The Analysis of $k$-Step Iterative Methods for Linear Systems from Summability Theory

W. Niethammer$^{1,3}$ and R.S. Varga$^{2,3}$

$^1$ Institut für Praktische Mathematik, Universität Karlsruhe, D-7500 Karlsruhe, West Germany
$^2$ Institute for Computational Mathematics, Kent State University, Kent, Ohio 44242, USA

Summary. Using the theory of Euler methods from summability theory, we investigate general iterative methods for solving linear systems of equations. In particular, for a given Euler method, a region $S$ of the complex plane is determined such that a $k$-step iterative method converges if the eigenvalues of an iteration operator $T$ are contained in $S$. For a given $S$, optimal methods are described, and upper and lower bounds are derived for the associated asymptotic rate of convergence. Special attention is given to two-step methods with complex parameters.

Subject Classifications: AMS(MOS): 65F10, 40G05; CR 5.14.

Contents

1. Introduction ................................................................. 177
2. Background Material from Summability Theory ......................... 182
3. Euler Functions and General Euler Methods .......................... 184
4. Rates of Convergence of Euler Methods ................................ 187
5. Computation of the Euler Transform .................................. 189
6. The Special Case $k=2$: Real Parameters ............................ 192
7. The Special Case $k=2$: Complex Parameters ........................ 193
8. Optimal Methods with Respect to a Given Set ....................... 196
9. Examples ................................................................... 198

§1. Introduction

Given the following system of $n$ linear equations in $n$ unknowns,

$$Ax = b, \quad (\text{where } A \in \mathbb{C}^{n \times n} \text{ is nonsingular}), \quad (1.1)$$

let $A = M - N$ be any splitting of $A$ (i.e., $M$ is nonsingular), so that (1.1) can be expressed as

$$x = Tx + c, \quad (\text{where } T = M^{-1}N \text{ and where } c = M^{-1}b). \quad (1.2)$$

$^3$ Research supported in part by the Air Force Office of Scientific Research

* Dedicated to Professor Johannes Weissinger (Universität Karlsruhe) on the occasion of his seventieth birthday, May 12, 1983
Now, (1.2) induces the following simple iterative method
\[ x_{m+1} := Tx_m + c, \quad m = 0, 1, \ldots, \] (1.3)
which, as is well known, converges, for any choice of the initial vector \( x_0 \) to the unique solution \( x \) of (1.1) if \( \rho(T) < 1 \), (where if \( \sigma(T) := \{ \lambda; \det(\lambda I - T) = 0 \} \) denotes the spectrum of \( T \), then \( \rho(T) = \max \{ |\lambda|; \lambda \in \sigma(T) \} \) denotes its spectral radius). Moreover, if \( e_m := x_m - x \), \( m = 0, 1, \ldots \), is the error vector associated with the iterate \( x_m \), it is further well-known (cf. Young [14, p. 87], Varga [13, p. 67]) that
\[ \rho(T) = \lim_{m \to \infty} (\|T^n\|)^{1/m} = \lim_{m \to \infty} (\sup_{\|z\|=1} \|T^m z\|)^{1/m}, \] (1.4)
for any vector norm \( \| \cdot \| \) on \( \mathbb{C}^n \), where \( \| T^n \| \) denotes the associated induced operator norm of \( T^n \). Consequently, for any \( e_0 \neq 0 \), we have from (1.4) that
\[ \lim_{m \to \infty} (\|e_m\|/\|e_0\|)^{1/m} \leq \rho(T), \] (1.5)
so that \( (\rho(T))^m \) is a rough measure of how \( \|e_m\|/\|e_0\| \) behaves asymptotically, as \( m \to \infty \).

For each positive integer \( k \), suppose we consider, in place of (1.3), a \( k \)-step stationary iterative method for (1.2), defined by
\[ y_{m} := \mu_0 (Ty_{m-1} + c) + \mu_1 y_{m-1} + \ldots + \mu_k y_{m-k}, \quad m = k, k + 1, \ldots, \] (1.6)
where \( y_0, y_1, \ldots, y_{k-1} \) are given starting vectors, and where \( \mu_0, \mu_1, \ldots, \mu_k \) are fixed complex numbers (independent of \( m \)) which are assumed to satisfy
\[ \mu_0 + \mu_1 + \ldots + \mu_k = 1. \] (1.7)

Obviously, on choosing the specific starting vectors as \( y_0 = \ldots = y_{k-1} = x \), we see from (1.6) and (1.7) that \( y_m = x \) for all \( m \geq 0 \). Next, we note that the choice \( \mu_0 = 1 \) and \( \mu_1 = \ldots = \mu_k = 0 \) for \( k \geq 1 \) is such that that iterative method of (1.3) reduces to that of (1.3). Moreover, the iterative methods of (1.3) and (1.6) can require just one matrix multiplication by the matrix \( T \), although more vector storage is in general required with (1.6), as compared with (1.3).

Because of the condition (1.7), the iterative method defined in (1.6) can be thought of as a type of averaging method of the matrix \( T \) and the identity matrix \( I \). For example, if \( k = 1 \) in (1.6), the condition (1.7) implies that (1.6) can be expressed as
\[ y_m = (1 - \mu_0) I + \mu_0 T \] (1.6)
which is known in the literature (cf. [14, p. 361]) both as an extrapolation iterative method, and a stationary first-order Richardson method, based on (1.2). Similarly, the case \( k = 2 \) of (1.6) becomes
\[ y_m = (\mu_1 I + \mu_0 T) y_{m-1} + (1 - \mu_0 - \mu_1) y_{m-2} + \mu_0 c, \quad m = 2, 3, \ldots, \]
which is known in the literature (cf. [14, p. 487]) both as a linear second-order iterative method, and as a stationary second-order Richardson iterative method, based on (1.2). It is interesting to note that optimized semiiterative methods (nonstationary second-order Richardson methods) degenerate, after sufficiently many iterations, precisely into the stationary 2-step methods of (1.6), when computers with finite word lengths are used (cf. Golub/Varga [3]).

Our aim here is to study the k-step stationary iterative method of (1.6), with the goal of understanding the theory for selecting the parameters \( \{\mu_j\}_{j=0}^k \), so as to make the associated error vectors

\[
\tilde{e}_m = y_m - x, \quad m = k, k+1, \ldots
\]

(1.8)
tend to zero as rapidly as possible, when \( m \to \infty \). It is obvious that any choice or optimization of the parameters \( \{\mu_j\}_{j=0}^k \) goes hand-in-hand with a knowledge of the spectrum \( \sigma(T) \). Usually, it is assumed that some information is given about \( \sigma(T) \), and the problem then is to find associated optimum parameters \( \{\mu_j\}_{j=0}^k \). Our point of view here, just the opposite of the above, can be posed as the following

**Question.** Given some parameters \( \{\mu_j\}_{j=0}^k \) satisfying (1.7), what are the resulting geometrical conclusions on the spectrum \( \sigma(T) \) of the matrix \( T \) such that the associated error vectors \( \tilde{e}_m \) of (1.8) decrease in norm as \( m^\kappa \), as \( m \to \infty \), where \( \kappa \) is assumed to satisfy \( 0 \leq \kappa < 1 \)?

If one can precisely describe these regions containing the spectrum of \( T \) as a function of \( \kappa \), then one has a better understanding of how to select the coefficients \( \{\mu_j\}_{j=0}^k \) in (1.6), so as to nearly "optimally" fit a specific spectrum \( \sigma(T) \). This has been recently considered by de Pillis [12] and Manteuffel [8] in a related context, where ellipses were used to fit (or "embrace") a spectrum \( \sigma(T) \) in the special case \( k = 2 \) in (1.6).

Our approach to the above posed Question comes from the **theory of summability**. Clearly, as (1.2) possesses a unique solution iff \( 1 \notin \sigma(T) \), then the unique solution of (1.2) is \( x = (I - T)^{-1} e \), assuming \( 1 \notin \sigma(T) \). If, moreover, \( \rho(T) < 1 \), then \( x \) can be computed from the convergent Neumann series

\[
x = (I - T)^{-1} e = \sum_{j=0}^{\infty} T^j e; \tag{1.9}
\]

and the iterations of (1.3) can be seen as a recursive computation of the partial sums \( x_m = \sum_{j=0}^{m} T^j e \) of the vector sum in (1.9), provided that \( x^{(0)} = e \). Now, if the Neumann series \( \sum_{j=0}^{\infty} T^j \) for \( (I - T)^{-1} \) can be transformed by a **general Euler method** (to be defined in §3), there results a polynomial series \( \sum_{j=0}^{\infty} v_j(T) \), where each \( v_j(T) \) is a polynomial in the matrix \( T \), of degree at most \( j \). If this series converges to \( (I - T)^{-1} \), then the solution \( x \) of (1.2) can be expressed as

\[
x = (I - T)^{-1} e = \sum_{j=0}^{\infty} v_j(T) e. \tag{1.10}
\]
For a special class of general Euler methods, which depend on the parameters \( \{\mu_j\}_{j=0}^{k} \), the partial sums \( y_m := \sum_{j=0}^{m} v_j(T) e \), for \( m = 0, 1, \ldots \), of the series of (1.10) can in fact be computed recursively from (1.6), provided that the initial vectors \( \{y_j\}_{j=0}^{k-1} \) are suitably chosen. This recursive computation is, of course, important in applications. The central idea is that general Euler methods are generated by a conformal mapping function \( p \), and it is the properties of this conformal mapping \( p \) which will give an answer to our Question above. This summability approach, which one can find briefly described in Niethammer [10], also reveals a connection with the methods of Kublanovskaja [7], described in Faddeev-Faddeeva ([1], p. 641ff).

We now briefly describe the remainder of our paper. In §2 we describe how a summability method, represented by an infinite triangular matrix \( \Psi \), can be applied to the Neumann series of \( T \). In §3 for a general Euler method, \( \Psi \) is generated by an Euler function \( p \). Such a \( p \) determines a region \( S(p) \) in \( \mathbb{C} \) such that (1.10) converges to \( x \) whenever \( \sigma(T) \subseteq S(p) \). In §4, for estimating the decay of the errors, it is useful to introduce, for some \( z \in S(p) \), the quantity

\[
\kappa(z,p) := \lim_{j \to \infty} |v_j(z)|^{1/j}.
\]

From the Jordan normal form of \( T \), it can be seen that an appropriate measure for the (asymptotic) decay of the terms \( v_j(T) \) of (1.10) is

\[
\kappa(T,p) := \max_{\tau \in \sigma(T)} \kappa(\tau, p),
\]

(1.11)
as similarly considered by Manteuffel [8]. If \( \tilde{p} \) is the identity mapping, then \( v_j(T) = T^j \) and \( \kappa(T, \tilde{p}) = \rho(T) \).

For each Euler function \( p \), there exists a maximal number \( \hat{\eta}(p) > 1 \), such that \( p \) is meromorphic and univalent in the disk with radius \( \hat{\eta}(p) \); we call \( \hat{\eta}(p) \) the maximal extension of \( p \) (cf. Def. 2 of §3). Then, for each \( \eta \) with \( 1 < \eta \leq \hat{\eta}(p) \), a closed subset \( S_\eta(p) \) of \( S(p) \) can be described such that

\[
\kappa(T, p) \leq \frac{1}{\eta},
\]

(1.11')

whenever \( \sigma(T) \subseteq S_\eta(p) \) (cf. Corollary 2).

Recursive formulas for the partial sums \( y_m \) of (1.10) are derived in §5; for computing \( y_m \), all preceding \( y_j \) \((j = 0, \ldots, m - 1)\) have to be stored. However, if \( p \) is given by

\[
p(\phi) = \frac{\mu_0 \phi}{1 - \mu_1 \phi - \cdots - \mu_k \phi^k}, \quad \mu_0 + \cdots + \mu_k = 1,
\]

(1.12)

then the central formula (1.6) results, so that a finite recursive formula is obtained (cf. §6).

As an example, we describe in §6 the regions \( S(p) \) and \( S_{\eta}(p) \) for a \( p \) of the form (1.12) with \( k = 2 \) and with real parameters. Then, \( p \) is an Euler function for \( \mu_0 > 0 \) and \(|\mu_2| < 1 \), and \( S(p) \) is the interior of an ellipse with center \(-\mu_1/\mu_0\), and semiaxes \((1 - \mu_2)/\mu_0\) and \((1 + \mu_2)/\mu_0\). \( S_{\eta}(p) \) is similarly shown to be the
closed interior of a confocal ellipse for \(1 \leq \eta < \hat{\eta}_i\), where \(\hat{\eta}_i = 1/\sqrt{\mu_i}\). For \(\eta = \hat{\eta}_i\), \(S_{\eta}(p)\) is the interval between the foci.

If \(\alpha\) and \(\beta\) are different complex numbers such that the line segment \([\alpha, \beta]\) joining \(\alpha\) and \(\beta\) doesn’t contain the point \(\gamma = 1\), then there exists a unique ellipse \(E\) in the complex plane such that \(E\) contains \(1\) and has foci \(\alpha\) and \(\beta\); in \(\S\), the optimal \(p_0\) for this segment \([\alpha, \beta]\) (and for all confocal elliptic regions \(E_{\eta} \subset \text{int}(E)\)) are constructed.

If one examines the problem of adjusting the parameters, it is useful to consider, rather than one special operator \(T\), the class of operators
\[
O_U := \{ T \in \mathcal{C}^* : \sigma(T) \subset U \}
\]
for some compact set \(U\), containing more than one point, whose complement contains \(1\) and is simply connected. For an Euler function \(p\) with \(S(p) \supset U\), we define from (1.11)
\[
\kappa(U, p) := \min \{1/\eta : 1 < \eta \leq \hat{\eta}(p) \text{ and } U \subset S_{\eta}(p)\}
\]
as the convergence factor of \(U\) with respect to \(p\). By definition (cf. (1.11) and (1.11'))
\[
\kappa(T, p) \leq \kappa(U, p) \quad \text{for all } T \in O_U.
\]
Now, the Euler function \(p_0\) is said to be optimal with respect to \(U\), if
\[
\kappa(U, p_0) \leq \kappa(U, p) \quad \text{for all Euler functions } p \text{ with } S(p) \supset U.
\]
From the Riemann Mapping Theorem, it follows (Theorem 8) that, for every \(U\), there always exists an optimal \(p_0\). Since the number \(\kappa(U, p_0)\) is unique, we set
\[
\kappa(U) := \kappa(U, p_0).
\]

Now, it can be seen (Corollary 10) that each Euler function \(p\) is optimal with respect to every \(S_{\eta}(p)\), where \(1 < \eta \leq \hat{\eta}(p)\); there holds
\[
\kappa(S_{\eta}(p)) = 1/\eta, \quad 1 < \eta \leq \hat{\eta}(p).
\]

Further, if \(U_1 \subsetneq U_2\) we have from Schwarz’s Lemma that (cf. Theorem 9)
\[
\kappa(U_1) < \kappa(U_2).
\]
Together with the result of (1.16), upper and lower bounds for \(\kappa(U)\) can easily be obtained, if Euler functions \(p_1\) and \(p_2\) are known to satisfy \(S_{\eta_1}(p_1) \supset U \subset S_{\eta_2}(p_2)\); thus (cf. Corollary 11),
\[
1/\eta_1 < \kappa(U) < 1/\eta_2.
\]

If an Euler function of the form (1.12) is given, then for \(m > k\), the recursive computation of the partial sums \(x_m\) of the Euler transform yields formula (1.6), where \(y_i\) is replaced by \(x_i (i= m-k, \ldots, m)\). If we want to interpret (1.6) as a
general $k$-step iteration formula, we have to admit arbitrary starting vectors $y_j$ ($j = 0, \ldots, k - 1$). Then, of course, the sequences $\{x_m\}_{m=0}^{\infty}$ and $\{y_m\}_{m=0}^{\infty}$ differ. But, in §9 it is shown that both sequences have the same asymptotic convergence behavior.

In the rest of §9, five examples are described. First, $\kappa([-v,v])$ for $0 < v < 1$, and $\kappa([-iv,iv])$ for arbitrary $v > 0$, are determined. The third example deals with the SOR-method (the successive overrelaxation method), applied to the special linear system

$$
\begin{bmatrix}
0 & B_1 \\
B_2 & 0
\end{bmatrix}
\begin{bmatrix}
x_m \\
y_m
\end{bmatrix} = 
\begin{bmatrix}
x_m + c_1 \\
y_m + c_2
\end{bmatrix},
$$

(1.19)

i.e., the iteration matrix in (1.19) is weakly two-cyclic in normal form (see [13], §4.1). Then, the SOR-method can be written as a two-step method for the partial vector $x$:

$$
x_m = \omega^2 (Tx_{m-1} + c) + 2(1-\omega)x_{m-2} - (1-\omega)^2 x_{m-2} \quad (m \geq 2)
$$

(1.20)

with $T = B_1 + B_2, c = c_1 + B_1 c_2$. Then, on applying the results of §6 and §7, the convergence behavior of the SOR-method with an arbitrary complex relaxation parameter $\omega$ satisfying $|\omega - 1| < 1$, can be determined.

Finally in Example 4, a special 4-step method is examined, whereas in Example 5, relation (1.18) is applied for the special case of the rectangle $R_v := \{z \in \mathbb{C} : -v \leq \Re z \leq v, -1 \leq \Im z \leq 1\}, 0 < v < 1$, i.e., upper and lower bounds for $\kappa(R_v)$ are derived.

§2. Background Material from Summability Theory

The sequence of vectors $\{x_m\}_{m=0}^{\infty}$, generated by (1.3), converges to the unique solution $x$ of (1.2), for any choice of $x^{(0)}$, iff $\rho(T) < 1$. Suggested by the classical theory of summability, the following sequence of vectors

$$
y_m := \sum_{j=0}^{m} y_{m,j} x_j \quad (m \geq 0)
$$

(2.1)

can be considered. The infinite triangular matrix $\Sigma := (y_{m,j})_{m \geq 0, 0 \leq j \leq m}$ yields the so-called sequence-to-sequence transformation. (For this notation and others appearing in the following, see e.g., Zeller-Beekmann [15]). If the condition

$$
\sum_{j=0}^{m} y_{m,j} = 1 \quad (m \geq 0)
$$

(2.2)

is imposed on $\Sigma$, then for the error vector $\epsilon_m := y_m - x$ of (1.8), it follows that

$$
\epsilon_m = q_m(T) \epsilon_0,
$$

(2.3)
where the polynomial $q_m(z)$ is defined by

$$q_m(z) = \sum_{j=0}^{m} \gamma_{m,j} z^j.$$  \hspace{1cm} (2.4)

If it is known that the eigenvalues of a Hermitian matrix $T$ are contained in the real interval $[\alpha, \beta]$ with $\beta < 1$, then, with respect to this information on the eigenvalues of $T$, the use of properly normalized Chebyshev polynomials $t_m(z)$ guarantees a maximal reduction of the Euclidean norm of the associated error vectors $\tilde{e}_m$. Moreover, the three-term recurrence relation for these polynomials $t_m(z)$ induces a corresponding three-term recurrence formula (cf. Chap. 5 of [13]) for the vectors $\gamma_m$, which by-passes the use of formula (2.1) which requires the storage of all intermediate vectors $x_j$. Methods of this type are usually called semi-iterative methods.

On the other hand, if the eigenvalues of $T$ are complex, but symmetric with respect to the real line and can be enclosed in an ellipse which doesn't contain the point $z=1$, then a corresponding iteration, again based on Chebyshev polynomials, in optimal, but only in an asymptotic sense (cf. [8]).

There is a second way, also suggested from summability theory, for transforming the vector iterates $\{x^{(m)}\}_{m=0}^{\infty}$ of (1.3). Instead of a sequence-to-sequence transformation (2.1), we consider the following series-to-series transformation. Given an infinite triangular matrix

$$\Psi = (\pi_{i,j})_{i \geq 0, \ j \leq i},$$  \hspace{1cm} (2.5)

the terms $T^j$ of the Neumann series (1.9) are transformed according to

$$v_j(T) = \sum_{i=0}^{j} \pi_{j,i} T^i, \quad j = 0, 1, \ldots,$$  \hspace{1cm} (2.6)

i.e.,

$$\begin{bmatrix}
  v_0(T) \\
  v_1(T) \\
  \vdots \\
  v_j(T)
\end{bmatrix} =
\begin{bmatrix}
  \pi_{0,0} \\
  \pi_{1,0} \quad \pi_{1,1} \quad 0 \\
  \vdots \\
  \pi_{j,0} \quad \pi_{j,1} \ldots \pi_{j,j}
\end{bmatrix}
\begin{bmatrix}
  I \\
  T \\
  \vdots \\
  T^j
\end{bmatrix}$$

The resulting series

$$\sum_{j=0}^{\infty} v_j(T)$$  \hspace{1cm} (2.7)

is considered as the transformed Neumann series. To this transformation, determined by $\Psi$, there corresponds a sequence-to-sequence transformation $\Sigma$ of the partial sums of the Neumann series to the partial sums of (2.7). Between $\Psi$
and \( \Phi \), the following relation holds (cf. [15, p. 7])

\[
\Phi = \Sigma \Phi \Sigma^{-1},
\]

where

\[
\Sigma = \begin{bmatrix}
1 \\
1 & 1 & 0 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
\cdots & \cdots & \cdots \\
1 & 1 & 1 \\
\end{bmatrix}, \quad \Sigma^{-1} = \begin{bmatrix}
1 \\
-1 & 1 & 0 \\
-1 & 1 & \cdots \\
0 & \cdots & \cdots \\
\end{bmatrix}.
\]

If the transformed Neumann series (2.7) converges to \((I - T)^{-1}\), then the solution \( x \) of (1.2) is given by

\[
x = (I - T)^{-1} e = \sum_{j=0}^{\infty} v_j(T) e. \tag{2.9}
\]

We note that if \( \Phi \) of (2.5) is chosen to be the identity matrix, then so is \( \Phi \) from (2.8), and from (2.6), we have \( v_j(T) = T^j \). Thus, when \( \rho(T) < 1 \), (1.9) is a special case of (2.9).

There are two reasons for applying such transformations. First the series (2.9) may converge more rapidly than that of (1.9), and second, the series (2.9) may even converge in cases where (1.9) diverges. This raises two obvious questions.

(i) Under what assumptions on \( \Phi \) and \( T \) does (2.9) hold, i.e., under what conditions on \( \Phi \) and \( T \) does the series (2.7) converge?

(ii) If (2.9) is valid, i.e., (2.7) converges to \((I - T)^{-1}\), does there exist a low-order recurrence relation for the partial sums of \( \sum_{j=0}^{\infty} v_j(T) e \) which allows one to by-pass the direct evaluation (and storage) required in (2.6)?

It should be mentioned that, in summability theory, the transformation (2.6) is not restricted to triangular matrices so that in general, an infinite series would result in (2.6). But since, in this case, no recurrence formula is known to exist, we confine ourselves to triangular matrices and especially to the so-called Euler methods which were examined in 1923 by Perron [11] and in 1926 by Knopp [5] as generalizations of Euler's series transformation (see [15], p. 132). We sketch below the derivation which reveals a close connection to known iterative methods for solving the equation \( x = Tx + c \); a more detailed treatment is given in Niethammer [9]. A further connection exists to the so-called "universal algorithms" of Kublanovskaya [7], described in [1].

§3. Euler Functions and General Euler Methods

For notation, let \( D_\eta = \{ \phi \in \mathbb{C} : |\phi| < \eta \} \) for any \( \eta > 0 \), and let \( \tilde{D}_\eta \) denote its closure. We make the following
Definition 1. A complex-valued function $p$ is said to be an Euler function (written $p \in \mathcal{E}$) if there exists an open neighborhood $\mathcal{D}$ of $D_1$, such that

a) $p$ is meromorphic and univalent (schlicht) in $\mathcal{D}$, and if
b) $p(0) = 0$, $p(1) = 1$.

We remark that if $p$ is an Euler function, then from a), $p$ can have at most one pole, of order one, in $\bar{D}_1$. Next, if $p$ is an Euler function, then from b), there exists a $v > 0$ such that $p$ is holomorphic in $D_v$. Thus, the following power series

$$ p(\phi) = \sum_{j=1}^{\infty} \pi_{j,1} \phi^j, $$

(3.1)

converges (uniformly and absolutely) for $|\phi| < v$. This implies that there exist power series for all powers of $p$, i.e.,

$$ [p(\phi)]^m = \sum_{j=m}^{\infty} \pi_{j,m} \phi^j, \quad m = 0, 1, 2, \ldots, \quad \text{for } |\phi| < v. $$

(3.2)

Hence, from an Euler function $p$, the coefficients $\{\pi_{j,m}\}_{j=m}^{\infty}$ of (3.2) can be used to define the $j$-th column ($0 \leq j < \infty$) of an infinite matrix $\mathfrak{B}$, and the series-to-series transformation (2.5) induced by $\mathfrak{B}$ is called the associated general Euler method. As we shall see, this introduction of a general Euler method is useful for the computation of the transformed series (2.9), which we call the Euler transform.

The following definition will be useful in the next section.

Definition 2. For each $p \in \mathcal{E}$, the maximal extension, $\hat{\eta}(p)$, is $\hat{\eta} = \hat{\eta}(p) := \max \{ \eta > 1 : p$ is meromorphic and univalent in $D_\eta \}$.\footnote{Note, from Definition 1, that $1 < \hat{\eta}(p) \leq \infty$ for each $p \in \mathcal{E}$.}

For the understanding of its analytical properties, the following derivation of the Euler transform is more appropriate. We introduce the resolvent operator for an $n \times n$ complex matrix $T$ with $1 \notin \sigma(T)$:

$$ r(\zeta, T) := (I - \zeta T)^{-1}; $$

(3.3)

$r$ is evidently matrix-valued and meromorphic, the poles $\zeta_i$ of $r$ and the eigenvalues $\tau_i$ of $T$ are related via $\zeta_i = 1/\tau_i$ ($i = 1, \ldots, n$). We are interested in $r(1, T) = (I - T)^{-1}$, which can be represented by the Neumann series $\sum_{j=0}^{\infty} T^j$ if there are no poles $\zeta_i$ with $|\zeta_i| \leq 1$, i.e., if there are no eigenvalues $\tau_i$ of $T$ with $|\tau_i| \leq 1$.

Since $r$ is holomorphic at $\zeta = 0$, then there is an $\eta > 0$ for which

$$ r(\zeta, T) = (I - \zeta T)^{-1} = \sum_{j=0}^{\infty} \zeta^j T^j, \quad \text{for } |\zeta| < \eta. $$

(3.4)

For an Euler function $p$, we now make the substitution $\zeta = p(\phi)$ in (3.4), use (3.2), and reorder according to the powers of $\phi$. Thus, there is an $\eta > 0$ for
which
\[ R(\phi, T) := r(p(\phi), T) = (I - p(\phi) T)^{-1} = \sum_{j=0}^{\infty} v_j(T) \phi^j, \quad \text{for } |\phi| < \tilde{\eta}, \] (3.5)
with
\[ v_0(T) := I, \quad \text{and} \quad v_j(T) := \sum_{i=1}^{j} \pi_{j,i} T^i \quad \text{for all } j = 1, 2, \ldots, \] (3.6)
which agrees with (2.6).

Now, our interest is in having the representation of \( R(\phi, T) \) in (3.5), valid for all \(|\phi| \leq 1\). If this were to be true, then the case \( \phi = 1 \) of (3.5) would give us that
\[ R(1, T) \mathbf{c} = (I - T)^{-1} \mathbf{c} = \sum_{j=0}^{\infty} v_j(T) \mathbf{c}, \]
so that the Euler transform of the Neumann series of \((I - T)^{-1}\) is convergent. This leads us to

**Theorem 1.** Assume \( p \in \mathcal{C} \). If
\[ \sigma(T) \subset \mathcal{C} \approx \tilde{p}(\tilde{D}_1) = : S(p), \]
(3.7)
where \( \tilde{D}_1 \) is the closed unit disk and \( \tilde{p} := 1/p \), then \( R(\phi, T) \) is nonsingular for all \( \phi \in \tilde{D}_1 \), and the expansion of (3.5) converges absolutely in \( \phi \) in \( \tilde{D}_1 \). Conversely, if the expansion in (3.5) is absolutely convergent for all \( \phi \) in \( \tilde{D}_1 \), then \( \sigma(T) \subset S(p) \).

**Proof.** From (3.5), we have that there exists an \( \tilde{\eta} > 0 \) for which the representation of (3.5) is valid. We claim, with the hypothesis of (3.7), that \( \tilde{\eta} > 1 \). To show this, assume \( \tilde{\eta} \leq 1 \), which implies that this representation has a singularity in some \( \tilde{\phi} \) with \(|\tilde{\phi}| = \tilde{\eta} \leq 1\), i.e., from (3.5) \((I - p(\tilde{\phi}) T)\) is singular. Hence, there is an eigenvalue \( \tau_i \) of \( T \) such that \( p(\tilde{\phi}) \tau_i = 1 \), so that
\[ \tau_i = \frac{1}{p(\tilde{\phi})} = \tilde{p}(-\tilde{\phi}). \]
But this contradicts (3.7).

Conversely, assume that the expansion in (3.5) is valid for all \( \phi \in \tilde{D}_1 \). Let \( \tau_i \) be any eigenvalue of \( T \), so that \( T \mathbf{x} = \tau_i \mathbf{x} \) for some \( \mathbf{x} \neq 0 \). Then, from (3.5),
\[ \left( \sum_{j=0}^{\infty} v_j(T) \phi^j \right) \mathbf{x} = \left( \sum_{j=0}^{\infty} v_j(\tau_i) \phi^j \right) \mathbf{x} = \frac{\mathbf{x}}{(1 - p(\phi) \tau_i)} \]
for all \( \phi \in \tilde{D}_1 \). Thus, \( \tau_i p(\phi) \neq 1 \) for any \( \phi \in \tilde{D}_1 \), which implies (3.7). \( \Box \)

The mappings \( p \) and \( \tilde{p} \), together with the open domain \( S(p) \), are shown in Fig. 1.

**Remark.** If \( T \) is nonsingular, i.e., \( T \) has no zero eigenvalue, then because of the relationship between the poles of the resolvent \( r \) and the eigenvalues of \( T \), it is evident that \( r \) is holomorphic in a neighborhood of \( \infty \). Thus, if \( \phi_0 \in \tilde{D}_1 \), is a simple pole of the Euler function \( p \), then \( R(\phi, T) = (I - p(\phi) T)^{-1} \) is holomorphic in a neighborhood of \( \phi_0 \), if \( T \) is nonsingular.
§4. Rates of Convergence of Euler Methods

It can be seen from the Jordan normal form of $T$ (cf. [8]) that an appropriate measure for the (asymptotic) decay of the terms $v_j(T)$ of the Euler transform is

$$
\kappa(T,p) := \max_{\tau \in \sigma(T)} \lim_{j \to \infty} |v_j(\tau)|^{1/j}. \tag{4.1}
$$

Thus, $\kappa(T,p)$ can be taken as a measure for the rate of convergence of the Euler transform. Note that for the identity mapping $\hat{p} \in \mathcal{E}$, we obtain $v_j(T) = T^j$ and hence $\kappa(T,\hat{p}) = \rho(T)$ (cf. (1.4)).

From the Cauchy-Hadamard formula, we conclude that $1/\kappa(T,p)$ is the radius of convergence of the power series (3.5) of $R(\cdot, T)$. Thus, if $R(\cdot, T)$ is holomorphic in $D_\eta := \{\phi: |\phi| < \eta\}$ for some $\eta > 1$, then

$$
\kappa(T,p) \leq 1/\eta. \tag{4.2}
$$

We now seek sufficient conditions on $p$ and $T$ so that $R(\cdot, T)$ is holomorphic in $D_\eta$, with $\eta > 1$. Thus, consider any $\eta$ with $1 < \eta \leq \hat{\eta}(p)$ (cf. Definition 2). Then, $R(\cdot, T)$ is holomorphic in $D_\eta$ if

$$
p(\phi) = \zeta_i \quad \text{for} \quad \phi \in D_\eta, \quad i = 1, \ldots, n,
$$

i.e.,

$$\tau_i = \hat{p}(\phi) \quad \text{for} \quad \phi \in D_\eta, \quad i = 1, \ldots, n,$$

i.e.,

$$\sigma(T) \subset S_\eta(p) := \mathbb{C} \setminus \hat{p}(D_\eta).$$

The above gives us the following

**Corollary 2.** If $p \in \mathcal{E}$, then for the associated Euler transform there holds

$$
\kappa(T,p) \leq 1/\eta \tag{4.3}
$$

for all $T$ with $\sigma(T) \subset S_\eta(p) := \mathbb{C} \setminus \hat{p}(D_\eta)$, where $\eta$ satisfies $1 < \eta \leq \hat{\eta}(p)$. Equality holds in (4.3) if there exists at least one eigenvalue of $T$ on the boundary of $S_\eta(p)$.

The region $S_\eta(p)$ is described in Fig. 2; for comparison, the region $S(p)$ from Fig. 1 is given by a dotted line.
We remark that inequality (4.3) was derived in Niethammer [10].

It follows from the above construction that, for \( p \in \mathfrak{C} \), the region \( S(p) \) is open with the point 1 on its boundary, and \( S(p) \) is closed for \( 1 < \eta \leq \hat{\eta}(p) \). If \( p \) is holomorphic in \( D_{\eta} \), then \( 0 \in S(p) \), while if \( p \) has a pole in \( D_{\eta} \), then \( 0 \notin S(p) \).

The following result will be very useful for comparing different Euler functions.

**Theorem 3.** (Comparison Theorem). Let \( p_1 \) and \( p_2 \) in \( \mathfrak{C} \) satisfy

\[
S_{\eta_2}(p_2) \subseteq S_{\eta_1}(p_1),
\]

where \( 1 < \eta_2 \leq \hat{\eta}(p_2) \), and \( 1 < \eta_1 \leq \hat{\eta}(p_1) \). Then

\[
1/\eta_2 \leq 1/\eta_1.
\]

Moreover, if equality does not hold in (4.4), then strict inequality holds in (4.5).

**Proof.** From the assumption (4.4), it follows that

\[
\tilde{p}_2(D_{\eta_2}) \supseteq \tilde{p}_1(D_{\eta_1}).
\]

Then, if \( q_2 \) denotes the inverse mapping of \( 1/p_2 \), it follows that the composition \( q_2 \circ \tilde{p}_1 \) maps \( D_{\eta_1} \) into \( D_{\eta_2} \). Thus,

\[
h(\phi) := \frac{1}{\eta_2} (q_2 \circ \tilde{p}_1)(\eta_1 \phi)
\]

is a mapping from \( D_{\eta_1} \) into \( D_{\eta_1} \) with \( h(0) = 0 \). By Schwarz's Lemma, we conclude that \( |h(\phi)| \leq |\phi| \) for all \( \phi \in D_{\eta_1} \). For \( \phi = 1/\eta_1 \), we obtain

\[
1/\eta_1 \geq |h(1/\eta_1)| = 1/\eta_2,
\]

the last equality coming from the fact \( p_1(1) = p_2(1) = 1 \). This establishes (4.5).

If equality does not hold in (4.4), then inequality is valid in (4.6), and, again from Schwarz's Lemma, we have, in analogy with (4.8),

\[
1/\eta_1 > |h(1/\eta_1)| = 1/\eta_2,
\]

which establishes strict inequality in (4.5). \( \square \)
It is interesting to remark that the univalence of the Euler function \( p \), as required in Definitions 1 and 2, is used in Theorem 3. However, on deleting the univalence hypothesis, one could work with a weaker definition of an Euler function, for which many of the results here, but not all (such as Theorem 3), would remain valid.

§5. Computation of the Euler Transform

Now, we come to the computation of the solution \( x = (I - T)^{-1} c \) from the series (3.7). If an Euler function \( p \) has the property that \( \sigma(T) \subset S(p) \), then the series (3.7) converges to \( x \), so that the partial sums

\[
    x_m := \sum_{j=0}^{m} v_j(T) c = \sum_{j=0}^{m} h_j
\]  

(5.1)

with \( h_0 = c \),

\[
    h_j := v_j(T) c = \sum_{i=1}^{j} \pi_{j,i} T^i c \quad (j > 0)
\]  

(5.2)

are convergent to \( x \) as \( m \to \infty \). (Note that since \([p(\phi)]^0 = 1\), we have

\[
    \pi_{0,0} = 1, \pi_{j,0} = 0 \quad (j > 0);
\]  

(5.3)

therefore in (5.2) the summation begins with \( l = 1 \).) It is not practical to evaluate the polynomials \( v_j(T) \) directly, so we seek for a recurrence formula for the \( x_m \)’s.

**Lemma 4.** Let \( p \in E \) be given by the power series

\[
    p(\phi) = \sum_{j=1}^{\infty} \pi_{j,1} \phi^j.
\]  

(5.4)

Then, the partial sums \( x_m \) in (5.1) can be recursively computed by

\[
    x_0 = h_0 = c,
\]

\[
    x_m = x_{m-1} + h_m, \quad (m > 0)
\]

with

\[
    h_m = T \left( \sum_{l=0}^{m-1} \pi_{m-l,1} h_l \right).
\]  

(5.5)

**Proof.** The coefficients \( \pi_{j,l} \) in (5.2) are the coefficients of the power series for \([p(\phi)]^l\). From the identity \( p(\phi) \cdot [p(\phi)]^{m-1} = [p(\phi)]^m \), we derive

\[
    \pi_{m,j} = \sum_{l=j}^{m} \pi_{l-1,j-1} \cdot \pi_{m-l+1,1} \quad (m = 1, 2, \ldots; 1 \leq j \leq m),
\]  

(5.6)
With (5.6) and (5.3), we obtain

\[ h_m = v_n(T)c = \sum_{j=1}^{m} \pi_{m,j} T^j c \]

\[ = \sum_{j=1}^{m} \left\{ \sum_{i=j}^{m} \pi_{i-1,j-1} \pi_{m-i+1,j} \right\} T^j c \]

\[ = T \sum_{i=1}^{m} \pi_{m-i+1,1} \left( \sum_{j=1}^{i} \pi_{i-1,j-1} T^{j-1} c \right) \]

\[ = T \sum_{i=0}^{m-1} \pi_{m-i,1} \left( \sum_{j=0}^{i} \pi_{1,j} T^j c \right) \]

\[ = T \sum_{i=0}^{m-1} \pi_{m-i,1} h_i. \]

The advantage of (5.5) is that, for each \( x_m \), the operator \( T \) has to be applied only once; further, the coefficients \( \pi_{m,j} \), for \( j > 1 \), do not appear in the recurrence formula, but only the coefficients \( \pi_{i,1} \) of the given power series (5.4). Yet, the disadvantage is that for computing \( x_m \), all preceding \( h_j \) (\( j = 0, \ldots, m-1 \)) must in general be stored.

The next theorem shows that there are special Euler functions \( p \) without the last mentioned drawback. For this, we make the following

**Definition 3.** For each positive integer \( k \), the subset \( \mathcal{E}_k \) of \( \mathcal{E} \) consists of all \( p \in \mathcal{E} \) of the form

\[ p(\varphi) = \frac{\mu_0 \varphi}{1 - \mu_1 \varphi - \cdots - \mu_k \varphi^k}. \]  

(5.7)

With this definition, we next establish

**Theorem 5.** If \( p \in \mathcal{E}_k \), then the partial sums \( x_m \) of (5.1) can be recursively computed by

\[ x_0 = h_0 = c, \]

\[ x_m = x_{m-1} + h_m \quad (m > 0) \]

with

\[ h_m = \mu_0 Th_{m-1} + \mu_1 h_{m-1} + \mu_2 h_{m-2} + \cdots + \mu_k h_{m-k} \]  

(5.8)

where

\[ h_i = \begin{cases} h_i, & i > 0 \\ 0, & i \leq 0 \end{cases} \]

A direct recurrence formula for \( x_m \) is

\[ x_0 = c, \]

\[ x_1 = c + \mu_0 Tx_0, \]

\[ x_m = (1 - \mu_1 - \cdots - \mu_{m-1}) c + \mu_0 Tx_{m-1} + \mu_1 x_{m-1} + \cdots + \mu_{m-1} x_1, \]

\[ (2 \leq m \leq k) \]

\[ x_m = \mu_0 (Tx_{m-1} + c) + \mu_1 x_{m-1} + \cdots + \mu_k x_{m-k} \quad (m > k). \]  

(5.10)
Proof. First, we derive a recurrence formula for the elements of the matrix $\Phi_{\pi_{j,m}}$ generated by a $p$ of the form (5.7). Again, we use the identity $p(\phi) [p(\phi)]^{p-1} = [p(\phi)]^{p}$ which gives

$$
\mu_0 \phi \left( \sum_{j_{m-1}}^{\infty} \pi_{j,m-1} \phi^j \right) = (1 - \mu_1 \phi - \ldots - \mu_k \phi^k) \left( \sum_{j_{m-1}}^{\infty} \pi_{j,m} \phi^j \right).
$$

A comparison of the coefficients of equal powers of $\phi$ yields

$$
\pi_{j,m} = \mu_0 \pi_{j-1,m-1} + \mu_1 \pi_{j-1,m} + \mu_2 \pi_{j-2,m} + \ldots + \mu_k \pi_{j-k,m}
$$

(j = 1, 2, \ldots; 1 \leq m \leq j), where $\pi_{j,m} = 0$ for all $j < m$. For the 0-th column, (5.3) holds.

Using (5.11) and $h_0 = e$, we get for $j > 0$ that

$$
h_j = v_j(T) e = \sum_{m=1}^{j} \pi_{j,m} T^m e
$$

$$
= \sum_{m=1}^{j} (\mu_0 \pi_{j-1,m-1} + \mu_1 \pi_{j-1,m} + \ldots + \mu_k \pi_{j-k,m}) T^m e
$$

$$
= (\mu_0 T \sum_{m=0}^{j-1} \pi_{j-1,m} T^m + \mu_1 \sum_{m=1}^{j-1} \pi_{j-1,m} T^m + \ldots + \mu_k \sum_{m=1}^{j-k} \pi_{j-k,m} T^m) e.
$$

The first term on the right side equals $h_{j-1}$, by definition; the terms with factor $\mu_l (l = 1, \ldots, k)$ equal $h_l$ as long as $j - l > 0$; otherwise, these terms are zero. Thus, we can write

$$
h_j = h_0 T h_{j-1} + \mu_1 h_{j-1} + \ldots + \mu_k h_{j-k}
$$

with $h_l := h_l$ if $l > 0$, and $h_l := 0$ if $l \leq 0$, which establishes (5.8).

Assuming first that $m > k$, we obtain with (5.8) that

$$
x_m = h_0 + \sum_{j=1}^{m} h_j
$$

$$
- \mu_0 T \sum_{j=1}^{m} h_{j-1} + \mu_1 \sum_{j=1}^{m} h_{j-1} + \ldots + \mu_k \sum_{j=1}^{m} h_{j-k},
$$

$$
= - \mu_0 T \sum_{j=0}^{m-1} h_j + \mu_1 \sum_{j=0}^{m-1} h_j + \ldots + \mu_k \sum_{j=0}^{m-k} h_j - \mu_1 h_0 - \ldots - \mu_k h_0,
$$

where the last terms can be considered as corrections since, in the terms with factor $\mu_l (1 \leq l \leq k)$, we have replaced $h_l$ by $h_l$. Now, since $(1 - \mu_1 - \ldots - \mu_k) = \mu_0$ and $h_0 = e$, we get (5.10) from $x_{m-l} = \sum_{j=0}^{m-l} h_l$ ($l = 1, \ldots, k$). The formula for $x_1$ follows from (5.8) for $m = 1$; for $2 \leq m \leq k$, the terms with factor $\mu_1, l \geq k - m$, disappear, i.e., no corrections have to be made in this case. This proves (5.9).
It is important to note that any Euler function in \( \mathfrak{E} \) yields, by virtue of (5.10) (with \( y_i \) replacing \( x_i \)) the \( k \)-step iterative method as introduced in (1.6).

§6. The Special Case \( k = 2 \): Real Parameters

As an example, we examine the particular Euler function \( p \) in (5.7) for the case \( k = 2 \), and determine the regions \( S(p) \) and \( S_1(p) \) according to Theorem 1. Let

\[
p(\phi) = \frac{\mu_0 \phi}{1 - \mu_1 \phi - \mu_2 \phi^2}, \quad \mu_0 \neq 0, \tag{6.1}
\]

where at first we confine ourselves to real parameters \( \mu_0, \mu_1, \mu_2 \) with \( \mu_0 + \mu_1 + \mu_2 = 1 \); from (6.1), \( p(0) = 0 \) and \( p(1) = 1 \) follow. We should mention that for the special case \( \mu_2 = 0 \), the associated Euler function \( p \) of (6.1) generates the classical series transformation of Euler-Knopp (cf. [15, p. 130ff.]).

Next, \( p \) of (6.1) is an Euler function if, for \( \bar{p} = 1/p \), it can be established that there exists an open set \( \mathcal{D} \) containing \( \bar{D}_1 \) for which \( \bar{p} \) is univalent in \( \mathcal{D} \). To show this, we introduce polar coordinates \( \phi = \eta e^{i\theta} \) and obtain

\[
\bar{p}(\eta e^{i\theta}) = \frac{\mu_1}{\mu_0} + \frac{1}{\mu_0} \left( \frac{1}{\eta} - \mu_2 \right) \cos \theta - i \left( \frac{1}{\eta} + \mu_2 \right) \sin \theta. \tag{6.2}
\]

From (6.2), we conclude that, for small \( \eta \), the image \( \bar{p}(D_\eta) \) is the exterior of an ellipse \( E_\eta \) with center \( -\mu_1/\mu_0 \) and semiaxes

\[
a_\eta = (1/\eta - \mu_2 \eta)/|\mu_0|, \quad b_\eta = (1/\eta + \mu_2 \eta)/|\mu_0|. \tag{6.3}
\]

For \( \eta = 1 \), one obtains the ellipse \( E \) with semiaxes

\[
a = (1 - \mu_2)/|\mu_0|, \quad b = (1 + \mu_2)/|\mu_0|. \tag{6.4}
\]

Since the semiaxes \( a_\eta \) and \( b_\eta \) are easily seen from (6.3) to be strictly monotone decreasing functions of \( \eta \), then there exists an open set \( \mathcal{D} \) containing \( D_1 \) such that \( \bar{p} \) is univalent in \( \mathcal{D} \) if, for \( 0 < \eta \leq 1 \), the semiaxes \( a_\eta \) and \( b_\eta \) remain positive. This is the case if

\[
\mu_0 = 0, \quad \text{and} \quad -1 < \mu_2 < 1. \tag{6.5}
\]

As a consequence of (6.5), \( 1 - \mu_2 > 0 \), or equivalently \( \mu_1 = 1 - \mu_2 - \mu_0 > -\mu_0 \). Thus, \( -\mu_1/\mu_0 < 1 \) for \( \mu_0 > 0 \) and \( -\mu_1/\mu_0 > 1 \) for \( \mu_0 < 0 \), i.e., the center of the ellipse \( E \) lies to the left (right) of unity if \( \mu_0 > 0 \) (\( \mu_0 < 0 \)).

Next, the above discussion shows that \( p \) is meromorphic and univalent for \( 1 < \eta \leq \hat{\eta} \) (where \( \hat{\eta} \) is the maximal extension (cf. Def. 2) of \( p \)) iff \( a_\eta \) and \( b_\eta \) of (6.3) are both positive for \( 1 < \eta < \hat{\eta} \). For \( \eta = \hat{\eta} \), either \( a_\eta \) or \( b_\eta \) is necessarily zero, from which it follows (cf. (6.3)) that

\[
\hat{\eta} = 1/|\mu_2|. \tag{6.6}
\]
Analysis of $k$-Step Iterative Methods

An easy calculation with (6.3) shows that the associated family of ellipses $E_\eta$, where $0 < \eta < \hat{\eta}$, are confocal with foci $F^-$ and $F^+$, where

$$F^\pm = \frac{(-\mu_\pm 2\sqrt{|\mu_2|})/\mu_0}{(-\mu_1 2i\sqrt{|\mu_2|})/\mu_0} \quad \text{if} \quad -1 < \mu_2 \leq 0,$$

$$\frac{(-\mu_1 2i\sqrt{|\mu_2|})/\mu_0}{0 \leq \mu_2 < 1}. \quad (6.7)$$

We note further that the limiting case $\eta = \hat{\eta}$ corresponds to the degenerate ellipse $E_\eta$, which is the line segment $[F^-, F^+]$.

As a consequence of these observations and Theorems 1 and 5, we next establish

**Corollary 6.** Let $p \in \mathcal{E}_2$, so that (6.5) is valid. Then,

(a) The Euler transform of the Neumann series of $T$ converges to $(I - T)^{-1}$ for all $T$ with $\sigma(T) \subset S(p)$, where $S(p)$ is the interior of the ellipse $E$ with center $-\mu_1/\mu_0$, horizontal semiaxes $a = (1 - \mu_2)/|\mu_0|$ and vertical semiaxes $b = (1 + \mu_2)/|\mu_0|$.

(b) For all $\eta$ with $1 < \eta < 1/|\mu_2|$, the region $S_\eta(p)$ is the closed interior of the ellipse $E_\eta$ (confocal to $E$) with semiaxis $a_\eta$ and $b_\eta$ given by (6.3). For $T$ with $\sigma(T) \subset S_\eta(p)$, there holds

$$\kappa(T, p) \leq 1/\eta.$$ 

(c) If $F^+$ and $F^-$ are the foci of $E$ (and $E_\eta$) according to (6.7), and if $\sigma(T) \subset [F^-, F^+] = \{F^- + t(F^+ - F^-) : 0 \leq t \leq 1\}$ for some $T$, then

$$\kappa(T, p) = 1/\hat{\eta} = 1/|\mu_2|.$$ 

(d) The partial sums $x_m$ of the series $x = \sum_{j=0}^{\infty} v_j(T) c$ can be computed according to

$$x_0 = c,$$

$$x_1 = c + \mu_0 Tc,$$

$$x_2 = (1 - \mu_1) c + \mu_0 Tc_1 + \mu_1 x_1,$$

$$x_m = \mu_0(Tx_{m-1} + c) + \mu_1 x_{m-1} + \mu_2 x_{m-2} \quad (m > 2). \quad (6.8)$$

**Proof.** Parts (a) and (b) follow from Theorem 1, part (d) from Theorem 5, and part (c) from Corollary 2. \qed

If an operator $T$, together with some information on $\sigma(T)$, is given, then one can try to adjust the real parameters $\mu_0$, $\mu_1$, $\mu_2$ such that $\sigma(T)$ is "embraced" by a fitting ellipse $E_\eta$; if $\sigma(T)$ is contained in the strip $-1 < \Re z < 1$, this is done by de Pillis in [12].

§ 7. The Special Case $k = 2$: Complex Parameters

Corollary 6 describes the situation when the parameters $\mu_0$, $\mu_1$, $\mu_2$ are real; in this case, $S(p)$ and $S_\eta(p)$ are elliptical regions, such that the foci $F^-$ and $F^+$ are on the real line ($\mu_2 < 0$) or on a vertical line ($\mu_2 > 0$). For studying the complex
case, let us begin with two different complex numbers \( z \) and \( \beta \), and let us again (cf. Corollary 6, (c)) denote by \([z, \beta]\) the complex line segment joining \( z \) and \( \beta \). If \( 1 \not\in [z, \beta] \), then \( z \) and \( \beta \) determine a unique ellipse \( E \) with foci \( z \) and \( \beta \) which passes through \( z = 1 \).

Now, we can try to find an Euler function \( p \) such that \( S(p) \) coincides with the interior of \( E \), and, such that \( p \) is of the form (6.1), with, eventually, complex parameters \( \mu_0, \mu_1, \mu_2 \). If there is a \( p \) with \( S(p) = \text{int}(E) \) then, by our derivation in the §3, \( \tilde{p} = 1/p \) would be a mapping from \( \mathcal{D}_1 \) onto the union of \( E \) and its exterior. We shall find this mapping \( p \) by constructing a mapping \( \tilde{p} \) from \( \mathcal{D}_{\eta} \), for some \( \eta > 1 \), onto \( \mathbb{C} \setminus [z, \beta] \). We do this in three steps.

First, we begin with a disk \( \mathcal{D}_{\eta} \), \( \eta = |s| > 1 \), where \( s \) is to be fixed later.

1. \( \phi \mapsto \tilde{\phi} = \phi / s \)
2. \( \tilde{\phi} \mapsto \psi = (\tilde{\phi} + 1/\tilde{\phi})/2 \)
3. \( \psi \mapsto z = \gamma \psi + \delta \) with \( \gamma = (\beta - x)/2, \delta = (x + \beta)/2 \).

In (1), the disk \( \mathcal{D}_{\eta} \) with \( \eta = |s| \) is mapped onto the unit disk \( \tilde{D} \) in the \( \tilde{\phi} \)-plane; by (II), \( \tilde{D} \) is mapped onto \( \mathbb{C} \setminus [-1, 1] \) in the \( \psi \)-plane, and by (III), \( \mathbb{C} \setminus [-1, 1] \) is mapped onto \( \mathbb{C} \setminus [z, \beta] \) in the \( z \)-plane.

The three steps are sketched in Fig. 3; the dotted line shows the images of the unit circle \( |\phi| = 1 \).

The mapping \( \phi \mapsto z = \tilde{p}(\phi) \) is given by

\[
z = p(\phi) = \delta + \frac{\gamma}{2} \left( \frac{1}{s} \phi + \frac{s}{\phi} \right)
= \frac{\gamma s^2 + 2s \delta \phi + \gamma \phi^2}{2s \phi}
= \frac{1 + \frac{2 \delta}{\gamma s} \phi + \frac{1}{s^2} \phi^2}{\frac{2}{\gamma s} \phi},
\]

i.e., \( p(\phi) = 1/\tilde{p}(\phi) \) is of the form (6.1) with

\[
\mu_0 = \frac{2}{\gamma s}, \quad \mu_1 = \frac{-2 \delta}{\gamma s}, \quad \mu_2 = \frac{-1}{s^2}.
\]
Now, \( s \) has to be determined from the condition \( 1 = p(1) = \bar{p}(1) \); from (7.1), we derive that
\[
1 = \delta + \frac{\gamma}{2} \left( \frac{1}{s} + s \right)
\]
or
\[
s^2 - \frac{2(1 - \delta)}{\gamma} s + 1 = 0. \tag{7.4}
\]
Denoting the two roots of (7.4) by \( s^\pm \), then
\[
s^\pm = \frac{(1 - \delta) \pm \sqrt{(1 - \delta)^2 - \gamma^2}}{\gamma}. \tag{7.5}
\]
By inserting \( \gamma := (\beta - \alpha)/2 \), \( \delta := (\alpha + \beta)/2 \) we get
\[
s^\pm = \frac{(1 - \alpha) + (1 - \beta) \pm 2 \sqrt{(1 - \alpha)(1 - \beta)}}{\beta - \alpha}
\]
or
\[
s^\pm = \frac{(\sqrt{1 - \alpha} \pm \sqrt{1 - \beta})^2}{\beta - \alpha}. \tag{7.6}
\]
From (7.4), we see \( s^+ = 1/s^- \); for use in (7.1)-(7.3), we take that particular solution which satisfies \( |s| > 1 \). Thus, the mapping \( p \) has been constructed. If we set
\[
\hat{\eta} = \max(|s^+|, |s^-|) = |s|, \tag{7.7}
\]
then by construction, \( \bar{p}(D_\eta) = \mathbb{C} - [\alpha, \beta] \) so that \( \hat{\eta} \) is indeed the maximal extension of \( p \) (cf. Def. 2). Thus, \( D_\eta \) is mapped onto the exterior of the ellipse \( E \) with foci \( \alpha \) and \( \beta \), passing through \( z = 1 \); the interior of \( E \) corresponds to \( S(p) \).

To any ellipse \( E_\eta \) from this family of ellipses with foci \( \alpha \) and \( \beta \) and with \( 1 < \eta < \hat{\eta} \), there corresponds a disk \( D_\eta \), \( 1 < \eta < \hat{\eta} \), such that \( D_\eta \) is mapped by \( \bar{p} \) onto the exterior of \( E_\eta \). Suppose \( z \) is an arbitrary point of \( \text{int} \, E \setminus [\alpha, \beta] \), and we wish to determine that value of \( \eta \) such that \( z \) lies in \( E_\eta \). To determine this \( \eta \), let \( \phi \) be the unique complex number with \( |\phi| < \eta \) such that \( z = p(\phi) \). Then, it is easy to verify that \( \eta = |\phi| \). Thus, we have established

**Theorem 7.** Let \( \alpha + \beta \in \mathbb{C} \) be such that \( 1 \not\in [\alpha, \beta] \), and let \( E \) be the ellipse with foci \( \alpha \) and \( \beta \) which passes through the point 1.

(a) Then by (7.3) and (7.6), there exists an Euler function \( p \) in \( \mathbb{E}_2 \) (cf. (6.1)) such that \( S(p) \) is the interior of \( E \), i.e., the Euler transform of the Neumann series of \( T \) converges to \( (I - T)^{-1} \) for all \( T \) with \( \sigma(T) \subset \text{int}(E) \).

(b) If \( \sigma(T) \subset [\alpha, \beta] \), then
\[
\kappa(T, p) = 1/\hat{\eta}, \tag{7.8}
\]
where \( \hat{\eta} \) is given by (7.7).
(c) If \( \sigma(T) \) is contained in the closed interior of a confocal ellipse \( E_\eta \) where \( 1 < \eta < \tilde{\eta} \), which passes through a point \( z \), then

\[
\kappa(T, p) \leq 1/\eta,
\]

where \( \eta = |\phi| \) and \( \phi \) is a solution of \( z = \tilde{p}(\phi) \) with \( 1 < |\phi| < \tilde{\eta} \).

(d) If \( \sigma(T) \subset \text{int}(E) \), then the partial sums \( x_n \) of

\[
x = (I - T)^{-1} e = \sum_{j=0}^{\infty} v_j(T) e
\]

can be computed by the formulas in Corollary 5, (d), where the parameters \( \mu_0 \), \( \mu_1 \), \( \mu_2 \) are given by (7.3).

Since \( \tilde{\eta} \) is the maximal extension of \( p \), equality holds in (7.8). Special cases of Theorem 7 are considered in Sect. 9.

§8. Optimal Methods with Respect to a Given Set

For convenience, we first let \( \mathfrak{M} \) denote the class of all compact sets \( U \) containing more than one point, whose complement contains \( 1 \) and is simply connected. Then, for each \( U \in \mathfrak{M} \), we define the set

\[
\mathfrak{U}_U := \{ p \in \mathfrak{C} : S(p) \supset U \}. \tag{8.1}
\]

(Later, we shall show that \( \mathfrak{U}_U \) is not empty.) Next, we define for each \( p \in \mathfrak{U}_U \)

\[
\kappa(U, p) := \min \{ 1/\eta : 1 < \eta \leq \tilde{\eta}(p) \text{ and } U \subset S_{\eta}(p) \}. \tag{8.2}
\]

Note that if \( \kappa := \kappa(U, p) \), then \( U \subset S_{1/\kappa}(p) \).

For each \( U \in \mathfrak{M} \), we can associate an asymptotic convergence factor \( \kappa(U) \) for \( U \), defined by

\[
\kappa(U) := \inf_{p \in \mathfrak{U}_U} \kappa(U, p). \tag{8.3}
\]

Any Euler function \( p_0 \in \mathfrak{U}_U \), for which

\[
\kappa(U, p_0) = \kappa(U) \tag{8.4}
\]

is then called optimal with respect to \( U \).

We next establish

**Theorem 8.** For each \( U \in \mathfrak{M} \), there exists a \( p_0 \in \mathfrak{U}_U \), which is optimal with respect to \( U \).

**Proof.** The complement \( \bar{U} := \mathfrak{C} \setminus U \) is open and simply connected in \( \mathfrak{C} \). Under our assumption on \( U \), the set \( \bar{U} \) has more than one boundary point. By the Riemann Mapping Theorem, there exists a univalent mapping \( \tilde{p} \) from the open unit disk \( D_1 \) onto \( \bar{U} \) with \( \tilde{p}(0) = \infty \). Since \( 1 \in \bar{U} \), there exists a unique \( s \neq 0 \), \( s \in D_1 \) with \( \tilde{p}(s) = 1 \). Setting \( q(\phi) = 1/\tilde{p}(s\phi) \), then it is evident that \( q(1) = 1 \) and \( q(0) = 0 \).
Moreover, by definition
\[ \tilde{q}(D_{1/|s|}) = \frac{1}{q(D_{1/|s|})} = \tilde{p}(s \cdot D_{1/|s|}) = \tilde{U}. \]
Thus, as \( S(q) = \overline{C} \setminus \tilde{q}(D_1) = \overline{C} \setminus \tilde{q}(D_{1/|s|}) = U \), then \( q \in \mathfrak{H}_U \). Moreover, as \( S_q(q) = \overline{C} \setminus \tilde{q}(D_0) \), then
\[ S_{1/|s|}(q) = \overline{C} \setminus \tilde{U} = U \), and \( \kappa(U, q) = |s|. \)

Next, assume that there exists a \( q_1 \) in \( \mathfrak{H}_U \) with \( \kappa(U, q_1) = |s_1| < |s|. \) Then it follows that
\[ U = S_{1/|s|}(q) \supset S_{1/|s|}(q_1). \]
From Comparison Theorem 3 of §4, we conclude that \( |s| \leq |s_1| \), which is a contradiction. Thus we have shown that \( q \equiv p_0 \) is optimal with respect to \( U \). \( \square \)

The essence of Theorem 8 is that we can replace the definition of \( \kappa(U) \) in (8.3) by
\[ \kappa(U) = \min_{p \in \mathfrak{H}_U} \kappa(U, p). \tag{8.5} \]
Next, as a consequence of Theorem 8 and Comparison Theorem 3, we establish

**Theorem 9.** Let the sets \( U_1 \) and \( U_2 \) in \( \mathfrak{H} \) satisfy \( U_1 \subseteq U_2 \). Then,
\[ \kappa(U_1) < \kappa(U_2). \tag{8.6} \]

**Proof.** From the proof of Theorem 8, there are \( p \in \mathfrak{C} \) and \( \eta_j \) with \( 1 < \eta_j \leq \hat{\eta}(p_j) \) such that \( S_{\eta_j}(p_j) = U_j \) \((j = 1, 2)\) with
\[ \kappa(U_j) = 1/\eta_j, \quad j = 1, 2. \tag{8.7} \]
Thus, from Comparison Theorem 3, \( 1/\eta_1 < 1/\eta_2 \) which, with (8.7), gives the desired result of (8.6). \( \square \)

As we have seen, the Riemann Mapping Theorem provides us with a tool for determining an optimal Euler function for each set \( U \) in \( \mathfrak{H} \). As is well known, finding the appropriate mapping function by means of the Riemann Mapping Theorem is not usually easy, so that upper and lower bounds for the asymptotic convergence factor \( \kappa(U) \) may be both useful and desirable. On the other hand, for every Euler function \( p \), there is a family of regions for which \( p \) is optimal (in the sense of (8.4)) for each of these regions. These two sets of ideas have been incorporated into the following results, whose proofs are respectively consequences of Theorems 8 and 9.

**Corollary 10.** If \( p \in \mathfrak{C} \), then \( p \) is optimal (cf. (8.4)) with respect to each set \( S_q(p) \) for any \( \eta \) satisfying \( 1 < \eta \leq \hat{\eta} \), and moreover
\[ \kappa(S_q(p)) = 1/\eta, \quad 1 < \eta \leq \hat{\eta} = \hat{\eta}(p). \tag{8.8} \]

**Corollary 11.** Let \( U \in \mathfrak{H} \). If \( p \in \mathfrak{C}, \) \( j = 1, 2\), with
\[ S_{\eta_j}(p_j) \subseteq U \subseteq S_{\eta_j}(p_j), \quad 1 < \eta_j \leq \hat{\eta}(p_j), \quad j = 1, 2, \tag{8.9} \]
then

\[ 1/\eta_1 < \kappa(U) < 1/\eta_2. \]

(8.10)

The next section is devoted to the study of various examples of sets \( U \) for which the exact asymptotic convergence factors \( \kappa(U) \) are found, and other sets \( U \) for which nontrivial upper and lower bounds for \( \kappa(U) \) are found.

§ 9. Examples

We now describe some examples, which make use of Corollary 10 and Corollary 11. As far as Corollary 10 is concerned, we can take two points of view. Initially, we can assume that \( U \in \mathfrak{M} \) is given, and we are looking for some Euler function \( p \) such that \( U = S_\eta(p) \) for some \( 1 < \eta \leq \tilde{\eta}(p) \). Or, we can start with an Euler function of the form (5.7). If \( \sigma(T) < S(p) \) for some operator \( T \), then the partial sums \( x_m \) of the Euler transform can be computed by the formulas in Theorem 5. For convenience, we repeat (5.7):

\[ x_m = \mu_0(Tx_{m-1} + e) + \mu_1 x_{m-1} + \ldots + \mu_k x_{m-k}, \quad (m > k). \]

(9.1)

Now, if we want to interpret (9.1) as a general \( k \)-step iteration formula, we have to admit arbitrary starting vectors \( x_j (j=0,\ldots,k-1) \). If we do this, of course, we get a sequence \( \{x_m\}_{m=0}^\infty \) which differs from the sequence \( \{x_m\}_{m=0}^\infty \) generated in (5.9) of Theorem 5. The question is, if the sequence \( \{\hat{x}_m\}_{m=0}^\infty \) has the same limit \( (I - T)^{-1} e \) and the same asymptotic convergence factor as the sequence \( \{x_m\}_{m=0}^\infty \).

Let us first assume that in Theorem 5, instead of \( x_0 = e \), we choose \( \hat{x}_0 = e \). Then, we will prove by induction that

\[ \hat{x}_m = x_m + \hat{w}_m \quad \text{with} \quad \hat{w}_m := v_m(T)(\hat{x}_0 - e) \]

(9.2)

and with \( v_m(T) \) defined by (5.1). Since \( v_0(T) = I \), we have

\[ \hat{x}_0 = e + (\hat{x}_0 - e) = x_0 + (\hat{x}_0 - e) = x_0 + v_0(T)(\hat{x}_0 - e). \]

Now, let (9.2) be true for all \( l \) with \( 1 \leq l \leq m - 1 \), \( 1 \leq m \leq k \). Then from (5.9), we have

\[ \hat{x}_m = (1 - \mu_1 - \ldots - \mu_{m-1})e + \mu_0 Tx_{m-1} + \mu_1 x_{m-1} + \ldots + \mu_{m-1} \hat{x}_1 
\]

\[ = (1 - \mu_1 - \ldots - \mu_{m-1})e + \mu_0 Tx_{m-1} + \mu_1 x_{m-1} + \ldots + \mu_{m-1} x_1 
\]

\[ + \mu_0 T \hat{x}_{m-1} + \mu_1 \hat{x}_{m-1} + \ldots + \mu_{m-1} \hat{x}_1. \]

Then, from (5.9), we conclude that the terms in first row on the right side equal \( x_m \); that the terms in the second row equal \( \hat{x}_m = v_m(T)(\hat{x}_0 - e) \) follows from (5.8), where \( e \) has to be replaced by \( (\hat{x}_0 - e) \). Similarly, from (5.10), it follows that (9.2) holds for \( m > k \).

Now, if we choose \( x_0 = e \), but \( \hat{x}_1 = x_1 \) for some \( l \) with \( 1 \leq l \leq k - 1 \), there results a sequence \( \{\hat{x}_m\}_{m=0}^\infty \); in the same way as above, one shows that \( \hat{x}_m \)
Analysis of k-Step Iterative Methods

\[ = x_m (0 < m < l - 1), \]
\[ \hat{x}_m = x_m + \hat{w}_m \quad \text{with} \quad \hat{w}_m := w_{m-1}(T)(\hat{x}_i - x_i) \quad (m \geq l). \]  

(9.3)

Finally, if we choose \( \hat{x}_i = x_i \) \( (j = 0, \ldots, l; l \leq k - 1) \), then the sequence \( \{\hat{x}_m\}_{m=0}^{\infty} \), resulting from (9.1), differs from \( \{x_m\}_{m=0}^{\infty} \) by \( l \) terms of the form (9.3).

Now, if \( \kappa := \kappa(T, p) < 1 \), then from the definition of (4.1), we know that for the sequence \( \{x_m\}_{m=0}^{\infty} \) given by Theorem 5, the difference sequence \( \{x_m - x_{m-1}\}_{m=0}^{\infty} = \{h_m\}_{m=0}^{\infty}, \ (x_{-1} = 0) \), tends to zero as \( \{k_m\}_{m=0}^{\infty} \). If we choose arbitrary starting vectors \( \hat{x}_j \) \( (j = 0, \ldots, k - 1) \) then, from above, we know that the resulting sequence \( \{\hat{x}_m\}_{m=0}^{\infty} \) and the sequence \( \{x_m\}_{m=0}^{\infty} \) differ by at most \( k \) sequences \( \{\hat{w}_m\}_{m=0}^{\infty} \) of the form (9.3), which tend to zero as \( \{k_m\}_{m=0}^{\infty} \), \( (l = 0, \ldots, k - 1) \), i.e., \( \{\hat{x}_m\}_{m=0}^{\infty} \) and \( \{x_m\}_{m=0}^{\infty} \) have the same limit and the same convergence behavior. Therefore, all statements concerning the convergence factor, which are derived by interpreting (9.1) as the recursive computation of an Euler transform, hold in the same way, if we interpret (9.1) as a k-step iteration formula with arbitrary starting vectors.

**Example 1.** Let \( U \) be the real interval \([-v, v]\) with \( 0 < v < 1 \). Applying Theorem 7 with \( \alpha := -v \) and \( \beta := v \), there is an Euler function \( p \in \mathcal{C}_2 \) such that (cf. (7.8) and (7.10))

\[ s = \frac{1 + v \pm \sqrt{1 - v^2}}{2v} = \frac{1 \pm \sqrt{1 - v^2}}{v}. \]  

(9.4)

In our special case, we have for the parameters \( \gamma \) and \( \delta \), appearing in (7.3), \( \gamma = (\beta - \alpha)/2 = v, \ \delta = (\beta + \alpha)/2 = 0 \); from (7.3) and (9.4) we get with \( s = s^+ \)

\[ \mu_0 = \frac{2}{\gamma s} = \frac{2}{1 + \sqrt{1 - v^2}}, \quad \mu_1 = \frac{2\delta}{\gamma s} = 0, \quad \mu_2 = \frac{1}{s^2} = \frac{v^2}{[1 + \sqrt{1 - v^2}]^2}. \]  

(9.5)

From Corollary 10 we conclude that the Euler function \( p \in \mathcal{C}_2 \), determined by the parameters in (9.5), is optimal with respect to \( U = [-v, v] \), and

\[ \kappa(U) = \frac{1}{s^+} = \frac{v}{1 + \sqrt{1 - v^2}}. \]  

(9.6)

From Theorem 7 and Corollary 10, we further conclude that \( p \) (determined by (9.5)) is optimal, too, with respect to the closed interior \( E_\eta \) of all ellipses with foci \( F_+ := (v, 0) \) and \( F_- := (-v, 0) \) and major semiaxis \( a_\eta, v < a_\eta < 1 \); we have \( \kappa(E_\eta) = 1/\eta \), where \( \eta \) is determined from (6.3); \( a_\eta = (1/\eta - \mu_2 \eta)/\mu_0 \). If the semiaxes \( a_\eta \) or \( b_\eta \) are not known, but some other point \( P \) on the boundary of \( E_\eta \) is known, then \( a_\eta \) is given by the well-known property of any ellipse: \( F_+ P + F_- P = 2a_\eta \).

It should be mentioned that if the Chebyshev semi-iterative method (see [13], § 5.1 and § 5.2) is applied to a linear system \( x = Bx + c \), where \( B \) has real eigenvalues with \( \rho(B) = v < 1 \), then, asymptotically, as \( m \to \infty \), the convergence rate is governed by (9.6), whereas the (nonstationary) parameters in the three-term recursive formulas tend to the values in (9.5), as \( m \to \infty \).
Example 2. In the same manner as in Example 1, we conclude in the case $U := \{ -ive, iv \}$, \( v > 0 \) arbitrary, that

\[
    s^\pm = \frac{\left( \sqrt{1 + iv} \pm \sqrt{1 - iv} \right)}{2iv} = \frac{1 \pm \sqrt{1 + v^2}}{iv}.
\]  

(9.7)

and with \( \gamma = iv \), \( \delta = 0 \), \( s = s^+ \)

\[
    \mu_0 = \frac{2}{\gamma s} = \frac{2}{1 + \sqrt{1 + v^2}}, \quad \mu_1 = 0, \quad \mu_2 = \frac{-v^2}{s^2} = \frac{-v^2}{(1 + \sqrt{1 + v^2})^2}.
\]  

(9.8)

Again, the Euler function \( p \in \mathcal{E}_2 \), determined by (9.8), is optimal with respect to \( U = \{ -ive, +ive \} \), and

\[
    \kappa(U) = \frac{1}{|s|} = \frac{v}{1 + \sqrt{1 + v^2}}.
\]  

(9.9)

Formulas (9.8) and (9.9) are derived, e.g., in de Pillis [12]. The statements of Example 1, concerning confocal ellipses, hold in a corresponding way.

Example 3. Let \( A\mathbf{z} = \mathbf{z} - B\mathbf{z} = \mathbf{c} \) be a nonsingular linear system, such that \( B \) is weakly two-cyclic (cf. [13]) with complex elements, and \( A\mathbf{z} = \mathbf{c} \) has the form

\[
    \begin{bmatrix}
    \mathbf{x} \\
    \mathbf{y}
    \end{bmatrix} = \begin{bmatrix}
    0 & B_1 \\
    B_2 & 0
    \end{bmatrix} \begin{bmatrix}
    \mathbf{x} \\
    \mathbf{y}
    \end{bmatrix} + \begin{bmatrix}
    \mathbf{c}_1 \\
    \mathbf{c}_2
    \end{bmatrix},
\]  

(9.10)

where the diagonal blocks are square and zero. Then, (9.10) can be written as two uncoupled systems of equations for \( \mathbf{x} \) and \( \mathbf{y} \) (see [13], §5.4). In a similar way, the SOR-method, applied to (9.10), can be written as a two-step method for the vector \( \mathbf{x} \) alone (see [9]): there results

\[
    \mathbf{x}_m = \omega^2 (B_1 B_2 \mathbf{x}_{m-1} + \mathbf{c}_1 + B_1 \mathbf{c}_2) + 2(1 - \omega) \mathbf{x}_{m-1} - (1 - \omega)^2 \mathbf{x}_{m-2}
\]  

\((m \geq 2)\). Writing

\[
    T := B_1 B_2, \quad \tilde{\mathbf{c}} = \mathbf{c}_1 + B_1 \mathbf{c}_2
\]  

(9.11)

we have

\[
    \mathbf{x}_m = \omega^2 (T \mathbf{x}_{m-1} + \tilde{\mathbf{c}}) + 2(1 - \omega) \mathbf{x}_{m-1} - (1 - \omega)^2 \mathbf{x}_{m-2}(m \geq 2),
\]  

(9.12)

i.e., (9.12) is of the form (9.1) with

\[
    \mu_0 = \omega^2, \quad \mu_1 = 2(1 - \omega), \quad \mu_2 = -(1 - \omega)^2.
\]  

(9.13)

The parameters \( \mu_0, \mu_1, \mu_2 \), which satisfy \( \mu_0 + \mu_1 + \mu_2 = 1 \) and which depend only on \( \omega \), determine an Euler function \( p = p(\omega) \in \mathcal{E}_2 \). From the considerations at the beginning of the section, we conclude that the sequence \( \{ \mathbf{x}_m \}_{m=0}^\infty \) of (9.12), generated by the SOR-method, has the same convergence behavior as the sequence of partial sums of the Euler transform induced by \( p(\omega) \). But for the latter, many new results, even for complex parameters, can be derived from our previous sections.
First, let us describe $S(p(\omega))$. Combining (7.3) and (9.13), we get

$$
\mu_0 = \omega^2 = \frac{2}{\gamma s},
$$

$$
\mu_1 = 2(1-\omega) = -\frac{2\delta}{\gamma s},
$$

$$
\mu_2 = -(1-\omega)^2 = -\frac{1}{s^2}.
$$

Solving for $\gamma$, $\delta$, $s$ yields

$$
\gamma = \frac{2(1-\omega)}{\omega^2}, \quad \delta = -\frac{2(1-\omega)}{\omega^2}, \quad s = \frac{1}{1-\omega}.
$$

(9.14)

From (7.7) we want to have $|s| > 1$, which gives

$$
|1-\omega| < 1,
$$

(9.15)

the well-known necessary condition for convergence of Kahan (see [13], Theorem 3.5). From §7, we know that $S(p(\omega))$ is the interior of the ellipse $E(\omega)$, with foci $\alpha$ and $\beta$, where $\gamma = (\beta - \alpha)/2$, $\delta = (\alpha + \beta)/2$, and which passes through the point $z = 1$. From (9.14), we get

$$
\alpha = 0, \quad \beta = -\frac{4(1-\omega)}{\omega^2}.
$$

(9.16)

If we now denote by $\bar{E}(\omega)$ the ellipse which passes through $\bar{z} = 1$ and has center 0 and foci

$$
F^\pm := \pm \frac{2\sqrt{1-\omega}}{\omega},
$$

(9.17)

then, by representing $\bar{E}(\omega)$ in polar coordinates, an easy calculation shows that, by the mapping $\bar{z} \to z^2$, the ellipse $\bar{E}(\omega)$ is mapped onto $E(\omega)$ (and $\bar{E}_o(\omega)$ onto $E_{p}(\omega)$). Further, since the eigenvalues $\tau_j$ of $T = B_1B_2$ are connected with the eigenvalues $\beta_j$ of the matrix $B$ of the given system $x - Bx = c$ via $\tau_j = \beta_j^2$, we have

$$
\tau_j \in E_o(\omega) \quad \text{iff} \quad \beta_j \in \bar{E}_o(\omega) \quad (j = 1, \ldots, n).
$$

This means that every result concerning the spectrum $\sigma(T)$ and the region $E_o(\omega)$, holds equally for $\sigma(B)$ and $\bar{E}_o(\omega)$. If $\rho(\mathcal{L}_o)$ denotes the spectral radius of the SOR-operator $\mathcal{L}_o$, then by comparing the asymptotic decrease of the error vectors of SOR and of the equivalent two-step method (9.11) (see §1), we have $\kappa(T, p(\omega)) = \rho(\mathcal{L}_o)$. From Theorem 7, it follows for the iteration (9.12) and therefore for the SOR-method, that the following is valid.

**Theorem 12.** Let $\omega \in \mathbb{C}$ be arbitrary with $|1-\omega| < 1$. Then, the SOR-method, applied to the nonsingular system $z - Bz = c$, where $B$ is weakly two-cyclic with elements from $\mathbb{C}$, is convergent, if $\sigma(B)$ is contained in the interior of the ellipse $\bar{E}(\omega)$ with foci $F^\pm(\omega) = \pm 2\sqrt{1-\omega}/\omega$, which passes through the point 1.
If \( \sigma(B) \) is contained in the closed interior of the confocal ellipse \( \tilde{E}_n(\omega) \subset \tilde{E}(\omega) \), then \( \rho(\Omega_{\omega}) \leq 1/\eta \), where equality holds if there is an eigenvalue of \( B \) on the boundary of \( \tilde{E}_n(\omega) \). In particular, if \( \sigma(B) = [F^{-}(\omega), F^{+}(\omega)] \), we have

\[
\rho(\Omega_{\omega}) = |1 - \omega|.
\]

Figure 4 shows the regions \( \tilde{E}(\omega) \) and the corresponding intervals \([F^{-}_j, F^{+}_j]\) for \( \omega_1 = 1 + 0.5 \), \( \omega_2 = 1 + 0.5i \), \( \omega_3 = 1 - 0.5 \); from (9.18) we conclude in each of the three cases, that \( \rho(\Omega_{\omega}) \leq 0.5 \) if \( \sigma(B) = [F^{-}_j, F^{+}_j] \), \( j = 1, 2, 3 \).

Since for arbitrary \( \omega \) with \( |1 - \omega| < 1 \), the interval \([1 - \omega, 1]\) with \( 0 < v < 1 \) is contained in \( \tilde{E}(\omega) \), we conclude that if \( I - B \) is symmetric and positive definite, then the SOR-method converges for every \( \omega \in \mathbb{C} \) with \( |1 - \omega| < 1 \).

Let \( B \) be given with \( \sigma(B) \subset \subset (-\infty, -1) \cup \{1, \infty\} \setminus \{1, -1\} \); in our notation of (1.13), this means \( B \in \mathcal{O}_{(-\infty, -1)} \). Then from Corollary 10, it follows that \( \rho(\omega) \) is optimal with respect to \([1 - \omega, 1]\) if \([F^{-}, F^{+}] = [1 - \omega, 1]\), where \( F^{\pm} \) are determined by (9.17). This yields the relation \( \pm v = \pm 2/1 - \omega_i/\omega \), which solved for \( \omega_i \) gives the well-known relation

\[
\omega_i = \omega_i = \frac{2}{1 + \sqrt{1 - v^2}}, \quad \kappa[1 - v, v] = |1 - \omega_i|; \tag{9.19}
\]

it should again be emphasized that (9.19) holds for every \( v \) with \( v \in \mathbb{C} \setminus \{1, \infty\} \cup \{1, -1\} \). Real two-cyclic matrices \( B \) (cf. (9.10)) with complex eigenvalues have been treated by Young [14, p. 191], while complex two-cyclic matrices \( B \) have been examined by Kjellberg [4] and Kredell [6].

Example 4. As an example of an Euler function from \( E_4 \), let us consider

\[
p(\phi) = \frac{\mu_0 \phi}{1 - \mu_4 \phi^4}, \tag{9.20}
\]
with real parameters $\mu_0$, $\mu_4$ and $\mu_0 + \mu_4 = 1$. Then, with (9.20), the following 4-step iteration formula can be formally deduced (cf. Theorem 5):

$$x_m = \mu_0 (Tx_{m-1} + c) + \mu_4 x_{m-4} \quad (m > 4).$$  \hspace{1cm} (9.21)

As in § 6, the function $p$ of (9.20) is an Euler function if, for $\tilde{p} = 1/p$, it can be established that there exists an open set $\mathcal{D}$ containing $D_1^*$ for which $\tilde{p}$ is univalent in $\mathcal{D}$. We introduce polar coordinates $\phi = \eta e^{i\theta}$ and obtain

$$z := \tilde{p}(\phi) = \frac{1}{\mu_0} \left[ \left( \frac{1}{\eta} \cos \theta - \mu_4 \eta^3 \cos 3\theta \right) - i \left( \frac{1}{\eta} \sin \theta + \mu_4 \eta^3 \sin 3\theta \right) \right].$$  \hspace{1cm} (9.22)

From (9.22) we conclude that, for small $\eta$ and $\mu_4$, the image $\tilde{p}(D_\eta^*)$ is the exterior of a simply closed curve $I_\eta$, which is nearly a circle with radius $1/\eta$: $I_\eta$ is a special type of a cycloid. For studying $I_\eta$, we form

$$\frac{dz}{d\theta} = \frac{1}{\mu_0} \left[ \left( -\frac{1}{\eta} \sin \theta + 3 \mu_4 \eta^3 \sin 3\theta \right) - i \left( \frac{1}{\eta} \cos \theta + 3 \mu_4 \eta^3 \cos 3\theta \right) \right].$$  \hspace{1cm} (9.23)

The following Fig. 5 shows the curves $I_\eta$ for $\mu_0 = 1.1$ (i.e., $\mu_4 = -0.1$), $\eta = 1.0$, 1.1, 1.2 and $\eta = \tilde{\eta}(p) = \sqrt[3]{1/3|\mu_4|} = \sqrt[3]{10}/3$.

![Fig. 5](image)

Now, let us assume $\mu_4 < 0$. Then, as can also be seen from Fig. 5, $\tilde{p}$ is univalent in $D_\eta$ iff the coefficient of $i$ in (9.23) remains positive for $\theta = 0$. This holds for $0 < \eta \leq 1$ if $1 + 3\mu_4 > 0$, i.e., if

$$-1/3 < \mu_4 < 0.$$  \hspace{1cm} (9.24)

Thus, a function $p$ of the form (9.20) is an Euler function iff (9.24) holds. Similarly, we conclude again from (9.23) for $\theta = 0$, that $\tilde{p}$ is univalent in $D_\eta$ iff $1/\eta + 3\mu_4 \eta^3 > 0$ or $\eta^3 < 1/3|\mu_4|$, i.e., we get

$$\tilde{\eta}(p) = \sqrt[3]{1/3|\mu_4|}.$$  \hspace{1cm} (9.25)
Figure 5 shows that $I_\delta$ has a cusp for $\theta=0$. From (9.22) and (9.25), it follows that

$$v = \bar{\rho}(\hat{\eta}) = \frac{1}{1 - \mu_4 \hat{\eta}^3} = \frac{\mu_4}{3} \frac{1}{1 + |\mu_4| \sqrt{3}}.$$  \hspace{1cm} (9.26)

Now, as again can be seen from Fig. 5, $S_\delta(p)$ is the closed interior of $I_\delta$ and has the four cusps $\pm v$, $\pm iv$, i.e. the "cross" $U_\delta := [ -v, v ] \cup [ -iv, iv ]$ is contained strictly in $S_\delta(p)$. From Corollary 11 it follows

$$\kappa(U_\delta) < \frac{1}{\tilde{\eta}} = \frac{1}{\sqrt{3}} |\mu_4|.$$ \hspace{1cm} (9.27)

On the other hand, if there is an operator $T$ such that $\sigma(T) \subset U_\delta$ for some $0 < v < 1$ (i.e., if $T \in O_\delta$) then, since $U_\delta$ is symmetric with respect to both coordinate axes, no $k$-step method of the form (9.1) with $k = 1$ or $k = 2$ yields an acceleration of convergence, the same holding for $k = 3$. Of course, since $\rho(T) \leq v < 1$, the simple iteration (1.2) converges. Now, since $v$ is given, (9.26) can be solved for $|\mu_4|$. Method (9.21) with $\mu_4 = -|\mu_4|$ and $\mu_0 = 1 - \mu_4$ has then a convergence factor, which is less than or equal $1/\tilde{\eta}$, where $\tilde{\eta} = (3 |\mu_4|)^{-1/4}$. A measure for the acceleration of convergence, which can be obtained by the 4-step method (9.20) is

$$\psi(v) = \frac{\ln(1/v)}{\ln(v)}.$$ \hspace{1cm} (9.28)

(see e.g. [13], p. 67). Table 1 shows some values of $\mu_4$ and the resulting values of $v$, $1/\tilde{\eta}$ and $\psi(v)$ according to (9.26), (9.25) and (9.28).

<table>
<thead>
<tr>
<th>$\mu_4$</th>
<th>$v$</th>
<th>$1/\tilde{\eta}$</th>
<th>$\psi(v)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.025</td>
<td>0.6807</td>
<td>0.5233</td>
<td>1.6838</td>
</tr>
<tr>
<td>-0.05</td>
<td>0.7903</td>
<td>0.6223</td>
<td>2.014</td>
</tr>
<tr>
<td>-0.1</td>
<td>0.8971</td>
<td>0.7401</td>
<td>2.771</td>
</tr>
<tr>
<td>-0.2</td>
<td>0.9779</td>
<td>0.8801</td>
<td>5.718</td>
</tr>
<tr>
<td>-0.3</td>
<td>0.9900</td>
<td>0.9740</td>
<td>25.77</td>
</tr>
</tbody>
</table>

To interpret the numbers of Table 1, the last row there indicates that the 4-step method (9.21) requires asymptotically 1/25 as many iterations, in this case, as does the one-step method (1.2).

Example 5. Given a nonsymmetric operator $T$, it is often possible to describe a rectangle which contains the spectrum of $T$. For simplicity, given some $v$ with $0 < v < 1$, let us consider the rectangle

$$R_v = \{ z \in \mathbb{C} : | -v \leq \text{Re} z \leq v, -1 \leq \text{Im} z \leq 1 \}.$$ \hspace{1cm} (9.29)

There exists, by Theorem 8, an Euler function $p_0$, optimal with respect to $R_v$, but the computation of the corresponding Euler transform would go very...
Lemma 4. On the other hand, by our considerations in §6, we understand very well how to operate with elliptic regions, usually associated with two-step iteration methods of type (9.1). The idea is as follows: By Corollary 11, every closed elliptic region $E$ — intervals included — which are contained in $R_\ast$, provides a lower bound $\kappa(E)$ for $\kappa(U_\ast)$, and $\kappa(E)$ is an upper bound, if $E \supseteq R_\ast$.

Since $R_\ast \supseteq [-v, v]$ and $R_\ast \supseteq [-i, +i]$, we get from Example 1 and Example 2, i.e., from (9.6) and (9.9), the first lower bound

$$
\kappa_1 := \max \left( \frac{v}{1 + \sqrt{1 - v^2}}, \frac{1}{1 + \sqrt{1/2}} \right) < \kappa(R_\ast),
$$

(9.30)

For our second lower bound $\kappa_2$, we use the fact that $R_\ast$ contains the elliptic region $E$, where the boundary is the ellipse $E_\ast$ with center 0 and semiaxes $v$ and 1. The foci are

$$
E^\pm := iE \quad \text{with} \quad \varepsilon := \sqrt{1 - v^2}.
$$

(9.31)

Now, we want to determine an Euler function $p$ from $\mathcal{E}_\ast$, such that $E = S_{\ast}(p)$ for some $\eta > 1$. Since the center of $E_\ast$ is 0, we have $\mu_1 = 0$. From (6.7) and (9.31), we get $\varepsilon = 2\sqrt{1 - \mu_2}/\mu_0$ and by solving for $\mu_0$,

$$
\mu_0 = \frac{2(\sqrt{1 + \varepsilon^2} - 1)}{\varepsilon^2}, \quad \mu_2 = 1 - \mu_0.
$$

(9.32)

Since one semiaxis of $E_\ast$ has endpoint 1, we get from (6.3) the equation $1 = (1/\eta + \mu_2 \eta)/\mu_0$, which we can solve for $\eta$; there result two solutions $\eta_1, \eta_2$ with $1 < \eta_1 < \eta = \sqrt{1/\mu_2} < \eta_2$. Thus, we have the second lower bound

$$
\kappa_2 := \kappa(E) = 1/\eta_1, \quad \text{where} \quad \eta_1 = \frac{\mu_0 - \sqrt{\mu_0^2 - 4 \mu_2}}{2 \mu_2},
$$

(9.33)

where $\mu_0, \mu_2$ are given by (9.32). Finally, by Corollary 11, we note that

$$
\kappa_1 < \kappa_2 < \kappa(R_\ast).
$$

(9.34)

For an upper bound $\kappa_3$ for $\kappa(U_\ast)$, we seek another elliptic region $E$ with $E \supseteq R_\ast$. Let $E_\ast$ be the ellipse with center 0 and semiaxes $a, b$ (where $v < a < 1$) such that $E_\ast$ passes through the corner $(v, i)$ of $R_\ast$. $E_\ast$ is not unique, but there holds $v^2/a^2 + 1/b^2 = 1$, i.e.,

$$
b^2 = \frac{a^2}{a^2 - v^2}.
$$

(9.35)

So, for $E_\ast$, we have a one-parameter family. The foci are $F^\pm := iE$, where $\varepsilon = \sqrt{b^2 - a^2}$; by (9.35), $\varepsilon$ is a function of $a$. As before, we can determine parameters $\mu_0$ and $\mu_2$ according to (9.32); from (6.7) we conclude $b = (1/\eta + \mu_2 \eta)/\mu_0$; solving for $\eta$ yields

$$
\kappa(R_\ast) < \kappa_3 = \kappa(E) = 1/\eta_1,
$$

where $\eta_1$ is a solution of

$$
\eta_1 = \frac{\mu_0 - \sqrt{\mu_0^2 - 4 \mu_2}}{2 \mu_2},
$$

(9.33)
where

$$\eta_1 = \frac{2\mu_2}{b\mu_0 - \sqrt{b^2\mu_0^2 - 4\mu_2}}.$$  (9.36)

Now, $b$, $\mu_0$, $\mu_2$ are all functions of $a$ and $v$, so $\kappa_3$ is a function of $a$ and $v$; we get the lowest upper bound by minimizing $\kappa_3$ as a function of $a$. Given $v$, we determine this minimum not analytically, but by a numerical search process.

Table 2 shows the values $\kappa_1$, $\kappa_2$ and $\kappa_3$ for various values of $v$, where $\kappa_1$ and $\kappa_2$ are lower bounds of $\kappa(R_v)$, and $\kappa_3$ is an upper bound of $\kappa(R_v)$.

<table>
<thead>
<tr>
<th>$v$</th>
<th>$\kappa_1$</th>
<th>$\kappa_2$</th>
<th>$\kappa_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>0.4142</td>
<td>0.5</td>
<td>0.6171</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4142</td>
<td>0.5941</td>
<td>0.7485</td>
</tr>
<tr>
<td>0.6</td>
<td>0.4142</td>
<td>0.7016</td>
<td>0.8605</td>
</tr>
<tr>
<td>0.8</td>
<td>0.5</td>
<td>0.8310</td>
<td>0.9503</td>
</tr>
</tbody>
</table>

Acknowledgement. We sincerely thank Dr. Martin Gutknecht (ETH, Zürich) for his very careful reading of our manuscript, and for his valuable suggestions.

References

12. de Pillis, J.: How to Embrace Your Spectrum for Faster Iterative Results. Linear Algebra and Applications 34, 125-143 (1980)

Received May 28, 1982 / November 18, 1982