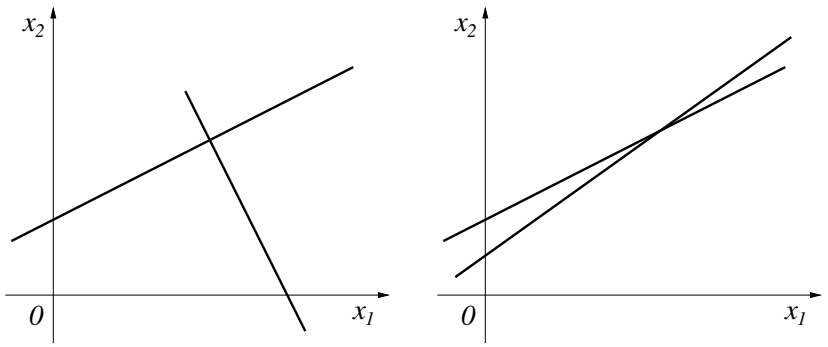
5.3 Conditioning of Linear Systems

Two linear systems that look quite similar at the first glance, may, in fact, have a very different degree of sensitivity of their solutions to the errors committed when specifying the input data. This phenomenon can be observed already for the systems $\boldsymbol{Ax} = \boldsymbol{f}$ of order two:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= f_1, \\ a_{21}x_1 + a_{22}x_2 &= f_2. \end{aligned} \tag{5.30}$$

With no loss of generality we will assume that the coefficients of system (5.30) are normalized: $a_{i1}^2 + a_{i2}^2 = 1$, $i = 1, 2$. Geometrically, each individual equation of system (5.30) defines a straight line on the Cartesian plane (x_1, x_2). Accordingly, the solution of system (5.30) can be interpreted as the intersection point of these two lines as shown in Figure 5.2.



(a) Weak sensitivity. (b) Strong sensitivity

FIGURE 5.2: Sensitivity of the solution of system (5.30) to perturbations of the data.

Let us qualitatively analyze two opposite situations. The straight lines that correspond to linear equations (5.30) can intersect "almost normally," as shown in Figure 5.2(a), or they can intersect "almost tangentially," as shown in Figure 5.2(b). If we slightly perturb the input data, i.e., the right-hand sides f_i, $i = 1, 2$, and/or the coefficients a_{ij}, $i, j = 1, 2$, then each line may move parallel to itself and/or tilt. In doing so, it is clear that the intersection point (i.e., the solution) on Figure 5.2(a) will only move slightly, whereas the intersection point on Figure 5.2(b) will move much more visibly. Accordingly, one can say that in the case of Figure 5.2(a) the sensitivity of the solution to perturbations of the input data is weak, whereas in the case of Figure 5.2(b) it is strong.

Quantitatively, the sensitivity of the solution to perturbations of the input data can be characterized with the help of the so-called *condition number* $\mu(\boldsymbol{A})$. We will later

see that not only does the condition number determine the aforementioned sensitivity, but also that it directly influences the performance of iterative methods for solving $\boldsymbol{Ax} = \boldsymbol{f}$. Namely, it affects the number of iterations and as such, the number of arithmetic operations, required for finding an approximate solution of $\boldsymbol{Ax} = \boldsymbol{f}$ within a prescribed tolerance, see Chapter 6.

5.3.1 Condition Number

The condition number of a linear operator $\boldsymbol{A}$ acting on a normed vector space $\mathbb{L}$ is defined as follows:

$$\mu(\boldsymbol{A}) = \|\boldsymbol{A}\| \cdot \|\boldsymbol{A}^{-1}\|. \tag{5.31}$$

The same quantity $\mu(\boldsymbol{A})$ given by formula (5.31) is also referred to as the condition number of a linear system $\boldsymbol{Ax} = \boldsymbol{f}$.

Recall that we have previously identified every matrix:

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$$

with a linear operator acting on an n-dimensional vector space $\mathbb{L}$ of the elements $\boldsymbol{x} = \begin{bmatrix} x_1 \\ \dots \\ x_n \end{bmatrix}$ (for example, $\mathbb{L} = \mathbb{R}^n$ or $\mathbb{L} = \mathbb{C}^n$). For a given $\boldsymbol{x} \in \mathbb{L}$, the operator yields $\boldsymbol{y} = \boldsymbol{Ax}$, where $\boldsymbol{y} = \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix}$ is computed as follows:

$$y_i = \sum_{j=1}^{n} a_{ij} x_j, \quad i = 1, 2, \dots, n.$$

Accordingly, the definition of the condition number $\mu(\boldsymbol{A})$ by formula (5.31) also makes sense for matrices, so that one can refer to the condition number of a matrix $\boldsymbol{A}$, as well as to the condition number of a system of linear algebraic equations specified in its canonical form (5.1) rather than only in the operator form.

The norms $\|\boldsymbol{A}\|$ and $\|\boldsymbol{A}^{-1}\|$ of the direct and inverse operators, respectively, in formula (5.31) are assumed induced by the vector norm chosen in $\mathbb{L}$. Consequently, the condition number $\mu(\boldsymbol{A})$ also depends on the choice of the norm in $\mathbb{L}$. If $\boldsymbol{A}$ is a matrix, and we use the maximum norm for the vectors from $\mathbb{L}$, $\|\boldsymbol{x}\|_\infty = \max_j |x_j|$, then we write $\mu = \mu_\infty(\boldsymbol{A})$; if we employ the first norm $\|\boldsymbol{x}\|_1 = \sum_j |x_j|$, then $\mu = \mu_1(\boldsymbol{A})$.

If $\mathbb{L}$ is a Euclidean (unitary) space with the scalar product $(\boldsymbol{x},\boldsymbol{y})$ given, for example, by formula (5.15) or (5.16), and the corresponding Euclidean (Hermitian) norm is defined by formula (5.17): $\|\boldsymbol{x}\|_2 = (\boldsymbol{x},\boldsymbol{x})^{1/2}$, then $\mu = \mu_2(\boldsymbol{A})$. If an alternative scalar product $[\boldsymbol{x},\boldsymbol{y}]_{\boldsymbol{B}} = (\boldsymbol{Bx},\boldsymbol{y})$ is introduced in $\mathbb{L}$ based on the original product $(\boldsymbol{x},\boldsymbol{y})$ and on the operator $\boldsymbol{B} = \boldsymbol{B}^* > 0$, see formula (5.22), and if a new norm is set up accordingly by formula (5.23): $\|\boldsymbol{x}\|_{\boldsymbol{B}} = ([\boldsymbol{x},\boldsymbol{y}]_{\boldsymbol{B}})^{1/2}$, then the corresponding condition number is denoted by $\mu = \mu_{\boldsymbol{B}}(\boldsymbol{A})$.

Let us now explain the geometric meaning of the condition number $\mu(\boldsymbol{A})$. To do so, consider the set $S \subset \mathbb{L}$ of all vectors with the norm equal to one, i.e., a unit sphere in the space $\mathbb{L}$. Among these vectors choose a particular two, $\boldsymbol{x}_{\max}$ and $\boldsymbol{x}_{\min}$, that satisfy the following equalities:

$$\|\boldsymbol{A}\boldsymbol{x}_{\max}\| = \max_{\boldsymbol{x}\in S} \|\boldsymbol{A}\boldsymbol{x}\| \quad \text{and} \quad \|\boldsymbol{A}\boldsymbol{x}_{\min}\| = \min_{\boldsymbol{x}\in S} \|\boldsymbol{A}\boldsymbol{x}\|.$$

It is easy to see that

$$\|\boldsymbol{A}\| = \|\boldsymbol{A}\boldsymbol{x}_{\max}\| \quad \text{and} \quad \|\boldsymbol{A}^{-1}\| = \frac{1}{\|\boldsymbol{A}\boldsymbol{x}_{\min}\|}.$$

Indeed, according to formula (5.24) we can write:

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x}\in\mathbb{L},\ \boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{x}\in\mathbb{L},\ \boldsymbol{x}\neq\boldsymbol{0}} \left\|\boldsymbol{A}\frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}\right\| = \max_{\tilde{\boldsymbol{x}}\in S} \|\boldsymbol{A}\tilde{\boldsymbol{x}}\| = \|\boldsymbol{A}\boldsymbol{x}_{\max}\|.$$

Likewise:

$$\|\boldsymbol{A}^{-1}\| = \max_{\boldsymbol{x}\in\mathbb{L},\ \boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}^{-1}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\tilde{\boldsymbol{x}}\in\mathbb{L},\ \tilde{\boldsymbol{x}}\neq\boldsymbol{0}} \frac{\|\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{A}\tilde{\boldsymbol{x}}\|} = \left[\min_{\tilde{\boldsymbol{x}}\in\mathbb{L},\ \tilde{\boldsymbol{x}}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\tilde{\boldsymbol{x}}\|}{\|\tilde{\boldsymbol{x}}\|}\right]^{-1} = \frac{1}{\|\boldsymbol{A}\boldsymbol{x}_{\min}\|}.$$

Substituting these expressions into the definition of $\mu(\boldsymbol{A})$ by means of formula (5.31), we obtain:

$$\mu(\boldsymbol{A}) = \frac{\max_{\boldsymbol{x}\in S} \|\boldsymbol{A}\boldsymbol{x}\|}{\min_{\boldsymbol{x}\in S} \|\boldsymbol{A}\boldsymbol{x}\|}. \tag{5.32}$$

This, in particular, implies that we always have:

$$\mu(\boldsymbol{A}) \geq 1.$$

According to formula (5.32), the condition number $\mu(\boldsymbol{A})$ is the ratio of magnitudes of the maximally stretched unit vector to the minimally stretched (i.e., maximally shrunk) unit vector. If the operator $\boldsymbol{A}$ is singular, i.e., if the inverse operator $\boldsymbol{A}^{-1}$ does not exist, then we formally set $\mu(\boldsymbol{A}) = \infty$.

The geometric meaning of the quantity $\mu(\boldsymbol{A})$ becomes particularly apparent in the case of an Euclidean space $\mathbb{L} = \mathbb{R}^2$ equipped with the norm $\|\boldsymbol{x}\|_2 = (\boldsymbol{x},\boldsymbol{x})^{1/2} = \sqrt{x_1^2 + x_2^2}$, and a self-adjoint operator $\boldsymbol{A} = \boldsymbol{A}^*$. In other words, let $\mathbb{L}$ be the Euclidean plane of the variables x_1 and x_2. In this case, S is a conventional unit circle: $x_1^2 + x_2^2 = 1$. A linear mapping by means of the operator $\boldsymbol{A}$ obviously transforms this circle into an ellipse. Formula (5.32) then implies that the condition number $\mu(\boldsymbol{A})$ is the ratio of the large semiaxis of this ellipse to its small semiaxis.

THEOREM 5.3

Let $\boldsymbol{A} = \boldsymbol{A}^*$ be an operator on the vector space $\mathbb{L}$ self-adjoint in the sense of a given scalar product $[\boldsymbol{x},\boldsymbol{y}]_B$. Then,

$$\mu_B(\boldsymbol{A}) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}, \tag{5.33}$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the eigenvalues of $\boldsymbol{A}$ with the largest and smallest moduli, respectively.

PROOF Let $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ be an orthonormal basis in the n-dimensional space $\mathbb{L}$ composed of the eigenvectors of $\boldsymbol{A}$. Orthonormality of the basis is understood in the sense of the scalar product $[\boldsymbol{x}, \boldsymbol{y}]_B$. Let λ_j be the corresponding eigenvalues: $\boldsymbol{A}\boldsymbol{e}_j = \lambda_j \boldsymbol{e}_j$, $j = 1, 2, \ldots, n$, which are all real. Also assume with no loss of generality that the eigenvalues are arranged in the descending order: $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$.

An arbitrary vector $\boldsymbol{x} \in \mathbb{L}$ can be expanded with respect to this basis:

$$\boldsymbol{x} = x_1 \boldsymbol{e}_1 + x_2 \boldsymbol{e}_2 + \ldots + x_n \boldsymbol{e}_n.$$

In doing so,

$$\boldsymbol{A}\boldsymbol{x} = \lambda_1 x_1 \boldsymbol{e}_1 + \lambda_2 x_2 \boldsymbol{e}_2 + \ldots + \lambda_n x_n \boldsymbol{e}_n,$$

and because of the orthonormality of the basis:

$$\|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{B}} = (|\lambda_1 x_1|^2 + |\lambda_2 x_2|^2 + \ldots + |\lambda_n x_n|^2)^{1/2}.$$

Consequently, $\|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{B}} \leq |\lambda_1| \|\boldsymbol{x}\|_{\boldsymbol{B}}$, and if $\|\boldsymbol{x}\|_{\boldsymbol{B}} = 1$, i.e., if $\boldsymbol{x} \in S$, then $\|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{B}} \leq |\lambda_1|$. At the same time, for $\boldsymbol{x} = \boldsymbol{e}_1 \in S$ we have $\|\boldsymbol{A}\boldsymbol{e}_1\|_{\boldsymbol{B}} = |\lambda_1|$. Therefore,

$$\max_{\boldsymbol{x} \in S} \|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{B}} = |\lambda_1| = |\lambda_{\max}|.$$

A similar argument immediately yields:

$$\min_{\boldsymbol{x} \in S} \|\boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{B}} = |\lambda_n| = |\lambda_{\min}|.$$

Then, formula (5.32) implies (5.33). □

Note that similarly to Theorem 5.2, the result of Theorem 5.3 can also be generalized to the case of normal matrices $\boldsymbol{A}\boldsymbol{A}^* = \boldsymbol{A}^*\boldsymbol{A}$.

5.3.2 Characterization of a Linear System by Means of Its Condition Number

THEOREM 5.4

Let $\mathbb{L}$ be a normed vector space (e.g., $\mathbb{L} = \mathbb{R}^n$ or $\mathbb{L} = \mathbb{C}^n$), and let $\boldsymbol{A} : \mathbb{L} \longmapsto \mathbb{L}$ be a non-singular linear operator. Consider a system of linear algebraic equations:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}, \tag{5.34}$$

where $\boldsymbol{x} \in \mathbb{L}$ is the vector of unknowns and $\boldsymbol{f} \in \mathbb{L}$ is a given right-hand side. Let $\Delta\boldsymbol{f} \in \mathbb{L}$ be a perturbation of the right-hand side that leads to the perturbation $\Delta\boldsymbol{x} \in \mathbb{L}$ of the solution so that:

$$\boldsymbol{A}(\boldsymbol{x}+\Delta\boldsymbol{x}) = \boldsymbol{f}+\Delta\boldsymbol{f}. \tag{5.35}$$

Then the relative error of the solution $\|\Delta\boldsymbol{x}\|/\|\boldsymbol{x}\|$ satisfies the inequality:

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le \mu(\boldsymbol{A})\frac{\|\Delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|}, \tag{5.36}$$

where $\mu(\boldsymbol{A})$ is the condition number of the operator $\boldsymbol{A}$, see formula (5.31). Moreover, there are particular $\boldsymbol{f} \in \mathbb{L}$ and $\Delta\boldsymbol{f} \in \mathbb{L}$ for which (5.36) transforms into a precise equality.

PROOF Formulae (5.34) and (5.35) imply that $\boldsymbol{A}\Delta\boldsymbol{x} = \Delta\boldsymbol{f}$, and consequently, $\Delta\boldsymbol{x} = \boldsymbol{A}^{-1}\Delta\boldsymbol{f}$. Let us also employ the expression $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{f}$ that defines the original system itself. Then,

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \frac{\|\boldsymbol{A}^{-1}\Delta\boldsymbol{f}\|}{\|\boldsymbol{x}\|} = \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\frac{\|\boldsymbol{A}^{-1}\Delta\boldsymbol{f}\|}{\|\Delta\boldsymbol{f}\|}\frac{\|\Delta\boldsymbol{f}\|}{\|\boldsymbol{A}\boldsymbol{x}\|} = \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\frac{\|\boldsymbol{A}^{-1}\Delta\boldsymbol{f}\|}{\|\Delta\boldsymbol{f}\|}\frac{\|\Delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|}. \tag{5.37}$$

According to the definition of the operator norm, see formula (5.24), we have:

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le \|\boldsymbol{A}\| \quad \text{and} \quad \frac{\|\boldsymbol{A}^{-1}\Delta\boldsymbol{f}\|}{\|\Delta\boldsymbol{f}\|} \le \|\boldsymbol{A}^{-1}\|. \tag{5.38}$$

Combining formulae (5.37) and (5.38), we obtain for any $\boldsymbol{f} \in \mathbb{L}$ and $\Delta\boldsymbol{f} \in \mathbb{L}$:

$$\frac{\|\Delta\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \le \|\boldsymbol{A}\|\|\boldsymbol{A}^{-1}\|\frac{\|\Delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|} = \mu(\boldsymbol{A})\frac{\|\Delta\boldsymbol{f}\|}{\|\boldsymbol{f}\|}, \tag{5.39}$$

which means that inequality (5.36) holds.

Furthermore, if $\Delta\boldsymbol{f}$ is the specific vector from the space $\mathbb{L}$ for which

$$\frac{\|\boldsymbol{A}^{-1}\Delta\boldsymbol{f}\|}{\|\Delta\boldsymbol{f}\|} = \|\boldsymbol{A}^{-1}\|,$$

and $\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}$ is the element of $\mathbb{L}$ for which

$$\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \|\boldsymbol{A}\|,$$

then expression (5.37) coincides with inequalities (5.39) and (5.36) that transform into precise equalities for these particular $\boldsymbol{f}$ and $\Delta\boldsymbol{f}$. □