Chapter 8
Semi-Iterative Methods

Abstract The semi-iteration comes in three formulations. The first one in Section 8.1 is the most general and associates each semi-iterate with a polynomial. Using the notion of Krylov spaces, we only require that the errors of the semi-iterates y^m be elements of the Krylov space $x^0 + N\mathcal{K}_m(AN, r^0)$. In the second formulation of Section 8.2, the polynomials p_m associated with y^m are related either by a two-term or by a three-term recursion. Section 8.3 tries to determine the optimal polynomials. Here the result depends on what quantity we want to minimise. Three minimisation problems are discussed. The last formulation is practically solvable and leads to (transformed) Chebyshev polynomials. The corresponding semi-iteration is called the Chebyshev method (cf. §8.3.4). The Chebyshev method improves the order of convergence. Its convergence speed corresponds to the square root of the spectral condition number (cf. §8.3.5). In Section 8.4 the Chebyshev method is applied to the iterations discussed in Part I. In Section 8.5 we describe the ADI method which is not really of the form discussed above, but it might be seen as a generalisation of semi-iterations (replacing scalar parameters by matrix-valued ones).

8.1 First Formulation

8.1.1 Notation

Let $\Phi \in \mathcal{L}$ be a linear and consistent (not necessarily convergent) iteration with an iteration matrix M. In the following Φ is also called the *basic iteration*. Assume that for a starting iterate x^0, the iterates

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

are computed. Up to now, the last computed iterate x^m is regarded as the result of the iterative process. The previously calculated x^j $(0 \le j \le m-1)$ are 'forgotten'. The semi-iterative method is based on a different view. Now, the result of m steps

W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*,
Applied Mathematical Sciences 95, DOI 10.1007/978-3-319-28483-5_8

of the basic iteration Φ is the complete sequence

$$X_m := (x^0, x^1, \ldots, x^m) \in (\mathbb{K}^I)^{m+1}. \tag{8.1}$$

We shall investigate whether a better result than x^m can be constructed from X_m. A semi-iterative method is a mapping

$$\Sigma : \bigcup_{m=0}^{\infty} (\mathbb{K}^I)^{m+1} \to \mathbb{K}^I.$$

The results

$$y^m := \Sigma(X_m) \qquad (m = 0, 1, 2, \ldots)$$

yield a new sequence: the semi-iterative sequence. We shall see that in many cases $\{y^m\}$ converges faster than $\{x^m\}$.

Remark 8.1. The simple example $y^m = \Sigma(x^0, x^1, \ldots, x^m) := x^m$ shows that an optimally chosen semi-iterative method cannot be worse than the basic iteration.

To simplify the notation of polynomials, we introduce the following definition.

Definition 8.2. For $m \in \mathbb{N}_0$, $\mathcal{P}_m$ is the linear space of polynomials of degree $\leq m$ with the underlying field $\mathbb{K}$. $\mathcal{P}_{-1} := \{0\}$ contains the zero polynomial.

8.1.2 Consistency and Asymptotic Convergence Rate

Similar as in Definition 2.5, a semi-iterative method Σ is called *consistent* if equation (8.2) holds for all solutions of $Ax = b$:

$$x = \Sigma(\underbrace{x, x, \ldots, x}_{m+1 \text{ arguments}}) \qquad (m = 0, 1, 2, \ldots). \tag{8.2}$$

The *convergence rate* $\rho = \rho(M)$ can be characterised as the minimal ρ satisfying

$$\lim_{m\to\infty} (\|x^m - x\| / \|x^0 - x\|)^{1/m} \leq \rho \text{ for all } x^0 \neq x \quad \text{(cf. Remark 2.22b).}$$

This characterisation can be transferred to the semi-iterative case.

Definition 8.3. The semi-iterative method has the asymptotic convergence rate ρ, if ρ is the smallest number with

$$\overline{\lim_{m\to\infty}} \left(\|y^m - x\| / \|y^0 - x\|\right)^{1/m} \leq \rho \qquad \left(x = A^{-1}b\right)$$

for all semi-iterative sequences $\{y^m\}$ corresponding to arbitrary starting iterates $y^0 = x^0$.

In the following, we restrict our considerations to linear semi-iterations. $\varSigma$ is called *linear* if $y^m = \varSigma(X_m)$ is a linear combination

$$y^m = \sum_{j=0}^{m} \alpha_{mj}\, x^j \tag{8.3}$$

with coefficients $\alpha_{mj} \in \mathbb{K}$ $(m \in \mathbb{N}_0,\ 1 \leq j \leq m)$. Obviously, a linear semi-iterative method is *consistent* if and only if

$$\sum_{j=0}^{m} \alpha_{mj} = 1 \qquad \text{for all } m = 0, 1, 2, \ldots\ . \tag{8.4}$$

Applying condition (8.4) to $m = 0$, we find that a consistent semi-iterative method satisfies the initial condition

$$y^0 = x^0. \tag{8.5}$$

8.1.3 Error Representation

Theorem 8.4. *Let x be a solution of $Ax = b$, while M denotes the iteration matrix of the basic iteration $\varPhi \in \mathcal{L}$. Then the error*

$$\eta^m := y^m - x \qquad (x = A^{-1}b) \tag{8.6a}$$

admits the representation

$$\eta^m = p_m(M)\, e^0 \qquad \textit{with } e^0 := x^0 - x, \tag{8.6b}$$

where $y^0 = x^0$ (cf. (8.5)) is the starting iterate and p_m is the polynomial

$$p_m(\zeta) = \sum_{j=0}^{m} \alpha_{mj}\, \zeta^j \in \mathcal{P}_m \tag{8.6c}$$

with the coefficients α_{mj} in (8.4).

Proof. Let $e^j = x^j - x$ be the iteration errors of the basic iteration. Subtracting $x = \sum_{j=0}^{m} \alpha_{mj} x$ from $y^m = \sum_{j=0}^{m} \alpha_{mj} x^j$ (cf. (8.2) and (8.4)), we obtain the semi-iterative error

$$\eta^m := y^m - x = \sum_{j=0}^{m} \alpha_{mj}(x^j - x) = \sum_{j=0}^{m} \alpha_{mj} e^j.$$

Inserting the representation $e^j = x^j - x = M^j e^0$ (cf. (2.16b)), we arrive at

$$\eta^m = \sum_{j=0}^{m} \alpha_{mj} \left(M^j e^0\right) = \Big(\sum_{j=0}^{m} \alpha_{mj} M^j\Big) e^0 = p_m(M) e^0. \qquad \square$$

Theorem 8.4 associates the linear semi-iteration Σ with a family of polynomials

$$\{p_m \in \mathcal{P}_m : m = 0, 1, \ldots\}.$$

Vice versa, any sequence $\{p_m \in \mathcal{P}_m\}$ of polynomials defines a semi-iterative method by means of its coefficients α_{mj}.

Remark 8.5. (a) A linear semi-iterative method Σ is uniquely described by the family of associated polynomial sequence $\{p_m \in \mathcal{P}_m\}$. Σ is consistent if and only if

$$p_m(1) = 1 \qquad \text{for } m = 0, 1, \ldots\ . \tag{8.6d}$$

(b) Let the basic iteration with iteration matrix M be consistent. Then the semi-iterates y^m have the representation[1]

$$y^m = M_m x^0 + N_m b \quad \text{with } M_m := p_m(M),\ N_m := (I - M_m) A^{-1}. \tag{8.7}$$

(c) The asymptotic convergence rate is equal to

$$\overline{\lim_{m\to\infty}}\ \rho(p_m(M))^{1/m}.$$

If M is diagonalisable, the quantity above coincides with $\overline{\lim}\, \|p_m(M)\|^{1/m}$. The equality

$$\overline{\lim}\, \rho(p_m(M))^{1/m} = \overline{\lim}\, \|p_m(M)\|^{1/m}$$

is not valid in general, but holds for many important polynomial sequences p_m (cf. Eiermann–Niethammer–Varga [120]).

From (8.6d), we derive an alternative characterisation of p_m.

Remark 8.6. (a) Any polynomial $p_m \in \mathcal{P}_m$ satisfying the consistency condition (8.6d) is uniquely associated with a polynomial $q_m \in \mathcal{P}_{m-1}$ so that

$$p_m(\zeta) = 1 - (1 - \zeta)\, q_m(1 - \zeta)\,. \tag{8.8}$$

(b) $M = I - NA$ (cf. (2.9′)) yields

$$p_m(M) = I - NA\, q_m(NA).$$

[1] The expression $I - M_m$ has the form $X_m A$ so that $(I - M_m)A^{-1} = X_m$ is well defined also for singular A.

8.1.4 Krylov Space

Definition 8.7. The Krylov space associated with a matrix $X \in \mathbb{K}^{I\times I}$ and with a vector $v \in \mathbb{K}^I$ is defined by

$$\mathcal{K}_m(X,v) := \operatorname{span}\{v, Xv, \ldots, X^{m-1}v\} \qquad \text{for } m \in \mathbb{N},$$

while $\mathcal{K}_0(X,v) := \{0\}$ (cf. Aleksey Nikolaevich Krylov [249]).

Exercise 8.8. Let $\mathcal{U} = \operatorname{span}\{u^1, \ldots, u^m\}$ be a subspace of $\mathbb{K}^I$.

(a) Prove that $\operatorname{span}\{\mathcal{U}, x\} = \operatorname{span}\{\mathcal{U}, y\}$ for any x, y with $x - y \in \mathcal{U}$.

(b) Let $A \in \mathbb{K}^{I\times I}$ be any matrix. $A\mathcal{U}$ abbreviates the subspace $\{Ax : x \in \mathcal{U}\}$. Prove that $A\mathcal{U} = \operatorname{span}\{Au^1, \ldots, Au^m\}$.

Since the monomials $\{1, x, \ldots, x^{m-1}\}$ span the space $\mathcal{P}_{m-1}$ of polynomials of degree $\leq m-1$, we obtain the first statement of the next remark. There we use the notation $v + \mathcal{U} := \{v + u : u \in \mathcal{U}\}$ for the *affine subspace* with a subspace $\mathcal{U} \subset \mathbb{K}^I$ and a vector $v \in \mathbb{K}^I$. The *residual* of an approximation $\tilde{x}$ is defined by $r := b - A\tilde{x}$ and is the negative defect (2.17).

Proposition 8.9. *(a) The connection with matrix polynomials is given by*

$$\mathcal{K}_m(X,v) = \{p(X)v : p \in \mathcal{P}_{m-1}\}.$$

(b) Assume that the iteration Φ with the iteration matrix $M = I - NA$ yields the iterates x^m with the errors $e^m = x^m - x$ and the residuals $r^m := b - Ax^m$. They satisfy

$$\begin{aligned}
x^m &\in x^0 + N\mathcal{K}_m(AN, r^0) &&= x^0 + NA\mathcal{K}_m(NA, e^0) \subset x + \mathcal{K}_{m+1}(NA, e^0),\\
e^m &\in e^0 + N\mathcal{K}_m(AN, r^0) &&= e^0 + NA\mathcal{K}_m(NA, e^0) \subset \mathcal{K}_{m+1}(NA, e^0),\\
r^m &\in r^0 + AN\mathcal{K}_m(AN, r^0) \subset \mathcal{K}_{m+1}(AN, r^0),
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{span}\{e^0, \ldots, e^{m-1}\} &= \mathcal{K}_m(M, e^0) = \mathcal{K}_m(NA, e^0),\\
\operatorname{span}\{r^0, \ldots, r^{m-1}\} &= \mathcal{K}_m(M, r^0) = \mathcal{K}_m(NA, r^0).
\end{aligned}$$

(c) The following identity holds for regular T:

$$T\mathcal{K}_m(X,v) = \mathcal{K}_m(TXT^{-1}, Tv).$$

(d) For $m \in \mathbb{N}_0$, we have

$$X\mathcal{K}_m(X,v) \subset v + X\mathcal{K}_m(X,v) \subset \operatorname{span}\{v\} + X\mathcal{K}_m(X,v) = \mathcal{K}_{m+1}(X,v). \tag{8.9}$$

Proof. The statements in part (b) follow by induction. Note that $\mathcal{K}_m(M, v) = \mathcal{K}_m(NA, v)$ holds for all v since a polynomial in $M = I - NA$ can be written as a polynomial of same degree in NA. The inclusions use part (d).

Part (c) is a consequence of Exercise A.16a. □

Definition 8.10. The degree of a vector $v \in \mathbb{K}^I$ (with respect to a matrix $X \in \mathbb{K}^{I \times I}$) is defined by

$$\deg_X(v) := \min\{m \in \mathbb{N}_0 : p(X)v = 0 \text{ for } p \in \mathcal{P}_m \text{ with } \mathrm{degree}(p) = m\}.$$

Exercise 8.11. For $m \in \mathbb{N}$, prove: (a) $\dim(\mathcal{K}_m(X, v)) = \min\{m, \deg_X(v)\} \leq m$.

(b) If $\dim(\mathcal{K}_{m+1}(X, v)) = \dim(\mathcal{K}_m(X, v))$, then $\mathcal{K}_{m+1}(X, v) = \mathcal{K}_m(X, v)$. If, in addition, X is regular, $X\mathcal{K}_m(X, v) = \mathcal{K}_m(X, v)$ also holds.

(c) $\deg_X(v) = 0$ holds if and only if $v = 0$, while $\deg_X(v) = 1$ characterises all eigenvectors of X.

(d) $\deg_X(v) \leq \mathrm{degree}(\mu_X) \leq \#I$, where μ_X is the minimum function (A.16c).

(e) Any $w \in \mathcal{K}_m(X, v)$ is characterised by a polynomial $p \in \mathcal{P}_{m-1}$ via $w = p(X)v$. If $\dim(\mathcal{K}_m(X, v)) = m$, this polynomial is unique.

Lemma 8.12. *For any $v \in \mathbb{K}^I$ and any regular matrix X, the polynomial p with $p(X)v = 0$ and $\mathrm{degree}(p) = \deg_X(v)$ satisfies $p(0) \neq 0$.*

Proof. If $p(0) = 0$, there is a polynomial $q \in P_{\deg_X(v)-1}$ with $p(\xi) = \xi\, q(\xi)$. Hence $0 = p(X)v = Xq(X)v$ implies that $q(X)v = 0$ in contradiction to the minimality of $\deg_X(v)$. □

Combining Proposition 8.9a with Theorem 8.4 and repeating the arguments of Proposition 8.9, we obtain the next statement.

Conclusion 8.13. *(a) The first formulation of a semi-iteration is equivalent to*

$$y^m \in x^0 + N\mathcal{K}_m(AN, r^0) \subset x + \mathcal{K}_{m+1}(NA, e^0),$$

where $x := A^{-1}b$. The polynomial (8.6c) coincides with the polynomial associated with the error $\eta^m = y^m - x \in \mathcal{K}_{m+1}(NA, e^0)$ in (8.6a) by Exercise 8.11e.

(b) If the polynomials in (8.6c) satisfy $\mathrm{degree}(p_\mu) = \mu$, the errors η^m span

$$\mathcal{K}_m(M, e^0) = \mathrm{span}\{\eta^0, \eta^1, \ldots, \eta^{m-1}\}.$$

(c) The residuals $r^m = -A\eta^m = b - Ay^m$ of the semi-iterates span the space $A\,\mathrm{span}\{\eta^0, \ldots, \eta^{m-1}\}$. Under the conditions of part (b), Proposition 8.9c yields

$$\mathrm{span}\{r^0, r^1, \ldots, r^{m-1}\} = A\,\mathcal{K}_m(M, e^0) = \mathcal{K}_m(AN, r^0).$$

8.2 Second Formulation of a Semi-Iterative Method

8.2.1 General Representation

The representation used in §8.1 requires storing all iterates $(x^0, x^1, \ldots, x^m)$, which is not desirable in the case of large m and high-dimensional systems. Since, in §8.1, the definition of $y^m = \Sigma(X_m)$ is completely independent of the previous iterates $y^j = \Sigma(X_j)$ $(0 \le j \le m-1)$, it is in general not possible to use the semi-iterative results $y^0, \ldots, y^{m-1}$ for computing y^m.

This situation changes in the second formulation. Let $\Phi \in \mathcal{L}$ be the basic iteration. After starting with

$$y^0 = x^0 \qquad \text{(cf. (8.5))}, \tag{8.10a}$$

we compute the iterates recursively by

$$y^m = \vartheta_m \Phi(y^{m-1}, b) + (1 - \vartheta_m) y^{m-1} \qquad (m \ge 1) \tag{8.10b}$$

with *extrapolation factors* $\vartheta_m \in \mathbb{K}$ $(m \in \mathbb{N})$ that may be chosen arbitrarily.

Exploiting the normal forms $\Phi(x, b) = Mx + Nb = x - N(Ax - b)$, equation (8.10b) can be written in the form (8.10b′) or (8.10b′/”):

$$\begin{aligned} y^m &= \vartheta_m \left(M y^{m-1} + Nb\right) + (1 - \vartheta_m)\, y^{m-1}, && \text{(8.10b}'\text{)} \\ y^m &= y^{m-1} - \vartheta_m N(A y^{m-1} - b) \;=\; \Phi_{\vartheta_m}(y^{m-1}, b). && \text{(8.10b}'/\text{”)} \end{aligned}$$

Formulae (8.10b′,b′′) represent one step of the damped version Φ_{ϑ_m} of the basic iteration (cf. §5.2.2), however with a parameter ϑ_m depending on m.

Below we state that recursion (8.10a,b) yields a semi-iterative method.

Theorem 8.14. *For arbitrary factors $\vartheta_m \in \mathbb{K}$ $(m \in \mathbb{N})$, algorithm (8.10a,b) defines a linear and consistent semi-iteration Σ. The polynomials $\{p_m \in \mathcal{P}_m\}$ describing Σ are recursively defined by*

$$p_0(\zeta) = 1, \qquad p_m(\zeta) = (\vartheta_m \zeta + 1 - \vartheta_m)\, p_{m-1}(\zeta) \qquad (m \in \mathbb{N}). \tag{8.11}$$

Proof. (i) One shows by induction that the polynomials p_m in (8.11) satisfy the consistency condition (8.6d): $p_m(1) = 1$. Also $\operatorname{degree}(p_m) \le m$ is obvious.

(ii) The basic iteration Φ is assumed to be consistent. By construction (8.10b′), the first matrix M_m in the representation $y^m = M_m x^0 + N_m b$ has the form $M_m = \vartheta_m M M_{m-1} + (1 - \vartheta_m) M_{m-1}$, where $M_0 = I$. According to (8.7), the polynomials in (8.11) lead to the same matrix $M_m = p_m(M)$. Since these matrices uniquely determine y^m because of $N_m := (I - M_m)A^{-1}$ (using the consistency of Φ), the method (8.10a,b) coincides with the semi-iteration defined by the polynomials (8.11). The case of an inconsistent basic iteration is left to the reader (proof by induction). □

The case $\vartheta_m = 0$ is uninteresting because of $y^m = y^{m-1}$. Therefore, we assume that $\vartheta_m \neq 0$. The set of all methods representable by (8.10a,b) is characterised next.

Lemma 8.15. *Let $\Phi \in \mathcal{L}$ be the basic iteration and assume $\vartheta_m \neq 0$ in (8.10b). Then the second formulation (8.10a,b) represents exactly those linear and consistent semi-iterations for which the associated polynomials p_m satisfy (8.6d) and*

$$\text{degree}(p_m) = m, \qquad p_{m-1} \textit{ is a divisor of } p_m \quad \textit{for all } m \geq 1. \tag{8.12a}$$

Given polynomials $\{p_m\}$ with (8.6d) and (8.12a), the extrapolation factors ϑ_m of the equivalent representation (8.10a,b) are determined by

$$\frac{p_m(\zeta)}{p_{m-1}(\zeta)} = 1 + \vartheta_m(\zeta - 1). \tag{8.12b}$$

Proof. In the case of $\vartheta_m \neq 0$, the method (8.10a,b) leads to polynomials (8.11) satisfying $\text{degree}(p_m) = m$; hence, (8.12a) is satisfied. Vice versa, under the assumption (8.12a), p_m/p_{m-1} must be a polynomial of the form (8.12b). □

The example of recursion (8.10a,b) shows that the mapping $X_m \mapsto y^m = \Sigma(X_m)$ does not need the iterates of X_m explicitly. Since X_m is uniquely determined by x^0, there is a mapping $\Xi : x^0 \mapsto y^m$ for $y^m = \Sigma(X_m)$. Recursion (8.10a,b) describes such a mapping Ξ.

By Lemma 8.15, the semi-iterate y^m for a fixed m can be produced as follows.

Remark 8.16. y^m is connected with a polynomial p_m. Let

$$p_m(\zeta) = c_m \prod\nolimits_{\nu=1}^{m} (\zeta - \zeta_\nu) \quad \text{with} \quad c_m = 1/\prod\nolimits_{\nu=1}^{m} (1 - \zeta_\nu) \tag{8.13a}$$

be a factorisation into linear factors (possibly with complex ζ_ν) and define auxiliary polynomials $\hat{p}_\mu$ for $0 \leq \mu \leq m$ by

$$\hat{p}_\mu(\zeta) = \prod\nolimits_{\nu=1}^{\mu} \frac{\zeta - \zeta_\nu}{1 - \zeta_\nu}. \tag{8.13b}$$

Set $\vartheta_\mu := \frac{1}{1-\zeta_\mu}$ for $0 < \mu \leq m$. Then all polynomials $\hat{p}_\mu$ satisfy (8.12a,b) and $\hat{p}_m = p_m$. The corresponding semi-iteration

$$\hat{y}^\mu = \Phi_{\vartheta_\mu}(\hat{y}^{\mu-1}, b) \qquad (1 \leq \mu \leq m; \ \text{cf. (8.10a,b)})$$

is as easy to perform and yields $\hat{y}^m = y^m$ (only for $\mu = m$, not for $\mu < m$). However, this approach has severe disadvantages.

1. To compute the next y^{m+1}, we have to perform (8.10a,b) again from $\mu = 0$ to $\mu = m + 1$, since then other auxiliary polynomials $\hat{p}_\mu$ are needed.
2. The second formulation (8.10a,b) may be unstable. For relative small m, the rounding error influence of the iteration errors $y^m - x$ can already predominate. It is possible to avoid instability by a suitable renumbering of the ϑ_ν. Concerning the stability analysis and the choice of an appropriate ordering, we refer to Lebedev–Finogenov [261, 262] (cf. also Samarskii–Nikolaev [330, §6.2.4]).

It will turn out that the three-term recursion described next is the best representation of the polynomials.

8.2.2 Three-Term Recursion

Algorithm (8.10b) determines y^m from y^{m-1}. Alternatively, a three-term recursion connects y^m with y^{m-1} and y^{m-2} (cf. §2.2.8):

$$y^0 = x^0, \tag{8.14a}$$
$$y^1 = (1-\tfrac{1}{2}\vartheta_1)\,x^1 + \tfrac{1}{2}\vartheta_1 x^0 \;=\; (1-\tfrac{1}{2}\vartheta_1)\,\Phi(x^0,b) + \tfrac{1}{2}\vartheta_1 x^0, \tag{8.14b}$$
$$y^m = \Theta_m\bigl[\Phi(y^{m-1},b) - y^{m-2}\bigr] + \vartheta_m\,(y^{m-1} - y^{m-2}) + y^{m-2}. \tag{8.14c}$$

From $\Phi(x,b) = Mx + Nb = x - N(Ax-b),$ we obtain the representations

$$\begin{aligned} y^1 &= (1-\tfrac{1}{2}\vartheta_1)(Mx^0+Nb) + \tfrac{1}{2}\vartheta_1 x^0 \\ &= x^0 - (1-\tfrac{1}{2}\vartheta_1)\,N(Ax^0-b), \\ y^m &= \Theta_m\left(My^{m-1}+Nb-y^{m-2}\right) + \vartheta_m(y^{m-1}-y^{m-2}) + y^{m-2} \\ &= (1+\vartheta_m+\Theta_m)y^{m-2} + (\vartheta_m+\Theta_m)(y^{m-1}-y^{m-2}) - \Theta_m N(Ay^{m-1}-b). \end{aligned}$$

Analogous to Theorem 8.14, one proves the next theorem.

Theorem 8.17. *For arbitrary factors Θ_m and ϑ_m, algorithm (8.14a–c) defines a linear and consistent semi-iteration Σ. The polynomials $\{p_m\}$ describing Σ are recursively defined by*

$$p_0(\zeta) = 1, \qquad p_1(\zeta) = \bigl(1-\tfrac{1}{2}\vartheta_1\bigr)\zeta + \tfrac{1}{2}\vartheta_1, \tag{8.15a}$$
$$p_m(\zeta) = \bigl(\Theta_m\zeta + \vartheta_m\bigr)p_{m-1}(\zeta) + \bigl(1-\Theta_m-\vartheta_m\bigr)p_{m-2}(\zeta). \tag{8.15b}$$

For the particular choice $\vartheta_m = 0,$ the recursion becomes

$$p_0(\zeta) = 1, \qquad p_1(\zeta) = \zeta, \tag{8.15c}$$
$$p_m(\zeta) = \Theta_m\bigl[\,\zeta\, p_{m-1}(\zeta) - p_{m-2}(\zeta)\,\bigr] + p_{m-2}(\zeta). \tag{8.15d}$$

We remark that all orthogonal polynomials can be generated by recursion of the form (8.15a,b) (cf. Quarteroni–Sacco–Saleri [314, §10.1]).

Exercise 8.18. Prove that the polynomials q_m in (8.8) associated with p_m and defined either in (8.15a,b) or (8.15c,d) can be determined by the recursion

$$\begin{aligned} &q_0(\xi) = 0, \qquad q_1(\xi) = 1-\tfrac{1}{2}\vartheta_1, \\ &q_m(\xi) = \Theta_m + (1-\Theta_m-\vartheta_m)\,q_{m-2}(\xi) + (\Theta_m\,(1-\xi)+\vartheta_m)\,q_{m-1}(\xi) \end{aligned}$$

or, respectively,

$$\begin{aligned} &q_0(\xi) = 0, \qquad q_1(\xi) = 1, \\ &q_m(\xi) = \Theta_m + (1-\Theta_m)\,q_{m-2}(\xi) + \Theta_m\,(1-\xi)\,q_{m-1}(\xi). \end{aligned}$$

8.3 Optimal Polynomials

Since the semi-iterates are completely determined by polynomials, we can ask for the best polynomials in the sense that the corresponding semi-iteration is as fast as possible. The quantity to be minimised is still to be specified. It might be a certain norm of the error (cf. Problem 8.19) or the convergence rate (cf. Problem 8.21) or an upper bound of the error (cf. Problem 8.20).

8.3.1 Minimisation Problem

Let Σ be a linear and consistent semi-iteration. By Theorem 8.4, the semi-iteration error $\eta^m = y^m - x$ has the representation (8.6b):

$$\eta^m = p_m(M)e^0.$$

Therefore, it seems reasonable to pose the following problem.

Problem 8.19 (first minimisation problem). Given $m \in \mathbb{N}$, determine a polynomial $p_m \in \mathcal{P}_m$ satisfying (8.6d), i.e.,

$$p_m(1) = 1, \tag{8.16}$$

such that

$$\|p_m(M)e^0\|_2 \overset{!}{=} \min, \tag{8.17}$$

i.e., $\|p_m(M)e^0\|_2 \le \|q_m(M)e^0\|_2$ for all admissible polynomials.

The solution of (8.17) seems hopeless, since the unknown error $e^0 = x^0 - x$ is involved in the problem (if e^0 were known, $x = x^0 - e^0$ already represents the solution). Nevertheless, we shall solve this problem with respect to the energy norm instead of $\|\cdot\|_2$ in §9.3 (cf. Remark 10.12).

Even if e^0 is unknown, $\|p_m(M)e^0\|_2$ can be estimated by

$$\|p_m(M)e^0\|_2 \le \|p_m(M)\|_2 \, \|e^0\|_2$$

and the factor $\|p_m(M)\|_2$ can be minimised separately.

Problem 8.20 (second minimisation problem). Given $m \in \mathbb{N}$, determine a polynomial $p_m \in \mathcal{P}_m$ with (8.16) such that

$$\|p_m(M)\|_2 \overset{!}{=} \min\,. \tag{8.18}$$

8.3.2 Discussion of the Second Minimisation Problem

A partial answer to the minimisation problems follows in Theorem A.37 (Cayley–Hamilton). Assume that M has no eigenvalue $\lambda = 1$ ($\rho(M) < 1$ is sufficient). For all $m \geq n := \#I$, the choice $p_m(\lambda) = \chi(\lambda) := \det(\lambda I - M)/\det(I - M)$ leads to a polynomial with the properties (8.16) and $\mathrm{degree}(p_m) \leq m$ solving problems (8.17) and (8.18). In particular, (8.19) holds:

$$p_m(M) = 0 \qquad \text{and} \qquad \|p_m(M)\| = 0. \tag{8.19}$$

The minimum function $p_m(\lambda) = \mu(\lambda)$ of M (cf. (A.16c)) already satisfies (8.19) for $m \geq m_\mu := \mathrm{degree}(\mu)$.

The solution given in (8.19) is unsatisfactory for two reasons. First, the characteristic polynomial χ (more precisely, its coefficients) is not easy to compute; second, the case $m \geq n$ is rather uninteresting.

Intermediately, we require that

$$M \text{ be normal,} \tag{8.20}$$

i.e., $MM^{\mathsf{H}} = M^{\mathsf{H}}M$ (M being Hermitian would be sufficient). Since then $p_m(M)$ is also normal, Theorem B.25 implies that

$$\|p_m(M)\|_2 = \rho(p_m(M)) = \max\{|p_m(\lambda)| : \lambda \in \sigma(M)\}.$$

Therefore, minimising (8.18) is equivalent to determining a polynomial whose absolute value is minimal on the set $\sigma(M)$. Even if the normality (8.20) does not hold, minimisation of $\max\{|p_m(\lambda)| : \lambda \in \sigma(M)\}$ makes sense. The new minimisation problem is

$$\rho(p_m(M)) = \max\{|p_m(\lambda)| : \lambda \in \sigma(M)\} \overset{!}{=} \min, \tag{8.21a}$$

i.e., the spectral radius is minimised over all admissible polynomial in $\mathcal{P}_m$ instead of the spectral norm $\|p_m(M)\|_2$.

For the next interpretation, we assume that $M = T^{-1}DT$ (D diagonal matrix) is diagonalisable. This leads to $p_m(M) = p_m(T^{-1}DT) = T^{-1}p_m(D)T$. Using the norm $|||\cdot|||_T$ defined in Exercise B.13c, we obtain

$$\begin{aligned} |||\, p_m(M)\, |||_T &= \|Tp_m(M)T^{-1}\|_2 = \|p_m(D)\|_2 = \rho(p_m(D)) \\ &= \rho(p_m(M)) = \max\{|p_m(\lambda)| : \lambda \in \sigma(M)\}. \end{aligned} \tag{8.21b}$$

Alternatively, we may estimate by

$$\|p_m(M)\|_2 \leq \mathrm{cond}_2(T)\|p_m(D)\| = \mathrm{cond}_2(T)\,\rho(p_m(M)). \tag{8.21c}$$

Hence minimising the spectral radius $\rho(p_m(M))$ in (8.21a) minimises the upper bound $\mathrm{cond}_2(T)\|p_m(D)\|$ in (8.21c).

According to §5.1.2, symmetric iterations have the property that $A > 0$ implies that $A^{1/2}MA^{-1/2}$ is also Hermitian. Then the energy norm of $p_m(M)$ is well defined and equal to

$$\|p_m(M)\|_A = \max\{|p_m(\lambda)| : \lambda \in \sigma(M)\} = \rho(p_m(M)). \tag{8.21d}$$

The minimisation of $\max\{|p_m(\lambda)| : \lambda \in \sigma(M)\}$ can only be solved with the knowledge of the spectrum $\sigma(M)$. Computing the complete spectrum, however, would be by far more expensive than the solution of the system.

As a remedy, we assume that there is an a priori known superset

$$\sigma_M \supset \sigma(M)$$

containing the spectrum. Then $\sigma(M)$ can be replaced with σ_M. An example for the larger set σ_M is the complex circle

$$\sigma_M = \{\lambda \in \mathbb{C} : |\lambda| \leq \bar{\rho}\} \qquad \text{with } \bar{\rho} \geq \rho(M).$$

Unfortunately, this circle is inappropriate for our purposes as we shall see in Theorem 8.32. If, however, M has only real eigenvalues, the interval

$$\sigma_M = [-\bar{\rho}, \bar{\rho}] \qquad \text{with } \bar{\rho} \geq \rho(M) \tag{8.22a}$$

is a candidate. In some cases, it is known that M has only nonnegative eigenvalues (cf. Theorem 3.34c). Then one may choose

$$\sigma_M = [0, \bar{\rho}] \qquad \text{with } \bar{\rho} \geq \rho(M). \tag{8.22b}$$

In all cases, it is sufficient to know an upper bound $\bar{\rho}$ of $\rho(M)$, where $\bar{\rho} = \rho(M)$ would be optimal and $\bar{\rho} < 1$ must hold. For instance, we may choose $\bar{\rho}$ as $\rho_{m+k,k}$ in (2.23b) for suitable m and k (cf. Remark 2.32).

Accordingly, the minimisation of $\|p_m(M)\|_2$ in Problem 8.20 is replaced with the following minimisation.

Problem 8.21 (third minimisation problem). Given $m \in \mathbb{N}$ and σ_M, determine a polynomial $p_m \in \mathcal{P}_m$ with (8.16) such that

$$\max\{|p_m(\lambda)| : \lambda \in \sigma_M\} \overset{!}{=} \min. \tag{8.23}$$

Finally, we briefly discuss the choice of alternative norms in (8.17) and (8.18). A non-Hilbert norm (as, e.g., the maximum or row-sum norm $\|\cdot\|_\infty$) leads to a considerably more complicated minimisation problem. It would be possible to replace the Euclidean norm $\|\cdot\|_2$ by $|||\ x\ |||_T = \|Tx\|_2$ or $\|x\|_K = \|K^{1/2}x\|_2$ (K positive definite) as already done in (8.21b,d). Examples for K would be A and the matrix W of the third normal form (cf. (3.35e) and (8.21d)).

8.3.3 Chebyshev Polynomials

As a preparation for the next section we discuss the Chebyshev polynomials.

Definition 8.22. The Chebyshev polynomials T_m are defined by

$$T_m(x) := \cos(m \arccos x) \qquad \text{for } m \in \mathbb{N}_0,\ -1 \le x \le 1. \tag{8.24}$$

Part (a) of the following theorem summarising all properties needed later shows that the functions T_m are in fact polynomials of degree m.

Lemma 8.23. *(a) The functions T_m in (8.24) fulfil the recursion*

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{m+1}(x) = 2xT_m(x) - T_{m-1}(x). \tag{8.25a}$$

(b) For $x \ge 1,$ the polynomials T_m have the representation

$$T_m(x) = \cosh(m \operatorname{arcosh} x) \qquad \textit{for } m \in \mathbb{N}_0,\ x > 1, \tag{8.25b}$$

where $\cosh(x) = \frac{e^x + e^{-x}}{2}$ *is the hyperbolic cosine, while* arcosh *(area-hyperbolic cosine) is its inverse function.*

(c) For all $x \in \mathbb{C}$, the representation (8.25c) holds:

$$T_m(x) = \frac{1}{2}\left[\left(x + \sqrt{x^2 - 1}\right)^m + \left(x + \sqrt{x^2 - 1}\right)^{-m}\right]. \tag{8.25c}$$

Proof. Eqs. (8.25a) follows from the cosine addition theorem. For (8.25b), it is sufficient to prove that the functions defined there also satisfy recursion (8.25a). Substituting $x = \cos\zeta$, we see that (8.25c) coincides with $\cos(m\zeta) = T_m(x)$. □

$\{T_m\}$ are orthogonal polynomials with respect to the weight function $\frac{1}{\sqrt{1-x^2}}$, i.e., $\int_{-1}^{1} \frac{T_m(x)\, T_n(x)}{\sqrt{1-x^2}}\, dx = 0$ for $n \ne m$ (cf. Quarteroni–Sacco–Saleri [314, §10.1.1]).

8.3.4 Chebyshev Method (Solution of the Third Minimisation Problem)

As in the examples (8.22a,b), we assume that σ_M is a real interval. The solution to the third minimisation problem (8.23) is given below.

Notation 8.24. In the following, the real numbers a, b with $-\infty < a \le b < 1$ define an interval with the property

$$\sigma_M = [a, b] \supset \sigma(M). \tag{8.26a}$$

Because $M = I - NA = I - W^{-1}A$ (cf. (2.9)), inclusion (8.26a) is equivalent to

$$[\gamma, \mathit{\Gamma}] \supset \sigma(NA) = \sigma(W^{-1}A) \tag{8.26b}$$

with

$$\gamma = 1 - b, \quad \mathit{\Gamma} = 1 - a. \tag{8.26c}$$

Note that $0 < \gamma \le \mathit{\Gamma} < \infty$. Often, the use of γ and $\mathit{\Gamma}$ leads to simpler formulae. In particular, the ratio

$$\kappa = \mathit{\Gamma}/\gamma \tag{8.26d}$$

is of interest. If the inclusion (8.26a) is strict, i.e., $a, b \in \sigma(M)$, $[\gamma, \mathit{\Gamma}] \supset \sigma(NA)$ is also strict and $\kappa = \kappa(NA)$ is the spectral number defined in (B.13).

Lemma 8.25. *Let* $[a, b]$ *be an interval with* $-\infty < a \le b < 1$. *The problem*

minimise $\max\{|p_m(\lambda)| : a \le \lambda \le b\}$
with respect to all polynomials $p_m \in \mathcal{P}_m$ *and* $p_m(1) = 1$

has the unique solution

$$p_m(\zeta) = T_m\left(\tfrac{2\zeta - a - b}{b - a}\right)/C_m \quad \textit{with } C_m := T_m\left(\tfrac{2-a-b}{b-a}\right) = T_m\left(\tfrac{\mathit{\Gamma}+\gamma}{\mathit{\Gamma}-\gamma}\right). \tag{8.27a}$$

Here, $\gamma, \mathit{\Gamma}$ *are as in (8.26c) and* T_m *is the Chebyshev polynomial defined in (8.24). The minimising polynomial* p_m *has the degree* m *and leads to the minimum*

$$\max\{|p_m(\lambda)| : a \le \lambda \le b\} = 1/C_m \qquad \textit{for } p_m \textit{ in (8.27a).} \tag{8.27b}$$

Proof. (i) The constant C_m does not vanish, since the argument $\frac{2-a-b}{b-a}$ lies outside of $[-1, 1]$ and the representation (8.25b) applies. By construction, $p_m(1) = 1$ and $\mathrm{degree}(p_m) = m$ hold. For $a \le \zeta \le b$, the argument $\frac{2\zeta-a-b}{b-a}$ belongs to $[-1, 1]$. Definition (8.24) shows that $|T_m| \le 1$ in $[-1, 1]$. Since T_m attains the bounds ± 1, the statement (8.27b) follows.

(ii) It remains to show that for any other polynomial the maximum in (8.27b) is larger than $1/C_m$. Let $q_m \in \mathcal{P}_m$ be a polynomial with $q_m(1) = 1$ and $\max\{|q_m(\lambda)| : \gamma \le \lambda \le \mathit{\Gamma}\} \le 1/C_m$. The Chebyshev polynomial $T_m(x) = \cos(m \arccos x)$ meets the values ± 1 in alternating ordering at $x = \cos \frac{n\pi}{m}$ for $n = -m, 1 - m, \ldots, 0$. The function p_m obtained from T_m by transforming $x \mapsto \zeta = \frac{1}{2}[a + b + x(b-a)]$ is $p_m(\frac{1}{2}[a + b + x(b-a)]) := T_m(x)$ and has the values

$$p_m(\zeta_\nu) = (-1)^\nu / C_m \qquad (-m \le \nu \le 0)$$

at $\zeta_\nu = \frac{1}{2}\left[a + b + (b - a)\cos\frac{\nu\pi}{m}\right]$. From $|q_m(\zeta_\nu)| \le 1/C_m = |p_m(\zeta_\nu)|$, we conclude that the difference $r := p_m - q_m$ satisfies

$$r(\zeta_\nu) \ge 0 \quad \text{for even } \nu, \qquad r(\zeta_\nu) \le 0 \quad \text{for odd } \nu.$$

By the intermediate value theorem, there exists at least one zero of r in each subinterval $[\zeta_{\nu-1}, \zeta_\nu]$ $(1 - m \le \nu \le 0)$. If the zeros in $[\zeta_{\nu-1}, \zeta_\nu]$ and $[\zeta_\nu, \zeta_{\nu+1}]$

coincide at the common point ζ_ν, this is a double zero. Hence, counted with respect to multiplicity, r has at least m zeros in $[a,b]$. By $p_m(1) = q_m(1) = 1$, the value $1 \notin [a,b]$ represents the $(m{+}1)$-th zero of r. Hence, $r = 0$ follows from $\operatorname{degree}(r) \le m$, proving that $p_m = q_m$ is unique. $\square$

Exercise 8.26. (a) Prove by means of (8.25a) that the polynomials p_m in (8.27a) can be obtained by the recursion

$$p_0(\zeta) = 1, \qquad p_1(\zeta) = \frac{2\zeta - a - b}{2 - a - b}, \tag{8.28a}$$

$$C_{m+1}\, p_{m+1}(\zeta) = 2\frac{2\zeta - a - b}{2 - a - b}\, C_m\, p_m(\zeta) - C_{m-1}\, p_{m-1}(\zeta). \tag{8.28b}$$

(b) Let $\vartheta_{\text{opt}} = \frac{2}{\Gamma+\gamma}$ (cf. (6.6a)). Prove that

$$p_m(I - NA) = \frac{1}{C_m} T_m\big(\tfrac{\Gamma+\gamma}{\Gamma-\gamma}\, I + \tfrac{2}{\Gamma-\gamma}\, NA\big) = \frac{1}{C_m} T_m\big(\tfrac{\Gamma+\gamma}{\Gamma-\gamma}\big[I + \vartheta_{\text{opt}}\, NA\big]\big).$$

To investigate the minimum $1/C_m = 1/T_m(\frac{2-a-b}{b-a})$ reached in (8.27b), we have to evaluate (8.25c) at

$$x_0 := \frac{2 - a - b}{b - a} = \frac{\Gamma + \gamma}{\Gamma - \gamma}$$

with γ and Γ defined in (8.26c). We use that $x_0^2 - 1 = 4\gamma\Gamma/(\Gamma - \gamma)^2 > 0$ and $x_0 + \sqrt{x_0^2 - 1} = \big(\sqrt{\Gamma} + \sqrt{\gamma}\big)^2/(\Gamma - \gamma)$. The representation (8.25c) shows that

$$C_m = \frac{1}{2}\left\{\left(\frac{(\sqrt{\Gamma} + \sqrt{\gamma})^2}{\Gamma - \gamma}\right)^m + \left(\frac{(\sqrt{\Gamma} + \sqrt{\gamma})^2}{\Gamma - \gamma}\right)^{-m}\right\}.$$

The bracket $\frac{(\sqrt{\Gamma}+\sqrt{\gamma})^2}{\Gamma-\gamma}$ can be rewritten as $\frac{\Gamma}{\Gamma-\gamma}\big(1 + \sqrt{\frac{\gamma}{\Gamma}}\big)^2$. To simplify the expression, we introduce

$$\kappa := \frac{\Gamma}{\gamma} \quad \text{and} \quad c := \left(1 - \frac{1}{\sqrt{\kappa}}\right) \Big/ \left(1 + \frac{1}{\sqrt{\kappa}}\right) \qquad \text{(cf. (8.26d))}.$$

Since $\frac{\Gamma-\gamma}{\Gamma} = 1 - \frac{1}{\kappa} = (1 - \frac{1}{\sqrt{\kappa}})(1 + \frac{1}{\sqrt{\kappa}})$ and $1 + \sqrt{\frac{\gamma}{\Gamma}} = 1 + \frac{1}{\sqrt{\kappa}}$, we arrive at

$$\frac{(\sqrt{\Gamma} + \sqrt{\gamma})^2}{\Gamma - \gamma} = \frac{1 + \frac{1}{\sqrt{\kappa}}}{1 - \frac{1}{\sqrt{\kappa}}} = \frac{1}{c}\,.$$

Hence, the expression for $1/C_m$ reduces to

$$\frac{1}{C_m} = \frac{2c^m}{1 + c^{2m}} \quad \text{with} \quad c = \frac{1 - 1/\sqrt{\kappa}}{1 + 1/\sqrt{\kappa}} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}, \quad \kappa = \frac{\Gamma}{\gamma}\,. \tag{8.28c}$$

For the interpretation of κ as a spectral condition number, compare with Notation 8.24.

Conclusion 8.27. *(a) For the case (8.22a), i.e.,* $\sigma_M = [-\bar{\rho}, \bar{\rho}]$ *with* $0 < \bar{\rho} < 1$, *the solution to the third minimisation problem (8.23) is:*

$$p_m(\zeta) = T_m(\zeta/\bar{\rho}) \,/\, C_m \qquad \textit{with } C_m := T_m(1/\bar{\rho})\,. \tag{8.29a}$$

(b) For the case (8.22b): $\sigma_M = [0, \bar{\rho}]$ *with* $0 < \bar{\rho} < 1$, *the respective solution becomes*

$$p_m(\zeta) = T_m\Big(\frac{2\zeta - \bar{\rho}}{\bar{\rho}}\Big) \,/\, C_m \qquad \textit{with } C_m := T_m\Big(\frac{2 - \bar{\rho}}{\bar{\rho}}\Big)\,. \tag{8.29b}$$

(c) The respective attained minima are

$$\frac{1}{C_m} = \frac{2c^m}{1 + c^{2m}} \qquad \textit{with} \qquad c = \begin{cases} \frac{2\bar{\rho}}{\left(\sqrt{1+\bar{\rho}}+\sqrt{1-\bar{\rho}}\right)^2} & \textit{for (8.29a)}, \\ \frac{\bar{\rho}}{\left(1+\sqrt{1-\bar{\rho}}\right)^2} & \textit{for (8.29b)}. \end{cases}$$

(d) If NA *is diagonalisable by a transformation* T *(cf. (8.21c)), the semi-iterates* y^m *satisfy the error estimate*

$$\|y^m - x\|_2 \le \eta_m \operatorname{cond}_2(T) \|x^0 - x\|_2 \qquad \textit{with} \tag{8.29c}$$
$$\eta_m = 2\left(1 - \tfrac{1}{\kappa}\right)^m / \Big[\Big(1 + \tfrac{1}{\sqrt{\kappa}}\Big)^{2m} + \Big(1 - \tfrac{1}{\sqrt{\kappa}}\Big)^{2m} \Big],$$

where κ *is defined by (8.28c). In the case of a symmetric iteration applied to* $A > 0$ *(cf. §3.5.2), an estimate analogous to (8.29c) holds with respect to the energy norm:*

$$\|y^m - x\|_A \le \eta_m \|x^0 - x\|_A.$$

Proof of (d). Use $c = (1 - 1/\kappa) \,/\, (1 + 1/\sqrt{\kappa})^2$. □

For the implementation of the Chebyshev method, one could in principle apply Remark 8.16 and use the second formulation. The Chebyshev polynomial T_m has the zeros

$$x_\nu = \cos\left([\nu + \tfrac{1}{2}]\pi/m\right) \qquad (1 \le \nu \le m).$$

Hence, the transformed polynomial p_m in (8.27a) admits the factorisation (8.13a) with $\zeta_\nu = \frac{1}{2}[a + b + (b - a)x_\nu]$. The auxiliary polynomials p_μ in (8.13b) lead to the damping factors $\vartheta_\mu := 1/(1 - \zeta_\mu)$ in (8.10b) and (8.12b). However, this approach suffers from numerical instabilities (cf. Lebedev–Finogenov [261, 262]).

The only elegant and practical implementation is the use of the *three-term recursion* (8.14a–c), since recursion (8.28a,b) is a particular case of (8.15a,b). The coefficients Θ_m and ϑ_m required in (8.14a–c) are provided by the next exercise.

Exercise 8.28. Prove: (a) For the case of $\sigma_M = [a, b]$ with $a < b < 1$, recursion (8.15a,b) for p_m in (8.28a,b) uses the factors

$$\Theta_m = 4C_{m-1}/[(b - a)C_m], \tag{8.30a}$$
$$\vartheta_m = -2(a + b)C_{m-1}/[(b - a)C_m]. \tag{8.30b}$$

(b) In the case of $\sigma_M = [-\rho, \rho]$ with $\rho > 0$, (8.28b) leads to recursion (8.15c,d) with

$$\Theta_m = 2C_{m-1}/(\rho C_m) = 1 + C_{m-2}/C_m.$$

(c) Which coefficients correspond to the case of $\sigma_M = [0, \rho]$?

(d) Use Eq. (8.28b) at $\zeta = 1$: $C_{m+1} = AC_m - C_{m-1}$ with $A := 2(2-a-b)/(b-a)$ and prove for the general case of $\sigma_M = [a, b]$ that

$$\Theta_m = \frac{16}{8\,(2-a-b) - (b-a)^2\Theta_{m-1}}, \quad \Theta_1 = \frac{4}{2-a-b}, \tag{8.30c}$$

$$\vartheta_m = -\frac{1}{2}(a+b)\Theta_m\,. \tag{8.30d}$$

(e) The coefficients converge monotonically to

$$\lim \Theta_m = \frac{4c}{b-a} \qquad \text{and} \qquad \lim \vartheta_m = \frac{-2c\,(a+b)}{b-a}$$

with c in (8.28c).

(f) The assumptions $a < b$ in (a) and $\rho > 0$ in (b) avoid the division by zero. Show that $a = b$ or $\rho = 0$ lead to a direct solution: the semi-iterate y^1 is already the exact solution.

Hint for (a): For $m > 2$, compare the coefficients in (8.15b) and (8.28b). For $m = 1$, compare (8.15a) with (8.28a), taking notice of $C_0 = 1$ and $C_1 = \frac{2-a-b}{b-a}$ according to (8.28c). Part (e): Insert (8.28c) into (8.30a,b).

Instead of Θ_m and ϑ_m, one can also compute the sum $\sigma_m := \Theta_m + \vartheta_m$ recursively from

$$\sigma_1 = 2, \qquad \sigma_m = 4\Big/\left\{4 - \left(\frac{1-1/\kappa}{1+1/\kappa}\right)^2 \sigma_{m-1}\right\}$$

(derived from (8.30c,d) with κ in (8.26d)). Equation (8.30d) yields the values

$$\Theta_m = 2\,\sigma_m/(2-a-b), \qquad \vartheta_m = -(a+b)\,\sigma_m/(2-a-b).$$

The coefficients σ_m can also be used directly for the three-term recursion. Given the matrix N of the second normal form of Φ, the formulae (8.14a–c) with the coefficients (8.30a,b) are equivalent to

$$\begin{aligned} y^0 &= x^0, \qquad y^1 = y^0 - \frac{2}{2-a-b} N\left(Ay^0 - b\right), \\ y^m &= \sigma_m\left\{y^{m-1} - \frac{2}{2-a-b}N(Ay^{m-1} - b)\right\} + (1-\sigma_m)y^{m-2}. \end{aligned} \tag{8.31}$$

The factor $\frac{2}{2-a-b}$ may also be written as $\frac{2}{\gamma+\Gamma}$ (cf. (8.26c)).

We recall the set $\mathcal{N}$ of nonlinear acceleration methods mentioned on page 173. The Chebyshev method is a first example.

Notation 8.29. We denote the Chebyshev method based on $\sigma_M = [a,b]$ by

$$\varUpsilon_{a,b}^{\mathrm{Cheb}} \in \mathcal{N}.$$

In principle, the Chebyshev method is well defined for all iterations $\varPhi \in \mathcal{L}$. However, the convergence statements only refer to algorithm $\varUpsilon_{a,b}^{\mathrm{Cheb}}[\varPhi]$ and matrices A such that $\sigma(M) \subset [a,b]$ holds for the iteration matrix $M = M_{\varPhi}[A]$ of $\varPhi$.

8.3.5 Order Improvement by the Chebyshev Method

Theorem 8.30. *(a) Assume that $\sigma(M) \subset \sigma_M = [a,b]$ holds with $a < b < 1$. The Chebyshev method has the asymptotic convergence rate $c = \lim_{m\to\infty} \sqrt[m]{\frac{1}{C_m}}$ with*

$$c = \frac{b-a}{2-a-b+2\sqrt{(1-a)\,(1-b)}} = \frac{\varGamma-\gamma}{(\sqrt{\varGamma}+\sqrt{\gamma}\,)^2} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}, \tag{8.32a}$$

where $\kappa = \varGamma/\gamma$ (cf. (8.26c)). Particular cases are

$$\lim_{m\to\infty} \sqrt[m]{\frac{1}{C_m}} = \frac{\rho}{1+\sqrt{1-\rho^2}} \qquad \text{for } \sigma_M = [-\rho,\rho],\ \rho<1, \tag{8.32b}$$

$$\lim_{m\to\infty} \sqrt[m]{\frac{1}{C_m}} = \frac{\rho}{\left(1+\sqrt{1-\rho}\right)^2} \qquad \text{for } \sigma_M = [0,\rho],\ \rho<1. \tag{8.32c}$$

(b) Let τ be the order of the basic iteration: $\rho(M) = 1 - Ch^{\tau} + \mathcal{O}(h^{2\tau})$. Then the Chebyshev method is of order $\tau/2$. The asymptotic convergence rate equals

$$\begin{aligned} &1-2\sqrt{\tfrac{C}{1-\gamma}}\,h^{\tau/2}+\mathcal{O}(h^{\tau}) && \text{for (8.32a) with } \varGamma=\rho(M),\\ &1-\sqrt{2C}\,h^{\tau/2}+\mathcal{O}(h^{\tau}) && \text{for } \sigma_M=[-\rho(M),\rho(M)],\\ &1-2\sqrt{C}\,h^{\tau/2}+\mathcal{O}(h^{\tau}) && \text{for } \sigma_M=[\,0,\rho(M)]. \end{aligned}$$

Proof. Since $0 \le c \le 1$, (8.28c) shows that $\sqrt[m]{\frac{1}{C_m}} = c\,\sqrt[m]{2/\left(1+c^{2m}\right)} \to c$. □

Therefore, the Chebyshev method achieves a halving of the order similar to the SOR iteration. Concerning the connection of both methods, we refer to §8.4.3 and Varga [375, §5.2].

8.3.6 Optimisation Over Other Sets

Up to now, we considered an interval $[a, b]$ with $a < b < 1$. If, for instance, no eigenvalue of M lies in $(a', b') \subset [a, b]$, we may replace σ_M by the smaller set

$$\sigma_M = [a, a'] \cup [b', b] \qquad (a \le a' < b' \le b). \tag{8.33}$$

Obviously, the minimum of $\{\max_{\zeta\in\sigma_M} |p_m(\zeta)| : p_m \in \mathcal{P}_m\}$ can only become smaller. In the case of $a' - a = b - b'$, it is easy to describe the optimal polynomial (cf. Axelsson–Barker [13, p. 26f]). Concerning the determination of optimal polynomials, we refer to de Boor–Rice [102] and Fischer [133, §3.3]. In particular, the case $\sigma_M = [a, a'] \cup [b', b]$ for $a \le a' < 1 < b' \le b$ is interesting. The latter situation occurs for indefinite matrices.

Remark 8.31. Consider discretisation of Helmholtz' equation $-\Delta u - cu = f$ with positive c, which leads to $A = A_\Delta - cI$, where A_Δ is the matrix of the Poisson model problem. Let $0 < \lambda_1 \le \lambda_2 \le \ldots \le \lambda_n$ be the eigenvalues of A_Δ. Assume for a suitable k that $\lambda_k < c < \lambda_{k+1}$. Then the spectrum of $A = A^{\mathrm{H}}$ is contained in

$$\sigma_A = [-\beta_-, -\alpha_-] \cup [\alpha_+, \beta_+] \qquad \text{with} \qquad -\beta_- \le -\alpha_- < 0 < \alpha_+ < \beta_+ ,$$

where $\beta_- := c - \lambda_1$, $\alpha_- := c - \lambda_k$, $\alpha_+ := \lambda_{k+1} - c$, $\beta_+ := \lambda_n - c$. The Richardson iteration with $0 < \Theta < 1/\beta_+$ leads to the iteration matrix $M = M_\Theta^{\mathrm{Rich}}$ whose spectrum is contained in the set σ_M described in (8.33), where

$$a = 1 - \Theta\beta_+ < a' = 1 - \Theta\alpha_+ < 1 < b' = 1 + \Theta\alpha_- \le b = 1 + \Theta\beta_- .$$

If one extreme eigenvalue b of M is known and the others are enclosed by $[a, b']$, we arrive at

$$\sigma_M = [a, b'] \cup \{b\} \quad \text{with} \quad b' < b, \quad b' < 1, \quad b \ne 1.$$

Let $q_{m-1} \in \mathcal{P}_{m-1}$ with $q_{m-1}(1) = 1$ be optimal for $[a, b']$. A simple but not optimal proposal for a polynomial p_m suited to σ_M is

$$p_m(\zeta) := q_{m-1}(\zeta)(\zeta - b)/(1 - b).$$

Concerning the construction of asymptotically optimal polynomials for arbitrary compact sets σ_M with $1 \notin \sigma_M$, we refer to Niethammer–Varga [294] and Eiermann–Niethammer–Varga [119]. The simplest set σ_M that is more general than the interval $[a, b]$ is the ellipse (cf. Fischer–Freund [134, 135], Niethammer–Varga [294], and Manteuffel [272]). Since, in general, a suitable ellipse enclosing the eigenvalues of M is not known a priori, one has to improve its parameters adaptively (cf. Manteuffel [272]). The fact that the ellipse lies in the complex plane does not imply that the optimal polynomial has also complex parameters. As long as $\sigma(M)$ is symmetric with respect to the real axis (i.e., all complex eigenvalues belong to conjugate pairs), one can find an optimal polynomial with real coefficients (cf. Opfer–Schober [297]).

In any case, the spectrum $\sigma(M)$ is enclosed by the complex circle

$$\sigma_M = \{z = x + \mathrm{i}y \in \mathbb{C} : x^2 + y^2 \leq \rho(M)^2\}.$$

Unfortunately, this choice does not lead to an interesting solution (cf. [297]).

Theorem 8.32. *Let σ_M be a circle around $z_0 \in \mathbb{C}\backslash\{1\}$ with radius $r < |1 - z_0|$. The optimal polynomial for σ_M is $p_m(\zeta) = [(\zeta - z_0)/(1 - z_0)]^m$. In particular, for $z_0 = 0$, the corresponding semi-iteration coincides with the basic iteration Φ. In the general case, the semi-iteration corresponds to the damped iteration Φ_ϑ with $\vartheta := 1/|1 - z_0|$.*

Proof. The absolute value of the polynomial p_m defined above takes its maximum

$$\rho := \max\{|p_m(\zeta)| : \zeta \in \sigma_M\} = r/|1 - z_0|$$

at all boundary points $\zeta \in \partial\sigma_M$. If p_m is not optimal, there is some polynomial $q_m \in \mathcal{P}_m$ with $q_m(1)=1$ and $\max\{|q_m(\zeta)| : \zeta \in \sigma_M\} < \rho$. $q_m(\zeta) < \rho = p_m(\zeta)$ holds for all boundary values $\zeta \in \partial\sigma_M$, so that the theorem of Rouché is applicable; i.e., the holomorphic functions p_m and $p_m - q_m$ have the same number of zeros in σ_M. Since p_m has an m-fold zero at z_0, $p_m - q_m$ has also m zeros in σ_M. Since $(p_m - q_m)(1) = p_m(1) - q_m(1) = 1 - 1 = 0$, the polynomial $p_m - q_m \in \mathcal{P}_m$ has even $m + 1$ zeros, implying $p_m = q_m$. Hence, p_m is already optimal. □

8.3.7 Cyclic Iteration

Following Conclusion 8.27, it has been mentioned that in principle it would be possible to apply the second formulation (8.10b) with the factors $\vartheta_\nu := 1/(1 - \zeta_\nu)$, $\zeta_\nu = \cos([\nu + \frac{1}{2}]\pi/m)$ for $\nu = 1, \ldots, m$. The result y^m (only for this fixed m) is the desired Chebyshev semi-iterate. However, by this approach the Chebyshev method cannot be continued. To obtain an infinite iterative process, we may repeat the extrapolation factors m-periodically:

$$\begin{aligned} &\vartheta_1, \vartheta_2, \ldots, \vartheta_m && \text{given,} \\ &\vartheta_i := \vartheta_{i-m} && \text{for } i > m. \end{aligned}$$

A semi-iterative method (8.10a,b) with these parameters is called a *cyclic iteration*. The restriction to the iterates y^0, y^m, y^{2m}, $y^{3m}, \ldots$ produces a proper linear iteration. The related iteration matrix is $p_m(M)$ with p_m generated by $\{\vartheta_i : 1 \leq i \leq m\}$. The convergence rate of the cyclic iteration is not described by $\rho(p_m(M))$ but by $\rho(p_m(M))^{1/m}$, since one cycle $y^0 \longmapsto y^m$ is thought to consist of m and not of one step. The cyclic iteration also runs the risk of numerical instabilities as already discussed after Conclusion 8.27.

Exercise 8.33. Prove: Viewing the cyclic iteration as a semi-iteration $\{y^0, y^1, \ldots\}$ of all iterates, the asymptotic convergence rate in Definition 8.3 also coincides with $\sqrt[m]{\rho(p_m(M))}$.

8.3.8 Two- and Multi-Step Iterations

Exercise 8.28e yields the limits $\Theta = \lim \Theta_m$ and $\vartheta = \lim \vartheta_m$. Hence, the three-term recursion (8.14c) converges to the (stationary) two-step iteration (2.27):

$$y^m = \Theta\left[\Phi(y^{m-1}, b) - y^{m-2}\right] + \vartheta(y^{m-1} - y^{m-2}) + y^{m-2}. \tag{8.35}$$

As described in §2.2.8, the convergence of iteration (2.27) can be reduced to the convergence of a one-step iteration with the iteration matrix

$$\mathbf{M} = \begin{bmatrix} \mu_0 M + \mu_1 I & \mu_2 I \\ I & 0 \end{bmatrix}, \quad \mu_0 = \frac{4c}{b-a}, \quad \mu_1 = -2\,c\,\frac{a+b}{b-a}, \quad \mu_2 = 1 - \mu_0 - \mu_1$$

(c defined in (8.28c)). From these coefficients, assuming that $\sigma(M) \subset \sigma_M$ and using Exercise 2.25, we obtain the value $\rho(\mathbf{M}) = c$, i.e., the (stationary) two-step iteration (8.35) achieves the same convergence rate as the semi-iterative method. Hence, the two-step iteration (8.35) also yield an improvement of the order of convergence.

More generally, one can consider the k-step iteration

$$x^m = \mu_0\,\Phi(x^m, b) + \sum_{i=1}^{k} \mu_i x^{m-i} \qquad \text{with } \sum_{i=1}^{k} \mu_i = 1.$$

The connection between k-step iterations and semi-iterative methods is described by Niethammer–Varga [294].

8.3.9 Amount of Work of the Semi-Iterative Method

We consider the realisation of the Chebyshev method by (8.31). There the call of the basic iteration $\Phi(x, b) = x - W^{-1}(Ax - b)$ is replaced by the call of $W^{-1}(b - Ay) = \Phi(x, b) - x$. Besides the call of the basic iteration Φ, the implementation (8.31) (for $m \geq 2$) requires six operations per grid point:

$$\text{semi-iterative Work}(\Phi) \;\leq\; \text{Work}(\Phi) + 6n$$

(cf. §2.3 and §3.4). Hence, the cost factor amounts to

$$C_{\Phi,\text{semi}} = C_\Phi + \tfrac{6}{C_A},$$

where $C_A n$ is defined in §2.3 as the number of nonzero elements of A.

Replacing in (2.31a) the convergence rate by the asymptotic value c in (8.32a), we obtain the effective amount of work

$$\text{Eff}_{\text{semi}}(\Phi) = -(C_\Phi + \tfrac{6}{C_A})/\log c\,.$$

If $\gamma/\Gamma \ll 1$ holds as in the examples discussed in §8.4, we can exploit the asymptotic behaviour $\log c = -2\sqrt{\gamma/\Gamma} + \mathcal{O}(\gamma/\Gamma)$:

$$\text{Eff}_{\text{semi}}(\Phi) \approx \left(\frac{C_\Phi}{2} + \frac{3}{C_A}\right)\sqrt{\frac{\Gamma}{\gamma}}\,. \tag{8.36}$$

Exercise 8.34. Assume that the iteration matrix of Φ fulfils $\sigma(M) \subset [a,b]$ with $b = 1 - \mathcal{O}(h^{-\tau})$ and $\tau > 0$. Prove the following comparison of $\mathrm{Eff}_{\mathrm{semi}}$ and Eff:

$$\mathrm{Eff}_{\mathrm{semi}}(\Phi) \approx \left(C_\Phi + \tfrac{6}{C_A}\right) \sqrt{\mathrm{Eff}(\Phi) / \left[(1-a)\, C_\Phi\right]}\,.$$

8.4 Application to Iterations Discussed Above

8.4.1 Preliminaries

The essential condition for the applicability[2] of the Chebyshev method is that the spectrum $\sigma(M)$ be real. This excludes the SOR method. Semi-iterative variants based on other supersets $\sigma_M \supset \sigma(M)$ are also not successful for the SOR method with $\omega \geq \omega_{\mathrm{opt}}$ (cf. §8.3.6)). The reason for this is statement (e) of Theorem 4.27. For $\omega \geq \omega_{\mathrm{opt}}$, all eigenvalues $\lambda \in \sigma(M_\omega^{\mathrm{SOR}})$ are situated on the boundary of the complex circle $|\zeta| = \omega - 1$, for which no convergence acceleration is possible, as stated in Theorem 8.32.

If A is positive definite, the following already mentioned iterations lead to a real spectrum: the Richardson, (block-)Jacobi, and (block-)SSOR methods. Numerical results for these choices of basic iterations will be presented for the Poisson model problem in the following sections.

Besides the iterations mentioned above, in §5.2 we constructed their damped variants. However, for a discussion of semi-iterative methods the damped variants are without any interest as stated next.

Lemma 8.35. *Let the iteration Φ have a real spectrum $\sigma(M)$. Then Φ and the corresponding damped iterations Φ_ϑ with $\vartheta \neq 0$ generate identical semi-iterative results y^m.*

Proof. By (8.6a,b), the semi-iterate y^m generated by Φ has the representation $y^m = x^0 + p_m(M)(x^0 - x)$. The damped iteration has the iteration matrix

$$M_\vartheta = I - \vartheta NA = I - N_\vartheta A \quad \text{with } N_\vartheta := \vartheta N.$$

For N_ϑ, inclusion (8.26b) can be written as $\sigma(N_\vartheta A) \subset [\gamma', \Gamma']$ with $\gamma' := \vartheta\gamma$ and $\Gamma' := \vartheta\Gamma$ (possibly a complex interval). $p_m(M) = T_m\big(\frac{\Gamma+\gamma}{\Gamma-\gamma}\, I + \frac{2}{\Gamma-\gamma}\, NA\big)$ (cf. Exercise 8.26b) is invariant with respect to the replacement of γ, Γ, N by γ', Γ', N_ϑ. Hence $p_{m,\vartheta}(M_\vartheta) = p_m(M)$, where $p_{m,\vartheta}$ is the polynomial adapted to the interval $[\gamma', \Gamma'] \supset \sigma(N_\vartheta A)$. The iterates $y_\vartheta^m = x^0 + p_{m,\vartheta}(M_\vartheta)(x^0 - x)$ of Φ_ϑ coincide with those of Φ. □

[2] Here 'applicability of the Chebyshev method' means that also the assumptions of the convergence statements hold. Otherwise, the Chebyshev method can be applied to any $A \in \mathfrak{D}(\Phi)$.

8.4.2 Semi-Iterative Richardson Method

According to Lemma 8.35, we may fix the factor of Richardson's method (3.4) by $\Theta = 1$, i.e., $x^{m+1} = x^m - (Ax - b)$. Then the matrix $N = I$ of the second normal form is as simple as possible and condition (8.26b) becomes $\sigma(A) \subset [\gamma, \Gamma]$.

Remark 8.36. (a) The Chebyshev method is applicable if A has only positive eigenvalues. For the estimation of γ and Γ in (8.26b), one has to use the respective bounds for the extreme eigenvalues of A.

(b) In particular, the assumptions are satisfied if A is positive definite. In this case, one has to choose $\gamma = 1/\|A^{-1}\|_2$ and $\Gamma = \|A\|_2$ (optimal choice) or at least $\gamma \leq 1/\|A^{-1}\|_2$ and $\Gamma \geq \|A\|_2$.

For the Poisson model problem, we obtain

$$\gamma = \lambda_{\min} = 8h^{-2}\sin^2(\pi h/2), \qquad \Gamma = \lambda_{\max} = 8h^{-2}\cos^2(\pi h/2)$$

according to (3.1b,c). Inserting these values into the asymptotic convergence rate (8.32a), we arrive at

$$\lim_{m\to\infty} \sqrt[m]{\frac{1}{C_m}} = c = \cos(\pi h)/(1 + \sin(\pi h)) = 1 - \pi h + \mathcal{O}(h^2).$$

For $h = 1/16$ and $h = 1/32$, we obtain $c = 0.82$ and $c = 0.906$. The numerical results in Table 8.1 show that the reduction factor approximates the convergence rate only for sufficiently large m. The ratios

$$\rho_m := \|y^m - x\|_2/\|y^{m-1} - x\|_2, \quad \hat{\rho}_m := (\|y^m - x\|_2/\|y^0 - x\|_2)^{1/m}$$

tend to c from above.

m	$\|y^m - x\|_2$	ρ_m	$\hat{\rho}_m$
1	6.44_{10}-1	9.09_{10}-1	9.09_{10}-1
10	2.44_{10}-1	8.91_{10}-1	8.99_{10}-1
20	6.35_{10}-2	8.59_{10}-1	8.86_{10}-1
30	1.29_{10}-2	8.48_{10}-1	8.75_{10}-1
40	2.36_{10}-3	8.41_{10}-1	8.67_{10}-1
50	4.07_{10}-4	8.36_{10}-1	8.61_{10}-1
60	6.75_{10}-5	8.34_{10}-1	8.57_{10}-1
70	1.08_{10}-5	8.32_{10}-1	8.53_{10}-1
80	1.72_{10}-6	8.31_{10}-1	8.50_{10}-1
90	2.67_{10}-7	8.29_{10}-1	8.48_{10}-1
100	4.11_{10}-8	8.28_{10}-1	8.46_{10}-1

m	$\|y^m - x\|_2$	ρ_m	$\hat{\rho}_m$
1	7.14_{10}-1	9.54_{10}-1	9.54_{10}-1
10	4.47_{10}-1	9.48_{10}-1	9.49_{10}-1
30	1.40_{10}-1	9.36_{10}-1	9.45_{10}-1
50	3.21_{10}-2	9.24_{10}-1	9.38_{10}-1
70	6.26_{10}-3	9.19_{10}-1	9.33_{10}-1
80	2.66_{10}-3	9.17_{10}-1	9.31_{10}-1
100	4.65_{10}-4	9.15_{10}-1	9.28_{10}-1
120	7.80_{10}-5	9.13_{10}-1	9.26_{10}-1
130	3.15_{10}-5	9.13_{10}-1	9.25_{10}-1
140	1.27_{10}-5	9.12_{10}-1	9.24_{10}-1
150	5.09_{10}-6	9.12_{10}-1	9.23_{10}-1

Table 8.1 Semi-iterative Richardson method for $h = 1/16$ (left) und $h = 1/32$ (right).

8.4.3 Semi-Iterative Jacobi and Block-Jacobi Method

Numerical examples are unnecessary, since in the Poisson model case, the Jacobi method coincides with the damped Richardson method and, according to Lemma 8.35, reproduces the results in Table 8.1.

Concerning the lower bound a of the spectrum $\sigma(M^{\mathrm{Jac}})$, Lemma 4.8 proves that $a = -b$ holds for a particular case.

Lemma 8.37. *If (A, D) is weakly 2-cyclic (cf. Definition 4.2), the Jacobi iteration matrix M^{Jac} has a symmetric spectrum: $\sigma(M^{\mathrm{Jac}}) = -\sigma(M^{\mathrm{Jac}})$. The smallest enclosing interval is $[\,a, b\,] = [-\rho(M^{\mathrm{Jac}}), \rho(M^{\mathrm{Jac}})]$.*

A comparison of the semi-iterative Jacobi iteration with the SOR method is possible. In the weakly 2-cyclic case, (8.32b) is applicable because of Lemma 8.37 and yields the asymptotic semi-iterative convergence rate

$$\beta/[1+\sqrt{1-\beta^2}\,] \qquad \text{with } \beta := \rho(M^{\mathrm{Jac}}).$$

This quantity coincides with the square root of the optimal SOR convergence rate $\omega_{\mathrm{opt}} - 1$; hence, the semi-iterative Jacobi iteration is half as fast as the SOR method. The order improvement by an optimal choice ω_{opt} in the SOR case and the order improvement by the Chebyshev method (cf. Theorem 8.30b) lead to very similar results.

m	$\Vert y^m - x\Vert_2$	ρ_m	$\hat\rho_m$
1	6.09_{10}-1	8.60_{10}-1	8.60_{10}-1
20	1.62_{10}-2	7.95_{10}-1	8.27_{10}-1
40	1.19_{10}-4	7.75_{10}-1	8.04_{10}-1
60	6.68_{10}-7	7.69_{10}-1	7.93_{10}-1
80	3.33_{10}-9	7.65_{10}-1	7.86_{10}-1
90	2.1_{10}-10	7.55_{10}-1	7.84_{10}-1

m	$\Vert y^m - x\Vert_2$	ρ_m	$\hat\rho_m$
1	6.94_{10}-1	9.28_{10}-1	9.28_{10}-1
20	1.53_{10}-1	9.12_{10}-1	9.23_{10}-1
40	1.84_{10}-2	8.92_{10}-1	9.11_{10}-1
60	1.70_{10}-3	8.85_{10}-1	9.03_{10}-1
80	1.40_{10}-4	8.81_{10}-1	8.98_{10}-1
100	1.08_{10}-5	8.78_{10}-1	8.94_{10}-1

Table 8.2 Semi-iterative column-block-Jacobi iteration for $h = 1/16$ (left) and $h = 1/32$ (right).

The block variants of the Jacobi iteration converge faster than the pointwise version. Correspondingly, the results of the semi-iterative column-block-Jacobi method in Table 8.2 are better than those in Table 8.1. The factors should tend to the asymptotic value 0.7565 for $h = 1/16$ and to 0.8702 for $h = 1/32$.

8.4.4 Semi-Iterative SSOR and Block-SSOR Iteration

As already mentioned in §8.4.1, the Gauss–Seidel and SOR methods are not suited for semi-iterative purposes, since, in general, the spectrum is not real. A remedy is offered by the symmetric Gauss–Seidel and SSOR iteration. Theorem 6.26 states that the spectrum of the SSOR method is real for Hermitian matrices A.

Theorem 6.28 gives an upper bound for the spectral radius $\rho(M_\omega^{\mathrm{SSOR}})$. Hence, under conditions (6.18a,b), the spectrum can be enclosed by the interval $[a,b]$ with

$$a=0, \quad b=1-2\varOmega/\Big[\frac{\varOmega^2}{\gamma}+\varOmega+\frac{\varGamma}{4}\Big], \quad \text{where } \varOmega:=\frac{2-\omega}{2\omega},\ 0<\omega<2. \tag{8.37}$$

Here, $\varGamma$ is defined by (6.18b). Corollary 3.45 helps to determine $\varGamma$. For the Poisson model problem, Lemma 3.62 yields the value $\varGamma=2$. Inequality (6.18a) states that γ coincides with λ in (3.35c) applied to the (block-)Jacobi method. In the Poisson model case, $\gamma=2\sin^2(\pi h/2)$ holds.

Theorem 8.38. *Let $A=D-E-E^{\mathsf{H}}>0$ and γ, $\varGamma$ satisfy the assumptions (6.18a,b). Assume, in addition, that $0<\omega\le 2/(\varGamma+1)$. Then*

$$a=\left(\frac{1-\xi}{1+\xi}\right)^2 \quad \textit{with} \quad \xi:=\frac{2-\omega}{\varGamma\omega} \tag{8.38}$$

is a lower bound of the spectrum $\sigma(M_\omega^{\mathrm{SSOR}})$.

Proof. Using the parameter $\varOmega$ in (3.46c), we can rewrite

$$W_\omega^{\mathrm{SSOR}}=\Big(\frac{1}{\omega}D-E\Big)\Big[\Big(\frac{2}{\omega}-1\Big)D\Big]^{-1}\Big(\frac{1}{\omega}D-E\Big)^{\mathsf{H}}$$

as

$$W_\omega^{\mathrm{SSOR}}=[\varOmega D+\varDelta](2\varOmega D)^{-1}[\varOmega D+\varDelta]^{\mathsf{H}} \quad \text{with} \quad \varDelta:=\frac{1}{2}D-E.$$

Defining $X:=\varOmega D+(1-\alpha)\varDelta$ for some real α, we have $[\varOmega D+\varDelta]=X+\alpha\varDelta$. The expansion of $[X+\alpha\varDelta](2\varOmega D)^{-1}[X+\alpha\varDelta]^{\mathsf{H}}$ yields

$$W_\omega^{\mathrm{SSOR}}=\frac{1}{2\varOmega}XD^{-1}X^{\mathsf{H}}+\frac{\alpha}{2}A+\frac{1}{2\varOmega}(2\alpha-\alpha^2)\varDelta D^{-1}\varDelta^{\mathsf{H}}.$$

because of $\varDelta+\varDelta^{\mathsf{H}}=A$. The factor $(2\alpha-\alpha^2)$ is negative for $\alpha>2$. Hence,

$$W_\omega^{\mathrm{SSOR}}\ge g(\alpha)A \qquad \text{with} \quad g(\alpha):=\frac{\alpha}{2}\left(1+\frac{\varGamma}{4}\frac{2-\alpha}{\varOmega}\right) \qquad \text{for } \alpha\ge 2.$$

The assumption $\omega\le 2/(\varGamma+1)$ implies $\alpha_0:=1+2\varOmega/\varGamma\ge 2$. Theorem 3.34a with $1-a=1/g(\alpha_0)$ yields the value (8.38). □

The statement is less interesting, since (because of $\varGamma=2$ for the Poisson model case) Theorem 8.38 only applies to strong underrelaxation: $\omega\le 2/3$.

There are two possibilities in improving (halving) the convergence order. First, this can be achieved by the optimal choice of ω in the SOR or SSOR method (cf. Conclusions 3.46 and 6.29). Second, the semi-iterative method leads to halving of the order compared with the basic iteration. In the case of SSOR as the basic iteration, both techniques can be applied simultaneously. First, the optimal SSOR relaxation parameter ω' is chosen as described in (3.47b). The hereby defined iteration $\Phi_{\omega'}^{\mathrm{SSOR}}$ is chosen as the basic iteration of the Chebyshev method. Together, we succeed in quartering the order. In the Poisson model case, we obtain the asymptotic convergence rate $1 - \mathcal{O}(h^{1/2})$.

The bound b in (8.37) becomes minimal for

$$\omega' = 2/\bigl(1 + \sqrt{\gamma\varGamma}\,\bigr).$$

The corresponding value is

$$b = \frac{\sqrt{\varGamma} - \sqrt{\gamma}}{\sqrt{\varGamma} + \sqrt{\gamma}} = \frac{1 - \sqrt{\gamma/\varGamma}}{1 + \sqrt{\gamma/\varGamma}} . \tag{8.39a}$$

Inserting this value into (8.32c) yields the asymptotic convergence rate

$$\lim_{m\to\infty} \sqrt[m]{\frac{1}{C_m}} = c = \frac{1 - \sqrt{1-b}}{1 + \sqrt{1-b}} \quad \text{with } b \text{ in (8.39a).} \tag{8.39b}$$

The spectral condition number $\kappa = \kappa((W^{\mathrm{SSOR}})^{-1}A)$ is equal to $\frac{1}{2}(1 + \sqrt{\gamma/\varGamma}\,)$. Using the inequality $\gamma \geq 1/\kappa(A)$ in Exercise 5.20, we end up with the result

$$\kappa((W^{\mathrm{SSOR}})^{-1}A) \leq \frac{1}{2}\left(1 + \sqrt{\varGamma\,\kappa(A)}\,\right). \tag{8.39c}$$

m	$\Vert y^m - x\Vert_2$	ρ_m	$\hat\rho_m$
1	4.673_{10}-1	6.24_{10}-1	6.24_{10}-1
2	2.761_{10}-1	5.90_{10}-1	6.07_{10}-1
3	1.359_{10}-1	4.92_{10}-1	5.66_{10}-1
4	7.681_{10}-2	5.65_{10}-1	5.66_{10}-1
5	3.801_{10}-2	4.94_{10}-1	5.51_{10}-1
20	2.080_{10}-6	5.08_{10}-1	5.27_{10}-1
21	1.007_{10}-6	4.84_{10}-1	5.25_{10}-1
22	5.195_{10}-7	5.15_{10}-1	5.24_{10}-1
23	2.541_{10}-7	4.89_{10}-1	5.23_{10}-1
29	3.395_{10}-9	4.82_{10}-1	5.15_{10}-1
30	1.628_{10}-9	4.79_{10}-1	5.14_{10}-1

N	ω'	c
2	0.8284	0.0470
4	1.1329	0.1467
8	1.4386	0.2727
16	1.6721	0.4059
32	1.8212	0.5315
64	1.9064	0.6408
128	1.9520	0.7305
256	1.9757	0.8010
512	1.9878	0.8549
1028	1.9939	0.8953
5000	1.9987	0.9511
10000	1.9993	0.9651

Table 8.3 *Left:* semi-iterative lexicographical SSOR for the parameters in (8.40); concerning ρ_m and $\hat\rho_m$ see Table 8.1. *Right:* optimal ω' and asymptotic rate c for $h = 1/N$.

For the values γ and $\varGamma$ in Lemma 3.62 (Poisson model case), the convergence rate (8.39b) is asymptotically equal to the value

$$c = 1 - Ch^{1/2} + \mathcal{O}(h) \quad \text{with } C = 2\sqrt{\pi}.$$

The results in Table 8.3 refer to the parameters

$$h = 1/32, \quad \omega = 1.8455, \quad a = 0, \quad b{=}0.878. \tag{8.40}$$

In §6.3.5 the value ω is proved to be optimal (note that ω' is optimal only for the bound in (6.18c)). We learn from Table 6.1 that $b = 0.878$ is an upper bound of the convergence rate. From (8.39b) with $b = 0.878$, one calculates the rate $c = 0.482$, which is numerically well confirmed (cf. Table 8.3). From $C_{\Phi}^{\mathrm{SSOR}} {=} 2 + 6/C_A {=} 3.2$ (according to Remark 6.27 and because of $C_A = 5$ for five-point formulae), we obtain the effective amount of work

$$\mathrm{Eff}_{\mathrm{semi}}(\Phi^{\mathrm{SSOR}}) = -3.2/\log c = 4.38 \tag{8.41}$$

for the semi-iterative SSOR method with $h = 1/32$, which can be compared, e.g., with $\mathrm{Eff}(\Phi^{\mathrm{SOR}}) = 7.05$ in Example 2.28.

If we use the values ω' in (3.47b), Eq. (8.39b) yields the asymptotic convergence rates c reported in Table 8.3. These values give an impression of the asymptotic value $c = 1 - \mathcal{O}(h^{1/2})$.

8.5 Method of Alternating Directions (ADI)

The *alternating-direction-implicit iteration* or shortly *ADI method* was first described in 1955 by Peaceman–Rachford [308] in connection with parabolic differential equations. ADI is not a semi-iterative method in the sense of the previous sections, but it can be considered as a generalisation using rational functions instead of polynomials (see also §8.5.4).

Further material can be found in Marchuk [274] and Wachspress [383].

8.5.1 Application to the Model Problem

For the model problem in §1.2, the matrix A can be split into

$$A = B + C, \qquad \text{where} \tag{8.42a}$$

$$(Bu)(x,y) = h^{-2}\left[-u(x-h,y) + 2u(x,y) - u(x+h,y)\right], \tag{8.42b}$$

$$(Cu)(x,y) = h^{-2}\left[-u(x,y-h) + 2u(x,y) - u(x,y+h)\right] \tag{8.42c}$$

for $(x,y) \in \Omega_h$ are the second differences of u with respect to the x and y direction. If we choose the rows (x direction) of Ω_h as blocks, $B + 2h^{-2}I$ represents the block diagonal of A. Similarly, $C + 2h^{-2}I$ is the block diagonal of A if the columns (y direction) are chosen as blocks.

Remark 8.39. For $A,\ B$, and C in (8.42a–c), the statements (8.43a,b) hold:

$$B > 0 \text{ and } C > 0\,, \tag{8.43a}$$

$$A, B, C \text{ are pairwise commutative.} \tag{8.43b}$$

The last statement is equivalent to

$$A,\ B,\ C \text{ can simultaneously be transformed to diagonal form.} \tag{8.43b$'$}$$

Proof. Lemma 3.58 analyses the block diagonal of A (with respect to the row-block structure). Because of the x–y symmetry, the same result holds for the column-block structure. Therefore, the spectrum of $B + 2h^{-2}I$ and $C + 2h^{-2}I$ is equal to

$$\left\{h^{-2}\left[2 + 4\sin^2\frac{jh\pi}{2}\right] : 1 \le j \le N-1\right\},$$

i.e., $4h^{-2}\sin^2\frac{jh\pi}{2}$ are the eigenvalues of B and C. Since these values are positive, (8.43a) is proved. By Lemma 3.58 the eigenvectors e^{ij} of A (cf. Lemma 3.2) are also the eigenvectors of $B + 2h^{-2}I$, $C + 2h^{-2}I$ and hence of B and C. This proves (8.43b′) and (8.43b). □

The first half-step of the ADI method corresponds to the additive splitting

$$A = W - R \qquad \text{with } W = \omega I + B \text{ and } R = \omega I - C \tag{8.44a}$$

and reads

$$x^{m+1/2} := \Phi_\omega^B(x^m, b) := (\omega I + B)^{-1}(b + \omega x^m - Cx^m), \tag{8.45a}$$

where ω is a (real) parameter. Interchanging the roles of B and C, i.e., *alternating* the directions, we generate the splitting (8.44b) of the second half-step (8.45b):

$$A = W - R \qquad \text{with } W = \omega I + C,\ R = \omega I - B, \tag{8.44b}$$

$$x^{m+1} := \Phi_\omega^C(x^{m+\frac12}, b) := (\omega I + C)^{-1}(b + \omega x^{m+\frac12} - Bx^{m+\frac12}). \tag{8.45b}$$

Remark 8.40. Each single half-step (8.45a,b) resembles a block-Jacobi method. For $\omega = 2h^{-2}$, iteration (8.45a) represents the row- and (8.45b) the column-block-Jacobi iteration. Because of (8.43a), the matrices $\omega I + B$ and $\omega I + C$ with $\omega \ge 0$ are positive definite and therefore regular; hence, the steps (8.45a,b) are well defined. Since, furthermore, $\omega I + B$ and $\omega I + C$ are tridiagonal matrices, the solution of $(\omega I + B)z = c$ or $(\omega I + C)z = c$ required in (8.45a,b) is easy to perform.

The complete ADI step $x^m \longmapsto x^{m+1}$ is the product iteration

$$\Phi_\omega^{\mathrm{ADI}} := \Phi_\omega^C \circ \Phi_\omega^B\,. \tag{8.45c}$$

8.5.2 General Representation

In the general case, we start from a splitting (8.42a): $A = B+C$ and assume (8.43a) in a weakened form. One of the matrices B or C may be only positive semidefinite. Without loss of generality, this might be C:

$$B > 0, \quad C \geq 0. \tag{8.46a}$$

Therefore, for

$$\omega > 0, \tag{8.46b}$$

the matrices $\omega I + B$ and $\omega I + C$ are positive definite and, in particular, regular. Hence, ADI iteration (8.45c) can be defined by (8.45a,b). To ensure practicability, we assume (8.46c):

$$\text{equations with } \omega I + B \text{ or } \omega I + C \text{ are easy to solve.} \tag{8.46c}$$

Theorem 8.41 (convergence). *(a) The iteration matrix of the ADI method is*

$$M_\omega^{\mathrm{ADI}} = (\omega I + C)^{-1}(\omega I - B)(\omega I + B)^{-1}(\omega I - C). \tag{8.47a}$$

(b) If (8.46a,b) holds, the ADI iteration converges.

Proof. M_ω^{ADI} is the product of the iteration matrices $(\omega I + C)^{-1}(\omega I - B)$ and $(\omega I + B)^{-1}(\omega I - C)$ of the respective half-steps Φ_ω^C and Φ_ω^B (cf. §5.4). Lemma A.20 allows a cyclic permutation of the factors in the argument of the spectral radius:

$$\begin{aligned}\rho(M_\omega^{\mathrm{ADI}}) &= \rho((\omega I - B)(\omega I + B)^{-1}(\omega I - C)(\omega I + C)^{-1}) \\ &\leq \left\|(\omega I - B)(\omega I + B)^{-1}(\omega I - C)(\omega I + C)^{-1}\right\|_2 \\ &\leq \left\|(\omega I - B)(\omega I + B)^{-1}\right\|_2 \left\|(\omega I - C)(\omega I + C)^{-1}\right\|_2 .\end{aligned} \tag{8.47b}$$

As B is Hermitian, $B_\omega := (\omega I - B)(\omega I + B)^{-1}$ is also. In particular, it is a normal matrix, implying that $\rho(B_\omega) = \|B_\omega\|_2$ (cf. Theorem B.25). Therefore, (8.47b) becomes

$$\rho(M_\omega^{\mathrm{ADI}}) \leq \rho(B_\omega)\,\rho(C_\omega) \tag{8.47c}$$

since analogous considerations also apply to $C_\omega := (\omega I - C)(\omega I + C)^{-1}$. By Remark A.15b, the spectrum of B_ω is equal to

$$\sigma(B_\omega) = \left\{\frac{\omega - \beta}{\omega + \beta} : \beta \in \sigma(B)\right\}, \quad \rho(B_\omega) = \max_{\beta \in \sigma(B)} \left|\frac{\omega - \beta}{\omega + \beta}\right|. \tag{8.47d}$$

By assumption (8.46a), β is positive. This fact implies that $|\omega - \beta| < |\omega + \beta|$ for all $\omega > 0$. This proves $\rho(B_\omega) < 1$. Since C is only positive semidefinite, a similar argument leads to $\rho(C_\omega) \leq 1$. (8.47c) proves $\rho(M_\omega^{\mathrm{ADI}}) < 1$. □

Exercise 8.42. Formulate a convergence statement in the case of normal matrices B and C under the condition that the splittings (8.44a,b) are regular. For which ω are (8.44a,b) regular splittings in the model case?

Next, we want to determine the optimal value ω_{opt} of the ADI method. Here, we restrict ourselves to the minimisation of $\rho(B_\omega)$. If, as for the model problem, $\rho(C_\omega) = \rho(B_\omega)$ holds, minimisation of $\rho(B_\omega)$ is equivalent to the minimisation of the bound $\rho(B_\omega)\rho(C_\omega)$ in (8.47c).

The extreme eigenvalues of B (or their bounds) are assumed to be

$$0 < \beta_{\min} \le \beta_{\max} \qquad \text{with } \sigma(B) \subset [\beta_{\min}, \beta_{\max}]. \tag{8.48a}$$

In the model case, as seen in the proof of Remark 8.39, the eigenvalues of B are $4h^{-2}\sin^2(jh\pi/2)$ for $1 \le j \le N-1$. This implies that

$$\beta_{\min} = 4h^{-2}\sin^2(h\pi/2), \qquad \beta_{\max} = 4h^{-2}\cos^2(h\pi/2).$$

For any $\beta \in [\beta_{\min}, \beta_{\max}]$ and therefore for any $\beta \in \sigma(B)$, we have

$$\left|\frac{\omega-\beta}{\omega+\beta}\right| \le \max\left\{\left|\frac{\omega-\beta_{\min}}{\omega+\beta_{\min}}\right|, \left|\frac{\omega-\beta_{\max}}{\omega+\beta_{\max}}\right|\right\} \qquad (\omega > 0) \tag{8.48b}$$

since $|\omega-\beta| / |\omega+\beta|$ as a function of β is decreasing in $[0,\omega]$ and increasing in $[\omega,\infty)$. To minimise the right-hand side in (8.48b), one has to determine ω from $\left|\frac{\omega-\beta_{\min}}{\omega+\beta_{\min}}\right| = \left|\frac{\omega-\beta_{\max}}{\omega+\beta_{\max}}\right|$. The result is given by

$$\omega_{\text{opt}} = \sqrt{\beta_{\min}\beta_{\max}}\,. \tag{8.48c}$$

Inserting this value into (8.47d), we obtain

$$\rho(B_{\omega_{\text{opt}}}) = \left(\sqrt{\beta_{\max}} - \sqrt{\beta_{\min}}\right) / \left(\sqrt{\beta_{\max}} + \sqrt{\beta_{\min}}\right).$$

Exercise 8.43. Prove for the Poisson model problem: (a) The following holds:

$$\begin{aligned}\omega_{\text{opt}} &= 2h^{-2}\sin h\pi,\\ \rho(B_{\omega_{\text{opt}}}) &= \left[\cos\frac{\pi h}{2} - \sin\frac{\pi h}{2}\right] / \left[\cos\frac{\pi h}{2} + \sin\frac{\pi h}{2}\right],\\ \rho(M^{\text{ADI}}_{\omega_{\text{opt}}}) &= [1-\sin(\pi h)] / [1+\sin(\pi h)]\,.\end{aligned}$$

(b) The convergence speed $\rho(M^{\text{ADI}}_{\omega_{\text{opt}}})$ coincides exactly with the optimal convergence rate (4.33) of the SOR iteration.

If we replace the definiteness in assumption (8.46a) by the M-matrix property, the convergence proof becomes much more difficult. A general convergence result of this kind (also for instationary ADI methods) is due to Alefeld [1]. Here, we call the method stationary if ω is constant during the iteration and instationary if it varies (as, e.g., it is assumed throughout the following section).

8.5.3 ADI in the Commutative Case

In addition to the assumptions (8.46a–c), we require that

$$BC = CB. \tag{8.49a}$$

Commutativity is equivalent to the simultaneous diagonalisability:

$$\begin{aligned} Q^{\mathsf{H}}BQ &= D_B = \operatorname{diag}\{\beta_\alpha : \alpha \in I\}, \\ Q^{\mathsf{H}}CQ &= D_C = \operatorname{diag}\{\gamma_\alpha : \alpha \in I\} \end{aligned} \tag{8.49b}$$

(cf. Theorem A.43), which here can be achieved by a unitary transformation Q, since B and C are Hermitian. Assumption (8.49b) implies that B_ω, C_ω, and the iteration matrix M_ω^{ADI} built from these matrices can also be transformed by Q to diagonal form (cf. (8.47a)):

$$Q^{\mathsf{H}}M_\omega^{\mathrm{ADI}}Q = \operatorname{diag}\left\{\frac{\omega-\gamma_\alpha}{\omega+\gamma_\alpha}\frac{\omega-\beta_\alpha}{\omega+\beta_\alpha} : \alpha \in I\right\}. \tag{8.49c}$$

In the following, we apply the ADI method with varying parameters $\omega = \omega_m$:

$$y^{m+1} = \Phi_{\omega_m}^{\mathrm{ADI}}(y^m, b) \qquad (m \in \mathbb{N}).$$

Exercise 8.44. Let x be the solution of $Ax = b$. Prove that the error $\eta^m = y^m - x$ has the representation

$$\eta^m = M_{\omega_m}^{\mathrm{ADI}} \cdot \ldots \cdot M_{\omega_1}^{\mathrm{ADI}}\eta^0.$$

We would like to choose the parameters $\omega_1, \omega_2, \ldots, \omega_m \geq 0$ such that the spectral norm of the matrix $M_{\omega_m}^{\mathrm{ADI}} \cdot \ldots \cdot M_{\omega_1}^{\mathrm{ADI}}$ becomes as small as possible:

$$\|M_{\omega_m}^{\mathrm{ADI}} \cdot \ldots \cdot M_{\omega_1}^{\mathrm{ADI}}\|_2 \overset{!}{=} \min. \tag{8.50a}$$

Multiplications by unitary matrices do not change the spectral norm:

$$\begin{aligned} \|Q^{\mathsf{H}}M_{\omega_m}^{\mathrm{ADI}} \cdot \ldots \cdot M_{\omega_1}^{\mathrm{ADI}}Q\|_2 &= \|Q^{\mathsf{H}}M_{\omega_m}^{\mathrm{ADI}}Q \cdot \ldots \cdot Q^{\mathsf{H}}M_{\omega_1}^{\mathrm{ADI}}Q\|_2 \\ &= \left\|\prod_{i=1}^{m} \operatorname{diag}\left\{\frac{\omega_i-\gamma_\alpha}{\omega_i+\gamma_\alpha}\frac{\omega_i-\beta_\alpha}{\omega_i+\beta_\alpha} : \alpha \in I\right\}\right\|_2. \end{aligned}$$

Together with (8.49c), we obtain

$$\left\|\operatorname{diag}_{\alpha\in I}\left\{\prod_{i=1}^{m}\frac{\omega_i-\gamma_\alpha}{\omega_i+\gamma_\alpha}\frac{\omega_i-\beta_\alpha}{\omega_i+\beta_\alpha}\right\}\right\|_2 = \max_{\alpha\in I}\left|\prod_{i=1}^{m}\frac{\omega_i-\gamma_\alpha}{\omega_i+\gamma_\alpha}\frac{\omega_i-\beta_\alpha}{\omega_i+\beta_\alpha}\right|.$$

Hence, the minimisation problem (8.50a) is equivalent to

$$\max_{\alpha\in I}\left|\prod_{i=1}^{m}\frac{\omega_i-\gamma_\alpha}{\omega_i+\gamma_\alpha}\frac{\omega_i-\beta_\alpha}{\omega_i+\beta_\alpha}\right| \overset{!}{=} \min\,. \tag{8.50b}$$

Remark 8.45. For $m \geq n := \#I$, as in §8.3.2, one finds parameters ω_i bringing the left-hand side in (8.50b) to the minimum 0. For this purpose, the values ω_i must be an enumeration of the eigenvalues $\{\gamma_\alpha : \alpha \in I\} \cup \{\beta_\alpha : \alpha \in I\}$.

Since, in general, γ_α or β_α are not known, we optimise over a larger set $[a, b]$ containing the spectra of B and C, as we did in the third minimisation problem (8.23):

$$0 < a \leq \gamma_\alpha, \beta_\alpha \leq b \qquad \text{for all } \alpha \in I.$$

Then, the minimisation problem takes the following form. Let

$$r_m(\zeta) := \prod_{i=1}^{m}\frac{\omega_i-\zeta}{\omega_i+\zeta}$$

be a rational function with a numerator and denominator of degree m replacing the previous polynomials. Substituting the discrete eigenvalues in (8.50b) by the interval $[a, b]$, we arrive at the problem

$$\begin{aligned}&\text{determine parameters } \{\omega_i : 1 \leq i \leq m\} \text{ so that}\\ &\max\{|r_m(\beta)r_m(\gamma)| : a \leq \beta, \gamma \leq b\} = \min\,.\end{aligned} \tag{8.51a}$$

Because of $\max_{\beta,\gamma}\{|r_m(\beta)r_m(\gamma)|\} = \max_\beta\{|r_m(\beta)|\}\max_\gamma\{|r_m(\gamma)|\}$, we may optimise each factor separately. Hence, problem (8.51a) simplifies to

$$\begin{aligned}&\text{determine parameters } \{\omega_i : 1 \leq i \leq m\} \text{ so that}\\ &\max\{|r_m(\zeta)| : a \leq \zeta \leq b\} = \min\,.\end{aligned} \tag{8.51b}$$

The following results are due to Wachspress [382] (see also Wachspress–Habetler [384] from 1960). We omit these proofs, since the derivation of Eqs. (8.52a–c) is presented in detail in the book of Varga [375, S. 224f].

Theorem 8.46 (optimal ADI parameters). *(a) For any $m \in \mathbb{N}$, the problem (8.51b) has a unique solution $\{\omega_1, \ldots, \omega_m\}$. The parameters ω_i are disjoint numbers in (a, b).*

(b) The increasingly ordered parameters $\omega_1 < \omega_2 < \ldots < \omega_m$ satisfy

$$\omega_{m+1-i} = ab/\omega_i \qquad \textit{for } 1 \leq i \leq m. \tag{8.52a}$$

(c) Denote the parameters $\omega_1 < \omega_2 < \ldots < \omega_m$ *belonging to* $m \in \mathbb{N}$ *and the interval* $[a, b]$ *with* $0 < a < b$ *by* $\omega_i(a, b, m)$ $(1 \le i \le m)$. *Then we have*

$$\omega_{2m+1-i}(a, b, 2m) = \omega_i\left(\sqrt{ab}, \tfrac{a+b}{2}, m\right) + \sqrt{\omega_i\left(\sqrt{ab}, \tfrac{a+b}{2}, m\right)^2 - ab} \quad (8.52b)$$

for $i = 1, \ldots, m$.

(d) The minimised quantities $\delta_m := \max\{|r_m(\zeta)| : a \le \zeta \le b\}$ *for* $m = 2^p$ *are*

$$\delta_m = \left(\sqrt{b_p} - \sqrt{a_p}\right) / \left(\sqrt{b_p} + \sqrt{a_p}\right), \quad (8.52c)$$

where $a_0 = a$, $b_0 = b$, $a_{i+1} = \sqrt{a_i b_i}$, $b_{i+1} = \frac{1}{2}(a_i + b_i)$ *for* $0 \le i \le p - 1$.

Determining the ADI parameters ω_i is very easy for binary powers $m = 2^p$. For $p = 0$ (i.e., $m = 1$), we conclude from (8.52a) that

$$\omega_1(a, b, 1) = \sqrt{ab}, \quad (8.52d)$$

repeating the result in (8.48c). As soon as the parameters for $m = 2^{p-1}$ are known, those for $2m = 2^p$ can be obtained from formula (8.52b) for the indices $2m+1-i \in [m + 1, \ldots, 2m]$. The parameters ω_i for $1 \le i \le m$ result from (8.48a).

Evidently, one may apply the calculated parameters ω_i in a cyclic manner: $\omega_{i+km} := \omega_i$ $(1 \le i \le m,\ k \in \mathbb{N})$. Different from the case in §8.3.7, the cyclic ADI process does not lead to stability problems.

δ_m in (8.52c) is the bound for $r_m(B_\omega)$ and $r_m(C_\omega)$. Therefore, the asymptotic rate is bounded by $\rho_m := \delta_m^{2/m}$. One recognises from (8.52c) that ρ_m depends only on the ratio a/b, which in the model case has the size $\mathcal{O}(h^2)$. The recursions $a_{i+1} = \sqrt{a_i b_i}$ and $b_{i+1} = \frac{1}{2}(a_i + b_i)$ prove the following remark.

Remark 8.47. Let $a/b = \mathcal{O}(h^\tau)$ and assume (8.49a). For the optimal choice of the parameters, the cyclic ADI method with m parameters has the order τ/m :

$$\rho_m = 1 - \mathcal{O}(h^{\tau/2m}) = 1 - C_m h^{\tau/2m} + \mathcal{O}(h^{\tau/m}).$$

Hence, the instationary ADI method permits not only halving of the order (for the case $m = 1$, compare also with Exercise 8.43b), but any arbitrarily small (and hence very favourable) order can be reached for sufficiently large m. However, we will see in §8.5.5 that the obvious conclusion of choosing a rather large number m leads to practical difficulties.

The construction of the parameters ω_i in Theorem 8.46d is restricted to $m = 2^p$. For other m, the description of ω_i requires elliptic integrals (cf. Wachspress [383], Samarskii–Nikolaev [330, page 276]). Lebedev [260] was the first suggesting that the solution to the approximation problem (8.51b) could be reformulated into another one for rational functions that is already solved in 1877 by Zolotarev. In this connection, we refer to the review paper of Todd [363] concerning the 'legacy of Zolotarev' (see also Todd [364]). Approximation problems appearing here also play an important role in the iterative solution of the Sylvester matrix equation $AX - XB = C$ $(A, B, C$ given, X unknown; cf. Starke [350] and Wachspress [383, §5]). Concerning the determination of the parameters in the case of nonsymmetric matrices, we refer to Starke–Niethammer [351].

Although the asymptotic convergence rates ρ_m in Remark 8.47 and the following Table 8.4 look quite favourable, the effective amount of work is less favourable because of the relatively expensive iteration (8.45a,b) (cf. Remark 8.49). Moreover, the assumption of commutativity (8.49a) is rarely satisfied in practice. As soon as it is violated, one is not able to achieve good convergence acceleration.

8.5.4 ADI Method and Semi-Iterative Methods

After choosing the Richardson method as the basic iteration, the half-steps (8.45a,b) have the representation (8.10b):

$$\begin{aligned} y^{m+\frac{1}{2}} &= \Theta_{m+\frac{1}{2}} \, (M_1^{\mathrm{Rich}} y^m \quad + N_1^{\mathrm{Rich}} b) + (1 - \Theta_{m+\frac{1}{2}}) \, y^m, \\ y^{m+1} &= \Theta_{m+1} \, (M_1^{\mathrm{Rich}} y^{m+\frac{1}{2}} + N_1^{\mathrm{Rich}} b) + (1 - \Theta_{m+1}) \, y^{m+\frac{1}{2}} \end{aligned}$$

with $M_1^{\mathrm{Rich}} = I - A$ and $N_1^{\mathrm{Rich}} = I,$ if we allow the matrix-valued factors

$$\Theta_{m+\frac{1}{2}} = (\omega I + B)^{-1}, \qquad \Theta_{m+1} = (\omega I + C)^{-1}.$$

These equations correspond to the second formulation in §8.2. If, as in the case of §8.5.3, B and C commute with A, we obtain the first formulation (8.3): $y^m = \sum \alpha_{mj} \, x^j$, where x^j are the Richardson iterates and α_{mj} are matrices commuting with A. In this sense, one might view the ADI method as a semi-iteration with matrix-valued coefficients.

On the other hand, the ADI method can function as a basic iteration of the Chebyshev method, as shown in the next exercise.

Exercise 8.48. Assume that B, C, and ω satisfy (8.46a,b) and (8.49a). Prove:
(a) The matrix of the third normal form of $\Phi_\omega^{\mathrm{ADI}}$ is

$$W_\omega = \tfrac{1}{2\omega}(\omega I + C)(\omega I + B) \qquad \text{(hint: (5.12c))}.$$

(b) $\Phi_\omega^{\mathrm{ADI}}$ is a positive definite iteration.

(c) Products $\Phi := \Phi_{\omega_1}^{\mathrm{ADI}} \circ \Phi_{\omega_2}^{\mathrm{ADI}} \circ \ldots \circ \Phi_{\omega_m}^{\mathrm{ADI}}$ with $\omega_j > 0$ form a positive definite iteration. Hint: Determine N_Φ in $\Phi(x, b) = x - N_\Phi(Ax - b)$ and show that $N_\Phi > 0$.

(d) In the stationary case, choose ω according to (8.52d). Determine the bounds in $\gamma W \leq A \leq \Gamma W$. What is the optimal damping factor ϑ_{opt} for $\Phi_\omega^{\mathrm{ADI}}$ (cf. Theorem 6.7)?

8.5.5 Amount of Work and Numerical Examples

The ADI method was already applied in §5.5.6 as a secondary iteration. In the following, we consider a general five-point formula ($C_A = 5$). The amount of work for solving the equations with the tridiagonal matrices $\omega I + C$, $\omega I + B$ amounts to $5n$ operations. Evaluating $b + \omega x - Cx$ and $b + \omega x - Bx$ requires $6n$ operations each. Because of $C_A = 5$, this leads to

$$C_\Phi^{\mathrm{ADI}} = 4.4$$

and, in the Poisson model case, even to $C_\Phi^{\mathrm{ADI}} = 4$.

m	$h = 1/32$	$1/64$	$1/128$
1	0.8215	0.9065	0.9521
2	0.5231	0.6373	0.7291
4	0.3735	0.4607	0.5365
8	0.3141	0.3874	0.4513
16	0.2880	0.3553	0.4139

Table 8.4 Asymptotic convergence rates ρ_m for ADI-cycle length m.

The asymptotic rates $\rho_m = \delta_m^{1/m}$ attainable by (8.52c) are reported in Table 8.4. We observe that for small step sizes, good rates are achieved. The concrete results for $h = \frac{1}{128}$ with $m = 4$ different parameters from Table 8.5 confirm that the factor 0.5365 in Table 8.4 is reached. The convergence behaves regularly modulo m. Each second ratio $\|e^k\|_2 / \|e^{k-1}\|_2$ is ≈ 1. However, since one cannot achieve the accuracy of $\|e^k\|_2 \approx \delta_m^k \|e^0\|_2$ with fewer than m iteration steps, the following dilemma arises:

(i) To exploit the good (asymptotic) convergence rate δ_m for large m, one must perform at least m iterations.

(ii) On the other hand, one would like to stop the iteration as soon as, e.g., the error becomes $\|e^k\|_2 \approx 1/1000$ (cf. Remark 2.34). The better the convergence rate, the fewer iterations one is willing to perform.

In the example of Table 8.5, about eight steps would be sufficient. Hence, one could still enlarge the cycle length from 4 to 8 (the corresponding result is $\|e^8\|_2 = 1.05_{10}$-3); a further increase to 16 or more parameters would not help. The last two columns in Table 8.5 correspond to $\rho_{k,k-1}$ and $\rho_{k,0}$.

k	value in the middle	$\|e^k\|_2$	$\frac{\|e^k\|_2}{\|e^{k-1}\|_2}$	$\sqrt[m]{\frac{\|e^k\|_2}{\|e^0\|_2}}$
1	-0.0320913257	5.02_{10}-1	0.6446	0.6446
2	-0.0342861024	4.62_{10}-1	0.9106	0.7699
3	0.3534506991	5.98_{10}-2	0.1295	0.4250
4	0.3538351873	5.49_{10}-2	0.9187	0.5154
5	0.4031600829	3.87_{10}-2	0.7055	0.5488
6	0.4063222547	3.68_{10}-2	0.9487	0.6012
7	0.4976831847	4.47_{10}-3	0.1215	0.4784
8	0.4976688610	4.26_{10}-3	0.9536	0.5215
9	0.4961625617	3.10_{10}-3	0.7284	0.5412
10	0.4961175712	2.97_{10}-3	0.9568	0.5729
11	0.4990489164	3.50_{10}-4	0.1179	0.4962
12	0.4990525844	3.37_{10}-4	0.9625	0.5244
13	0.4993912351	2.51_{10}-4	0.7454	0.5388
14	0.4994041549	2.41_{10}-4	0.9612	0.5615
15	0.4999776724	2.79_{10}-5	0.1154	0.5053
16	0.4999776614	2.69_{10}-5	0.9671	0.5262

Table 8.5 ADI results (Poisson model problem, four parameters ω_i, $h = 1/128$).

Remark 8.49. Good convergence rates are combined with a relatively high cost factor $C_\Phi^{\mathrm{ADI}} = 4$ in the Poisson model case. For the example in Table 8.5, the effective amount of work is equal to $\mathrm{Eff}(\Phi^{\mathrm{ADI}}) = \frac{-4}{\log 0.5365} = 6.42$. For $h = 1/32$ and four parameters, we obtain $\mathrm{Eff}(\Phi^{\mathrm{ADI}}) = 4.06$ (for comparison: $\mathrm{Eff}(\Phi^{\mathrm{SOR}}) = 7.05$ in Example 2.28 and $\mathrm{Eff}(\Phi_{\mathrm{semi}}^{\mathrm{SSOR}}) = 4.38$ in (8.41)).