A})$. Consequently, for one and the same value of p we will have $\rho_{\boldsymbol{P}}^p \ll \rho^p$.

A typical situation when preconditioners of type (6.46) prove efficient arises in the context of discrete approximations for elliptic boundary value problems. It was first identified and studied by D'yakonov in the beginning of the sixties; the key results have then been summarized in a later monograph [D'y96].

Consider a system of linear algebraic equations:

$$\boldsymbol{A}_n\boldsymbol{x} = \boldsymbol{f}, \quad \boldsymbol{f}\in\mathbb{R}^n, \quad \boldsymbol{x}\in\mathbb{R}^n,$$

obtained as a discrete approximation of an elliptic boundary value problem. For example, it may be a system of finite-difference equations introduced in Section 5.1.3. In doing so, the better the operator $\boldsymbol{A}_n$ approximates the original elliptic differential operator, the higher the dimension n of the space $\mathbb{R}^n$ is. As such, we are effectively dealing with a sequence of approximating spaces $\mathbb{R}^n$, $n \longrightarrow \infty$, that we will assume Euclidean with the scalar product $(\boldsymbol{x},\boldsymbol{y})^{(n)}$.

Let $\boldsymbol{A}_n : \mathbb{R}^n \longmapsto \mathbb{R}^n$ be a sequence of operators such that $\boldsymbol{A}_n = \boldsymbol{A}_n^* > 0$, and let $\boldsymbol{A}_n\boldsymbol{x} = \boldsymbol{f}$, where $\boldsymbol{x},\boldsymbol{f}\in\mathbb{R}^n$, be a sequence of systems to be solved in the respective spaces. Suppose that the condition number, $\mu(\boldsymbol{A}_n)$, increases when the dimension

n increases so that $\mu(\boldsymbol{A}_n) \sim n^s$, where $s > 0$ is a constant. Then, according to formulae (6.38), it will take $\mathcal{O}(-n^s \ln \sigma)$ iterations to find the solution $\boldsymbol{x} \in \mathbb{R}^n$ with the guaranteed accuracy $\sigma > 0$.

Next, let $\boldsymbol{P}_n : \mathbb{R}^n \longmapsto \mathbb{R}^n$ be a sequence of operators equivalent to the respective operators $\boldsymbol{A}_n$ by energy, with the equivalence constants γ_1 and γ_2 *that do not depend on* n. Then, for $\boldsymbol{C}_n = \boldsymbol{P}_n^{-1}\boldsymbol{A}_n$ we obtain:

$$\mu_{\boldsymbol{P}_n}(\boldsymbol{C}_n) \leq \gamma_2/\gamma_1 = \text{const.} \tag{6.48}$$

Hence we can replace the original system (6.44): $\boldsymbol{A}_n\boldsymbol{x} = \boldsymbol{f}$ by its equivalent (6.45):

$$\boldsymbol{C}_n\boldsymbol{x} = \boldsymbol{g}, \quad C_n = \boldsymbol{P}_n^{-1}\boldsymbol{A}_n, \quad \boldsymbol{g} = \boldsymbol{P}_n^{-1}\boldsymbol{f}. \tag{6.49}$$

In doing so, because of a uniform boundedness of the condition number with respect to n, see formula (6.48), the number of iterations required for reducing the $\|\cdot\|_{\boldsymbol{P}_n}$ norm of the initial error by a predetermined factor of σ:

$$\|\boldsymbol{x} - \boldsymbol{x}^{(p)}\|_{\boldsymbol{P}_n} \leq \sigma \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\|_{\boldsymbol{P}_n}, \tag{6.50}$$

will not increase when the dimension n increases, and will remain $\mathcal{O}(\ln \sigma)$.

Furthermore, let the norms:

$$\|\boldsymbol{x}\| = \sqrt{(\boldsymbol{x},\boldsymbol{x})^{(n)}} \quad \text{and} \quad \|\boldsymbol{x}\|_{\boldsymbol{P}_n} = \sqrt{[\boldsymbol{x},\boldsymbol{x}]_{\boldsymbol{P}_n}} \equiv \sqrt{(\boldsymbol{P}_n\boldsymbol{x},\boldsymbol{x})^{(n)}}$$

be related to one another via the inequalities:

$$n^{-l}\|\boldsymbol{x}\|_{\boldsymbol{P}_n} \leq \|\boldsymbol{x}\| \leq n^l \|\boldsymbol{x}\|_{\boldsymbol{P}_n}, \quad \text{where} \quad l \geq 0, \;\; l = \text{const.}$$

Then, in order to guarantee the original error estimate (6.37):

$$\|\boldsymbol{x} - \boldsymbol{x}^{(p)}\| \leq \sigma \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\| \tag{6.51}$$

when iterating system (6.49), it is sufficient that the following inequality hold:

$$\|\boldsymbol{x} - \boldsymbol{x}^{(p)}\|_{\boldsymbol{P}_n} \leq \frac{\sigma}{n^l} \|\boldsymbol{x} - \boldsymbol{x}^{(0)}\|_{\boldsymbol{P}_n}. \tag{6.52}$$

Inequality (6.52) is obtained from (6.50) by replacing σ with σn^{-l}, so that solving the preconditioned system (6.49) by the Richardson method will only require $\mathcal{O}(-\ln(\sigma n^{-l})) = \mathcal{O}(\ln n - \ln \sigma)$ iterations, as opposed to $\mathcal{O}(-n^s \ln \sigma)$ iterations required for solving the original non-preconditioned system (6.44).

Of course, the key question remains of how to design a preconditioner equivalent by spectrum to the operator $\boldsymbol{A}$, see (6.46). In the context of elliptic boundary value problems, good results can often be achieved when preconditioning a discretized operator with variable coefficients, such as the one from equation (6.40), with the discretized Laplace operator. On a regular grid, a preconditioner of this type $\boldsymbol{P} = -\Delta^{(h)}$, see Section 5.1.3, can be easily inverted with the help of the FFT, see Section 5.7.3.

Overall, the task of designing an efficient preconditioner is highly problem-dependent. One general approach is based on availability of some a priori knowledge of where the matrix $\boldsymbol{A}$ originates from. A typical example here is the aforementioned spectrally equivalent elliptic preconditioners. Another approach is purely algebraic and only uses the information contained in the structure of a given matrix $\boldsymbol{A}$. Examples include incomplete factorizations ($\boldsymbol{LU}$, Cholesky, modified unperturbed and perturbed incomplete $\boldsymbol{LU}$), polynomial preconditioners (e.g., truncated Neumann series), and various ordering strategies, foremost the multilevel recursive orderings that are conceptually close to the idea of multigrid (Section 6.4). For further detail, we refer the reader to specialized monographs [Axe94, Saa03, vdV03].

6.1.5 Scaling

One reason for a given matrix $\boldsymbol{A}$ to be poorly conditioned, i.e., to have a large condition number $\mu(\boldsymbol{A})$, may be large disparity in the magnitudes of its entries. If this is the case, then scaling the rows of the matrix so that the largest magnitude among the entries in each row becomes equal to one often helps improve the conditioning.

Let $\boldsymbol{D}$ be a non-singular diagonal matrix ($d_{ii} \neq 0$), and instead of the original system $\boldsymbol{Ax} = \boldsymbol{f}$ let us consider its equivalent:

$$\boldsymbol{DAx} = \boldsymbol{Df}. \tag{6.53}$$

The entries d_{ii}, $i = 1, 2, \ldots, n$ of the matrix $\boldsymbol{D}$ are to be chosen so that the maximum absolute value of the entry in each row of the matrix $\boldsymbol{DA}$ be equal to one.

$$\max_j |d_{ii} a_{ij}| = 1, \quad i = 1, 2, \ldots, n. \tag{6.54}$$

The transition from the matrix $\boldsymbol{A}$ to the matrix $\boldsymbol{DA}$ is known as *scaling* (of the rows of $\boldsymbol{A}$). By comparing equations (6.53) and (6.41) we conclude that it can be interpreted as a particular approach to preconditioning with $\boldsymbol{P}^{-1} = \boldsymbol{D}$. Note that different strategies of scaling can be employed; instead of (6.54) we can require, for example, that all diagonal entries of $\boldsymbol{DA}$ have the same magnitude.

Scaling typically reduces the condition number of a system: $\mu(\boldsymbol{DA}) < \mu(\boldsymbol{A})$. To solve the system $\boldsymbol{Ax} = \boldsymbol{f}$ by iterations, we can first transform it to an equivalent system $\boldsymbol{Cx} = \boldsymbol{g}$ with a self-adjoint positive definite matrix $\boldsymbol{C} = \boldsymbol{A}^*\boldsymbol{A}$ and the right-hand side $\boldsymbol{g} = \boldsymbol{A}^*\boldsymbol{f}$, and then apply the Richardson method. According to Theorem 6.3, the rate of convergence of the Richardson iteration will be determined by the condition number $\mu(\boldsymbol{C})$. It is possible to show that for the Euclidean condition numbers we have: $\mu(\boldsymbol{C}) = \mu^2(\boldsymbol{A})$ (see Exercise 7 after Section 5.3). If the matrix $\boldsymbol{A}$ is scaled ahead of time, see formula (6.53), then the convergence of the iterations will be faster, because $\mu((\boldsymbol{DA})^*(\boldsymbol{DA})) = \mu^2(\boldsymbol{DA}) < \mu^2(\boldsymbol{A})$.

Note that the transition from a given $\boldsymbol{A}$ to $\boldsymbol{C} = \boldsymbol{A}^*\boldsymbol{A}$ is almost never used in practice as a means of enabling the solution by iterations that require a self-adjoint matrix, because an additional matrix multiplication may eventually lead to large errors when computing with finite precision. Therefore, the foregoing example shall only be

regarded as a simple theoretical illustration. However, scaling can also help when solving the system $\boldsymbol{Ax} = \boldsymbol{f}$ by a direct method rather than by iterations. For a matrix $\boldsymbol{A}$ with large disparity in the magnitudes of entries it may improve stability of the Gaussian elimination algorithm (Section 5.4). Besides, the system $\boldsymbol{Ax} = \boldsymbol{f}$ with a general matrix $\boldsymbol{A}$ can be solved by an iterative method that does not require a self-adjoint matrix, e.g., by a Krylov subspace iteration (see Section 6.3). In this case, scaling may be very helpful in reducing the condition number $\mu(\boldsymbol{A})$.

Exercises

1. Assume that the eigenvalues of the operator $\boldsymbol{A} : \mathbb{R}^{100} \longmapsto \mathbb{R}^{100}$ are known:

$$\lambda_k = k^2, \quad k = 1, 2, \ldots, 100. \tag{6.55}$$

   The system $\boldsymbol{Ax} = \boldsymbol{f}$ is to be solved by the non-stationary Richardson iterative method:

$$\boldsymbol{x}^{(p+1)} = (\boldsymbol{I} - \tau_p \boldsymbol{A})\boldsymbol{x}^{(p)} + \tau_p \boldsymbol{f}, \quad p = 0, 1, 2, \ldots, \tag{6.56}$$

   where τ_p, $p = 0, 1, 2, \ldots$, are some positive parameters.

   Find a particular set of parameters $\{\tau_0, \tau_1, \ldots, \tau_{99}\}$ that would guarantee $\boldsymbol{x}^{(100)} = \boldsymbol{x}$, where $\boldsymbol{x}$ is the exact solution of the system $\boldsymbol{Ax} = \boldsymbol{f}$.

   Hint. First make sure that $\boldsymbol{x} - \boldsymbol{x}^{(p+1)} \equiv \boldsymbol{\varepsilon}^{(p+1)} = (\boldsymbol{I} - \tau_p \boldsymbol{A})\boldsymbol{\varepsilon}^{(p)} \equiv (\boldsymbol{I} - \tau_p \boldsymbol{A})(\boldsymbol{x} - \boldsymbol{x}^{(p)})$, $p = 0, 1, 2, \ldots$ Then expand the initial error:

$$\boldsymbol{\varepsilon}^{(0)} = \varepsilon_1^{(0)} \boldsymbol{e}_1 + \varepsilon_2^{(0)} \boldsymbol{e}_2 + \ldots + \varepsilon_{100}^{(0)} \boldsymbol{e}_{100}, \tag{6.57}$$

   where $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{100}$ are the eigenvectors of $\boldsymbol{A}$ that correspond to the eigenvalues (6.55). Finally, as the eigenvalues (6.55) are given explicitly, choose the iteration parameters $\{\tau_0, \tau_1, \ldots, \tau_{99}\}$ in such a way that each iteration will eliminate precisely one term from the expansion of the error (6.57).

2. Let the iteration parameters in Exercise 1 be chosen as follows:

$$\tau_p = \frac{1}{(p+1)^2}, \quad p = 0, 1, 2, \ldots, 99. \tag{6.58}$$

   a) Show that in this case $\boldsymbol{x}^{(100)} = \boldsymbol{x}$.

   b) Implementation of algorithm (6.56) with the iteration parameters (6.58) on a real computer encounters a critical obstacle. Very large numbers are generated in the course of computation; they ruin the accuracy and make the computation practically impossible. Explain the mechanism of the foregoing phenomenon.

      Hint. Take expansion (6.57) and operate on it with the matrices $(\boldsymbol{I} - \tau_p \boldsymbol{A})$, where τ_p are chosen according to (6.58). Components of the error with the indexes close to 100 become excessively large before they get canceled. Cancellation of a given component means that a very large number is subtracted from the current iterate $\boldsymbol{x}^{(p)}$ to generate the next iterate $\boldsymbol{x}^{(p+1)}$ and, eventually, the solution $\boldsymbol{x}$. This leads to the loss of significant digits and ruins the accuracy of the solution.

3. Let the iteration parameters in Exercise 1 be chosen as follows:

$$\tau_p = \frac{1}{(100-p)^2}, \quad p = 0, 1, 2, \ldots, 99. \tag{6.59}$$

a) Show that in this case also $\boldsymbol{x}^{(100)} = \boldsymbol{x}$.

b) Implementation of algorithm (6.56) with the iteration parameters (6.59) on a real computer encounters another critical obstacle. Small round-off errors rapidly increase and destroy the overall accuracy. This, again, makes the computation practically impossible. Explain the mechanism of the aforementioned phenomenon.
Hint. When expansion (6.57) is operated on by the matrices $(\boldsymbol{I} - \tau_p \boldsymbol{A})$ with τ_p of (6.59), components of the error with large indexes are canceled first. The cancellation, however, is not exact, its accuracy is determined by the machine precision. Show that the corresponding round-off errors will subsequently grow.

6.2 Chebyshev Iterations and Conjugate Gradients

For the linear system:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}, \quad \boldsymbol{A} = \boldsymbol{A}^* > 0, \quad \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{f} \in \mathbb{R}^n, \tag{6.60}$$

we will describe two iterative methods of solution that offer a better performance (faster convergence) compared to the Richardson method of Section 6.1. We will also discuss the conditions that may justify preferring one of these methods over the other. The two methods are known as the Chebyshev iterative method and the method of conjugate gradients, both are described in detail, e.g., in [SN89b].

As we require that $\boldsymbol{A} = \boldsymbol{A}^* > 0$, all eigenvalues λ_j, $j = 1, 2, \ldots, n$, of the operator $\boldsymbol{A}$ are strictly positive. With no loss of generality, we will assume that they are arranged in the ascending order. We will also assume that two numbers $a > 0$ and $b > 0$ are known such that:

$$0 < a \leq \lambda_1 \leq \ldots \leq \lambda_n \leq b. \tag{6.61}$$

The two numbers a and b in formula (6.61) are called boundaries of the spectrum of the operator $\boldsymbol{A}$. If $a = \lambda_1$ and $b = \lambda_n$ these boundaries are referred to as sharp. As in Section 6.1.3, we will also introduce their ratio:

$$\xi = \frac{a}{b} < 1.$$

If the boundaries of the spectrum are sharp, then clearly $\xi = \mu(\boldsymbol{A})^{-1}$, where $\mu(\boldsymbol{A})$ is the Euclidean condition number of $\boldsymbol{A}$ (Theorem 5.3).

6.2.1 Chebyshev Iterations

Let us specify the initial guess $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ arbitrarily, and let us then compute the iterates $\boldsymbol{x}^{(p)}$, $p = 1, 2, \ldots$, according to the following formulae:

$$\begin{gathered} \boldsymbol{x}^{(1)} = (\boldsymbol{I} - \tau \boldsymbol{A})\boldsymbol{x}^{(0)} + \tau \boldsymbol{f}, \\ \boldsymbol{x}^{(p+1)} = \alpha_{p+1}(\boldsymbol{I} - \tau \boldsymbol{A})\boldsymbol{x}^{(p)} + (1 - \alpha_{p+1})\boldsymbol{x}^{(p-1)} + \tau \alpha_{p+1} \boldsymbol{f}, \\ p = 1, 2, \ldots, \end{gathered} \tag{6.62a}$$