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:

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

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:

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

where the parameters τ and α_p are given by:

$$\tau = \frac{2}{a+b}, \quad \alpha_1 = 2, \quad \alpha_{p+1} = \frac{4}{4-\rho_0^2\alpha_p}, \quad \rho_0 = \frac{1-\xi}{1+\xi}, \quad p = 1,2,\ldots \tag{6.62b}$$

We see that the first iteration of (6.62a) coincides with that of the Richardson method, and altogether formulae (6.62) describe a second order linear non-stationary iteration scheme. It is possible to show that the error $\boldsymbol{\varepsilon}^{(p)} = \boldsymbol{x} - \boldsymbol{x}^{(p)}$ of the iterate $\boldsymbol{x}^{(p)}$ satisfies the estimate:

$$\|\boldsymbol{\varepsilon}^{(p)}\| \le \frac{2\rho_1^p}{1+\rho_1^{2p}}\|\boldsymbol{\varepsilon}^{(0)}\|, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \quad p = 1,2,\ldots \tag{6.63}$$

where $\|\cdot\| = \sqrt{(\cdot,\cdot)}$ is a Euclidean norm on $\mathbb{R}^n$. Based on formula (6.63), the analysis very similar to the one performed in the end of Section 6.1.3 will lead us to the conclusion that in order to reduce the norm of the initial error by a predetermined factor σ, i.e., in order to guarantee the following estimate:

$$\|\boldsymbol{\varepsilon}^{(p)}\| \le \sigma\|\boldsymbol{\varepsilon}^{(0)}\|, \tag{6.64}$$

it is sufficient to choose the number of iterations p so that:

$$p \ge -\frac{1}{2}\left(\ln\frac{\sigma}{2}\right)\frac{1}{\sqrt{\xi}} = -\frac{1}{2}\left(\ln\frac{\sigma}{2}\right)\sqrt{\frac{b}{a}}. \tag{6.65}$$

This number is about $\sqrt{\frac{b}{a}}$ times smaller than the number of iterations required for achieving the same error estimate (6.64) using the Richardson method. In particular, when the sharp boundaries of the spectrum are available, we have:

$$\frac{b}{a} = \frac{1}{\xi} = \mu(\boldsymbol{A}).$$

Then, by comparing formulae (6.38a) and (6.65) we can see that whereas for the Richardson method the number of iterations p is proportional to the condition number $\mu(\boldsymbol{A})$ itself, for the new method (6.62) it is only proportional to the square root $\sqrt{\mu(\boldsymbol{A})}$. Therefore, the larger the condition number, the more substantial is the relative economy offered by the new approach. It is also important to mention that the iterative scheme (6.62) is computationally stable.

Construction of the iterative method (6.62) per se, as well as the proof of its computational stability and other properties, including error estimate (6.63), are based on the analysis of Chebyshev polynomials (see Section 3.2.3), and that of some related polynomials. We omit the corresponding discussion here and rather refer the reader, e.g., to [SN89b] for detail. We only mention that because of its relation to Chebyshev polynomials, iterative method (6.62) is often referred to as the second order Chebyshev method in the literature.

Let us also note that along with the second order Chebyshev method (6.62), there is also a first order method of a similar design and with similar properties. It is