Numerical Algorithms 6 (1994) 229–244

Convergence of intermediate rows of minimal polynomial and reduced rank extrapolation tables

Avram Sidi

Computer Science Department, Technion – Israel Institute of Technology, Haifa 32000, Israel

Received 18 January 1993
Communicated by C. Brezinski

Let $\{x_m\}_{m=0}^{\infty}$ be a vector sequence obtained from a linear fixed point iterative technique in a general inner product space. In two previous papers [6,9] the convergence properties of the minimal polynomial and reduced rank extrapolation methods, as they are applied to the vector sequence above, were analyzed. In particular, asymptotically optimal convergence results pertaining to some of the rows of the tables associated with these two methods were obtained. In the present work we continue this analysis and provide analogous results for the remaining (intermediate) rows of these tables. In particular, when $\{x_m\}_{m=0}^{\infty}$ is a convergent sequence, the main result of this paper says, roughly speaking, that all of the rows converge, and it also gives the rate of convergence for each row. The results are demonstrated numerically through an example.

1. Introduction

Let $\boldsymbol{B}$ be an inner product space over $\boldsymbol{C}$, the field of complex numbers, and let (x, y) and $\|x\| = \sqrt{(x, x)}$ be, respectively, the inner product and norm associated with $\boldsymbol{B}$. The homogeneity property of the inner product is such that, for $\alpha, \beta \in \boldsymbol{C}$ and $x, y \in \boldsymbol{B}$, $(\alpha x, \beta y) = \bar{\alpha}\beta(x, y)$. Let $x_0, x_1, x_2, \ldots$, be a given sequence of vectors in $\boldsymbol{B}$. In case this sequence converges denote its limit by s, otherwise, let s stand for its antilimit. Whether this sequence converges or not, we can apply to it vector extrapolation methods in order to obtain good approximations to s. In the present work we shall concentrate on two such methods that have proved to be especially successful in many cases. These are the minimal polynomial extrapolation (MPE) of [1] and the reduced rank extrapolation (RRE) of [2] and [5]. For a method almost identical to RRE, see also [4]. For a detailed survey of these and other related methods, see [10]. For their efficient and stable numerical implementation, see [8].

When applied to the sequence $x_0, x_1, x_2, \ldots$, each of the methods MPE and RRE produces a two-dimensional array of approximations to s. These approximations, which we denote $s_{n,k}$, are of the form

$$s_{n,k} = \sum_{j=0}^{k} \gamma_j^{(n,k)} x_{n+j}\,, \tag{1.1}$$

with

$$\sum_{j=0}^{k} \gamma_j^{(n,k)} = 1\,. \tag{1.2}$$

The $\gamma_j^{(n,k)}$, in addition to (1.2), satisfy the k linear equations

$$\sum_{j=0}^{k} u_{i,j}^{(n)} \gamma_j^{(n,k)} = 0, \quad 0 \leqslant i \leqslant k-1\,, \tag{1.3}$$

where

$$u_{i,j}^{(n)} = \begin{cases} (u_{n+i}, u_{n+j}) & \text{for MPE}\,, \\ (w_{n+i}, u_{n+j}) & \text{for RRE}\,, \end{cases} \tag{1.4}$$

with

$$u_i = \Delta x_i = x_{i+1} - x_i \quad \text{and} \quad w_i = \Delta u_i = \Delta^2 x_i, \quad i = 0, 1, 2, \ldots\,. \tag{1.5}$$

Using (1.1)–(1.4), $s_{n,k}$ can be expressed as a quotient of two determinants. For all these developments, see [6].

Let us order the $s_{n,k}$ in a table akin to the Padé table as follows:

$$\begin{array}{cccc} s_{0,0} & s_{1,0} & s_{2,0} & \cdots \\ s_{0,1} & s_{1,1} & s_{2,1} & \cdots \\ s_{0,2} & s_{1,2} & s_{2,2} & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{array} \tag{1.6}$$

In [6] and [9] the problem of convergence of the rows of this table was addressed for sequences $\{x_m\}_{m=0}^{\infty}$ in $\boldsymbol{B}$ that satisfy

$$x_m \sim s + \sum_{j=1}^{\infty} P_j(m) \lambda_j^m \quad \text{as} \quad m \to \infty\,. \tag{1.7}$$

Here s is the limit or antilimit of the sequence $\{x_m\}_{m=0}^{\infty}$ as already mentioned, and λ_j are scalars that satisfy

$$|\lambda_1| \geqslant |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots\,, \tag{1.8a}$$

$$\lambda_i \neq \lambda_j \text{ if } i \neq j, \quad \lambda_i \neq 0 \text{ and } \lambda_i \neq 1 \text{ for all } i\,. \tag{1.8b}$$

In addition, we assume that there can be only a finite number of λ_j having the same modulus. $P_j(m)$ are polynomials in m with coefficients in $\boldsymbol{B}$, which we write in the form

$$P_j(m) = \sum_{l=0}^{p_j} y_{jl} \binom{m}{l}, \tag{1.9}$$

where $\binom{m}{l}$ are binomial coefficients and $y_{jl}, l = 0, 1, \ldots, p_j, j = 1, 2, \ldots$, form a linearly independent set of vectors in $\boldsymbol{B}$. We agree to order the λ_j such that

$$\text{if } |\lambda_j| = |\lambda_{j+1}|, \quad \text{then } p_j \geqslant p_{j+1}. \tag{1.10}$$

The meaning of (1.7) is that for any integer N there exist a positive constant K and a positive integer m_0 that depend only on N, such that for every $m \geqslant m_0$,

$$\left\| x_m - s - \sum_{j=1}^{N-1} P_j(m) \lambda_j^m \right\| \leqslant K m^{p_N} |\lambda_N|^m. \tag{1.11}$$

Sequences of vectors generated from nonsingular linear systems of equations by using stationary fixed point iterative techniques are exactly of the form described above. For this case the λ_j are *distinct nonzero* eigenvalues of the matrix of iteration, and, for each j, the vectors $y_{jl}, 0 \leqslant l \leqslant p_j$, are in the invariant subspace of this matrix that corresponds to λ_j. The vector s, which is the limit or antilimit of the sequence $\{x_m\}_{m=0}^{\infty}$, now is simply the solution of the linear system. Note that when the matrix of iteration is diagonalizable, $p_j = 0$ for all j, i.e., $P_j(m)$ are all constant in m. For a defective matrix of iteration $p_j \neq 0$ for some j, in general. A short description of all this will be given in the beginning of the next section. For the detailed derivation, see [9, section 2].

The following result concerning the rows of the MPE and RRE tables was proved in [9, theorem 3.1], see also [6, theorem 3.1].

THEOREM 1.1

Let the sequence $\{x_m\}_{m=0}^{\infty}$ be exactly as described above. Let

$$|\lambda_t| > |\lambda_{t+1}| \tag{1.12}$$

for some integer t, and let

$$k = \sum_{j=1}^{t} (p_j + 1). \tag{1.13}$$

Then, for both MPE and RRE, the approximations $s_{n,k}$ exist for all sufficiently large n, and satisfy

$$s_{n,k} - s = \Gamma(n) n^{p_{t+1}} |\lambda_{t+1}|^n, \tag{1.14}$$

where

$$\sup_n \|\Gamma(n)\| < \infty. \tag{1.15}$$

Furthermore, the dominant part of $\Gamma(n)$ as $n \to \infty$ is the same for both MPE and RRE, as a consequence of which, we also have

$$s_{n,k}^{\mathrm{MPE}} - s \sim s_{n,k}^{\mathrm{RRE}} - s \quad \text{as } n \to \infty . \tag{1.16}$$

Under the conditions imposed, this result is optimal and cannot be improved upon.

Judging from (1.12) and (1.13), we see that theorem 1.1 can cover all the rows of the extrapolation tables provided both of the following are satisfied:

(i) $p_j = 0$ for all j,

(ii) $|\lambda_1| > |\lambda_2| > |\lambda_3| > \ldots$.

Otherwise, theorem 1.1 covers only part of the rows dictated by (1.12) and (1.13), i.e., only those rows k for which k is as given by (1.13) for some t, and (1.12) is satisfied for this t. For instance, when $p_{t+1} > 0$, the rows k, with $k = \sum_{j=1}^{t}(p_j + 1) + i$, $1 \leqslant i \leqslant p_{t+1}$, are not covered by theorem 1.1.

The purpose of the present work is to treat the problem of these intermediate rows. We shall do this under the additional assumption that the sequence $\{x_m\}_{m=0}^{\infty}$ is generated by a stationary linear fixed point iterative technique. In section 2 we consider a sequence obtained from the iterative solution of a linear system of equations in $\boldsymbol{C}^N$, and show that a result very similar to (1.14) and (1.15) holds for all intermediate rows of the RRE and MPE tables, unconditionally for the former and under some mild conditions for the latter. The main result of section 2 is theorem 2.3. In section 3 we give the solution to an integer programming problem that arises in theorem 2.3, the main result of this section being theorem 3.2. In section 4 we verify numerically the results of theorems 1.1 and 2.3 by applying MPE and RRE to vector sequences obtained from finite difference discretization of a two-dimensional convection-diffusion equation. Now theorem 1.1 actually holds for a slightly generalized form of MPE and RRE that is described in [9, p.37, eq. (1.16)]. In section 5 we describe this generalization, and show that theorem 2.3 holds for this generalized form as well. In section 6 we go back to the general inner product space $\boldsymbol{B}$, and show that the results of section 2 can easily be extended to this case with no substantive changes.

2. Theory for finite dimensional spaces

2.1. GENERAL CONSIDERATIONS

We assume in this section that $\boldsymbol{B}$ is the N-dimensional space $\boldsymbol{C}^N$. Let us denote by s the solution of the nonsingular linear system of equations

$$x = Ax + b . \tag{2.1}$$

Starting with an arbitrary vector x_0, if we now generate the sequence $\{x_m\}_{m=0}^{\infty}$ by the iterative technique

$$x_{j+1} = Ax_j + b, \quad j = 0, 1, \ldots , \tag{2.2}$$

then, as mentioned in the previous section, the vector x_m is of the form

$$x_m = s + \sum_{j=1}^{\nu} P_j(m)\lambda_j^m, \quad \text{for all } m \geqslant d, \text{ some } d \geqslant 0. \tag{2.3}$$

Here λ_j are some or all of the distinct nonzero eigenvalues of A, which can be ordered as in (1.8a), and satisfy (1.8b) automatically. The $P_j(m)$ are precisely as described in the previous section, and the vectors y_{jl} in (1.9) satisfy

$$(A - \lambda_j I)^i y_{jl} = 0, \quad p_j - i + 1 \leqslant l \leqslant p_j, 1 \leqslant i \leqslant p_j + 1,$$

$$(A - \lambda_j I)^i y_{j,p_j - i} \neq 0, \quad 0 \leqslant i \leqslant p_j. \tag{2.4}$$

Actually, (2.4) is a result of the fact that y_{jp_j} is an eigenvector of A corresponding to λ_j, while, for $p_j > 0$, y_{j,p_j-i} is a linear combination of eigenvectors and 1st, ..., ith principal vectors that correspond to λ_j. If r_j is the dimension of the largest Jordan block corresponding to the eigenvalue λ_j, then $p_j \leqslant r_j - 1$. For an arbitrary initial vector x_0, in general, $p_j = r_j - 1$ may hold. Finally, the integer d in (2.3) is determined by the zero eigenvalues of A. We have $d = 0$ when A is nonsingular, and d is equal to the index of A when A is singular.

2.2. EXISTENCE OF APPROXIMATIONS

Since we are going to be discussing the convergence of the sequence $\{s_{n,k}\}_{n=0}^{\infty}$ for arbitrary fixed k, we first have to address the question of existence of the $s_{n,k}$.

THEOREM 2.1.

Let k_0 be the degree of the minimal polynomial of A with respect to the vector $x_n - s$. Then (i) $s_{n,k}$ for RRE exists and is unique unconditionally for all $k < k_0$, and (ii) $s_{n,k}$, for MPE exists and is unique for all $k < k_0$, provided the hermitian part of the matrix $\alpha(I - A)$ is positive definite for some $\alpha \in C$, $|\alpha| = 1$. Also s_{n,k_0} exists and is equal to s unconditionally for both MPE and RRE.

The part of theorem 2.1 pertaining to RRE follows from [7, theorem 2.1], while that pertaining to MPE follows from [7, theorem 2.2] and is a slight improvement of the latter. We also note that the minimal polynomials of A with respect to the vectors $x_m - s$, $m \geqslant d$, are identical to each other.

We note that the sufficient condition concerning the existence of $s_{n,k}$ for MPE that is given in theorem 1.1 is different from the one in theorem 2.1. They do not contradict each other, but are supplementary. In the sequel we shall assume the existence of $s_{n,k}$ for MPE and RRE precisely under the conditions stated in theorem 2.1.

2.3. THEORETICAL UPPER BOUNDS FOR ERROR NORMS

We define the residual vector $r(x)$ associated with an arbitrary vector x by

$$r(x) = Ax + b - x = (A - I)(x - s)\,. \tag{2.5}$$

Consequently, $\|r(x)\|$ is a true norm for $x - s$.

We also define the matrix C and its hermitian part C_H by

$$C = \alpha(I - A) \quad \text{and} \quad C_H = \tfrac{1}{2}(C + C^*)\,, \tag{2.6}$$

for some $\alpha \in C, |\alpha| = 1$. In addition, in case C_H is positive definite, we define the vector norm $\|\cdot\|'$ by

$$\|x\|' = \sqrt{(x, C_H x)}\,. \tag{2.7}$$

Theorem 2.2 below will be the starting point of our treatment of the intermediate rows of the MPE and RRE tables.

THEOREM 2.2

With $k < k_0$, where k_0 is the degree of the minimal polynomial of A with respect to $x_n - s$, we have, for RRE

$$\|r(s_{n,k})\| \leqslant \|Cq(A)(x_n - s)\|\,, \tag{2.8}$$

while for MPE, assuming that C_H is positive definite,

$$\|s_{n,k} - s\|' \leqslant \|C_H^{-1/2} C^* q(A)(x_n - s)\|'\,, \tag{2.9}$$

where $q(\lambda)$ is an arbitrary polynomial of degree at most k that satisfies $q(1) = 1$.

The part of theorem 2.2 pertaining to RRE follows by combining the result

$$\|r(s_{n,k})\| \leqslant \|q(A)r(x_n)\|\,, \tag{2.10}$$

see [7, theorem 4.2], with (2.5), while that pertaining to MPE is a slight improvement of [7, theorem 4.4].

2.4. MAIN RESULT FOR CONVERGENCE OF INTERMEDIATE ROWS

We now state the main result of the present work.

THEOREM 2.3

Assume that x_m is exactly as described in section 2.1, and that the conditions of theorem 2.2 hold as well. Let

$$|\lambda_t| > |\lambda_{t+1}| = \ldots = |\lambda_{t+r}| > |\lambda_{t+r+1}| \tag{2.11}$$

for some $t \geqslant 0$ and $r \geqslant 1$. (Here we set $|\lambda_0| = \infty$ and $\lambda_{\nu+1} = 0$.) Denote for convenience

$$\omega_j = p_j + 1, \quad j = 1, 2, \ldots . \tag{2.12}$$

Let the integer k satisfy

$$\sum_{j=1}^{t} \omega_j < k < \sum_{j=1}^{t+r} \omega_j , \tag{2.13a}$$

and set

$$\tau = k - \sum_{j=1}^{t} \omega_j . \tag{2.13b}$$

Define the set $S(\tau)$ of integer r-tuples $(\sigma_1, \ldots, \sigma_r)$ by

$$S(\tau) = \left\{ (\sigma_1, \ldots, \sigma_r) : 0 \leqslant \sigma_i \leqslant \omega_{t+i}, 1 \leqslant i \leqslant r, \text{ and } \sum_{i=1}^{r} \sigma_i = \tau \right\}. \tag{2.14}$$

Define the nonnegative integer β by

$$\beta = \min_{(\sigma_1, \ldots, \sigma_r) \in S(\tau)} \max_{1 \leqslant i \leqslant r} (p_{t+i} - \sigma_i) . \tag{2.15}$$

Then, for both MPE and RRE,

$$\|s_{n,k} - s\| = O(n^{\beta} |\lambda_{t+1}|^n) \quad \text{as } n \to \infty . \tag{2.16}$$

NOTE

As will be shown in the next section, β is nonincreasing in τ, and takes on all values between p_{t+1} and 0 as τ takes on all values between 0 and $\sum_{i=1}^{r} \omega_{t+i} - 1$. Also, $\beta = 0$ when $p_{t+i} = 0, 1 \leqslant i \leqslant r$, as is seen from (2.15).

Proof

We start by analyzing the vector $q(A)(x_n - s)$ that appears in both (2.8) and (2.9), recalling at the same time that $q(\lambda)$ is an arbitrary polynomial of degree at most k that satisfies $q(1) = 1$. Let us pick

$$q(\lambda) = \prod_{j=1}^{t} \left(\frac{\lambda - \lambda_j}{1 - \lambda_j} \right)^{\omega_j} \prod_{i=1}^{r} \left(\frac{\lambda - \lambda_{t+i}}{1 - \lambda_{t+i}} \right)^{\sigma_i}, \quad (\sigma_1, \ldots, \sigma_r) \in S(\tau) . \tag{2.17}$$

Now, by (1.9) and (2.4),

$$(A - \lambda_j I)^{\omega_j} P_j(n) = 0 \quad \text{for all } n . \tag{2.18}$$

Employing (2.3), (2.17), and (2.18), we obtain

$$q(A)(x_n - s) = \sum_{j=t+1}^{\nu} [q(A) P_j(n)] \lambda_j^n . \tag{2.19}$$

Again, from (1.9) and (2.4) and the fact that $\binom{n}{l} \sim n^l / l!$ as $n \to \infty$, for $1 \leqslant i \leqslant r$,

$$(A-\lambda_{t+i}I)^{\sigma_i}P_{t+i}(n)=\sum_{l=0}^{p_{t+i}-\sigma_i}[(A-\lambda_{t+i}I)^{\sigma_i}y_{jl}]\binom{n}{l}$$
$$=\begin{cases}O(n^{p_{t+i}-\sigma_i}) & \text{as } n\to\infty, \quad 0\leqslant\sigma_i\leqslant p_{t+i}\,,\\ 0, & \sigma_i=\omega_{t+i}\,,\end{cases} \tag{2.20}$$

as a consequence of which we have

$$q(A)P_{t+i}(n)=\begin{cases}O(n^{p_{t+i}-\sigma_i}) & \text{as } n\to\infty, \quad 0\leqslant\sigma_i\leqslant p_{t+i}\,,\\ 0, & \sigma_i=\omega_{t+i}\,.\end{cases} \tag{2.21}$$

This, along with (2.11), results in

$$\sum_{j=t+1}^{t+r}[q(A)P_j(n)]\lambda_j^n=O(n^{\delta}|\lambda_{t+1}|^n) \quad \text{as } n\to\infty\,, \tag{2.22}$$

where

$$\delta=\max_{1\leqslant i\leqslant r}(p_{t+i}-\sigma_i)\,. \tag{2.23}$$

From (1.8a) and (1.10) we also have

$$\sum_{j=t+r+1}^{\nu}[q(A)P_j(n)]\lambda_j^n=O(n^{p_{t+r+1}}|\lambda_{t+r+1}|^n) \quad \text{as } n\to\infty\,. \tag{2.24}$$

Combining (2.22) and (2.24) in (2.19), and recalling (2.11) again, we obtain

$$q(A)(x_n-s)=O(n^{\delta}|\lambda_{t+1}|^n) \quad \text{as } n\to\infty\,. \tag{2.25}$$

Finally, we recall that $(\sigma_1,\dots,\sigma_r)$ in (2.23) is in $S(\tau)$, but is arbitrary otherwise. This means that δ can be minimized over the set $S(\tau)$. As a result of this minimization process we obtain $\delta=\beta$, with β as defined in (2.15). This completes the proof. □

Remarks

(1) Comparing (2.11) and (2.13a) in theorem 2.3 with (1.12) and (1.13) in theorem 1.1, we realize that theorem 2.3 indeed covers *all* of the intermediate rows of the MPE and RRE tables. This is true whether the matrix A is diagonalizable or not.

(2) Combining theorems 1.1 and 2.3, we conclude that, for all values of k, the convergence of row k is at least as rapid as that of row $k-1$.

(3) Again, combining theorems 1.1 and 2.3, we see that, for n sufficiently large, $\|s_{n,k}-s\|$ may not decrease substantially when k corresponds to consecutive intermediate rows, i.e., when k satisfies (2.13a). However, provided $|\lambda_{t+1}|$ is substantially larger than $|\lambda_{t+r+1}|$, a large jump from $\|s_{n,k-1}-s\|$ to $\|s_{n,k}-s\|$ may take place for $k=\sum_{j=1}^{t+r}\omega_j$.

2.5. THEOREM 1.1 REVISITED

We would like to note that the technique that we used in the proof of theorem 2.3 can also be used in the proof of theorem 1.1 under the additional conditions of theorem 2.3. By (1.13) and (2.13b) we see first that $\tau = 0$ for theorem 1.1. This, of course, forces $\sigma_i = 0, i = 1, \ldots, r$, everywhere. As a result, we have $\beta = \delta = p_{t+1}$, which, along with (2.16), produces (1.14) with (1.15).

We should also note, however, that the result produced by this technique is not optimal as it is based on the inequalities of theorem 2.2. In particular, it does not give the dominant behavior of the vector $\Gamma(n)$ in (1.14), and thus it does not enable us to obtain a very refined result such as (1.16).

3. Solution of the optimization problem in (2.15)

In this section we would like to tackle the integer programming problem in (2.15). For simplicity, and without loss of generality, we set $t = 0$ in (2.14) and (2.15), and denote the min-max problem there by $Q(\tau)$, and β by $\beta(\tau)$. That is, $Q(\tau)$ stands for the integer programming problem

$$\beta(\tau) = \min_{(\sigma_1,\ldots,\sigma_r)\in S(\tau)} \max_{1\leqslant i\leqslant r} (p_i - \sigma_i). \tag{3.1}$$

THEOREM 3.1

For $0 \leqslant \tau < \sum_{i=1}^{r} \omega_i$, $\beta(\tau)$ is a nonincreasing function of τ.

Proof

Let $(\sigma_1^*, \ldots, \sigma_r^*) \in S(\tau)$ be a solution to $Q(\tau)$, and consider $Q(\tau+1)$. Now at least one of the σ_i^*, σ_q^* say, satisfies $0 \leqslant \sigma_q^* \leqslant p_q$. Let us set $\sigma_q = \sigma_q^* + 1$, and $\sigma_i = \sigma_i^*, i \neq q$. Then $(\sigma_1, \ldots, \sigma_r) \in S(\tau+1)$, and

$$\beta(\tau+1) \leqslant \max_{1\leqslant i\leqslant r} (p_i - \sigma_i) = \max \left[\max_{\substack{1\leqslant i\leqslant r \\ i\neq q}} (p_i - \sigma_i^*), \quad p_q - \sigma_q^* - 1 \right] \leqslant \beta(\tau). \tag{3.2}$$

This completes the proof. □

To understand the nature of the solution to $Q(\tau)$, let us first consider small values of τ. We shall adhere to our convention $p_1 \geqslant p_2 \geqslant \ldots \geqslant p_r$ established in (1.10).

(i) When $\tau = 0, \sigma_i^* = 0, 1 \leqslant i \leqslant r$, and $\beta(0) = p_1$, trivially.

(ii) When $\tau = 1$, $\sigma_1^* = 1$, $\sigma_i^* = 0$, $2 \leqslant i \leqslant r$, and $\beta(1) = p_1$ if $p_1 = p_2$ or $\beta(1) = p_1 - 1$ if $p_1 > p_2$.

(iii) When $\tau = 2$, $\sigma_1^* = \sigma_2^* = 1$, $\sigma_i^* = 0$, $3 \leqslant i \leqslant r$, if $p_1 = p_2$, or $\sigma_1^* = 2$, $\sigma_i^* = 0$, $2 \leqslant i \leqslant r$, if $p_1 > p_2$. In the former case $\beta(2) = p_1$ if $p_1 = p_2 = p_3$ or

$\beta(2) = p_1 - 1$ if $p_1 = p_2 > p_3$, while in the latter $\beta(2) = p_1 - 1$ if $p_1 = p_2 + 1$ or $\beta(2) = p_1 - 2$ if $p_1 \geqslant p_2 + 2$.

THEOREM 3.2.

Let the integers $r_1, r_2, \ldots$, be such that

$$p_1 = \ldots = p_{r_1} > p_{r_1+1} = \ldots = p_{r_1+r_2} > \ldots . \tag{3.3}$$

(It may be possible that $r_1 = r$ already, so that $r_2, r_3, \ldots$, are all zero). Let us also define

$$R_j = \sum_{i=1}^{j} r_i \quad \text{and} \quad c_j = p_{R_j} - p_{R_{j+1}}, \quad j = 1, 2, \ldots . \tag{3.4}$$

Let τ, $1 \leqslant \tau < \sum_{j=1}^{r} \omega_j$, be given. Then there exist unique integers q, e, and ρ, satisfying $q \geqslant 0$, $0 \leqslant e \leqslant c_{q+1}$, and $0 \leqslant \rho < R_{q+1}$, such that

$$\tau = \sum_{j=1}^{q} c_j R_j + eR_{q+1} + \rho . \tag{3.5}$$

With this, an optimal solution $(\sigma_1^*, \ldots, \sigma_r^*)$ to $Q(\tau)$ is given by

$$\sigma_i^* = \sum_{j=l}^{q} c_j + e + [i \leqslant \rho] \quad \text{if } R_{l-1} + 1 \leqslant i \leqslant R_l , \tag{3.6}$$

where $[i \leqslant \rho] = 1$ if $i \leqslant \rho$, and $[i \leqslant \rho] = 0$ otherwise, and $R_0 = 0$. Also

$$\beta(\tau) = \begin{cases} p_1 - \sigma_1^* & \text{if } \rho = 0 , \\ p_1 - \sigma_1^* + 1 & \text{if } \rho > 0 . \end{cases} \tag{3.7}$$

As a result, as τ increases from 0 to $\sum_{j=1}^{r} \omega_j - 1$, $\beta(\tau)$ is nonincreasing and takes on all the values $p_1, p_1 - 1, \ldots, 0$, in this order.

Instead of giving a formal proof of theorem 3.2, we will show graphically how the optimal solution is constructed. This construction will also show that the solution given in (3.6) is an optimal one. We shall do this through an example.

EXAMPLE

$r = 8$, $p_1 = p_2 = 4$, $p_3 = p_4 = p_5 = 2$, $p_6 = p_7 = 1$, $p_8 = 0$. Thus $r_1 = 2$, $r_2 = 3$, $r_3 = 2, r_4 = 1, R_1 = 2, R_2 = 5, R_3 = 7, R_4 = 8$, and $c_1 = 2, c_2 = c_3 = 1$.

We now form the $i - p_i$ histogram, see fig. 1. The integers within the squares are all possible values of τ between 1 and $\sum_{i=1}^{8} (p_i + 1)$. Note the order in which the squares are numbered.

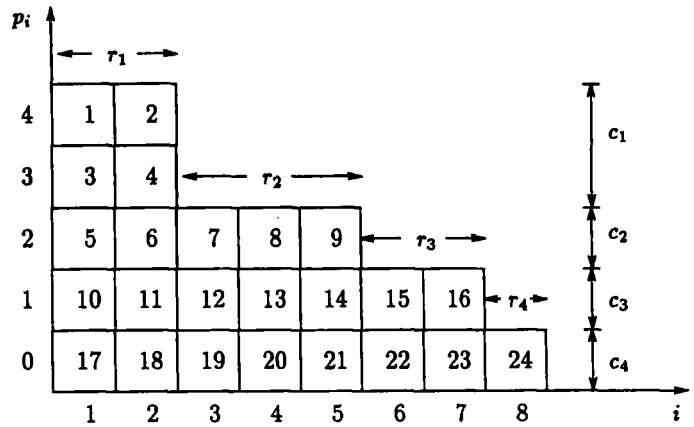

Fig. 1.

Let us consider the optimal solution for $\tau = 13$. We start by putting an X in all the squares numbered 1 through 13, see fig. 2. Then the number of X's in the $i = 1$ column is σ_1^*, the number of X's in the $i = 2$ column is σ_2^*, etc. Thus $\sigma_1^* = \sigma_2^* = 4$, $\sigma_3^* = \sigma_4^* = 2, \sigma_5^* = 1, \sigma_6^* = \sigma_7^* = \sigma_8^* = 0$. As a result, $\beta(13) = 1$.

We also see from this that the solution to $Q(\tau + 1)$ can be obtained from that of $Q(\tau)$ by increasing only one of the σ_i^* by 1. For example, the solution to $Q(14)$ is obtained from that of $Q(13)$ by increasing σ_5^* from 1 to 2, while the rest of the σ_i^* are kept unchanged. In addition, $\beta(14) = 1$.

By moving the X's to the other squares, while their total number is kept fixed, we can see that $\max_i(p_i - \sigma_i)$ either increases or remains unchanged, as should happen at the minimum. We also see that the solution to $Q(\tau)$ is not necessarily unique.

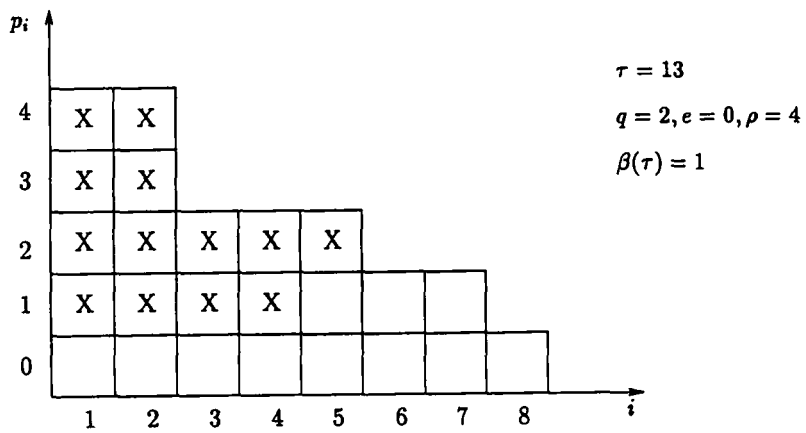

Fig. 2.

4. A numerical example

Consider the two-dimensional convection-diffusion equation

$$-\frac{\partial^2 u}{\partial x^2}-\frac{\partial^2 u}{\partial y^2}+\gamma\left(x\frac{\partial u}{\partial x}+y\frac{\partial u}{\partial y}\right)+\beta u=f \quad \text{in } \Omega,$$

$$u=g \quad \text{on } \partial\Omega, \tag{4.1}$$

where Ω is the unit square. This equation has been used as a test problem for vector extrapolation methods and Krylov subspace methods on nonsymmetric and/or indefinite systems. See, e.g., [3].

Let $x_i = i\delta x$, $0 \leqslant i \leqslant M_x + 1$, and $y_j = j\delta y$, $0 \leqslant j \leqslant M_y + 1$, where $\delta x = 1/(M_x + 1)$ and $\delta y = 1/(M_y + 1)$ for some positive integers M_x and M_y. We discretize this equation by replacing all the partial derivatives at (x_i, y_j) by central differences. If we now order the unknowns $u_{i,j}$, which are the approximations to the corresponding $u(x_i, y_j)$, in the form $u_{11}, u_{12}, \ldots, u_{1M_y}, u_{21}, u_{22}, \ldots, u_{2M_y}, \ldots, u_{M_x 1}, u_{M_x 2}, \ldots, u_{M_x M_y}$, then we obtain a linear system of equations with a block tridiagonal matrix. If $\beta = 0 = \gamma$, then we have the usual 5-point discretization scheme for Poisson's equation, in which case the matrix of the linear system is symmetric and positive definite. By increasing β in the negative direction we can make the matrix less and less positive definite and ultimately cause it to become indefinite. By picking $\gamma \neq 0$ we make the matrix nonsymmetric, the amount of asymmetry being directly related to the size of γ.

In our computations we pick $M_x = M_y = 30$ so that the number of unknowns is $N = M_x M_y = 900$. We also take $g = 0$ as our boundary condition and $f = 0$, causing the solution (both of the partial differential equation and of the difference equations) to be zero everywhere. We use the Jacobi iteration technique starting with $(1, 1/\sqrt{2}, 1/\sqrt{3}, \ldots, 1/\sqrt{N})^{\mathrm{T}}$ as the initial vector.

We now apply MPE and RRE in conjunction with the Jacobi iterative technique for the linear system above. We first recall that the matrix of this system is consistently ordered. This implies that if μ is an eigenvalue of the Jacobi iterative matrix, then so is $-\mu$. As a result, the number of distinct eigenvalues having the same modulus is always even. This, of course, implies that theorem 1.1 applies only for some or all of the even k, and theorem 2.3 applies for the rest.

The numerical results reported in this section were all obtained by using the FORTRAN 77 code given in [8]. The computations were performed in extended double precision arithmetic on an IBM-370 machine.

Case 1: Let us take $\gamma = 0$ and $\beta = 0$. Then we are dealing with the Poisson equation with Dirichlet boundary conditions on the unit square. The Jacobi iteration matrix is real symmetric and all its eigenvalues are in $(-1, 1)$ and are symmetri-

cally distributed about 0. Thus theorem 1.1 applies to $s_{n,k}$ with $k = 2, 4, 6, \ldots$, since $p_j = 0$ for all j; (1.14) holds with $p_{l+1} = 0$, and (1.16) holds too.

Table 1 gives $\|s_{n,k} - s\|$, the Euclidean norm of the error, for $n = 500$ and $k = 1, 2, \ldots, 20$, for both MPE and RRE. We observe that $\|s^{\mathrm{MPE}}_{500,k} - s\|$ is almost identical to $\|s^{\mathrm{RRE}}_{500,k} - s\|$ for $k = 2, 4, \ldots, 20$, in accordance with (1.16). We also observe that, for both MPE and RRE, $\|s_{500,2q+1} - s\|$ is not much smaller than $\|s_{500,2q} - s\|$, in accordance with remark 3 following the proof of theorem 2.3.

Case 2: Let us now take $\gamma = 100$ and $\beta = 0$. The resulting system of linear equations is real nonsymmetric. As a result of this the Jacobi iteration matrix is also real nonsymmetric. We do not know for sure whether the eigenvalues of the Jacobi iteration matrix are real or not for this case. As mentioned above, however, we know that theorem 1.1 applies for some or all of the even k's.

Table 2 gives $\|s_{n,k} - s\|$ for $n = 200$ and $k = 1, 2, \ldots, 20$, for both MPE and RRE. It seems from this table that theorem 1.1 applies with $k = 2, 6, 8, 12, 14$, at least. This implies first that the Jacobi iteration matrix has two distinct real eigenvalues $\pm\mu$ of largest size. Then there are either two distinct real eigenvalues $\pm\nu$ each

Table 1
l_2-norm of the errors in $s_{500,k}, k = 0, 1, \ldots, 20$ where $s_{n,k}$ is computed by applying MPE and RRE in conjunction with the Jacobi iteration method for the finite difference equations obtained from (4.1) with $f = 0$ and $g = 0$ and $\beta = 0$ and $\gamma = 0$. The solution for this case is $s = 0$. The initial vector is taken to be $(1, 1/\sqrt{2}, 1/\sqrt{3}, \ldots, 1/\sqrt{N})^{\mathrm{T}}$, where N is the number of unknowns. Recall that $s_{n,0} = x_n$ for all n.

k	$\|s^{\mathrm{MPE}}_{500,k} - s\|$	$\|s^{\mathrm{RRE}}_{500,k} - s\|$
0	1.03D − 01	1.03D − 01
1	6.07D − 02	1.03D − 01
2	1.19D − 03	1.19D − 03
3	1.01D − 03	1.18D − 03
4	8.57D − 06	8.58D − 06
5	8.05D − 06	8.46D − 06
6	5.09D − 07	5.10D − 07
7	4.55D − 07	5.04D − 07
8	1.37D − 08	1.38D − 08
9	1.28D − 08	1.35D − 08
10	4.31D − 10	4.33D − 10
11	2.23D − 10	4.15D − 10
12	7.09D − 12	7.10D − 12
13	5.73D − 12	6.90D − 12
14	1.29D − 13	1.30D − 13
15	1.22D − 13	1.26D − 13
16	1.03D − 14	1.05D − 14
17	7.20D − 15	1.00D − 14
18	1.22D − 16	1.22D − 16
19	7.50D − 17	1.16D − 16
20	3.65D − 18	3.67D − 18

Table 2
l_2-norm of the errors in $s_{200,k}, k = 0, 1, \ldots, 20$ where $s_{n,k}$ is computed by applying MPE and RRE in conjunction with the Jacobi iteration method for the finite difference equations obtained from (4.1) with $f = 0$ and $g = 0$ and $\beta = 0$ and $\gamma = 100$. The solution for this case is $s = 0$. The initial vector is taken to be $(1, 1/\sqrt{2}, 1/\sqrt{3}, \ldots, 1/\sqrt{N})^{\mathrm{T}}$, where N is the number of unknowns. Recall that $s_{n,0} = x_n$ for all n.

k	$\|S_{200,k}^{\mathrm{MPE}} - s\|$	$\|S_{200,k}^{\mathrm{RRE}} - s\|$
0	2.32D − 03	2.32D − 03
1	2.18D − 03	2.26D − 03
2	8.78D − 07	8.78D − 07
3	7.38D − 07	8.09D − 07
4	5.25D − 08	2.01D − 07
5	3.33D − 08	9.50D − 08
6	5.31D − 11	5.31D − 11
7	3.88D − 11	4.46D − 11
8	1.01D − 15	1.01D − 15
9	6.29D − 16	7.57D − 16
10	2.78D − 17	4.71D − 18
11	1.38D − 17	1.07D − 17
12	3.00D − 21	3.00D − 21
13	1.60D − 21	1.99D − 21
14	5.29D − 27	5.29D − 27
15	3.15D − 27	3.90D − 27
16	2.72D − 27	3.02D − 27
17	1.36D − 27	1.79D − 27
18	8.74D − 29	1.12D − 28
19	5.72D − 29	8.20D − 29
20	4.05D − 31	3.91D − 31

having one corresponding eigenvector and one corresponding principal vector or four distinct complex eigenvalues $\pm(\alpha \pm \mathrm{i}\beta)$ having, of course, the same modulus. We can continue this way and come to conclusions about the rest of the eigenvalues.

5. A further development

For given n and k let the $\gamma_j^{(n,k)}$ for MPE and RRE be exactly as defined through (1.2)–(1.5) in section 1. Let us now replace the approximation $s_{n,k}$ of section 1 by a corresponding approximation $s_{n,k}^{(q)}$, where

$$s_{n,k}^{(q)} = \sum_{j=0}^{k} \gamma_j^{(n,k)} x_{n+q+j}, \quad q \geqslant 0 \text{ a fixed integer}. \tag{5.1}$$

As $s_{n,k}^{(0)} = s_{n,k}$, $s_{n,k}^{(q)}$ is, in fact, a (slight) generalization of $s_{n,k}$. When the $\gamma_j^{(n,k)}$ are those obtained from RRE, $s_{0,k}^{(1)}$ turns out to be precisely the approximation given in [4].

Obviously, for given n and k, *all* $s_{n,k}^{(q)}, q = 0, 1, 2, \ldots$, use the *same* set of $\gamma_j^{(n,k)}$. The reader should be cautioned not to confuse $s_{n,k}^{(q)}$ with $s_{n+q,k}$. The simplest difference between the two is that $s_{n,k}^{(q)}$ is constructed from the vectors x_i, $n \leqslant i \leqslant n + k + \max(1, q)$, whereas $s_{n+q,k}$ is constructed from $x_i, n + q \leqslant i \leqslant n + q + k + 1$, and these two sets of vectors are different from each other when $q > 0$. Actually, the vector $s_{n,k}^{(q)}$ with $q > 0$ does *not* belong in the table given in (1.6) whenever $k < k_0$, cf. theorem 2.1.

The treatment of the row convergence problem in [9] was actually achieved for the $s_{n,k}^{(q)}$. In particular, theorem 1.1 holds in its entirety with $s_{n,k}$ replaced by $s_{n,k}^{(q)}$. In view of this fact, it is natural to ask whether the results for the intermediate rows hold with $s_{n,k}$ replaced by $s_{n,k}^{(q)}$. The question is of interest also since theorem 2.2, which forms the starting point of the proof of theorem 2.3, is not satisfied by $s_{n,k}^{(q)}$ when $q > 0$. Theorem 5.1 below provides the answer to this question.

THEOREM 5.1

Theorems 2.1 and 2.3 hold without any changes when $s_{n,k}$ is replaced by $s_{n,k}^{(q)}$.

Proof

First, we recall that, under the conditions of theorem 2.1, $s_{n,k}$ exists and is unique if and only if the $\gamma_j^{(n,k)}$ exist and are unique. This and (5.1) are enough to conclude that theorem 2.1 about the existence of the $s_{n,k}$ holds with $s_{n,k}$ replaced by $s_{n,k}^{(q)}$. Next, we recall that (2.2) implies

$$x_m - s = A^m(x_0 - s), \quad m = 0, 1, \ldots . \tag{5.2}$$

Combining this with (5.1) and with $\sum_{j=0}^{k} \gamma_j^{(n,k)} = 1$, we obtain

$$\begin{aligned} s_{n,k}^{(q)} - s &= \sum_{j=0}^{k} \gamma_j^{(n,k)} (x_{n+q+j} - s) \\ &= A^q \left[\sum_{j=0}^{k} \gamma_j^{(n,k)} (x_{n+j} - s) \right] = A^q (s_{n,k} - s) . \end{aligned} \tag{5.3}$$

From this we immediately have

$$\| s_{n,k}^{(q)} - s \| \leqslant \| A^q \| \, \| s_{n,k} - s \| \tag{5.4}$$

in *any* norm whatsoever. Invoking now theorem 2.3 on the right hand side of (5.4), we see that it holds with $s_{n,k}$ replaced by $s_{n,k}^{(q)}$. □

6. Extension to general inner product spaces

In section 2 we restricted ourselves to vector sequences obtained from iterative solution of linear systems of equations of the form (2.1) in C^N. In this section we consider vector sequences obtained from iterative solution of linear operator equa-

tions of the form (2.1) in an infinite dimensional inner product space $\boldsymbol{B}$ described in the first paragraph of section 1. We assume that the λ_j and y_{jl} are precisely as in section 1, and that the vectors x_m satisfy (1.7) the summation there being indeed infinite. With this last condition the concept of the minimal polynomial of A with respect to a vector loses its meaning and relevance, and we have $k_0 = \infty$ formally.

By going through the proofs of theorems 2.1 and 2.2, we see that their results remain unchanged whether A is an $N \times N$ matrix or a bounded linear operator in a general inner product space $\boldsymbol{B}$. The same argument applies to theorem 2.3 as well. We only need to keep in mind that (i) in (2.19) and (2.24) ν is to be replaced by ∞, (ii) the equality sign $=$ in (2.19) is to be replaced by the asymptotic equivalence sign $\sim$, and (iii) the interpretation of (1.7) through (1.11) should be recalled. Similarly, theorem 5.1 remains valid.

Acknowledgement

This research was supported by the Fund for the Promotion of Research at the Technion.

References

[1] S. Cabay and L.W. Jackson, A polynomial extrapolation method for finding limits and antilimits of vector sequences, SIAM J. Numer. Anal. 13 (1976) 734–752.

[2] R.P. Eddy, Extrapolating to the limit of a vector sequence, in: *Information Linkage between Applied Mathematics and Industry*, ed. P.C.C. Wang (Academic Press, New York, 1979) 387–396.

[3] W. Gander, G.H. Golub and D. Gruntz, Solving linear equations by extrapolation, Manuscript NA-89-11, Stanford University, Stanford, CA (October 1989).

[4] S. Kaniel and J. Stein, Least-square acceleration of iterative methods for linear equations, J. Optim. Theory Appl. 14 (1974) 431–437.

[5] M. Mesina, Convergence acceleration for the iterative solution of the equations $X = AX + f$, Comp. Meth. Appl. Mech. Eng. 10 (1977) 165–173.

[6] A. Sidi, Convergence and stability properties of minimal polynomial and reduced rank extrapolation algorithms, SIAM J. Numer. Anal. 23 (1986) 197–209.

[7] A. Sidi, Extrapolation vs. projection methods for linear systems of equations, J. Comp. Appl. Math. 22 (1988) 71–88.

[8] A. Sidi, Efficient implementation of minimal polynomial and reduced rank extrapolation methods, J. Comp. Appl. Math. 36 (1991) 305–337.

[9] A. Sidi and J. Bridger, Convergence and stability analyses for some vector extrapolation methods in the presence of defective iteration matrices, J. Comp. Appl. Math. 22 (1988) 35–61.

[10] D.A. Smith, W.F. Ford and A. Sidi, Extrapolation methods for vector sequences, SIAM Rev. 29 (1987) 199–233.