Numer. Math. 41, 177–206 (1983)

Numerische Mathematik
© Springer-Verlag 1983

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

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

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

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

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

[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

$$\mathbf{x}_{m+1} := T\mathbf{x}_m + \mathbf{c}, \qquad m = 0, 1, \ldots, \tag{1.3}$$

which, as is well known, converges, for any choice of the initial vector $\mathbf{x}_0$ to the unique solution $\mathbf{x}$ of (1.1) iff $\rho(T) < 1$, (where if $\sigma(T) := \{\lambda\colon \det(\lambda I - T) = 0\}$ denotes the *spectrum* of T, then $\rho(T) := \max\{|\lambda|\colon \lambda \in \sigma(T)\}$ denotes its *spectral radius*). Moreover, if $\boldsymbol{\varepsilon}_m := \mathbf{x}_m - \mathbf{x}$, $m = 0, 1, \ldots$, is the error vector associated with the iterate $\mathbf{x}_m$, it is further well-known (cf. Young [14, p. 87], Varga [13, p. 67]) that

$$\rho(T) = \lim_{m \to \infty} (\|T^m\|)^{1/m} = \lim_{m \to \infty} \Big(\sup_{\boldsymbol{\varepsilon}_0 \neq 0} \|\boldsymbol{\varepsilon}_m\| / \|\boldsymbol{\varepsilon}_0\|\Big)^{1/m}, \tag{1.4}$$

for any vector norm $\|\cdot\|$ on $\mathbb{C}^n$, where $\|T^m\|$ denotes the associated induced operator norm of T^m. Consequently, for any $\boldsymbol{\varepsilon}_0 \neq 0$, we have from (1.4) that

$$\overline{\lim_{m \to \infty}} \, (\|\boldsymbol{\varepsilon}_m\| / \|\boldsymbol{\varepsilon}_0\|)^{1/m} \leqq \rho(T), \tag{1.5}$$

so that $(\rho(T))^m$ is a rough *measure* of how $\|\boldsymbol{\varepsilon}_m\| / \|\boldsymbol{\varepsilon}_0\|$ behaves *asymptotically*, a $m \to \infty$.

For each positive integer k, suppose we consider, in place of (1.3), a *k-ste stationary iterative method* for (1.2), defined by

$$\mathbf{y}_m := \mu_0(T\mathbf{y}_{m-1} + \mathbf{c}) + \mu_1 \mathbf{y}_{m-1} + \ldots + \mu_k \mathbf{y}_{m-k}, \qquad m = k, k+1, \ldots, \tag{1.6}$$

where $\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$ are given starting vectors, and where $\mu_0, \mu_1, \ldots, \mu_k$ a fixed complex numbers (independent of m) which are assumed to satisfy

$$\mu_0 + \mu_1 + \ldots + \mu_k = 1. \tag{1.7}$$

Obviously, on choosing the specific starting vectors as $\mathbf{y}_0 = \ldots = \mathbf{y}_{k-1} = \mathbf{x}$, see from (1.6) and (1.7) that $\mathbf{y}_m = \mathbf{x}$ for all $m \geqq 0$. Next, we note that the choi $\mu_0 = 1$ and $\mu_1 = \ldots = \mu_k = 0$ for $k \geqq 1$ is such that that iterative method of (1 reduces to that of (1.3). Moreover, the iterative methods of (1.3) and (1.6) ea require just *one* matrix multiplication by the matrix T, although more vect 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 thought of as a type of *averaging method* of the matrix T and the ident matrix I. For example, if $k = 1$ in (1.6), the condition (1.7) implies that (1.6) c be expressed as

$$\mathbf{y}_m = \{(1 - \mu_0) I + \mu_0 T\} \mathbf{y}_{m-1} + \mu_0 \mathbf{c}, \qquad m = 1, 2, \ldots,$$

which is known in the literature (cf. [14, p. 361]) both as an extrapola iterative method, and a stationary first-order Richardson method, based (1.2). Similarly, the case $k = 2$ of (1.6) becomes

$$\mathbf{y}_m = (\mu_1 I + \mu_0 T) \mathbf{y}_{m-1} + (1 - \mu_0 - \mu_1) \mathbf{y}_{m-2} + \mu_0 \mathbf{c}, \qquad m = 2, 3, \ldots,$$

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 (*non*stationary 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

$$\tilde{\boldsymbol{\varepsilon}}_m := \mathbf{y}_m - \mathbf{x}, \qquad m = k, k+1, \ldots \tag{1.8}$$

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 $\{\tilde{\boldsymbol{\varepsilon}}_m\}_{m=0}^{\infty}$ of (1.8) decrease in norm as κ^m, as $m \to \infty$, where κ is assumed to satisfy $0 \leqq \kappa < 1$?

If one can precisely describe these regions containing the spectrum of T as a function of κ, 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 $\mathbf{x} = (I-T)^{-1}\mathbf{c}$, assuming $1 \notin \sigma(T)$. If, moreover, $\rho(T) < 1$, then $\mathbf{x}$ can be computed from the convergent Neumann series

$$\mathbf{x} = (I-T)^{-1}\mathbf{c} = \sum_{j=0}^{\infty} T^j \mathbf{c}, \tag{1.9}$$

and the iterations of (1.3) can be seen as a recursive computation of the partial sums $\mathbf{x}_m := \sum_{j=0}^{m} T^j \mathbf{c}$ of the vector sum in (1.9), provided that $\mathbf{x}^{(0)} := \mathbf{c}$. 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 $\mathbf{x}$ of (1.2) can be expressed as

$$\mathbf{x} = (I-T)^{-1}\mathbf{c} = \sum_{j=0}^{\infty} v_j(T)\,\mathbf{c}. \tag{1.10}$$

For a *special* class of general Euler methods, which depend on the parameters $\{\mu_j\}_{j=0}^{k}$, the partial sums $\mathbf{y}_m := \sum_{j=0}^{m} v_j(T)\,\mathbf{c}$, 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 $\{\mathbf{y}_j\}_{j=0}^{k-1}$ 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 $\mathfrak{P}$, can be applied to the Neumann series of T. In §3 for a general Euler method, $\mathfrak{P}$ is generated by an Euler function p. Such a p determines a region $S(p)$ in $\mathbb{C}$ such that (1.10) converges to $\mathbf{x}$ whenever $\sigma(T)\subset S(p)$. In §4, for estimating the decay of the errors, it is useful to introduce, for some $z\in S(p)$, the quantity

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

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

$$\kappa(T,p) := \max_{\tau_i\in\sigma(T)} \kappa(\tau_i,p), \tag{1.11}$$

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

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 η with $1<\eta\leqq\hat{\eta}(p)$, a closed subset $S_\eta(p)$ of $S(p)$ can be described such that

$$\kappa(T,p)\leqq 1/\eta, \tag{1.11'}$$

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

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

$$p(\phi)=\frac{\mu_0\phi}{1-\mu_1\phi-\ldots-\mu_k\phi^k}, \qquad \mu_0+\ldots+\mu_k=1, \tag{1.12}$$

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 $1 \leqq \eta < \hat{\eta}$, where $\hat{\eta} := 1/\sqrt{|\mu_2|}$. For $\eta = \hat{\eta}$, $S_{\hat{\eta}}(p)$ is the interval between the foci.

If α and β are different complex numbers such that the line segment $[\alpha, \beta]$ joining α and β doesn't contain the point $z=1$, then there exists a unique ellipse E in the complex plane such that E contains 1 and has foci α and β; in §8, the optimal p_0 for this segment $[\alpha, \beta]$ (and for all confocal elliptic regions $E_\eta \subset \operatorname{int}(E)$) 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

$$O_U := \{T \in \mathbb{C}^{n,n} \colon \sigma(T) \subset U\} \tag{1.13}$$

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)

$$\kappa(U,p) := \min\{1/\eta \colon 1 < \eta \leqq \hat{\eta}(p) \quad \text{and} \quad U \subset S_\eta(p)\} \tag{1.14}$$

as the *convergence factor* of U with respect to p. By definition (cf. (1.11) and (1.11′),

$$\kappa(T,p) \leqq \kappa(U,p) \qquad \text{for all} \quad T \in O_U.$$

Now, the Euler function p_0 is said to be *optimal with respect to* U, if

$$\kappa(U,p_0) \leqq \kappa(U,p),$$

for all Euler functions p 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

$$\kappa(U) := \kappa(U,p_0). \tag{1.15}$$

Now, it can be seen (Corollary 10) that each Euler function p is optimal with respect to every $S_\eta(p)$, where $1 < \eta \leqq \hat{\eta}(p)$; there holds

$$\kappa(S_\eta(p)) = 1/\eta, \qquad 1 < \eta \leqq \hat{\eta}(p). \tag{1.16}$$

Further, if $U_1 \subsetneqq U_2$ we have from Schwarz's Lemma that (cf. Theorem 9)

$$\kappa(U_1) < \kappa(U_2). \tag{1.17}$$

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 $S_{\eta_1}(p_1) \subsetneqq U \subsetneqq S_{\eta_2}(p_2)$; thus (cf. Corollary 11),

$$1/\eta_1 < \kappa(U) < 1/\eta_2. \tag{1.18}$$

If an Euler function of the form (1.12) is given, then for $m > k$, the recursive computation of the partial sums $\mathbf{x}_m$ of the Euler transform yields formula (1.6), where $\mathbf{y}_i$ is replaced by $\mathbf{x}_i$ $(i = m-k, \ldots, m)$. If we want to interpret (1.6) as a general k-step iteration formula, we have to admit *arbitrary* starting vectors $\mathbf{y}_j$ $(j=0,\ldots,k-1)$. Then, of course, the sequences $\{\mathbf{x}_m\}_{m=0}^{\infty}$ and $\{\mathbf{y}_m\}_{m=0}^{\infty}$ 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

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}=\begin{bmatrix}0 & B_1\\ B_2 & 0\end{bmatrix}\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}+\begin{bmatrix}\mathbf{c}_1\\ \mathbf{c}_2\end{bmatrix}, \tag{1.19}$$

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 $\mathbf{x}$:

$$\mathbf{x}_m=\omega^2(T\mathbf{x}_{m-1}+\tilde{\mathbf{c}})+2(1-\omega)\mathbf{x}_{m-1}-(1-\omega)^2\mathbf{x}_{m-2} \quad (m\geqq 2) \tag{1.20}$$

with $T:=B_1B_2$, $\tilde{\mathbf{c}}:=\mathbf{c}_1+B_1\mathbf{c}_2$. Then, on applying the results of §6 and §7, the convergence behavior of the SOR-method with an arbitrary *complex* relaxation parameter ω 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 $R_v:=\{z\in\mathbb{C}\colon -v\leqq \operatorname{Re} z\leqq v,\ -1\leqq \operatorname{Im} z\leqq 1\}$, $0<v<1$, i.e., upper and lower bounds for $\kappa(R_v)$ are derived.

§2. Background Material from Summability Theory

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

$$\mathbf{y}_m:=\sum_{j=0}^{m}\gamma_{m,j}\mathbf{x}_j \quad (m\geqq 0) \tag{2.1}$$

can be considered. The infinite triangular matrix $\mathfrak{Q}:=(\gamma_{m,j})_{m\geqq 0,\,0\leqq j\leqq m}$ 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

$$\sum_{j=0}^{m}\gamma_{m,j}=1 \quad (m\geqq 0) \tag{2.2}$$

is imposed on $\mathfrak{Q}$, then for the error vector $\tilde{\mathbf{e}}_m:=\mathbf{y}_m-\mathbf{x}$ of (1.8), it follows that

$$\tilde{\mathbf{e}}_m=q_m(T)\tilde{\mathbf{e}}_0, \tag{2.3}$$

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

$$q_m(z) := \sum_{j=0}^{m} \gamma_{m,j} z^j. \tag{2.4}$$

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 Chebsyhev polynomials $t_m(z)$ guarantees a *maximal reduction* of the Euclidean norm of the associated error vectors $\tilde{\varepsilon}_m$. 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 $\mathbf{y}_m$, which by-passes the use of formula (2.1) which requires the storage of all intermediate vectors $\mathbf{x}_j$. 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 $\{\mathbf{x}^{(m)}\}_{m=0}^{\infty}$ 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

$$\mathfrak{P} := (\pi_{i,j})_{i \geqq 0,\, 0 \leqq j \leqq i}, \tag{2.5}$$

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

$$v_j(T) := \sum_{l=0}^{j} \pi_{j,l} T^l, \qquad j = 0, 1, \ldots, \tag{2.6}$$

i.e.,

$$\begin{bmatrix} v_0(T) \\ v_1(T) \\ \vdots \\ v_j(T) \\ \vdots \end{bmatrix} = \begin{bmatrix} \pi_{0,0} & & & \\ \pi_{1,0} & \pi_{1,1} & 0 & \\ \vdots & & \ddots & \\ \pi_{j,0} & \pi_{j,1} \cdots & \pi_{j.j} & \\ \vdots & & & \ddots \end{bmatrix} \begin{bmatrix} I \\ T \\ \vdots \\ T^j \\ \vdots \end{bmatrix}.$$

The resulting series

$$\sum_{j=0}^{\infty} v_j(T) \tag{2.7}$$

is considered as the *transformed Neumann series.* To this transformation, determined by $\mathfrak{P}$, there corresponds a sequence-to-sequence transformation $\mathfrak{Q}$ of the partial sums of the Neumann series to the partial sums of (2.7). Between $\mathfrak{P}$

and $\mathfrak{Q}$, the following relation holds (cf. [15, p. 7])

$$\mathfrak{Q} = \Sigma \mathfrak{P} \Sigma^{-1}, \tag{2.8}$$

where

$$\Sigma := \begin{bmatrix} 1 & & & & \\ 1 & 1 & & 0 & \\ 1 & 1 & 1 & & \\ 1 & 1 & 1 & 1 & \\ \cdots & \cdots & \cdots & & \ddots \\ \cdots & \cdots & \cdots & \cdots & \ddots \end{bmatrix}; \quad \Sigma^{-1} := \begin{bmatrix} 1 & & & \\ -1 & 1 & & 0 \\ & -1 & 1 & \\ & & \ddots & \ddots \\ 0 & & & \ddots \end{bmatrix}.$$

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

$$\mathbf{x} = (I-T)^{-1}\mathbf{c} = \sum_{j=0}^{\infty} v_j(T)\,\mathbf{c}. \tag{2.9}$$

We note that if $\mathfrak{P}$ of (2.5) is chosen to be the identity matrix, then so is $\mathfrak{Q}$ 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 $\mathfrak{P}$ and T does (2.9) hold, i.e., under what conditions on $\mathfrak{P}$ 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 $\sum_{j=0}^{\infty} v_j(T)\mathbf{c}$ 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 $\mathbf{x} = T\mathbf{x} + \mathbf{c}$; 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 $\bar{D}_\eta$ denote its closure. We make the following

Definition 1. A complex-valued function p is said to be an *Euler function* (written $p \in \mathfrak{E}$) if there exists an open neighborhood $\mathfrak{D}$ of $\bar{D}_1$, such that

a) p is meromorphic and univalent (schlicht) in $\mathfrak{D}$, 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

$$p(\phi) := \sum_{j=1}^{\infty} \pi_{j,1} \phi^j, \tag{3.1}$$

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

$$[p(\phi)]^m := \sum_{j=m}^{\infty} \pi_{j,m} \phi^j, \qquad m=0,1,2,\ldots, \quad \text{for } |\phi|<v. \tag{3.2}$$

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 $(0 \leqq j < \infty)$ of an infinite matrix $\mathfrak{P}$, and the series-to-series transformation (2.5) induced by $\mathfrak{P}$ 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 $p \in \mathfrak{E}$, the *maximal extension,* $\hat{\eta}(p)$, is $\hat{\eta}=\hat{\eta}(p) := \max\{\eta>1 : p$ is meromorphic and univalent in $D_\eta\}$.

Note, from Definition 1, that $1<\hat{\eta}(p) \leqq \infty$ for each $p \in \mathfrak{E}$.

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 $1 \notin \sigma(T)$:

$$r(\zeta, T) := (I-\zeta T)^{-1}; \tag{3.3}$$

r is evidently matrix-valued and meromorphic, the poles ζ_i of r and the eigenvalues τ_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 ζ_i with $|\zeta_i| \leqq 1$, i.e., if there are no eigenvalues τ_i of T with $|\tau_i| \geqq 1$.

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

$$r(\zeta, T) = (I-\zeta T)^{-1} = \sum_{j=0}^{\infty} \zeta^j T^j, \qquad \text{for } |\zeta|<\eta. \tag{3.4}$$

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 ϕ. Thus, there is an $\tilde{\eta}>0$ for which

$$R(\phi, T) := r(p(\phi), T) = (I - p(\phi)\, T)^{-1} = \sum_{j=0}^{\infty} v_j(T)\, \phi^j, \qquad \text{for } |\phi| < \tilde{\eta}, \tag{3.5}$$

with

$$v_0(T) := I, \quad \text{and} \quad v_j(T) := \sum_{l=1}^{j} \pi_{j,l}\, T^l \qquad \text{for all } j = 1, 2, \ldots, \tag{3.6}$$

which agrees with (2.6).

Now, our interest is in having the representation of $R(\phi, T)$ in (3.5), valid for *all* $|\phi| \leqq 1$. 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* $p \in \mathfrak{E}$. *If*

$$\sigma(T) \subset \bar{\mathbb{C}} \setminus \tilde{p}(\bar{D}_1) =: S(p), \tag{3.7}$$

where $\bar{D}_1$ *is the closed unit disk and* $\tilde{p} := 1/p$, *then* $R(\phi, T)$ *is nonsingular for all* $\phi \in \bar{D}_1$, *and the expansion of* (3.5) *converges absolutely in* ϕ *in* $\bar{D}_1$. *Conversely, if the expansion in* (3.5) *is absolutely convergent for all* ϕ *in* $\bar{D}_1$, *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 $\tilde{\eta} \leqq 1$, which implies that this representation has a singularity in some $\hat{\phi}$ with $|\hat{\phi}| = \tilde{\eta} \leqq 1$, i.e., from (3.5) $(I - p(\hat{\phi})\, T)$ is singular. Hence, there is an eigenvalue τ_i of T such that $p(\hat{\phi})\,\tau_i = 1$, so that

$$\tau_i = \frac{1}{p(\hat{\phi})} = \tilde{p}(\hat{\phi}).$$

But this contradicts (3.7).

Conversely, assume that the expansion in (3.5) is valid for all $\phi \in \bar{D}_1$. Let τ_i be any eigenvalue of T, so that $T\mathbf{x} = \tau_i \cdot \mathbf{x}$ for some $\mathbf{x} \neq \mathbf{0}$. 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 $\phi \in \bar{D}_1$. Thus, $\tau_i\, p(\phi) \neq 1$ for any $\phi \in \bar{D}_1$, which implies (3.7). □

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 ∞. Thus, if $\phi_0 \in \bar{D}_1$, is a simple pole of the Euler function p, then $R(\phi, T) = (I - p(\phi)\, T)^{-1}$ is holomorphic in a neighborhood of ϕ_0, if T is nonsingular.

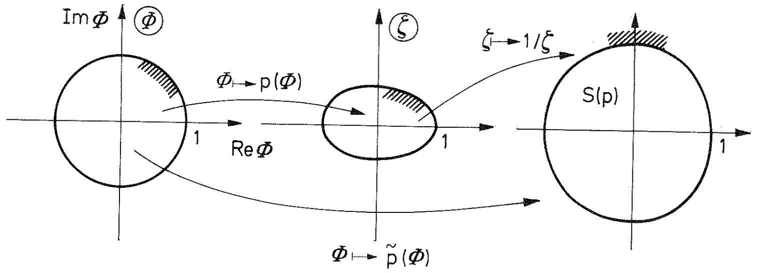

Fig. 1

§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

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

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 $\breve{p} \in \mathfrak{E}$, we obtain $v_j(T)=T^j$ and hence $\kappa(T,\breve{p})=\rho(T)$ (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 $D_\eta := \{\phi \colon |\phi| < \eta\}$ for some $\eta > 1$, then

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

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

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

i.e.,

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

i.e.,

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

The above gives us the following

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

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

for all T with $\sigma(T) \subset S_\eta(p) := \bar{\mathbb{C}} \setminus \tilde{p}(D_\eta)$, where η satisfies $1 < \eta \leqq \hat{\eta}(p)$. Equality holds in (4.3) *if there exists at last 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.

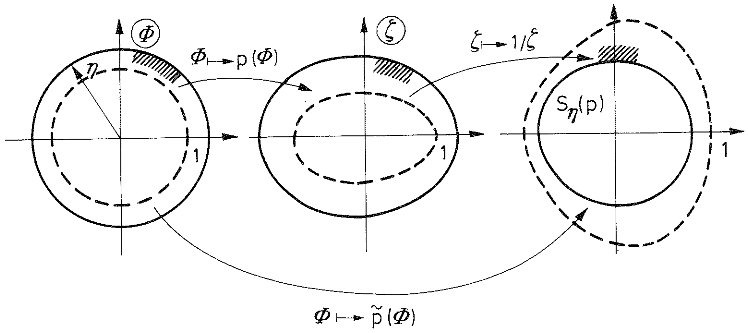

Fig. 2

We remark that inequality (4.3) was derived in Niethammer [10].

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

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

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

$$S_{\eta_2}(p_2) \subseteq S_{\eta_1}(p_1), \tag{4.4}$$

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

$$1/\eta_2 \leqq 1/\eta_1. \tag{4.5}$$

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

$$\tilde{p}_2(D_{\eta_2}) \supseteq \tilde{p}_1(D_{\eta_1}). \tag{4.6}$$

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_{η_1} into D_{η_2}. Thus,

$$h(\phi) := \frac{1}{\eta_2} (q_2 \circ \tilde{p}_1)(\eta_1 \phi) \tag{4.7}$$

is a mapping from D_1 into D_1 with $h(0) = 0$. By Schwarz's Lemma, we conclude that $|h(\phi)| \leqq |\phi|$ for all $\phi \in D_1$. For $\phi = 1/\eta_1$, we obtain

$$1/\eta_1 \geqq |h(1/\eta_1)| = 1/\eta_2, \tag{4.8}$$

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). □

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 $\mathbf{x}=(I-T)^{-1}\mathbf{c}$ 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 $\mathbf{x}$, so that the partial sums

$$\mathbf{x}_m:=\sum_{j=0}^{m} v_j(T)\,\mathbf{c}=:\sum_{j=0}^{m}\mathbf{h}_j \tag{5.1}$$

with $\mathbf{h}_0=\mathbf{c}$,

$$\mathbf{h}_j:=v_j(T)\,\mathbf{c}=\sum_{l=1}^{j}\pi_{j,l}\,T^l\,\mathbf{c} \qquad (j>0) \tag{5.2}$$

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

$$\pi_{0,0}=1,\,\pi_{j,0}=0 \qquad (j>0); \tag{5.3}$$

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 $\mathbf{x}_m$'s.

Lemma 4. *Let $p\in\mathfrak{E}$ be given by the power series*

$$p(\phi)=\sum_{j=1}^{\infty}\pi_{j,1}\,\phi^j. \tag{5.4}$$

Then, the partial sums $\mathbf{x}_m$ in (5.1) *can be recursively computed by*

$$\mathbf{x}_0=\mathbf{h}_0=\mathbf{c},$$

$$\mathbf{x}_m=\mathbf{x}_{m-1}+\mathbf{h}_m, \qquad (m>0)$$

with

$$\mathbf{h}_m=T\left(\sum_{l=0}^{m-1}\pi_{m-l,1}\,\mathbf{h}_l\right). \tag{5.5}$$

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

$$\pi_{m,j}=\sum_{l=j}^{m}\pi_{l-1,j-1}\cdot\pi_{m-l+1,1} \qquad (m=1,2,\ldots;\ 1\leqq j\leqq m). \tag{5.6}$$

With (5.6) and (5.3), we obtain

$$\begin{aligned}\mathbf{h}_m = v_m(T)\,\mathbf{c} &= \sum_{j=1}^{m} \pi_{m,j}\, T^j \mathbf{c}\\ &= \sum_{j=1}^{m} \left\{ \sum_{l=j}^{m} \pi_{l-1,\,j-1}\, \pi_{m-l+1,\,1} \right\} T^j \mathbf{c}\\ &= T \sum_{l=1}^{m} \pi_{m-l+1,\,1} \left(\sum_{j=1}^{l} \pi_{l-1,\,j-1}\, T^{j-1}\mathbf{c} \right)\\ &= T \sum_{l=0}^{m-1} \pi_{m-l,\,1} \left(\sum_{j=0}^{l} \pi_{l,\,j}\, T^j \mathbf{c} \right)\\ &= T \sum_{l=0}^{m-1} \pi_{m-l,\,1}\, \mathbf{h}_l. \quad \square\end{aligned}$$

The advantage of (5.5) is that, for each $\mathbf{x}_m$, 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 $\pi_{l,1}$ of the given power series (5.4). Yet, the disadvantage is that for computing $\mathbf{x}_m$, all preceding $\mathbf{h}_j$ $(j=0,\ldots,m-1)$ 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 $\mathfrak{E}_k$ of $\mathfrak{E}$ consists of all $p\in\mathfrak{E}$ of the form

$$p(\phi) = \frac{\mu_0 \phi}{1-\mu_1\phi-\ldots-\mu_k\phi^k}. \tag{5.7}$$

With this definition, we next establish

Theorem 5. *If* $p\in\mathfrak{E}_k$, *then the partial sums* $\mathbf{x}_m$ *of* (5.1) *can be recursively computed by*

$$\mathbf{x}_0 = \mathbf{h}_0 = \mathbf{c},$$

$$\mathbf{x}_m = \mathbf{x}_{m-1} + \mathbf{h}_m \qquad (m>0)$$

with

$$\mathbf{h}_m := \mu_0\, T\mathbf{h}_{m-1} + \mu_1 \tilde{\mathbf{h}}_{m-1} + \mu_2 \tilde{\mathbf{h}}_{m-2} + \ldots + \mu_k \tilde{\mathbf{h}}_{m-k} \tag{5.8}$$

where

$$\tilde{\mathbf{h}}_l := \begin{cases} \mathbf{h}_l, & l>0\\ \mathbf{0}, & l\leqq 0 \end{cases} .$$

A direct recurrence formula for $\mathbf{x}_m$ *is*

$$\begin{aligned}&\mathbf{x}_0 = \mathbf{c},\\ &\mathbf{x}_1 = \mathbf{c} + \mu_0 T\mathbf{x}_0,\\ &\mathbf{x}_m = (1-\mu_1-\ldots-\mu_{m-1})\,\mathbf{c} + \mu_0 T\mathbf{x}_{m-1} + \mu_1 \mathbf{x}_{m-1} + \ldots + \mu_{m-1}\mathbf{x}_1,\\ &\qquad (2\leqq m\leqq k)\end{aligned} \tag{5.9}$$

$$\mathbf{x}_m = \mu_0(T\mathbf{x}_{m-1} + \mathbf{c}) + \mu_1 \mathbf{x}_{m-1} + \ldots + \mu_k \mathbf{x}_{m-k} \qquad (m>k). \tag{5.10}$$

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

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

A comparison of the coefficients of equal powers of ϕ yields

$$\pi_{j,m} = \mu_0 \pi_{j-1,m-1} + \mu_1 \pi_{j-1,m} + \mu_2 \pi_{j-2,m} + \ldots + \mu_k \pi_{j-k,m} \tag{5.11}$$

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

Using (5.11) and $\mathbf{h}_0 = \mathbf{c}$, we get for $j > 0$ that

$$\begin{aligned}
\mathbf{h}_j = v_j(T)\mathbf{c} &= \sum_{m=1}^{j} \pi_{j,m} T^m \mathbf{c} \\
&= \sum_{m=1}^{j} (\mu_0 \pi_{j-1,m-1} + \mu_1 \pi_{j-1,m} + \ldots + \mu_k \pi_{j-k,m})\, T^m \mathbf{c} \\
&= (\mu_0 T \sum_{m=0}^{j-1} \pi_{j-1,m} T^m + \mu_1 \sum_{m=1}^{j-1} \pi_{j-1,m} T^m + \ldots + \mu_k \sum_{m=1}^{j-k} \pi_{j-k,m} T^m)\, \mathbf{c}.
\end{aligned}$$

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

$$\mathbf{h}_j = \mu_0 T \mathbf{h}_{j-1} + \mu_1 \tilde{\mathbf{h}}_{j-1} + \ldots + \mu_k \tilde{\mathbf{h}}_{j-k}$$

with $\tilde{\mathbf{h}}_l := \mathbf{h}_l$ if $l > 0$, and $\tilde{\mathbf{h}}_l := \mathbf{0}$ if $l \leqq 0$, which establishes (5.8).

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

$$\begin{aligned}
\mathbf{x}_m &= \mathbf{h}_0 + \sum_{j=1}^{m} \mathbf{h}_j \\
&= \mathbf{h}_0 + \mu_0 T \sum_{j=1}^{m} \mathbf{h}_{j-1} + \mu_1 \sum_{j=1}^{m} \tilde{\mathbf{h}}_{j-1} + \ldots + \mu_k \sum_{j=1}^{m} \tilde{\mathbf{h}}_{j-k} \\
&= \mathbf{h}_0 + \mu_0 T \sum_{j=0}^{m-1} \mathbf{h}_j + \mu_1 \sum_{j=0}^{m-1} \mathbf{h}_j + \ldots + \mu_k \sum_{j=0}^{m-k} \mathbf{h}_j - \mu_1 \mathbf{h}_0 - \ldots \\
&\qquad - \mu_k \mathbf{h}_0,
\end{aligned}$$

where the last terms can be considered as corrections since, in the terms with factor μ_l $(1 \leqq l \leqq k)$, we have replaced $\tilde{\mathbf{h}}_l$ by $\mathbf{h}_l$. Now, since $(1 - \mu_1 - \ldots - \mu_k) = \mu_0$ and $\mathbf{h}_0 = \mathbf{c}$, we get (5.10) from $\mathbf{x}_{m-l} = \sum_{j=0}^{m-l} \mathbf{h}_l$ $(l=1,\ldots,k)$. The formula for $\mathbf{x}_1$ follows from (5.8) for $m=1$; for $2 \leqq m \leqq k$, the terms with factor μ_l, $l \geqq k-m$, 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 $\mathfrak{E}_k$ yields, by virtue of (5.10) (with $\mathbf{y}_l$ replacing $\mathbf{x}_l$) 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 $S_\eta(p)$ according to Theorem 1. Let

$$p(\phi):=\frac{\mu_0\phi}{1-\mu_1\phi-\mu_2\phi^2}, \qquad \mu_0\neq 0, \tag{6.1}$$

where at first we confine ourselves to *real* parameters μ_0, μ_1, μ_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 $\tilde{p}:=1/p$, it can be established that there exists an open set $\mathfrak{D}$ containing $\bar{D}_1$ for which $\tilde{p}$ is univalent in $\mathfrak{D}$. To show this, we introduce polar coordinates $\phi=\eta e^{i\theta}$ and obtain

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

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

$$a_\eta:=(1/\eta-\mu_2\eta)/|\mu_0| \qquad b_\eta:=(1/\eta+\mu_2\eta)/|\mu_0|. \tag{6.3}$$

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

$$a:=(1-\mu_2)/|\mu_0|, \qquad b:=(1+\mu_2)/|\mu_0|. \tag{6.4}$$

Since the semiaxes a_η and b_η are easily seen from (6.3) to be strictly monotone decreasing functions of η, then there exists an open set $\mathfrak{D}$ containing $\bar{D}_1$ such that $\tilde{p}$ is univalent in $\mathfrak{D}$ iff, for $0<\eta\leqq 1$, the semiaxes a_η and b_η remain *positive.* This is the case iff

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

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 $\mu_0>0$ (<0).

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

$$\hat{\eta}=1/\sqrt{|\mu_2|}. \tag{6.6}$$

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

$$F^{\pm}:=\begin{cases}(-\mu_1\pm 2\sqrt{|\mu_2|})/\mu_0 & \text{if } -1<\mu_2\leqq 0,\\ (-\mu_1\pm 2i\sqrt{|\mu_2|})/\mu_0 & \text{if } \quad 0\leqq\mu_2<1.\end{cases} \tag{6.7}$$

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

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

Corollary 6. *Let $p\in\mathfrak{C}_2$, 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 η with $1<\eta<1/\sqrt{|\mu_2|}$, the region $S_\eta(p)$ is the closed interior of the ellipse E_η (confocal to E) with semiaxis a_η and b_η given by* (6.3). *For T with $\sigma(T)\subset S_\eta(p)$, there holds*

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

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

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

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

$$\begin{aligned}&\mathbf{x}_0=\mathbf{c},\\ &\mathbf{x}_1=\mathbf{c}+\mu_0 T\mathbf{x}_0,\\ &\mathbf{x}_2=(1-\mu_1)\mathbf{c}+\mu_0 T\mathbf{x}_1+\mu_1\mathbf{x}_1,\\ &\mathbf{x}_m=\mu_0(T\mathbf{x}_{m-1}+\mathbf{c})+\mu_1\mathbf{x}_{m-1}+\mu_2\mathbf{x}_{m-2}\qquad(m>2).\end{aligned} \tag{6.8}$$

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

If an operator T, together with some information on $\sigma(T)$, is given, then one can try to *adjust* the real parameters μ_0, μ_1, μ_2 such that $\sigma(T)$ is "embraced" by a fitting ellipse E_η; if $\sigma(T)$ is contained in the strip $-1<\operatorname{Re} z<1$, this is done by de Pillis in [12].

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

Corollary 6 describes the situation when the parameters μ_0, μ_1, μ_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 α and β, and let us again (cf. Corollary 6, (c)) denote by $[\alpha,\beta]$ the complex line segment joining α and β. If $1\notin[\alpha,\beta]$, then α and β determine a unique ellipse E with foci α and β 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 μ_0, μ_1, μ_2. If there is a p with $S(p)=\operatorname{int}(E)$ then, by our derivation in the §3, $\tilde{p}=1/p$ would be a mapping from $\bar{D}_1$ onto the union of E and its exterior. We shall find this mapping p by constructing a mapping $\tilde{p}$ from $D_{\hat{\eta}}$, for some $\hat{\eta}>1$, onto $\bar{\mathbb{C}}\setminus[\alpha,\beta]$. We do this in three steps.

First, we begin with a disk D_η, $\eta=|s|>1$, where s is to be fixed later.

(I) $\phi\mapsto\tilde{\phi}:=\phi/s$

(II) $\tilde{\phi}\mapsto\psi:=(\tilde{\phi}+1/\tilde{\phi})/2$

(III) $\psi\mapsto z:=\gamma\psi+\delta$ with $\gamma:=(\beta-\alpha)/2,\ \delta:=(\alpha+\beta)/2.$

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

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

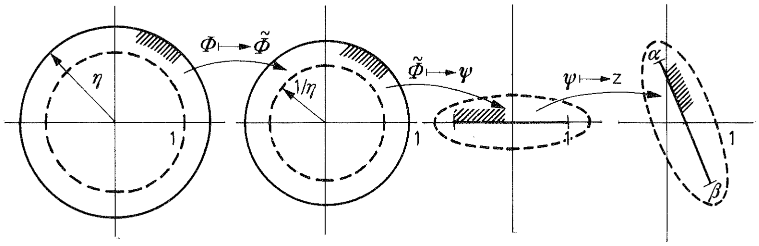

Fig. 3

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

$$z=\tilde{p}(\phi)=\delta+\frac{\gamma}{2}\left(\frac{1}{s}\phi+\frac{s}{\phi}\right) \tag{7.1}$$

$$=\frac{\gamma s^2+2s\delta\phi+\gamma\phi^2}{2s\phi}$$

$$=\frac{1+\frac{2\delta}{\gamma s}\phi+\frac{1}{s^2}\phi^2}{\frac{2}{\gamma s}\phi}, \tag{7.2}$$

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

$$\mu_0:=\frac{2}{\gamma s},\qquad \mu_1:=\frac{-2\delta}{\gamma s},\qquad \mu_2:=\frac{-1}{s^2}. \tag{7.3}$$

Now, s has to be determined from the condition $1=p(1)=\tilde{p}(1)$; 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, $\tilde{p}(D_{\hat{\eta}})=\bar{\mathbb{C}}\setminus[\alpha,\beta]$ so that $\hat{\eta}$ is indeed the *maximal extension* of p (cf. Def. 2). Thus, D_1 is mapped onto the exterior of the ellipse E with foci α and β, passing through $z=1$; the interior of E corresponds to $S(p)$.

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

Theorem 7. *Let $\alpha\neq\beta\in\mathbb{C}$ be such that $1\notin[\alpha,\beta]$, and let E be the ellipse with foci α and β which passes through the point* 1.

(a) *Then by* (7.3) *and* (7.6), *there exists an Euler function p in $\mathfrak{E}_2$ (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 $\sigma(T)\subset\operatorname{int}(E)$.*

(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_η *where* $1<\eta<\hat{\eta}$, *which passes through a point* z, *then*

$$\kappa(T, p) \leqq 1/\eta,$$

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

(d) *If* $\sigma(T) \subset \operatorname{int}(E)$, *then the partial sums* $\mathbf{x}_m$ *of*

$$\mathbf{x}=(I-T)^{-1}\mathbf{c}=\sum_{j=0}^{\infty} v_j(T)\mathbf{c}$$

can be computed by the formulas in Corollary 5, (d), *where the parameters* μ_0, μ_1, μ_2 *are given by* (7.3).

Since $\hat{\eta}$ 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

$$\mathfrak{A}_U := \{p \in \mathfrak{E} \colon S(p) \supset U\}. \tag{8.1}$$

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

$$\kappa(U, p) := \min\{1/\eta \colon 1<\eta \leqq \hat{\eta}(p) \quad \text{and} \quad U \subset S_\eta(p)\}. \tag{8.2}$$

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

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

Any Euler function $p_0 \in \mathfrak{A}_U$ 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* $p_0 \in \mathfrak{A}_U$, *which is optimal with respect to* U.

Proof. The complement $\tilde{U} := \bar{\mathbb{C}} \setminus U$ is open and simply connected in $\bar{\mathbb{C}}$. Under our assumption on U, the set $\tilde{U}$ 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 $\tilde{U}$ with $\tilde{p}(0)=\infty$. Since $1 \in \tilde{U}$, 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 $S(q){:}=\bar{\mathbb{C}}\setminus\tilde{q}(D_1)\supset\bar{\mathbb{C}}\setminus\tilde{q}(D_{1/|s|})=U$, then $q\in\mathfrak{A}_U$. Moreover, as $S_\eta(q){:}=\bar{\mathbb{C}}\setminus\tilde{q}(D_\eta)$, then

$$S_{1/|s|}(q)=\bar{\mathbb{C}}\setminus\tilde{U}=U, \quad \text{and} \quad \kappa(U,q)=|s|.$$

Next, assume that there exists a q_1 in $\mathfrak{A}_U$ with $\kappa(U,q_1)=|s_1|<|s|$. Then it follows that

$$U=S_{1/|s|}(q)\subset S_{1/|s|}(q_1).$$

From Comparison Theorem 3 of §4, we conclude that $|s|\leqq|s_1|$, which is a contradiction. Thus we have shown that $q\equiv p_0$ is optimal with respect to U. □

The essence of Theorem 8 is that we can replace the definition of $\kappa(U)$ in (8.3) by

$$\kappa(U)=\min_{p\in\mathfrak{A}_U}\kappa(U,p). \tag{8.5}$$

Next, as a consequence of Theorem 8 and Comparison Theorem 3, we establish

Theorem 9. *Let the sets U_1 and U_2 in $\mathfrak{M}$ satisfy $U_1\subsetneqq U_2$. Then,*

$$\kappa(U_1)<\kappa(U_2). \tag{8.6}$$

Proof. From the proof of Theorem 8, there are $p_j\in\mathfrak{E}$ and η_j with $1<\eta_j\leqq\hat{\eta}(p_j)$ 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). □

As we have seen, the Riemann Mapping Theorem provides us with a tool for determining an optimal Euler function for *each* set U in $\mathfrak{M}$. 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 $p\in\mathfrak{E}$, then p is optimal* (*cf.* (8.4)) *with respect to each set $S_\eta(p)$ for any η satisfying $1<\eta\leqq\hat{\eta}$, and moreover*

$$\kappa(S_\eta(p))=1/\eta, \quad 1<\eta\leqq\hat{\eta}=\hat{\eta}(p). \tag{8.8}$$

Corollary 11. *Let $U\in\mathfrak{M}$. If $p_j\in\mathfrak{E}$, $j=1,2$, with*

$$S_{\eta_1}(p_1)\subsetneqq U\subsetneqq S_{\eta_2}(p_2), \quad 1<\eta_j\leqq\hat{\eta}(p_j),\ j=1,2, \tag{8.9}$$

then

$$1/\eta_1 < \kappa(U) < 1/\eta_2. \tag{8.10}$$

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 $1 < \eta \leqq \hat{\eta}(p)$. Or, we can start with an Euler function of the form (5.7). If $\sigma(T) \subset S(p)$ for some operator T, then the partial sums $\mathbf{x}_m$ of the Euler transform can be computed by the formulas in Theorem 5. For convenience, we repeat (5.7):

$$\mathbf{x}_m = \mu_0(T\mathbf{x}_{m-1} + \mathbf{c}) + \mu_1 x_{m-1} + \dots + \mu_k x_{m-k}, \quad (m > k). \tag{9.1}$$

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

Let us first assume that in Theorem 5, instead of $\mathbf{x}_0 = \mathbf{c}$, we choose $\hat{\mathbf{x}}_0 \neq \mathbf{c}$. Then, we will prove by induction that

$$\hat{\mathbf{x}}_m = \mathbf{x}_m + \hat{\mathbf{w}}_m \quad \text{with} \quad \hat{\mathbf{w}}_m := v_m(T)(\hat{\mathbf{x}}_0 - \mathbf{c}) \tag{9.2}$$

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

$$\hat{\mathbf{x}}_0 = \mathbf{c} + (\hat{\mathbf{x}}_0 - \mathbf{c}) = \mathbf{x}_0 + (\hat{\mathbf{x}}_0 - \mathbf{c}) = \mathbf{x}_0 + v_0(T)(\hat{\mathbf{x}}_0 - \mathbf{c}).$$

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

$$\begin{aligned}\hat{\mathbf{x}}_m &= (1 - \mu_1 - \dots - \mu_{m-1})\mathbf{c} + \mu_0 T\hat{\mathbf{x}}_{m-1} + \mu_1 \hat{\mathbf{x}}_{m-1} + \dots + \mu_{m-1}\hat{\mathbf{x}}_1\\ &= (1 - \mu_1 - \dots - \mu_{m-1})\mathbf{c} + \mu_0 T\mathbf{x}_{m-1} + \mu_1 \mathbf{x}_{m-1} + \dots + \mu_{m-1}\mathbf{x}_1\\ &\qquad + \mu_0 T\hat{\mathbf{w}}_{m-1} + \mu_1 \hat{\mathbf{w}}_{m-1} + \dots + \mu_{m-1}\hat{\mathbf{w}}_1.\end{aligned}$$

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

Now, if we choose $\mathbf{x}_0 = \mathbf{c}$, but $\hat{\mathbf{x}}_l \neq \mathbf{x}_l$ for some l with $1 \leqq l \leqq k-1$, there results a sequence $\{\hat{\mathbf{x}}_m\}_{m=0}^\infty$; in the same way as above, one shows that $\hat{\mathbf{x}}_m$

$=\mathbf{x}_m (0<m<l-1)$,

$$\hat{\mathbf{x}}_m=\mathbf{x}_m+\hat{\mathbf{w}}_m \quad \text{with} \quad \hat{\mathbf{w}}_m:=v_{m-l}(T)(\hat{\mathbf{x}}_l-\mathbf{x}_l) \qquad (m \geqq l). \tag{9.3}$$

Finally, if we choose $\hat{\mathbf{x}}_j \neq \mathbf{x}_j$ $(j=0,\ldots,l;\ l \leqq k-1)$, then the sequence $\{\hat{\mathbf{x}}_m\}_{m=0}^{\infty}$, resulting from (9.1), differs from $\{\mathbf{x}_m\}_{m=0}^{\infty}$ 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 $\{\mathbf{x}_m\}_{m=0}^{\infty}$ given by Theorem 5, the difference sequence $\{\mathbf{x}_m-\mathbf{x}_{m-1}\}_{m=0}^{\infty}=\{\mathbf{h}_m\}_{m=0}^{\infty}$, $(\mathbf{x}_{-1}=\mathbf{0})$, tends to zero as $\{\kappa^m\}_{m=0}^{\infty}$. If we choose arbitrary starting vectors $\hat{\mathbf{x}}_j$ $(j=0,\ldots,k-1)$ then, from above, we know that the resulting sequence $\{\hat{\mathbf{x}}_m\}_{m=0}^{\infty}$ and the sequence $\{\mathbf{x}_m\}_{m=0}^{\infty}$ differ by at most k sequences $\{\hat{\mathbf{w}}_m\}_{m=0}^{\infty}$ of the form (9.3), which tend to zero as $\{\kappa^{m-l}\}_{m=l}^{\infty}$ $(l=0,\ldots,k-1)$, i.e., $\{\hat{\mathbf{x}}_m\}_{m=0}^{\infty}$ and $\{\mathbf{x}_m\}_{m=0}^{\infty}$ 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 $p \in \mathfrak{E}_2$ such that (cf. (7.8) and (7.10))

$$s^{\pm}=\frac{(\sqrt{1+v}\pm\sqrt{1-v})^2}{2v}=\frac{1\pm\sqrt{1-v^2}}{v}. \tag{9.4}$$

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

$$\mu_0=\frac{2}{\gamma s}=\frac{2}{1+\sqrt{1-v^2}}, \qquad \mu_1=\frac{-2\delta}{\gamma s}=0, \qquad \mu_2=-\frac{1}{s^2}=-\frac{v^2}{[1+\sqrt{1-v^2}]^2}. \tag{9.5}$$

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

$$\kappa(U)=\frac{1}{s^+}=\frac{v}{1+\sqrt{1-v^2}}. \tag{9.6}$$

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_η of all ellipses with foci $F_1^+:=(v,0)$ and $F_2^-:=(-v,0)$ and major semiaxis a_η, $v<a_\eta<1$; we have $\kappa(E_\eta)=1/\eta$, where η is determined from (6.3): $a_\eta=(1/\eta-\mu_2\eta)/\mu_0$. If the semiaxes a_η or b_η are not known, but some other point P on the boundary of E_η is known, then a_η is given by the well-known property of any ellipse: $\overline{F^+P}+\overline{F^-P}=2a_\eta$.

It should be mentioned that if the Chebyshev semi-iterative method (see [13], § 5.1 and § 5.2) is applied to a linear system $\mathbf{x}=B\mathbf{x}+\mathbf{c}$, 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 $U := [-iv, iv]$, $v > 0$ arbitrary, that

$$s^{\pm} = \frac{(\sqrt{1+iv} \pm \sqrt{1-iv})}{2iv} = \frac{1 \pm \sqrt{1+v^2}}{iv} \tag{9.7}$$

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

$$\mu_0 = \frac{2}{\gamma s} = \frac{2}{1+\sqrt{1+v^2}}, \qquad \mu_1 = 0, \qquad \mu_2 = \frac{-1}{s^2} = \frac{v^2}{(1+\sqrt{1+v^2})^2}. \tag{9.8}$$

Again, the Euler function $p \in \mathfrak{E}_2$, determined by (9.8), is *optimal* with respect to $U = [-iv, +iv]$, and

$$\kappa(U) = \frac{1}{|s|} = \frac{v}{1+\sqrt{1+v^2}}. \tag{9.9}$$

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

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} 0 & B_1 \\ B_2 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} + \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \end{bmatrix}, \tag{9.10}$$

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

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

$(m \geqq 2)$. Writing

$$T := B_1 B_2, \qquad \tilde{\mathbf{c}} = \mathbf{c}_1 + B_1 \mathbf{c}_2 \tag{9.11}$$

we have

$$\mathbf{x}_m = \omega^2(T\mathbf{x}_{m-1} + \tilde{\mathbf{c}}) + 2(1-\omega)\mathbf{x}_{m-1} - (1-\omega)^2 \mathbf{x}_{m-2} (m \geqq 2), \tag{9.12}$$

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

$$\mu_0 := \omega^2, \qquad \mu_1 := 2(1-\omega), \qquad \mu_2 := -(1-\omega)^2. \tag{9.13}$$

The parameters μ_0, μ_1, μ_2, which satisfy $\mu_0 + \mu_1 + \mu_2 = 1$ and which depend only on ω, determine an Euler function $p = p(\omega) \in \mathfrak{E}_2$. 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 γ, δ, s yields

$$\gamma=\frac{2(1-\omega)}{\omega^2},\qquad \delta=\frac{-2(1-\omega)}{\omega^2},\qquad s=\frac{1}{1-\omega}. \tag{9.14}$$

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

$$|1-\omega|<1, \tag{9.15}$$

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 α and β, where $\gamma=(\beta-\alpha)/2$, $\delta=(\alpha+\beta)/2$, and which passes through the point $z=1$. From (9.14), we get

$$\alpha=0,\qquad \beta=\frac{-4(1-\omega)}{\omega^2}. \tag{9.16}$$

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

$$F^{\pm}:=\pm\frac{2\sqrt{1-\omega}}{\omega}, \tag{9.17}$$

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

$$\tau_j\in E_\eta(\omega)\quad \text{iff}\quad \beta_j\in\tilde{E}_\eta(\omega)\quad (j=1,\dots,n).$$

This means that every result concerning the spectrum $\sigma(T)$ and the region $E_\eta(\omega)$, holds equally for $\sigma(B)$ and $\tilde{E}_\eta(\omega)$. If $\rho(\mathfrak{L}_\omega)$ denotes the spectral radius of the SOR-operator $\mathfrak{L}_\omega$, 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 $\kappa(T,p(\omega))=\rho(\mathfrak{L}_\omega)$. 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 $\mathbf{z}-B\mathbf{z}=\mathbf{c}$, where B is weakly two-cyclic with elements from $\mathbb{C}$, is convergent, if $\sigma(B)$ is contained in the interior of the ellipse $\tilde{E}(\omega)$ 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 $\tilde{E}_\eta(\omega)\subset\tilde{E}(\omega)$, then $\rho(\mathfrak{L}_\omega)\leqq 1/\eta$, where equality holds if there is an eigenvalue of B on the boundary of $\tilde{E}_\eta(\omega)$. In particular, if $\sigma(B)\subset[F^-(\omega), F^+(\omega)]$, we have

$$\rho(\mathfrak{L}_\omega)=|1-\omega|. \tag{9.18}$$

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 $\rho(\mathfrak{L}_{\omega_j})\leqq 0.5$ if $\sigma(B)\subset[F_j^-, F_j^+]$, $j=1,2,3$.

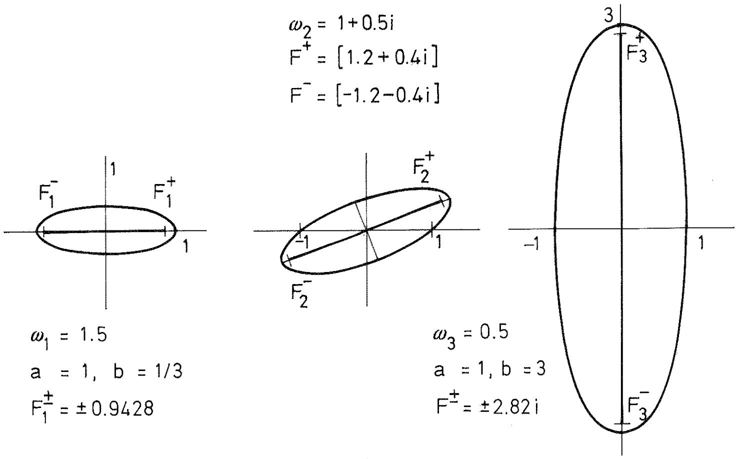

Fig. 4

Since for arbitrary ω with $|1-\omega|<1$, the interval $[-v, v]$ 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 $\sigma(B)\subset[-v, v]$, $v\in\mathbb{C}\setminus\{[1,\infty)\cup(-\infty,-1)\}$; in our notation of (1.13), this means $B\in O_{[-v,v]}$. Then from Corollary 10, it follows that $p(\omega)$ is optimal with respect to $[-v, v]$ if $[F^-, F^+]=[-v, v]$, where $F^\pm$ are determined by (9.17). This yields the relation $\pm v=\pm 2\sqrt{1-\omega}/\omega$, which solved for ω, gives the wellknown relation

$$\omega=\omega_b=\frac{2}{1+\sqrt{1-v^2}}, \qquad \kappa[-v, v]=|1-\omega_b|; \tag{9.19}$$

it should again be emphasized that (9.19) holds for every v with $v\in\mathbb{C}\setminus\{[1,\infty)\cup(-\infty,-1]\}$. 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 $\mathfrak{E}_4$, let us consider

$$p(\phi):=\frac{\mu_0\phi}{1-\mu_4\phi^4} \tag{9.20}$$

with real parameters μ_0, μ_4 and $\mu_0+\mu_4=1$. Then, with (9.20), the following 4-step iteration formula can be formally deduced (cf. Theorem 5):

$$\mathbf{x}_m=\mu_0(T\mathbf{x}_{m-1}+\mathbf{c})+\mu_4\mathbf{x}_{m-4} \qquad (m>4). \tag{9.21}$$

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 $\mathfrak{D}$ containing $\bar{D}_1$ for which $\tilde{p}$ is univalent in $\mathfrak{D}$. We introduce polar coordinates $\phi=\eta e^{i\theta}$ and obtain

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

From (9.22) we conclude that, for small η *and* μ_4, the image $\tilde{p}(D_\eta)$ is the exterior of a simply closed curve Γ_η, which is nearly a circle with radius $1/\eta$; Γ_η is a special type of a cycloid. For studying Γ_η, we form

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

The following Fig. 5 shows the curves Γ_η for $\mu_0=1.1$ (i.e., $\mu_4=-0.1$), $\eta=1.0$, 1.1, 1.2 and $\eta=\hat{\eta}(p)=\sqrt[4]{1/3|\mu_4|}=\sqrt[4]{10/3}$.

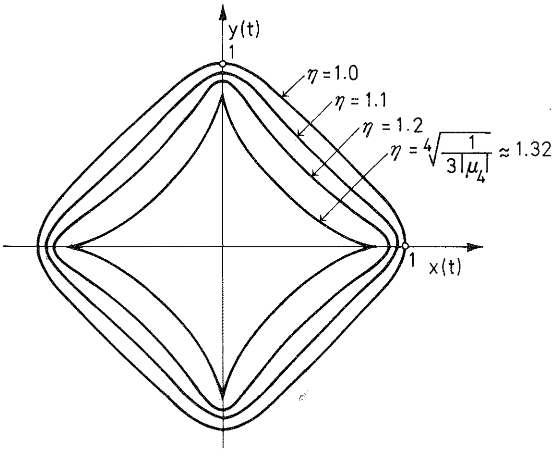

Fig. 5

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

$$-1/3<\mu_4<0. \tag{9.24}$$

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_η iff $1/\eta+3\mu_4\eta^3>0$ or $\eta^4<1/3|\mu_4|$, i.e., we get

$$\hat{\eta}(p)=\sqrt[4]{1/3|\mu_4|}. \tag{9.25}$$

Figure 5 shows that $\Gamma_{\hat\eta}$ has a cusp for $\theta=0$. From (9.22) and (9.25), it follow that

$$v=\tilde{p}(\hat\eta)=\frac{1}{1-\mu_4}\left[\frac{1}{\hat\eta}-\mu_4\hat\eta^3\right]=\frac{4}{3}\cdot\frac{1}{1+|\mu_4|}\sqrt[4]{3|\mu_4|}. \qquad (9.26$$

Now, as again can be seen from Fig. 5, $S_{\hat\eta}(p)$ is the closed interior of $\Gamma_{\hat\eta}$ and ha the four cusps $\pm v$, $\pm iv$, i.e. the "cross" $U_v:=[-v,v]\cup[-iv,iv]$ is containe strictly in $S_{\hat\eta}(p)$. From Corollary 11 it follows

$$\kappa(U_v)<\frac{1}{\hat\eta}=\sqrt[4]{3|\mu_4|}. \qquad (9.27$$

On the other hand, if there is an operator T such that $\sigma(T)\subset U_v$ for some with $0<v<1$ (i.e., if $T\in O_{U_v}$) then, since U_v is symmetric with respect to bot coordinate axes, *no* k-step method of the form (9.1) with $k=1$ or $k=2$ yields a acceleration of convergence, the same holding for $k=3$. Of course, sinc $\rho(T)\leqq v<1$, the simple iteration (1.2) converges. Now, since v is given, (9.2 can be solved for $|\mu_4|$. Method (9.21) with $\mu_4:=-|\mu_4|$ and $\mu_0=1-\mu_4$ has the a convergence factor, which is less than or equal $1/\hat\eta$, where $\hat\eta=(3|\mu_4|)^{-1/4}$. measure for the acceleration of convergence, which can be obtained by the 4 step method (9.20) is

$$\psi(v):=\frac{\ln(1/\eta)}{\ln(v)} \qquad (9.2$$

(see e.g. [13], p. 67). Table 1 shows some values of μ_4 and the resulting value of v, $1/\hat\eta$ and $\psi(v)$ according to (9.26), (9.25) and (9.28).

Table 1

μ_4	v	$1/\hat\eta$	$\psi(v)$
−0.025	0.6807	0.5233	1.6838
−0.05	0.7903	0.6223	2.014
−0.1	0.8971	0.7401	2.771
−0.2	0.9779	0.8801	5.715
−0.3	0.9990	0.9740	25.77

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

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

$$R_v:=\{z\in\mathbb{C}: |-v\leqq \operatorname{Re} z\leqq v,\ -1\leqq \operatorname{Im} z\leqq 1\}. \qquad (9.2$$

There exists, by Theorem 8, an Euler function p_0, optimal with respect to R but the computation of the corresponding Euler transform would go v

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_v, provides a lower bound $\kappa(E)$ for $\kappa(U_v)$, and $\kappa(E)$ is an upper bound, if $E \supset R_v$.

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

$$\kappa_1 := \max\left(\frac{v}{1+\sqrt{1-v^2}}, \frac{1}{1+\sqrt{2}}\right) < \kappa(R_v). \tag{9.30}$$

For our second lower bound κ_2, we use the fact that R_v contains the elliptic region E, where the boundary is the ellipse E_η with center 0 and semiaxes v and 1. The foci are

$$F^{\pm} := i\varepsilon \quad \text{with} \quad \varepsilon := \sqrt{1-v^2}. \tag{9.31}$$

Now, we want to determine an Euler function p from $\mathfrak{E}_2$, such that $E = S_\eta(p)$ for some $\eta > 1$. Since the center of E_η is 0, we have $\mu_1 = 0$. From (6.7) and (9.31), we get $\varepsilon = 2\sqrt{\mu_2}/\mu_0$ and by solving for μ_0,

$$\mu_0 = \frac{2(\sqrt{1+\varepsilon^2}-1)}{\varepsilon^2}, \qquad \mu_2 = 1 - \mu_0. \tag{9.32}$$

Since one semiaxes of E_η has endpoint 1, we get from (6.3) the equation $1 = (1/\eta + \mu_2\eta)/\mu_0$, which we can solve for η; there result two solutions η_1, η_2 with $1 < \eta_1 < \hat{\eta} = \sqrt{\mu_2} < \eta_2$. Thus, we have the second lower bound

$$\kappa_2 := \kappa(E) = 1/\eta_1, \qquad \text{where } \eta_1 = \frac{\mu_0 - \sqrt{\mu_0^2 - 4\mu_2}}{2\mu_2}, \tag{9.33}$$

where μ_0, μ_2 are given by (9.32). Finally, by Corollary 11, we note that

$$\kappa_1 < \kappa_2 < \kappa(R_v). \tag{9.34}$$

For an upper bound κ_3 for $\kappa(U_v)$, we seek another elliptic region E with $E \supset R_v$. Let E_η be the ellipse with center 0 and semiaxes a, b (where $v < a < 1$) such that E_η passes through the corner (v, i) of R_v. E_η is not unique, but there holds $v^2/a^2 + 1/b^2 = 1$, i.e.,

$$b^2 = \frac{a^2}{a^2 - v^2}. \tag{9.35}$$

So, for E_η, we have a one-parameter family. The foci are $F^{\pm} := i\varepsilon$, where $\varepsilon := \sqrt{b^2 - a^2}$; by (9.35), ε is a function of a. As before, we can determine parameters μ_0 and μ_2 according to (9.32); from (6.7) we conclude $b = (1/\eta + \mu_2\eta)/\mu_0$; solving for η yields

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

where

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

Now, b, μ_0, μ_2 are all functions of a and v, so κ_3 is a function of a and v; we get the lowest upper bound by minimizing κ_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 κ_1, κ_2 and κ_3 for different values of v, where κ_1 and κ_2 are *lower bounds* of $\kappa(R_v)$, and κ_3 is an *upper bound* of $\kappa(R_v)$.

Table 2

v	κ_1	κ_2	κ_3
0.2	0.4142	0.5	0.6171
0.4	0.4142	0.5941	0.7485
0.6	0.4142	0.7016	0.8605
0.8	0.5	0.8310	0.9503

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

References

1. Faddeev, D.K., Faddeeva, V.N.: Computational Methods of Linear Algebra. Freeman, San Francisco, 1963
2. Golub, G.H., Overton, M.: Convergence of a Two-Stage Richardson Iterative Procedure for Solving Systems of Linear Equations. Computer Science Dept. Manuscript NA-81-17, Stanford: Stanford University, California, 1981
3. Golub, G.H., Varga, R.S.: Chebyshev Semiiterative Methods, Successive Overrelaxation Iterative Methods, and Second Order Richardson Iterative Methods. Numer. Math. **3**, 147–156 (1961)
4. Kjellberg, K.: On the Convergence of Successive Over-Relaxation Applied to a Class of Linear Systems of Equations with Complex Eigenvalues. Ericsson Technics **2**, 245–258 (1958)
5. Knopp, K.: Über Polynomentwicklungen im Mittag-Lefflerschen Stern durch Anwendung der Eulerschen Reihentransformation. Acta Math. **47**, 313–335 (1926)
6. Kredell, B.: On Complex Successive Overrelaxation. BIT **2**, 143–152 (1962)
7. Kublanovskaya, V.N.: Application of analytic continuation in numerical analysis by means of change of variables. Trudy Mat. Inst. Steklov **53**, 145–185 (1959)
8. Manteuffel, T.A.: The Tchebychev Iteration for Nonsymmetric Linear Systems. Numer. Math. **28**, 307–327 (1977)
9. Niethammer, W.: Iterationsverfahren und allgemeine Euler-Verfahren. Math. Zeit. **102**, 288–317 (1967)
10. Niethammer, W.: Konvergenzbeschleunigung bei einstufigen Iterationsverfahren durch Summierungsmethoden. Iterationsverfahren, Numerische Mathematik, Approximationstheorie, Birkhäuser 1970, pp. 235–243
11. Perron, O.: Über eine Verallgemeinerung der Eulerschen Reihentransformation. Math. Zeit. **18**, 157–172 (1923)
12. de Pillis, J.: How to Embrace Your Spectrum for Faster Iterative Results. Linear Algebra Appl. **34**, 125–143 (1980)
13. Varga, R.S.: Matrix Iterative Analysis. Prentice Hall, Englewood Cliffs, N.J., 1962
14. Young, D.M.: Iterative Solution of Large Linear Systems. New York: Academic Press 1971
15. Zeller, K., Beekmann, W.: Theorie der Limitierungsverfahren. Berlin, Heidelberg, New York: Springer, 1970

Received May 28, 1982 / November 18, 1982