# Chapter 2
# Iterative Methods

**Abstract** In this chapter we consider general properties of iterative methods. Such properties are *consistency*, ensuring the connection between the iterative method and the given system of equations, as well as *convergence*, guaranteeing the success of the iteration. The most important result of this chapter is the characterisation of the convergence of linear iterations by the spectral radius of the iteration matrix (cf. §2.1.4). Since we only consider iterative methods for systems with regular matrices, iterative methods for singular systems or those with rectangular matrices will not be studied.[1] The quality of a linear iteration depends on both the cost and the convergence speed. The resulting efficacy is discussed in Section 2.3. Finally, Section 2.4 explains how to test iterative methods numerically.

## 2.1 Consistency and Convergence

### *2.1.1 Notation*

We want to solve the *system of linear equations*

$$Ax = b \qquad (A \in \mathbb{K}^{I\times I} \text{ and } b \in \mathbb{K}^I \text{ given}) \tag{2.1}$$

(cf. (1.10)). To guarantee solvability for all $b \in \mathbb{K}^I$, we generally assume:

$$A \text{ is regular.} \tag{2.2}$$

An iterative method producing iterates $x^1, x^2, \ldots$ from the starting value $x^0$ can be characterised by a prescription $x^{m+1} := \Phi(x^m)$. $\Phi$ depends on the data $A$ and $b$ in (2.1). These parameters are explicitly expressed by the notation

[1] Concerning this topic, we refer, e.g., to Björck [47], Marek [275], Kosmol–Zhou [241], Berman–Plemmons [46], and Remark 5.17.



W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*,
Applied Mathematical Sciences 95, DOI 10.1007/978-3-319-28483-5_2

$$x^{m+1} := \Phi(x^m, b, A) \qquad (m \geq 0,\ b \text{ in (2.1)}). \tag{2.3}$$

Since in most of the cases the matrix $A$ is fixed, we usually write

$$x^{m+1} := \Phi(x^m, b)$$

instead of $\Phi(x^m, b, A)$. By $\Phi(\cdot, \cdot, A)$ we express the fact that we consider the iteration (2.3) exclusively for the matrix $A$.

**Definition 2.1.** An *iterative method* is a (in general nonlinear) mapping

$$\Phi : \mathbb{K}^I \times \mathbb{K}^I \times \mathbb{K}^{I\times I} \to \mathbb{K}^I.$$

By $x^m = x^m(x^0, b, A)$ we denote the *iterates* of the sequence generated by the prescription (2.3) with a starting value $x^0 = y \in \mathbb{K}^I$:

$$\begin{aligned} x^0(y, b, A) &:= y\,, \\ x^{m+1}(y, b, A) &:= \Phi(x^m(y, b, A), b, A) \quad \text{for } m \geq 0. \end{aligned} \tag{2.4}$$

If $A$ is fixed, we write $x^m(y, b)$ instead of $x^m(y, b, A)$. If all parameters $y, b, A$ are fixed, we write $x^m$.

If $\Phi$ is called an *iteration method*, we expect that the method is applicable to a whole class of matrices $A$. Here 'applicable' means that $\Phi$ is well defined (including the case that the sequence $x^m$ diverges).

**Definition 2.2.** (a) $\mathfrak{D}(\Phi) := \{A : \Phi(\cdot, \cdot, A) \text{ well defined}\}$ is the *domain of* $\Phi$.
(b) An iteration is called *algebraic* if the definition of $\Phi(\cdot, \cdot, A)$ can be based exclusively on the data of $A \in \mathfrak{D}(\Phi)$.

In the case of the Gauss–Seidel iteration $\Phi^{\mathrm{GS}}$ in (1.15), the domain is defined by $\mathfrak{D}(\Phi^{\mathrm{GS}}) = \{A \in \mathbb{K}^{I\times I} : a_{ii} \neq 0 \text{ for all } i \in I,\ I \text{ finite}\}$. Another extreme case is $\mathfrak{D}(\Phi) = \{A\}$, i.e., the iteration can only be applied to one particular matrix $A$.

### *2.1.2 Fixed Points*

**Definition 2.3.** $x^* = x^*(b, A)$ is called a fixed point of the iteration $\Phi$ corresponding to $b \in \mathbb{K}^I$ and $A \in \mathfrak{D}(\Phi)$ (or shortly: a fixed point of $\Phi(\cdot, b, A)$) if

$$x^* = \Phi(x^*, b, A).$$

If the sequence $\{x^m\}$ of the iterates generated by (2.3) converges, we may form the limit in (2.3) and obtain the next lemma.

**Lemma 2.4.** *Let the iteration $\Phi$ be continuous with respect to the first argument. If*

$$x^* := \lim_{m\to\infty} x^m(y, b, A) \qquad \textit{(cf. (2.4))}$$

*exists, $x^*$ is a fixed point of $\Phi(\cdot, b, A)$.*

### 2.1.3 Consistency

Lemma 2.4 states that possible results of the iteration method have to be sought in the set of fixed points. Therefore, a minimum condition is that the solution of system (2.1) with the right-hand side $b \in \mathbb{K}^I$ be a fixed point with respect to $b$. This property is the subject of the following definition.

**Definition 2.5 (consistency).** The iterative method $\Phi$ is called *consistent* to the system (2.1) with $A \in \mathfrak{D}(\Phi)$ if, for all right-hand sides $b \in \mathbb{K}^I$, any solution of $Ax = b$ is a fixed point of $\Phi(\cdot, b, A)$.

According to Definition 2.5, consistency means: For all $b, x \in \mathbb{K}^I$ and all matrices $A \in \mathfrak{D}(\Phi)$, the implication $Ax = b \Rightarrow x = \Phi(x, b, A)$ holds. The reverse implication would yield an *alternative* (nonequivalent) form of consistency:

$$Ax = b \quad \text{for all fixed points } x \text{ of } \Phi(\cdot, b, A) \text{ and for all } b \in \mathbb{K}^I,\ A \in \mathfrak{D}(\Phi). \quad (2.5)$$

Note that both variants of consistency do not require the regularity assumption (2.2). Even without (2.2), there may be a solution of $Ax = b$ for certain $b$. Then Definition 2.5 implies the existence of a fixed point of $\Phi(\cdot, b)$. Vice versa, (2.5) states the existence of a solution of $Ax = b$ as soon as $\Phi(\cdot, b, A)$ has a fixed point. The regularity of $A$ will be discussed in Theorem 2.8.

### 2.1.4 Convergence

A natural definition of the convergence of an iterative method $\Phi$ seems to be

$$\lim_{m\to\infty} x^m(y, b, A) \quad \text{exists for all } y, b \in \mathbb{K}^I, \quad (2.6)$$

where $x^m(y, b, A)$ are the iterates defined in (2.4) corresponding to the starting value $x^0 := y$, while $A \in \mathfrak{D}(\Phi)$ is a fixed matrix. Since the starting value may be chosen arbitrarily, it may happen that an iteration satisfying (2.6) converges, but to different limits *depending* on the starting value. Therefore, the independence of the limit has to be incorporated into the definition of convergence. This yields the following definition, which is stronger than (2.6).

**Definition 2.6.** Fix $A \in \mathfrak{D}(\Phi)$. An iterative method $\Phi(\cdot, \cdot, A)$ is called *convergent* if for all $b \in \mathbb{K}^I$, there is a limit $x^*(b, A)$ of the iterates (2.4) independent of the starting value $x^0 = y \in \mathbb{K}^I$.

Note that consistency is a property of $\Phi$ for *all* $A \in \mathfrak{D}(\Phi)$, whereas convergence is required for a particular $A \in \mathfrak{D}(\Phi)$. Therefore $\Phi(\cdot, \cdot, A)$ may be convergent for some $A$, while $\Phi(\cdot, \cdot, A')$ diverges for another $A'$.

### 2.1.5 Convergence and Consistency

**Remark 2.7.** In the following, we shall often assume that the iterative method $\Phi$ is *convergent and consistent*. The term 'convergent and consistent' refers to a matrix $A \in \mathfrak{D}(\Phi)$ and means precisely: $\Phi$ is consistent and, for $A \in \mathfrak{D}(\Phi)$, the particular iteration $\Phi(\cdot, \cdot, A)$ is convergent.

It will turn out that the chosen definitions of the terms 'convergence' and 'consistency' of $\Phi$ are almost equivalent to the combination of the alternative definitions in (2.5) and (2.6).

**Theorem 2.8.** *Let $\Phi$ be continuous in the first argument. Then $\Phi$ is consistent and convergent if and only if $A$ is regular and $\Phi$ fulfils the conditions (2.5) and (2.6).*

*Proof.* (i) Assume $\Phi$ to be consistent and convergent. (2.6) follows from Definition 2.6. If $A$ is singular, the equation $Ax = 0$ would have a nontrivial solution $x^{**} \neq 0$ besides $x^* = 0$. By consistency, both are fixed points of $\Phi$ with respect to $b = 0$. Therefore, choosing the starting values $x^0 = x^*$ and $x^0 = x^{**}$, we obtain the constant sequences $x^m(x^*, 0) = x^*$ and $x^m(x^{**}, 0) = x^{**}$. The convergence definition states that the limits $x^*$ and $x^{**}$ coincide contrary to the assumption. Hence, $A$ is regular. It remains to prove (2.5). The preceding argument shows that a convergent iterative method can have only *one* fixed point with respect to $b$. Because of the regularity of $A$, there is a solution of $Ax = b$ that, thanks to consistency, is the unique fixed point of $\Phi$ with $b$. Hence, (2.5) is proved.

(ii) Assume $\Phi(x, b)$ to be continuous in $x$ and that (2.5) and (2.6) are fulfilled. Furthermore, let $A$ be regular. Due to Lemma 2.4, $x^* := \lim x^m(y, b)$ is a fixed point of $\Phi$ with respect to $b$ and therefore, by (2.5), a solution of $Ax = b$. Because of the regularity of $A$, the solution of the system is unique and hence also the limit of $x^m(y, b)$, which thereby cannot depend on $y$. Hence, $\Phi$ is convergent in the sense of Definition 2.6. Convergence leads to the uniqueness of the fixed point with respect to $b$ (cf. part (i)). Since, by (2.5), this fixed point is the uniquely determined solution of $Ax = b$, $\Phi$ is consistent. □

### 2.1.6 Defect Correction as an Example of an Inconsistent Iteration

In this monograph, all iterations will be assumed to be consistent. Usually, inconsistent iterations are an involuntary consequence of a bug in the implementation. However, there are examples where inconsistent iterations are of practical relevance. Assume that both $Ax = b$ and $By = c$ are discretisations of the same partial differential equation. Assume further that $Ax = b$ is simpler to solve than $By = c$, but the error of the discretisation by $B$ is smaller than the discretisation error of $A$. Then there are combinations of both discretisations so that the overall treatment is as simple as for $A$ but yielding the accuracy of $B$.

The standard defect correction $x^{m+1} = x^m - A^{-1}(Bx^m - c)$ can be stopped after a few iteration steps since the desired discretisation accuracy is reached (cf. [194, §14.2.2], [197, §7.5.9.2]). This is even true if the matrix $B$ is singular or almost singular (this is the case of an unstable but consistent[2] discretisation). An extreme case of solving a problem with an unstable discretisation of high consistency order is demonstrated in [178].

Another mixing of both discretisation is described in [194, §14.3.3], where parts of the multigrid iteration for $Ax = b$ use $B$ in the smoothing step. The limit $x^*$ of the iterates solves neither $Ax^* = b$ nor $Bx^* = c$.

## 2.2 Linear Iterative Methods

One would expect iterative methods to be linear in $x, b$, since they solve linear equations. In fact, most of the methods described in this book are linear, but there are also important nonlinear iterations as, e.g., discussed in Part II.

### *2.2.1 Notation, First Normal Form*

**Definition 2.9 (linear iteration, iteration matrix).** An iterative method $\Phi$ is called *linear* if $\Phi(x, b)$ is linear in $(x, b)$, i.e., if there are matrices $M$ and $N$ such that

$$\Phi(x, b, A) = M[A]\, x + N[A]\, b.$$

In most of the cases, $A$ is fixed and we use the shorter form

$$\Phi(x, b) = Mx + Nb. \tag{2.7}$$

Here, the matrix $M = M[A]$ is called the *iteration matrix* of the iteration $\Phi$.

Iteration (2.3) takes the form (2.8), which represents the *first normal form* of the iteration $\Phi$:

$$x^{m+1} := Mx^m + Nb \qquad (m \geq 0,\ b \text{ in (2.1)}). \tag{2.8}$$

Whenever possible, we shall denote the iteration matrix of a specific iteration method 'xyz' by $M^{xyz}$; e.g., $M^{\mathrm{GS}}$ belongs to the Gauss-Seidel method. Similarly for $N^{xyz}$. When we refer to the mapping $\Phi$, we write $M_\Phi, N_\Phi$, etc.

**Remark 2.10.** Assume (2.2). If $N = N[A]$ is singular, there is some $x^* \neq 0$ with $Nx^* = 0$ and $b := Ax^* \neq 0$. Starting iteration (2.8) with $x^0 = 0$ yields $x^m = 0$ and hence $\lim x^m = 0$. In Corollary 2.17b we shall state that, in this case, the iteration is not convergent.

The iteration $\Phi(\cdot, \cdot, A)$ is algebraic in the sense of Definition 2.2b if and only if the matrices $M$ and $N$ are explicit functions of $A$.

[2] Concerning the terms 'consistent' and 'consistency order', we refer to Hackbusch [197, §§6,7].

### 2.2.2 Consistency and Second Normal Form

For a linear and consistent iteration $\Phi$, each solution of $Ax = b$ must be a fixed point with respect to $b$: $x = Mx + Nb$. Each $x \in \mathbb{K}^I$ can be the solution of $Ax = b$ (namely, for $b := Ax$). Hence,

$$x = Mx + Nb = Mx + NAx$$

holds for all $x$ and leads to the matrix equation

$$M[A] + N[A]\,A = I, \tag{2.9}$$

or in short,

$$M + NA = I,$$

establishing a relation between $M$ and $N$ in (2.8). This proves the next theorem.

**Theorem 2.11 (consistency).** *A linear iteration $\Phi$ is consistent if and only if the iteration matrix $M$ can be determined from $N$ by*

$$M[A] = I - N[A]\,A \qquad \textit{for all } A \in \mathfrak{D}(\Phi). \tag{2.9$'$}$$

*If, in addition, $A$ is regular, $N$ can be represented as a function of $M$:*

$$N[A] = (I - M[A])A^{-1}. \tag{2.9$''$}$$

Combining formulae (2.8) and (2.9$'$), we can represent linear and consistent iterations in their *second normal form*:

$$x^{m+1} := x^m - N[A]\,(Ax^m - b) \qquad (m > 0,\ A, b \text{ in (2.1)}). \tag{2.10}$$

In the sequel, the matrix

$$N = N[A] = N_\Phi = N_\Phi[A]$$

will be called the 'matrix of the second normal form of $\Phi$'. Equation (2.10) shows that $x^{m+1}$ is obtained from $x^m$ by a correction which is the *defect* $Ax^m - b$ of $x^m$ multiplied by $N$. The fact that the defect of $x^m$ vanishes if and only if it is a solution of $Ax = b$, proves the next remark.

**Remark 2.12.** The second normal form (2.10) with arbitrary $N \in \mathbb{K}^{I\times I}$ represents all linear and consistent iterations.

Since consistent linear iterations are the standard case, we introduce the following notation for the set of these iterations:

$$\mathcal{L} := \{\Phi : \mathbb{K}^I \times \mathbb{K}^I \times \mathbb{K}^{I\times I} \to \mathbb{K}^I \text{ consistent linear iteration}, \#I < \infty\}. \tag{2.11}$$

### 2.2.3 Third Normal Form

The *third normal form* of a linear iteration reads as follows:

$$W[A]\,(x^m - x^{m+1}) = Ax^m - b \qquad (m > 0,\ A, b \text{ in (2.1)}). \tag{2.12}$$

$W = W[A] = W_\Phi = W_\Phi[A]$ is called the 'matrix of the third normal form of $\Phi$'. Equation (2.12) can be understood in the following algorithmic form:

$$\text{solve } W\delta = Ax^m - b \qquad \text{and define} \qquad x^{m+1} := x^m - \delta. \tag{2.12$'$}$$

This represents a definition of $x^{m+1}$ as long as $W$ is regular. Under this assumption, one can solve for $x^{m+1}$. A comparison with (2.10) proves the following.

**Remark 2.13.** If $W$ in (2.12) is regular, iteration (2.12) coincides with the second normal form (2.10), where $N$ is defined by

$$N = W^{-1}. \tag{2.13}$$

Vice versa, the representation (2.10) with regular $N$ can be rewritten as (2.12) with $W = N^{-1}$.

We shall see that for the interesting cases, $N$ must be regular (cf. Remark 2.18). Combining (2.9$'$) and (2.13) yields

$$M[A] = I - W[A]^{-1}A. \tag{2.13$'$}$$

### 2.2.4 Representation of the Iterates $x^m$

By the notation $x^m(x^0, b, A)$ in (2.4) we express the dependency on the starting value $x^0$ and on the the data $b$, $A$ of the system (2.1). The explicit representation of $x^m$ in terms of $x^0$ and $b$ is given in (2.14).

**Theorem 2.14.** *The linear iteration (2.7) produces the iterates*

$$x^m(x^0, b, A) = M[A]^m x^0 + \sum_{k=0}^{m-1} M[A]^k N[A]\, b \tag{2.14}$$

*for* $m \geq 0$ *and* $A \in \mathfrak{D}(\Phi)$.

*Proof.* For the induction start at $m = 0$, Eq. (2.14) takes the form $x^0(x^0, b) = x^0$ in accordance with (2.4). Assuming (2.14) for $m - 1$, we obtain from (2.7) that

$$\begin{aligned} x^m(x^0, b) &= Mx^{m-1} + Nb = M\left(M^{m-1}x^0 + \sum_{k=0}^{m-2} M^k Nb\right) + Nb \\ &= M^m x^0 + \sum_{k=1}^{m-1} M^k Nb + Nb = M^m x^0 + \sum_{k=0}^{m-1} M^k Nb. \end{aligned}$$

□

In the following, $e^m$ denotes the (iteration) error of $x^m$:

$$e^m := x^m - x, \qquad \text{where } x \text{ solves } Ax = b. \tag{2.15}$$

Assuming consistency, we have $x = Mx + Nb$ for the solution $x$ in (2.15). Forming the difference with (2.8): $x^{m+1} = Mx^m + Nb$, we attain the simple relation

$$e^{m+1} = Me^m \quad (m \geq 0), \qquad e^0 = x^0 - x, \tag{2.16a}$$

between two successive errors. Therefore the iteration matrix is the amplification matrix of the error. A trivial conclusion is

$$e^m = M^m e^0 \qquad (m \geq 0). \tag{2.16b}$$

The expression $Ax - b$ is called the *defect* of a vector $x$. In particular,

$$d^m := Ax^m - b \tag{2.17}$$

denotes the defect of the $m$-th iterate $x^m$.

**Exercise 2.15.** Prove: (a) The defect $\bar{d} = A\bar{x} - b$ and the error $\bar{e} = \bar{x} - x$ fulfil the equation $A\,\bar{e} = \bar{d}$.
(b) Let $\Phi \in \mathcal{L}$ (cf. (2.11)) and assume that $A$ is regular. Then the defects satisfy

$$d^{m+1} = AMA^{-1}d^m, \quad d^0 := Ax^0 - b, \quad d^m = (AMA^{-1})^m d^0.$$

### 2.2.5 Convergence

A necessary and sufficient convergence criterion can be formulated by the spectral radius $\rho(M)$ of the iteration matrix (cf. Definition A.17).

**Theorem 2.16 (convergence theorem, convergence rate).** *A linear iteration (2.7) with the iteration matrix $M = M[A]$ is convergent if and only if*

$$\rho(M) < 1. \tag{2.18}$$

*$\rho(M)$ is called the* convergence rate *of the iteration $\Phi(\cdot, \cdot, A)$.*

In the sequel, the terms *convergence rate*, *convergence speed*, and *iteration speed* are used synonymously for $\rho(M)$. Some authors define the convergence rate as the negative logarithm $-\log(\rho(M))$ (cf. (2.30a) and Varga [375], Young [412]).

*Proof.* (i) Let iteration (2.7) be convergent. In Definition 2.6 we may choose $b := 0$ and exploit the representation (2.14): $x^m = M^m x^0$. The starting value $x^0 := 0$ yields the limit $x^* = 0$, which by the convergence definition must hold for any starting value. If $\rho(M) \geq 1$, one could choose $x^0 \neq 0$ as the eigenvector corresponding to an eigenvalue $\lambda$ with $|\lambda| = \rho(M) \geq 1$. The resulting sequence $x^m = \lambda^m x^0$ cannot converge to $x^* = 0$. Hence, inequality (2.18) is necessary for convergence.

(ii) Now let (2.18) be valid: $\rho(M) < 1$. By Lemma B.28, $M^m x^0$ converges to zero, while Theorem B.29 proves $\sum_{k=0}^{m-1} M^k \to (I - M)^{-1}$. Thanks to the representation (2.14), $x^m$ tends to

$$x^* := (I - M)^{-1} N b. \tag{2.19}$$

Since this limit does not depend on the starting value, the iteration is convergent. □

The proof already contains the first statement of the following corollary.

**Corollary 2.17.** (a) If the iterative method (2.7) is convergent, the iterates converge to $(I - M)^{-1} N b$.
(b) If the iteration is convergent, then $A$ and $N = N[A]$ are regular.
(c) If, in addition, the iteration is consistent, the iterates $x^m$ converge to the unique solution $x = A^{-1} b$.

*Proof.* (b) If either $A$ or $N$ are singular, the product $AN$ is singular and $ANx = 0$ holds for some $x \neq 0$. As $M = I - NA$, $x$ is an eigenvector of $M$ with the eigenvalue 1. Hence $\rho(M) \geq 1$ proves the divergence of the iteration. This proves part (b).

(c) By consistency and part (b), there is a representation (2.10) with regular $N$ and $A$, so that $(I - M)^{-1} N = A^{-1}$ follows from (2.9). (2.19) proves part (c). □

**Remark 2.18.** Since only convergent and consistent iterations are of interest and since in this case, by Corollary 2.17b, $A$ and $N$ are regular, the representation (2.9″) of $N$ and the third normal form (2.4) hold with the matrix $W = N^{-1}$.

The convergence $x^m \to x$ is an asymptotic statement for $m \to \infty$ that allows no conclusion concerning the error $e^m = x^m - x$ for some fixed $m$. The values of $u^m_{16,16}$ given in Tables 1.1–1.2 even deteriorate during the first steps before they converge monotonically to the limit $\frac{1}{2}$. Often, one would like to have a statement for a *fixed* iteration number $m$. In this case, the convergence criterion (2.18) has to be replaced with a norm estimate.

**Theorem 2.19.** *Let $\|\cdot\|$ be a corresponding matrix norm. A sufficient condition for convergence of an iteration is the estimate*

$$\|M\| < 1 \tag{2.20}$$

*of the iteration matrix $M$. If the iteration is consistent, the error estimates (2.21) hold:*

$$\|e^{m+1}\| \leq \|M\| \, \|e^m\|, \qquad \|e^m\| \leq \|M\|^m \, \|e^0\|. \tag{2.21}$$

*Proof.* (2.20) implies (2.18) (cf. (B.20b)). (2.21) is a consequence of (2.16a,b). □

$\|M\|$ is called the *contraction number* of the iteration (with respect to the norm $\|\cdot\|$). In the case of (2.20), the iteration is called *monotonically convergent* with respect to the norm $\|\cdot\|$, since $\|e^{m+1}\| < \|e^m\|$. If the norm $\|\cdot\|$ fulfils the equality $\rho(M) = \|M\|$, the terms 'convergence' and 'monotone convergence' coincide.

### 2.2.6 Convergence Speed

Inequality (2.21), i.e., $\|e^{m+1}\| \le \zeta \|e^m\|$ with $\zeta := \|M\| < 1$, describes linear convergence. Faster convergence than linear convergence is only attainable by nonlinear methods (cf. §10.2.3). The contraction number $\zeta$ depends on the choice of the norm. According to (B.20b), the contraction number $\zeta$ is always larger or equal to the convergence rate $\rho(M)$. On the other hand, Lemma B.26 ensures that for a suitable choice of the norm, the contraction number $\zeta$ approximates the convergence rate $\rho(M)$ arbitrarily well.

The contraction number as well as the convergence rate determine the quality of an iterative method. Both quantities can be determined from the errors $e^m$ as follows.

**Remark 2.20.** The contraction number is the maximum of the ratios $\|e^1\|/\|e^0\|$ taken over all starting values $x$.

*Proof.* Use (2.16b) for $m = 1$ and Exercise B.10d. □

**Exercise 2.21.** Prove: (a) In general, Remark 2.20 becomes wrong if $\|e^1\|/\|e^0\|$ is replaced with $\|e^{m+1}\|/\|e^m\|$ for some $m > 0$.
(b) The latter quotient takes the maximum

$$\zeta_{m+1} := \begin{cases} \max\{\|Mx\| \,/\, \|x\| : x \in \operatorname{range}(M^m)\backslash\{0\}\} & \text{if } M^m \neq 0, \\ 0 & \text{otherwise,} \end{cases}$$

which can be interpreted as the matrix norm of the mapping $x \mapsto Mx$ restricted to the subspace $V_m := \operatorname{range}(M^m) := \{M^m x : x \in \mathbb{K}^I\}$.
(c) The inclusion $V_{m+1} \subset V_m$ holds with an equality sign at least for $m \ge \#I$.
(d) $\rho(M) \le \zeta_{m+1} \le \zeta_m \le \zeta_0 = \zeta := \|M\|$ holds for $m \ge 0$.
(e) For regular $M$, one has $\zeta_m = \zeta$ for all $m$.

Exercise 2.21 demonstrates that the contraction number is a somewhat too coarse term: It may happen that the contraction number gives a too pessimistic prediction of the convergence speed. A more favourable estimate can be obtained by the numerical radius $r(\cdot)$ of the matrix $M^m$ (cf. §B.3.4). The inequalities

$$\|M^m\|_2 \le 2\, r(M^m) \qquad \text{(cf. (B.28d))} \tag{2.22a}$$

and (2.16b) yield the error estimate

$$\|e^m\|_2 \le 2\, r(M^m)\, \|e^0\|_2 \qquad (m \ge 0) \tag{2.22b}$$

with respect to the Euclidean norm. If $\|\cdot\|_C$ is the norm defined by (C.5a) with a positive definite matrix $C$, one analogously proves the inequality

$$\|e^m\|_C \le 2\, r(C^{1/2} M^m C^{-1/2}) \|e^0\|_C \qquad (m \ge 0). \tag{2.22c}$$

For the practical judgment of the convergence speed from 'experimental data', i.e., from a sequence of errors $e^m$ belonging to a special starting value $x^0$, one may use the *reduction factors*

$$\rho_{m+1,m} := \|e^{m+1}\|/\|e^m\| \,. \tag{2.23a}$$

These numbers can, e.g., be found in the last column of Tables 1.1–1.2. More interesting than a single value $\rho_{m+1,m}$ is the geometric mean

$$\rho_{m+k,m} := \left[\rho_{m+k,m+k-1} \cdot \rho_{m+k-1,m+k-2} \cdot \ldots \cdot \rho_{m+1,m}\right]^{1/k},$$

which due to definition (2.23a) can more easily be represented by

$$\rho_{m+k,m} := \left[\|e^{m+k}\|/\|e^m\|\right]^{1/k}. \tag{2.23b}$$

The properties of $\rho_{m+k,m}$ are summarised below.

**Remark 2.22.** (a) Denote the dependence of the magnitude $\rho_{m+k,m}$ on the starting value $x^0$ by $\rho_{m+k,m}(x^0)$. Then

$$\lim_{k\to\infty} \max\{\rho_{m+k,m}(x^0) : x^0 \in \mathbb{K}^I\} = \rho(M) \qquad \text{for all } m.$$

(b) Even without maximisation over all $x^0 \in \mathbb{K}^I$,

$$\lim_{k\to\infty} \rho_{m+k,m}(x^0) = \rho(M) \qquad \text{for all } m \tag{2.23c}$$

holds, provided that $x^0$ does not lie in the subspace $U \subset \mathbb{K}^I$ of dimension $< \#I$ spanned by all eigenvectors and possibly existing principal vectors of the matrix $M$ corresponding to eigenvalues $\lambda$ with $|\lambda| < \rho(M)$. (2.23c) holds almost always because a stochastically chosen starting value $x^0$ lying in a fixed lower dimensional subspace has probability zero.
(c) The reduction factors $\rho_{m+1,m}(x^0)$ tend to the spectral radius of $M$:

$$\lim_{m\to\infty} \rho_{m+1,m}(x^0) = \rho(M) \tag{2.23d}$$

for all $x^0 \notin U$ with $U$ in part (b) if and only if there is exactly one eigenvalue $\lambda \in \sigma(M)$ with $|\lambda| = \rho(M)$, and if, for this eigenvalue, the geometric and algebraic multiplicities coincide. Sufficient conditions are: (i) $\lambda \in \sigma(M)$ with $|\lambda| = \rho(M)$ is a single eigenvalue, or (ii) $M$ is a positive matrix (cf. (C.11a)).
(d) Choose a norm $\|\cdot\| = \|\cdot\|_C$ with $C > 0$ (cf. (2.22c)) in (2.23a). If $C^{\frac{1}{2}} M C^{-\frac{1}{2}}$ is Hermitian, $\rho_{m+1,m}(x^0)$ ($x^0 \notin U$) converges monotonically increasing to $\rho(M)$.

*Proof.* (i) Use

$$\rho(M) \le \max_{x^0\in\mathbb{K}^I} \rho_{m+k,m}(x^0) \le \max_{x^0\in\mathbb{K}^I} \rho_{k,0}(x^0) \le \|M^k\|^{1/k}$$

and $\|M^k\|^{1/k} \to \rho(M)$ according to Theorem B.27. This proves part (a).

(ii) Let $I_0 \subset I$ be the nonempty index subset $I_0 := \{i \in I : |J_{ii}| = \rho(M)\}$, where $J_{ii}$ are the diagonal elements of the Jordan normal form $M = TJT^{-1}$ (cf. (A.15a,b)). The subspace $U := \{x : (T^{-1}x)_i = 0 \text{ for all } i \in I_0\}$ is the maximal subspace with the property $\lim_{m\to\infty} [\,\|M^m x\| \,/\, \|x\|\,]^{1/m} < \rho(M)$. Its dimension is $\dim(U) = \#I - \#I_0 < \#I$.

(iii) Define $\hat{M} = C^{1/2}MC^{-1/2}$ and $\hat{e}^m := C^{1/2}e^m$. Since the norms are related by $\|e^m\|_C = \|\hat{e}^m\|_2$, we obtain for $m \geq 1$ that

$$\begin{aligned}\|\hat{e}^m\|_2^2 = \|\hat{M}^m\hat{e}^0\|_2^2 &= \left\langle \hat{M}^m\hat{e}^0, \hat{M}^m\hat{e}^0 \right\rangle = \left\langle \hat{M}^{m+1}\hat{e}^0, \hat{M}^{m-1}\hat{e}^0 \right\rangle \\ &= \left\langle \hat{e}^{m+1}, \hat{e}^{m-1} \right\rangle \leq \|\hat{e}^{m+1}\|_2 \|\hat{e}^{m-1}\|_2.\end{aligned}$$

Hence it follows that $\rho_{m+1,m} = \frac{\|e^{m+1}\|}{\|e^m\|} = \frac{\|\hat{e}^{m+1}\|_2}{\|\hat{e}^m\|_2} \geq \frac{\|\hat{e}^m\|_2}{\|\hat{e}^{m-1}\|_2} = \rho_{m,m-1}$. □

Remark 2.22 allows us to view the value $\rho_{m+k,m}$ and possibly also $\rho_{m+1,m}$ for sufficiently large $m$ as a good approximation of the spectral radius. This viewpoint can be reversed.

**Remark 2.23.** The convergence rate $\rho(M)$ is a suitable measure for judging (asymptotically) the convergence speed. This holds even if convergence is required with respect to a specific norm.

*Proof.* By Theorem B.27, for each $\varepsilon > 0$ there is some $m_0$ such that $m \geq m_0$ implies that $\rho(M) \leq \|M^m\|^{1/m} \leq \rho(M) + \varepsilon$ and $\|e^m\| \leq (\rho(M) + \varepsilon)^m \|e^0\|$. □

### *2.2.7 Remarks Concerning the Matrices M, N, and W*

Considerations in §§2.2.5–2.2.6 show the close connection between the iteration matrix $M$ and the convergence speed. $M$ directly describes the *error reduction* or amplification (cf. (2.16a)). Roughly speaking, the convergence is better the smaller $M$ is. $M = 0$ would be optimal. However, then $\Phi$ is a direct method, since $x^1$ is already the exact solution (its error is $e^1 = Me^0 = 0$).

The matrix $N$ transforms the defect $Ax^m - b$ into the correction $x^m - x^{m+1}$. The optimal case[3] $M = 0$ mentioned above corresponds to $N[A] = A^{-1}$. Therefore, one may regard $N[A]$ as an *approximate inverse* of $A$.

Concerning implementation, often the matrix $W$ of the third normal form (2.12) is the important one. By the relation $W = N^{-1}$ (cf. (2.13)), $W = A$ would be optimal. However, then computing the correction $x^m - x^{m+1}$ is equivalent to the direct solution of the original equation. Therefore, one has to find approximations $W$ of $A$, so that the solution of the system $W\delta = d$ is sufficiently easy.

In the case of some of the classical iterations discussed in §3, we have explicit expressions for $N$ or $W$ and may use these matrices for the computation. On the other hand, there will be iterative methods, for which the algorithm is implemented differently without reference to the matrices $M$, $N$, $W$ (see, e.g., Propositions 3.13 or 5.25).

[3] Consistent linear iterations with $M = 0$ can be called direct solvers. Vice versa, any direct solver defines a linear iteration with $M = 0$.

### 2.2.8 Three-Term Recursions, Two- and Multi-Step Iterations

So far we considered *one-step iterations*, i.e., $x^{m+1}$ is computed in one step from $x^m$. Sometimes linear iterations occur, in which computing $x^{m+1}$ involves $x^m$ and $x^{m-1}$:

$$x^{m+1} = M_0\, x^m + M_1\, x^{m-1} + N_0\, b \qquad (m \geq 1). \tag{2.24}$$

For the starting procedure, one needs two initial values $x^0$ and $x^1$. Such *two-step iterations* are also called *three-term recursions* since they involved the three terms $x^{m+1}$, $x^m$, $x^{m-1}$. Formally, a three-term recursion can be reduced to a standard one-step iteration acting in the space $\mathbb{K}^I \times \mathbb{K}^I$:

$$\begin{bmatrix} x^{m+1} \\ x^m \end{bmatrix} = \mathbf{M} \begin{bmatrix} x^m \\ x^{m-1} \end{bmatrix} + \begin{bmatrix} N_0 b \\ 0 \end{bmatrix} \qquad \text{with } \mathbf{M} := \begin{bmatrix} M_0 & M_1 \\ I & 0 \end{bmatrix}. \tag{2.25}$$

Now the convergence condition

$$\rho(\mathbf{M}) < 1 \tag{2.26a}$$

ensures that recursion (2.25) has a limit that is also the fixed point. The consistency condition takes the form

$$I - M_0 - M_1 = N_0 A\,. \tag{2.26b}$$

**Exercise 2.24.** The limit of the iteration (2.25) has the general form $\left[\begin{smallmatrix}\xi\\ \eta\end{smallmatrix}\right] \in \mathbb{K}^I \times \mathbb{K}^I$. Show that the conditions (2.26a,b) imply $\xi = \eta = A^{-1}b$.

**Exercise 2.25.** Given an iteration $x^{m+1} = Mx^m + Nb$, define the matrices $M_0$, $M_1$, $N_0$ in (2.24) by

$$\begin{aligned} M_0 &:= \Theta M + \vartheta I, \\ M_1 &:= (1 - \Theta - \vartheta)\, I, \\ N_0 &:= \Theta N \end{aligned}$$

with $\Theta, \vartheta \in \mathbb{R}$. The three-term recursion (2.24) takes the form

$$x^{m+1} = \Theta\left[(Mx^m + Nb) - x^{m-1}\right] + \vartheta(x^m - x^{m-1}) + x^{m-1}. \tag{2.27}$$

Prove that (a) $\mathbf{M}$ has the spectrum

$$\sigma(\mathbf{M}) = \left\{ \frac{1}{2}\,(\Theta\lambda + \vartheta) \pm \sqrt{1 - \Theta - \vartheta + \frac{1}{4}\,(\Theta\lambda + \vartheta)^2} :\ \lambda \in \sigma(M) \right\}.$$

(b) Conclude from $\rho(M) < 1$ and $\Theta > 0,\ \vartheta \geq 0,\ \Theta + \vartheta \leq 1$ that $\rho(\mathbf{M}) < 1$.

## 2.3 Efficacy of Iterative Methods

The convergence rate cannot be the only criterion for the quality of an iterative method because one has also to take into account the amount of computational work of $\Phi$.

### *2.3.1 Amount of Computational Work*

The representation (2.12′) suggests that any iteration requires at least computing the defect $Ax^m - b$. For a general $n \times n$ matrix $A \in \mathbb{K}^{I \times I}$ ($n = \#I$), multiplying $Ax^m$ would require $2n^2$ operations. However, as discussed in §1.7, it is more realistic to assume that $A$ is sparse; i.e., the number $s(n)$ of the nonzero elements of $A$ is distinctly smaller than $n^2$. For matrices arising from discretisations of partial differential equations, one has

$$s(n) \leq C_A n, \tag{2.28}$$

where $C_A$ is a constant with respect to $n$, but depends on the matrix $A$. For the five-point formula (1.4a) of the model problem, inequality (2.28) holds with $C_A = 5$. Under assumption (2.28), one can perform matrix-vector multiplication in $2C_A n$ operations.

After evaluating $d := Ax^m - b$, one has still to solve the system $W\delta = d$ in (2.12′). For any practical iterative method, we should require that this part consumes only $\mathcal{O}(n)$ operations, so that the total amount of work is also of the order $\mathcal{O}(n)$. We relate the constant in $\mathcal{O}(n)$ to $C_A$ in (2.28) and obtain the following formulation:

$$\begin{array}{l}\text{The number of arithmetic operations per iteration}\\ \text{step of the method } \Phi \text{ is } \operatorname{Work}(\Phi, A) \leq C_\Phi C_A n\,.\end{array} \tag{2.29}$$

Here, $\operatorname{Work}(\Phi, A)$ is the amount of work of the $\Phi$ iteration applied to $Ax = b$. Note that $C_\Phi$ depends on the iteration $\Phi$ but not on $A$, whereas $C_A n$ indicates the degree of sparsity of $A$. Therefore, the constant $C_\Phi$ may be called the *cost factor* of the iteration $\Phi$.

So far we only discussed the cost arising by performing one iteration step of $\Phi$. Depending on the method, some *initialisation* may be necessary for precomputing some quantities required by $\Phi$. Let $\operatorname{Init}(\Phi, A)$ be the corresponding cost.

**Remark 2.26.** If $m$ iteration steps are performed, the effective cost per iteration is

$$\operatorname{Work}(\Phi, A) + \operatorname{Init}(\Phi, A)/m.$$

In the standard case, the initialisation uses only the data of $A$. Therefore it pays if many systems $Ax^i = b^i$ are solved with different right-hand sides $b^i$ but the same matrix $A$.

### 2.3.2 *Efficacy*

An iteration $\Phi$ can be called 'more effective' than $\Psi$ if for the same amount of work $\Phi$ is faster, or if $\Phi$ has the same convergence rate, but consumes less work than $\Psi$. To obtain a common measure, we ask for the amount of work that is necessary to reduce the error by a fixed factor. This factor is chosen as $1/\mathrm{e}$, since the natural logarithm is involved. According to Remark 2.23, we use the convergence rate $\rho(M)$ for the (asymptotic) description of the error reduction per iteration step. After $m$ iteration steps, the asymptotic error reduction is $\rho(M)^m$. In order to ensure $\rho(M)^m \le 1/\mathrm{e}$, we have to choose $m \ge -1/\log(\rho(M))$, provided that convergence holds: $\rho(M) < 1 \Leftrightarrow \log(\rho(M)) < 0$. Therefore, we define

$$\mathrm{It}(\Phi) := -1/\log(\rho(M)). \tag{2.30a}$$

$\mathrm{It}(\Phi)$ represents the (asymptotic) number of the iteration steps for an error reduction by the factor of $1/\mathrm{e}$. Note that, in general, $\mathrm{It}(\Phi)$ is not an integer.

**Remark 2.27.** (a) Convergence of $\Phi$ is equivalent to $0 \le \mathrm{It}(\Phi) < \infty$. The value $\mathrm{It}(\Phi) = 0$ corresponds to $\rho(M) = 0$, i.e., to a direct method.

(b) Let $\Phi \in \mathcal{L}$. To reduce the iteration error (asymptotically) by a factor of $\varepsilon < 1$, we need the following number of iteration steps:

$$\mathrm{It}(\Phi, \varepsilon) := -\mathrm{It}(\Phi)\log(\varepsilon) \tag{2.30b}$$

(c) If $\rho(M) = \|M\|$ or $\rho(M)$ in (2.30a) is replaced with $\|M\| < 1$, one can guarantee (not only asymptotically) that

$$\|e^{m+k}\| \le \varepsilon \|e^m\| \qquad \text{for } k \ge \mathrm{It}(\Phi, \varepsilon). \tag{2.30c}$$

(d) If $r(M) < 1$ holds for the numerical radius of $M$ introduced in §B.3.4, definition (2.30b) can be replaced with $\mathrm{It}(\Phi, \varepsilon) := \log(\varepsilon/2)/\log(r(M))$. Then, inequality (2.30c) holds with respect to the Euclidean norm.

The amount of work corresponding to the error reduction by $1/\mathrm{e}$ is the product $\mathrm{It}(\Phi)\,\mathrm{Work}(\Phi, A) \le \mathrm{It}(\Phi) C_\Phi C_A n$ (cf. (2.29)). As a characteristic quantity we choose the *effective amount of work*

$$\mathrm{Eff}(\Phi) := \mathrm{It}(\Phi) C_\Phi = -C_\Phi/\log(\rho(M)). \tag{2.31a}$$

$\mathrm{Eff}(\Phi)$ measures the amount of work for an error reduction by $1/\mathrm{e}$ in the unit '$C_A n$ arithmetic operations'. Correspondingly, the effective amount of work for the error reduction by the factor of $1/\mathrm{e}$ is given by

$$\mathrm{Eff}(\Phi, \varepsilon) := -\mathrm{It}(\Phi) C_\Phi \log(\varepsilon) = C_\Phi \log(\varepsilon)/\log(\rho(M)). \tag{2.31b}$$

**Example 2.28.** In the case of the model problem, the cost factor of the Gauss–Seidel iteration is $C_\Phi = 1$ (because of $C_A = 5$, cf. Remark 1.14). The numerical values in Table 1.1 suggest $\rho(M) = 0.99039$ for the grid size $h = 1/32$. Thus, the effective amount of work equals $\mathrm{Eff}(\Phi) = 103.6$. Using $\rho(M) = 0.82$ for the SOR method and $C_\Phi = 7/5$, we deduce an effective amount of work of $\mathrm{Eff}(\Phi) = 7.05$ for the SOR method with $h = 1/32$.

### 2.3.3 Order of Linear Convergence

The convergence rates $\rho(M)$ in Example 2.28 are typically close to one; i.e., the convergence is rather slow. Therefore, we may use the ansatz

$$\rho(M) = 1 - \eta \qquad (\eta \text{ small}). \tag{2.32a}$$

The Taylor expansion yields $\log(1-\eta) = -\eta + \mathcal{O}(\eta^2)$ and $\frac{-1}{\log(1-\eta)} = \frac{1}{\eta(1+\mathcal{O}(\eta))} = 1/\eta + \mathcal{O}(1)$, since $1/(1-\zeta) = 1 + \zeta + \mathcal{O}(\zeta^2)$. Assuming (2.32a), we obtain the following effective amount of work:

$$\mathrm{Eff}(\Phi) = C_\Phi/\eta + \mathcal{O}(1). \tag{2.32b}$$

For instance, the respective numbers in Example 2.28 yield $C_\Phi/\eta = 104$ for the Gauss–Seidel iteration and 7.8 for SOR.

For most of the methods we are going to discuss, assumption (2.32a) holds in the case of the model problem. More precisely, $\eta$ is related to the grid size $h = 1/N = 1/(1+\sqrt{n})$ by (2.32c) with some exponent $\tau > 0$ and a constant $C_\eta$:

$$\eta = C_\eta h^\tau + \mathcal{O}(h^{2\tau}), \quad \text{i.e., } \rho(M) = 1 - C_\eta h^\tau + \mathcal{O}(h^{2\tau}) \quad \text{with } \tau > 0 \tag{2.32c}$$

Inserting this relation into (2.32b), we obtain

$$\mathrm{Eff}(\Phi) = C_{\mathrm{eff}} h^{-\tau} + \mathcal{O}(1) \qquad \text{with } C_{\mathrm{eff}} := C_\Phi/C_\eta. \tag{2.32d}$$

**Remark 2.29.** (a) The exponent $\tau$ in (2.32c) is called the *order of convergence rate*. If an iteration $\Phi$ has a higher order than an iteration $\Psi$, $\Phi$ is more expensive than $\Psi$ for sufficiently small step size $h$. The smaller the order, the better the method.
(b) If $\Phi_1$ and $\Phi_2$ have the same order but different constants $C_{\mathrm{eff},1} < C_{\mathrm{eff},2}$, then $\Phi_2$ is more expensive by a factor of $C_{\mathrm{eff},2}/C_{\mathrm{eff},1}$.

## 2.4 Test of Iterative Methods

In later chapters numerous iterative methods will be defined. For the judgement and presentation of numerical results, one may ask how iterations should be tested.

### 2.4.1 Consistency Test

Because of a bug in the implementation, it may happen that an iterative method is nicely converging, but to a wrong solution. The reason is a violation of consistency. For that reason, one should choose some nontrivial vector $x \in \mathbb{K}^I$ (e.g., defined by random) and compute $b := Ax$. In that case, the solution $x$ of $Ax = b$ is known and one can observe the errors $e^m = x^m - x$.

### 2.4.2 Convergence Test

The quality of an iteration is (at least asymptotically) determined by the effective amount of work $\mathrm{Eff}(\Phi)$. The amount of computational work per iteration is obtained by counting the operations.[4] It remains to determine the convergence speed experimentally. The following trivial remark emphasises the fact that one need not test the method with different right-hand sides $b$ (and thereby with different solutions $x$).

**Remark 2.30.** A linear iteration applied to the two systems $Ax = b$ and $Ax' = b'$ results in the same errors $x^m - x$ and $x'^m - x'$ if the starting values $x^0$ and $x'^0$ are related by $x^0 - x = x'^0 - x'$.

**Conclusion 2.31.** *Without loss of generality, one may always choose* $x = b = 0$, *together with an arbitrary starting value* $x^0 \neq 0$.

According to Remark 2.30, the test of an iteration can be based on the errors $e^m = x^m - x$ and the ratio of their norms,

$$\rho_{m+1,m} := \|e^{m+1}\| / \|e^m\| \qquad \text{(cf. (2.23a))},$$

for one or more starting vectors $e^0$.

Different starting values yield different errors. However, since the geometrical mean $\rho_{m+k,m} = (\|e^{m+k}\|/\|e^m\|)^{1/k}$ (cf. (2.23b)) converges to $\rho(M)$ for $k \to \infty$, the ratios can show remarkable deviations only during the first iteration steps. However, note the following remark.

**Remark 2.32.** In the exceptional case that the starting error $e^0 = x^0 - x$ lies in the subspace $U$ defined in Remark 2.22b, the numbers $\rho_{m+k,m}$ approximate a value smaller than $\rho(M)$.

In practice, meeting this exceptional case is unlikely, in particular, when the solution $x$ is unknown. Furthermore, the usual floating-point errors prevent the iterate $x^m$ from staying in the described subspace.

Computing $\rho_{m+1,m} = \|e^{m+1}\|/\|e^m\|$ requires the knowledge of the exact solution. If we choose $b = 0$ and $x = 0$ according to Conclusion 2.31, $\rho_{m+1,m} = \|x^{m+1}\|/\|x^m\|$ holds. If one wishes to estimate the convergence rate during the iterative computation of an unknown solution $x$, one may use

$$\hat{\rho}_{m+1,m} = \|x^{m+1} - x^m\| / \|x^m - x^{m-1}\|$$

and $\hat{\rho}_{m+k,m} := (\hat{\rho}_{m+k,m+k-1} \cdot \ldots \cdot \hat{\rho}_{m+1,m})^{1/k}$ instead of $\rho_{m+k,m}$.

**Exercise 2.33.** Prove: In spite of $1 \in \sigma(M)$, $\hat{\rho}_{m+k,m} \to \rho < 1 \leq \rho(M)$ may happen for $k \to \infty$. If $1 \notin \sigma(M)$, $\hat{\rho}_{m+k,m} \to \rho(M)$ is valid for all starting errors $e^0 \notin U$ with $U$ defined in Remark 2.22b.

[4] Alternatively, the number of iterations may be replaced with the CPU time.

### 2.4.3 Test by the Model Problem

Deviating from the proposal $x = b = 0$ but according to the choice in §1.6, we define the solution $x$ of the Poisson model problem as the grid function with the components

$$u_{ij} = (ih)^2 + (jh)^2 \qquad (1 \le i, j \le N-1) \tag{2.33a}$$

corresponding to the right-hand side (2.33b) (cf. Remark 1.15):

$$b \text{ defined by (1.6a) with } f = -4. \tag{2.33b}$$

We recall that $u$ and $x$ are different representations of the same quantity (1.6b). The vector $b$ coincides with $f$ in grid points not neighboured to the boundary; otherwise boundary data are added in (1.6a).

### 2.4.4 Stopping Criterion

A comment has to be added concerning the desirable size of the (unavoidable) iteration error $\|e^m\|$. For an unlimited iterative process, the rounding errors prevent the iteration error from converging to zero. Instead, the error will oscillate around $\mathrm{const} \cdot \|x\| \cdot \mathrm{eps}$ ($\mathrm{eps}$ : relative machine precision). For testing an iteration, one may approach this lower limit; in practice, however, there is almost never a reason for such high accuracy.

**Remark 2.34.** The (exact) solution $x$ of the Poisson model problem in §1.2 is only approximating the true solution of the boundary with a discretisation error, which in this case has the order $\mathcal{O}(h^2)$ (cf. Hackbusch [193, §4.5]). Therefore, an additional iteration error of the same order $\mathcal{O}(h^2)$ is acceptable.

The algorithm in §11.5 will automatically yield an approximation for which the discretisation and iteration errors are similar in size.

A more accurate approximation $x^m$ is needed if, e.g., $x^m$ is the starting point of an error estimation (cf. Verfürth [379]) or for the extrapolation to the limit $h \to 0$ ('Richardson extrapolation', cf. Richardson–Gaunt [325], [194, §14.1.1]).

Often, the stopping criterion is based on the defect $Ax^m - b$ (or the residual $b - Ax^m$). Here caution must be exercised: $\|b - Ax^m\|_2 \le 10^{-16}$ might hold, in spite of $\|e^m\|_2 \approx 1$.

**Remark 2.35.** In general, the sizes of $\|b - Ax^m\|_2$ and $\|e^m\|_2$ are not comparable. Their ratio depends not only on the condition $\mathrm{cond}_2(A)$ (cf. §5.6.5.2 and Proposition B.14) but also on the scaling of the vectors $x$ and $b$.