Numerical Algorithms (2020) 84:957–982
https://doi.org/10.1007/s11075-019-00788-6

ORIGINAL PAPER

A convergence study for reduced rank extrapolation on nonlinear systems

Avram Sidi[1]

Received: 31 January 2019 / Accepted: 24 July 2019 / Published online: 20 August 2019

Abstract
Reduced Rank Extrapolation (RRE) is a polynomial type method used to accelerate the convergence of sequences of vectors $\{\boldsymbol{x}_m\}$. It is applied successfully in different disciplines of science and engineering in the solution of large and sparse systems of linear and nonlinear equations of very large dimension. If $\boldsymbol{s}$ is the solution to the system of equations $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})$, first, a vector sequence $\{\boldsymbol{x}_m\}$ is generated via the fixed-point iterative scheme $\boldsymbol{x}_{m+1} = \boldsymbol{f}(\boldsymbol{x}_m)$, $m = 0, 1, \ldots$, and next, RRE is applied to this sequence to accelerate its convergence. RRE produces approximations $\boldsymbol{s}_{n,k}$ to $\boldsymbol{s}$ that are of the form $\boldsymbol{s}_{n,k} = \sum_{i=0}^{k} \gamma_i \boldsymbol{x}_{n+i}$ for some scalars γ_i depending (nonlinearly) on $\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_{n+k+1}$ and satisfying $\sum_{i=0}^{k} \gamma_i = 1$. The convergence properties of RRE when applied in conjunction with linear $\boldsymbol{f}(\boldsymbol{x})$ have been analyzed in different publications. In this work, we discuss the convergence of the $\boldsymbol{s}_{n,k}$ obtained from RRE with nonlinear $\boldsymbol{f}(\boldsymbol{x})$ (i) when $n \to \infty$ with fixed k, and (ii) in two so-called *cycling* modes.

Keywords Vector extrapolation methods · Minimal polynomial extrapolation (MPE) · Reduced rank extrapolation (RRE) · Krylov subspace methods · Nonlinear equations · Cycling mode

Mathematics Subject Classification (2010) Primary 65B05 · 65H10; Secondary 65F10

1 Introduction

Consider a system of nonlinear algebraic equations of dimension N, which we choose to write as

$$\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x}), \quad \boldsymbol{f} : \mathbb{C}^N \to \mathbb{C}^N; \quad \boldsymbol{s} \text{ solution}, \tag{1.1}$$

✉ Avram Sidi
asidi@cs.technion.ac.il
http://www.cs.technion.ac.il/∼asidi

[1] Computer Science Department, Technion - Israel Institute of Technology, Haifa 32000, Israel

where

$$\boldsymbol{x} = \left[x^{(1)}, \ldots, x^{(N)}\right]^T, \quad \boldsymbol{s} = \left[s^{(1)}, \ldots, s^{(N)}\right]^T; \quad x^{(i)},\ s^{(i)} \quad \text{scalars}, \tag{1.2}$$

and

$$\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), \ldots, f_N(\boldsymbol{x})]^T; \quad f_i(\boldsymbol{x}) = f_i\left(x^{(1)}, \ldots, x^{(N)}\right) \quad \text{scalar functions}. \tag{1.3}$$

One immediate way of solving this system is via the fixed-point iterative scheme

$$\boldsymbol{x}_{m+1} = \boldsymbol{f}(\boldsymbol{x}_m), \quad m = 0, 1, \ldots; \quad \text{for some } \boldsymbol{x}_0, \tag{1.4}$$

provided the sequence $\{\boldsymbol{x}_m\}$ converges. Let $\boldsymbol{f}(\boldsymbol{x})$ be twice continuously differentiable in a neighborhood of $\boldsymbol{s}$, and let $\boldsymbol{F}(\boldsymbol{x})$ be the Jacobian matrix of $\boldsymbol{f}$ evaluated at $\boldsymbol{x}$, that is,

$$\boldsymbol{F}(\boldsymbol{x}) = \begin{bmatrix} f_{1,1}(\boldsymbol{x}) & f_{1,2}(\boldsymbol{x}) & \cdots & f_{1,N}(\boldsymbol{x}) \\ f_{2,1}(\boldsymbol{x}) & f_{2,2}(\boldsymbol{x}) & \cdots & f_{2,N}(\boldsymbol{x}) \\ \vdots & \vdots & & \vdots \\ f_{N,1}(\boldsymbol{x}) & f_{N,2}(\boldsymbol{x}) & \cdots & f_{N,N}(\boldsymbol{x}) \end{bmatrix}; \quad f_{i,j}(\boldsymbol{x}) = \frac{\partial f_i}{\partial x^{(j)}}(\boldsymbol{x}). \tag{1.5}$$

It is known that (see Ortega and Rheinboldt [22], for example) if $\rho(\boldsymbol{F}(\boldsymbol{x}))$, the spectral radius of $\boldsymbol{F}(\boldsymbol{x})$, is such that $\rho(\boldsymbol{F}(\boldsymbol{s})) < 1$ and if $\boldsymbol{x}_0$ is sufficiently close to $\boldsymbol{s}$, then the sequence $\{\boldsymbol{x}_m\}$ converges to $\boldsymbol{s}$. The closer $\rho(\boldsymbol{F}(\boldsymbol{s}))$ is to one, the slower is the convergence of $\{\boldsymbol{x}_m\}$ to $\boldsymbol{s}$; this is the case in most practical engineering applications.

The convergence of $\{\boldsymbol{x}_m\}$ to $\boldsymbol{s}$ can be accelerated substantially by applying to it a vector extrapolation method. When applied to $\{\boldsymbol{x}_m\}$, an extrapolation method produces approximations $\boldsymbol{s}_{n,k}$ to $\boldsymbol{s}$ that are, either directly or indirectly, of the form

$$\boldsymbol{s}_{n,k} = \sum_{i=0}^{k} \gamma_i\, \boldsymbol{x}_{n+i}; \quad \gamma_i \text{ some scalars}, \quad \sum_{i=0}^{k} \gamma_i = 1, \tag{1.6}$$

the γ_i depending nonlinearly on the $\boldsymbol{x}_m$ used in constructing $\boldsymbol{s}_{n,k}$. Let M be the number of the $\boldsymbol{x}_m$ needed to construct $\boldsymbol{s}_{n,k}$. (Of course, M is not necessarily the same for all vector extrapolation methods.)[1]

For the sake of completeness, here we mention briefly those vector extrapolation methods that have been shown to be useful in applications.

1. **Polynomial type methods:** These are *minimal polynomial extrapolation (MPE)*, *reduced rank extrapolation (RRE)*, *modified minimal polynomial extrapolation*

[1] It is clear that the integers n and k are chosen by the user and that M is determined by n, k, and the extrapolation method being used.

(MMPE), and the most recent *singular value decomposition-based minimal polynomial extrapolation (SVD-MPE)*. MPE was introduced by Cabay and Jackson [9], RRE was introduced independently by Kaniel and Stein [18], Eddy [11], and Mešina [21].[2] MMPE was introduced independently by Brezinski [6], Pugachev [23], and Sidi, Ford, and Smith [35]. SVD-MPE is a new method by Sidi [31].

2. **Epsilon algorithms:** These are the *scalar epsilon algorithm (SEA)*, the *vector epsilon algorithm (VEA)*, and the *topological epsilon algorithm (TEA)*. SEA is a method that is based entirely on the famous *epsilon algorithm* of Wynn [45] that implements the transformation of Shanks [24] for scalar sequences. VEA was introduced by Wynn [46]. TEA was introduced by Brezinski [6].

For an earlier account of the epsilon algorithms, see the book by Brezinski [7]. For a comprehensive survey covering the developments that took place until the 1980s, see the survey paper by Smith, Ford, and Sidi [39] and the book by Brezinski and Redivo Zaglia [8]. For a geometric approach to the treatment of vector extrapolation methods as these are being applied to linear systems, see Jbilou and Sadok [16]. For a more recent review of MPE and RRE, see Sidi [30]. For a detailed and up-to-date treatment, including development, analysis, numerical implementation, and various applications, of all these methods, see the recent book of Sidi [33].

Numerically stable and efficient algorithms for implementing polynomial methods have been proposed by Sidi [27], [31] for MPE, RRE, and SVD-MPE and by Jbilou and Sadok [17] for MMPE. The epsilon algorithms are normally implemented via their definitions, which involve recursion relations. When applied to sequences $\{\boldsymbol{x}_m\}$ generated via fixed-point iterative schemes from systems of linear equations, MPE, RRE, and TEA turn out to be equivalent to known Krylov subspace methods for linear systems. This is explored in Sidi [26]. Yet another recent paper by Sidi [32] shows that MPE and RRE are very closely related in more than one way.

Now, all the methods mentioned above have interesting convergence and convergence acceleration properties that concern the precise asymptotic behavior of the sequences $\{\boldsymbol{s}_{n,k}\}_{n=0}^{\infty}$, with fixed k, when the sequences $\{\boldsymbol{x}_m\}$ are generated via fixed-point iterative schemes from systems of linear equations; see Sidi [25], [28], Sidi, Ford, and Smith [35], and Sidi and Bridger [34], and also Sidi [33, Chapter 6] for the methods MPE, RRE, MMPE, and TEA, Wynn [47] and Sidi [29] for SEA, and Graves-Morris and Saff [14] for VEA. We shall call this mode of usage of vector extrapolation methods the *n-Mode*.

Unfortunately, the n-Mode convergence theories that apply to the case in which $\boldsymbol{f}(\boldsymbol{x})$ is linear do not apply to the case in which $\boldsymbol{f}(\boldsymbol{x})$ is nonlinear. This is one of the topics we would like to study here, RRE being the extrapolation method used. That is, we would like to investigate the convergence properties of the sequences $\{\boldsymbol{s}_{n,k}\}_{n=0}^{\infty}$, with fixed k, obtained by applying RRE to $\{\boldsymbol{x}_m\}$ generated as in (1.4), where $\boldsymbol{f}(\boldsymbol{x})$ is nonlinear.

[2]The approaches of [18] and [21] to RRE are almost identical, in the sense that $\boldsymbol{s}_{n,k} = \sum_{i=0}^{k} \gamma_i \boldsymbol{x}_{n+i}$ in [21], while $\boldsymbol{s}_{n,k} = \sum_{i=0}^{k} \gamma_i \boldsymbol{x}_{n+i+1}$ in [18], the γ_i being the same for both. The approaches of [11] and [21] are completely different, however; their equivalence was proved in the review paper of Smith, Ford, and Sidi [39].

The numerical implementations of polynomial extrapolation methods and of epsilon algorithms, when generating the vectors $\boldsymbol{s}_{n,k}$, necessitate the keeping of resp. $k+2$ and $2k+1$ vectors in core memory simultaneously. In case we would like to increase k to improve the quality of the $\boldsymbol{s}_{n,k}$, this may pose a serious problem when we are dealing with very high dimensional vectors, which is the case in most large scale applications. Within the context described via (1.1)–(1.4) in the first paragraph of this section, it is best to apply vector extrapolation methods in the so-called *cycling* mode, and this has been the usual practice. This mode of usage of vector extrapolation methods, which we shall call the *C-Mode*, can be described via the following steps:

C-Mode

C0. Choose integers $n \geq 0$ and $k \geq 1$ and an initial vector $\boldsymbol{x}_0$.
C1. Compute the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M$ [via $\boldsymbol{x}_{m+1} = \boldsymbol{f}(\boldsymbol{x}_m)$].[3]
C2. Apply the extrapolation method to the vectors $\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_M$, and compute $\boldsymbol{s}_{n,k}$.
C3. If $\boldsymbol{s}_{n,k}$ satisfies the accuracy test, stop.
Otherwise, set $\boldsymbol{x}_0 = \boldsymbol{s}_{n,k}$ and go to step C1.

We call each application of steps C1–C3 a *cycle* and denote by $\boldsymbol{s}^{(r)}$ the $\boldsymbol{s}_{n,k}$ computed in the rth cycle. We will also denote the initial vector $\boldsymbol{x}_0$ in step C0 by $\boldsymbol{s}^{(0)}$. Under suitable conditions, it has been shown rigorously for MPE and RRE that the sequence $\{\boldsymbol{s}^{(r)}\}_{r=0}^{\infty}$ has very good convergence properties when $\boldsymbol{f}(\boldsymbol{x})$ is linear. See [36], [37]. See also [33, Chapter 7]. The case in which $\boldsymbol{f}(\boldsymbol{x})$ is nonlinear has proved to be complicated and has not been resolved till the present.

In some cases, RRE stalls if applied in the C-Mode with $n = 0$, in the sense that it takes too many iterations until one sees meaningful convergence; in such cases, even a moderate $n > 0$ can be very helpful to accelerate convergence effectively. See the numerical examples in [36], [37].

A different cycling procedure involving the minimal polynomial of the (constant) Jacobian matrix $\boldsymbol{F}(\boldsymbol{s})$ with respect to a nonzero vector[4] has been considered in various publications. The description of this procedure, which we shall call the *MC-Mode*, is as follows:

MC-Mode

MC0. Choose an integer $n \geq 0$ and an initial vector $\boldsymbol{x}_0$.

[3]Note that $M = n + k + 1$ for MPE, RRE, MMPE, and SVD-MPE, while $M = n + 2k$ for SEA, VEA, and TEA.

[4]Given a nonzero vector $\boldsymbol{u} \in \mathbb{C}^N$, the monic polynomial $P(\lambda)$ is said to be a *minimal polynomial of the matrix* $\boldsymbol{T} \in \mathbb{C}^{N\times N}$ *with respect to* $\boldsymbol{u}$ if $P(\boldsymbol{T})\boldsymbol{u} = \boldsymbol{0}$ and if $P(\lambda)$ has smallest degree.

The polynomial $P(\lambda)$ exists and is unique. Moreover, if $P_1(\boldsymbol{T})\boldsymbol{u} = \boldsymbol{0}$ for some polynomial $P_1(\lambda)$ with $\deg P_1 > \deg P$, then $P(\lambda)$ divides $P_1(\lambda)$. In particular, $P(\lambda)$ divides the minimal polynomial of $\boldsymbol{T}$, which in turn divides the characteristic polynomial of $\boldsymbol{T}$. [Thus, the degree of $P(\lambda)$ is at most N and its zeros are some or all of the eigenvalues of $\boldsymbol{T}$.]

MC1. Compute the vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M$ [via $\boldsymbol{x}_{m+1} = \boldsymbol{f}(\boldsymbol{x}_m)$], M being as explained in footnote[3], with k there being the degree of the minimal polynomial of $\boldsymbol{F}(\boldsymbol{s})$ with respect to $\boldsymbol{\epsilon}_n = \boldsymbol{x}_n - \boldsymbol{s}$.[5]

MC2. Apply the extrapolation method to the vectors $\boldsymbol{x}_n, \boldsymbol{x}_{n+1}, \ldots, \boldsymbol{x}_M$, and compute $\boldsymbol{s}_{n,k}$.

MC3. If $\boldsymbol{s}_{n,k}$ satisfies the accuracy test, stop.
Otherwise, set $\boldsymbol{x}_0 = \boldsymbol{s}_{n,k}$ and go to step MC1.

As before, we call each application of steps MC1–MC3 a *cycle* and denote by $\boldsymbol{s}^{(r)}$ the $\boldsymbol{s}_{n,k}$ computed in the rth cycle.[6] We will also denote the initial vector $\boldsymbol{x}_0$ in step MC0 by $\boldsymbol{s}^{(0)}$. It is observed in many numerical examples that the sequence $\{\boldsymbol{s}^{(r)}\}_{r=0}^{\infty}$ converges quadratically to the solution $\boldsymbol{s}$ of the system $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})$ when $\boldsymbol{f}(\boldsymbol{x})$ is nonlinear.[7] The first papers dealing with this topic (that is the MC-Mode with $\boldsymbol{s}_{0,k}$ only) are those by Brezinski [4], [5], Gekeler [12], and Skelboe [38]. Of these, [4], [5], and [12] consider the application of the epsilon algorithms, while [38] also considers the application of MPE and RRE. The quadratic convergence proofs in all of these papers have a gap in that they all end up with the relation

$$\|\boldsymbol{s}^{(r+1)} - \boldsymbol{s}\|_2 \leq K_r \|\boldsymbol{s}^{(r)} - \boldsymbol{s}\|_2^2,$$

from which they conclude that $\{\boldsymbol{s}^{(r)}\}_{r=0}^{\infty}$ converges quadratically. However, K_r is a scalar that depends on r through $\boldsymbol{s}^{(r)}$, and the proofs do not show how it depends on r. In particular, they do not show whether K_r is bounded in r or how it grows with r if it is not bounded. This gap was disclosed in the review paper of Smith, Ford, and Sidi [39].

A more recent paper by Jbilou and Sadok [15] deals with the same MC-Mode cycling via MPE and RRE. Yet another paper by Le Ferrand [20] treats TEA. Both these works provide proofs of quadratic convergence by imposing some global conditions on the whole sequence $\{\boldsymbol{s}^{(r)}\}_{r=0}^{\infty}$ as well as on $\boldsymbol{f}(\boldsymbol{x})$. (See also Laurens and Le Ferrand [19].)

In this work, we present a new convergence study of RRE when it is being applied to nonlinear systems. Specifically, we treat the convergence of RRE (i) in the n-Mode, and (ii) in the two cycling modes mentioned above. By making a *global* assumption, we are able to prove convergence in all cases. We can justify heuristically the plausibility of this assumption; we do not have a rigorous justification for it, however. This difficulty is inherent to all studies. We explore the source of this difficulty here. It must be mentioned that the difficulties that exist in the previous papers mentioned above are similar to ours, although they take different forms. Whether and how we can circumvent these difficulties is not clear at this time.

[5] It is clear that to apply any of the extrapolation methods in this mode, one needs to know the matrix $\boldsymbol{F}(\boldsymbol{s})$, for which one also needs to know the solution $\boldsymbol{s}$.

[6] Note that k is not necessarily fixed in this mode of cycling; it may vary from one cycle to the next. It always satisfies $k \leq N$, however.

[7] Quadratic convergence is relevant only when $\boldsymbol{f}(\boldsymbol{x})$ is nonlinear. When $\boldsymbol{f}(\boldsymbol{x})$ is linear, that is, $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{T}\boldsymbol{x} + \boldsymbol{d}$, where $\boldsymbol{T}$ is a fixed $N \times N$ matrix and $\boldsymbol{d}$ is a fixed vector, hence $\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{T}$, the solution $\boldsymbol{s}$ is obtained already at the end of step MC2 of the first cycle, that is, we have $\boldsymbol{s}^{(1)} = \boldsymbol{s}$. Therefore, there is nothing to analyze when $\boldsymbol{f}(\boldsymbol{x})$ is linear.

The plan of this paper is as follows: In Section 2, we give a brief description of RRE, which is needed throughout. In Section 3, we derive a formula for the error vector $\boldsymbol{s}_{n,k} - \boldsymbol{s}$ when the vectors $\boldsymbol{x}_m$ are generated via (1.4) with a nonlinear $\boldsymbol{f}(\boldsymbol{x})$. In Section 4, we use this error formula to derive an upper bound on $\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\|$. In Section 5, we complete the convergence studies of RRE in the different modes mentioned above. We mention that our results concerning the convergence of RRE in the n-Mode and the C-Mode are the first ones in the literature of extrapolation methods. In our study, we make much use of the results presented in Sidi [26] throughout these studies. In Section 6, we discuss the nature of the problem/difficulty mentioned above and compare our global assumption with that of [15]. In the appendix, we review some known theorems concerning Moore–Penrose generalized inverses of perturbed matrices, which we use in Section 4. (For generalized inverses, see Ben-Israel and Greville [3] and Campbell and Meyer [10], for example.)

Throughout this work, we will use lowercase boldface italic letters to denote vectors and we will use uppercase boldface italic letters to denote matrices.

Finally, we mention that in our study of RRE, we employ two different vector norms:

(i) The standard l_2 vector norm defined via $\|z\|_2 = \sqrt{z^*z}$ and the matrix norm induced by it, namely, $\|\boldsymbol{A}\|_2 = \sigma_{\max}(\boldsymbol{A})$, where $\sigma_{\max}(\boldsymbol{A})$ is the largest singular value of the matrix $\boldsymbol{A}$.

(ii) The $\boldsymbol{G}$ norm defined via $\|z\| = \|\boldsymbol{G}z\|_2$ and the matrix norm $\|\boldsymbol{A}\|$ induced by it, where $\boldsymbol{G} = \boldsymbol{I} - \boldsymbol{F}(\boldsymbol{s})$. Note that $\boldsymbol{G}$ is nonsingular since $\boldsymbol{F}(\boldsymbol{s})$ does not have unity as an eigenvalue; therefore, the $\boldsymbol{G}$ norm is a true vector norm.

Of course, the two vector norms are equivalent and we have

$$\frac{1}{\|\boldsymbol{G}^{-1}\|_2}\|z\|_2 \le \|z\| \le \|\boldsymbol{G}\|_2\,\|z\|_2. \tag{1.7}$$

When $\boldsymbol{A}$ is an $N \times N$ (square) matrix, we have $\|\boldsymbol{A}\| = \|\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^{-1}\|_2$. We make extensive use of these connections between the two norms, $\|\cdot\|$ and $\|\cdot\|_2$, in the sequel.

2 Description of RRE

Consider the system of equations given in (1.1)–(1.3), and let the sequence $\{\boldsymbol{x}_m\}$ be generated via the fixed-point iterative scheme in (1.4).

Define the first and second order differences of the $\boldsymbol{x}_m$ as in

$$\boldsymbol{u}_m = \boldsymbol{x}_{m+1} - \boldsymbol{x}_m, \quad \boldsymbol{w}_m = \boldsymbol{u}_{m+1} - \boldsymbol{u}_m = \boldsymbol{x}_{m+2} - 2\boldsymbol{x}_{m+1} + \boldsymbol{x}_m, \quad m = 0, 1, \ldots, \tag{2.1}$$

and, for some fixed $n \ge 0$, form the $N \times (j+1)$ matrices

$$\boldsymbol{U}_j = [\,\boldsymbol{u}_n \,|\, \boldsymbol{u}_{n+1} \,|\, \cdots \,|\, \boldsymbol{u}_{n+j}\,], \quad \boldsymbol{W}_j = [\,\boldsymbol{w}_n \,|\, \boldsymbol{w}_{n+1} \,|\, \cdots \,|\, \boldsymbol{w}_{n+j}\,], \quad j = 0, 1, \ldots. \tag{2.2}$$

Then the γ_i in (1.6) for RRE are the solution to the constrained standard l_2 minimization problem

$$\min_{\gamma_0,\gamma_1,\ldots,\gamma_k} \left\| \sum_{i=0}^{k} \gamma_i \boldsymbol{u}_{n+i} \right\|_2 \quad \text{subject to } \sum_{i=0}^{k} \gamma_i = 1, \tag{2.3}$$

which can also be expressed in matrix terms as

$$\min_{\boldsymbol{\gamma}} \|\boldsymbol{U}_k \boldsymbol{\gamma}\|_2 \quad \text{subject to } \sum_{i=0}^{k} \gamma_i = 1; \quad \boldsymbol{\gamma} = [\gamma_0, \gamma_1, \ldots, \gamma_k]^T \in \mathbb{C}^{k+1}. \tag{2.4}$$

Then, with the solution $\boldsymbol{\gamma}$ of this problem, the RRE approximation $\boldsymbol{s}_{n,k}$ is given as in

$$\boldsymbol{s}_{n,k} = \sum_{i=0}^{k} \gamma_i \, \boldsymbol{x}_{n+i}. \tag{2.5}$$

Noting that

$$\boldsymbol{x}_{n+m} = \boldsymbol{x}_n + \sum_{j=0}^{m-1} \boldsymbol{u}_{n+j}, \quad \boldsymbol{u}_{n+m} = \boldsymbol{u}_n + \sum_{j=0}^{m-1} \boldsymbol{w}_{n+j}, \quad m = 0, 1, \ldots,$$

we can reexpress $\boldsymbol{s}_{n,k}$ and $\boldsymbol{U}_k \boldsymbol{\gamma}$ as

$$\boldsymbol{s}_{n,k} = \boldsymbol{x}_n + \sum_{j=0}^{k-1} \xi_j \, \boldsymbol{u}_{n+j} = \boldsymbol{x}_n + \boldsymbol{U}_{k-1} \boldsymbol{\xi}, \quad \boldsymbol{U}_k \boldsymbol{\gamma} = \boldsymbol{u}_n + \sum_{j=0}^{k-1} \xi_j \boldsymbol{w}_{n+j} = \boldsymbol{u}_n + \boldsymbol{W}_{k-1} \boldsymbol{\xi}, \tag{2.6}$$

where

$$\boldsymbol{\xi} = [\xi_0, \xi_1, \ldots, \xi_{k-1}]^T \in \mathbb{C}^k; \quad \xi_j = \sum_{i=j+1}^{k} \gamma_i, \quad j = 0, 1, \ldots, k-1. \tag{2.7}$$

The (constrained) minimization problem for the vector $\boldsymbol{\gamma}$ in (2.4) can now be replaced by the following (unconstrained) minimization problem for the vector $\boldsymbol{\xi}$ in (2.6):

$$\min_{\boldsymbol{\xi}} \|\boldsymbol{u}_n + \boldsymbol{W}_{k-1} \boldsymbol{\xi}\|_2, \quad \boldsymbol{\xi} = [\xi_0, \xi_1, \ldots, \xi_{k-1}]^T \in \mathbb{C}^k. \tag{2.8}$$

Now, the solution to this problem (for $\boldsymbol{\xi}$) is simply $-\boldsymbol{W}_{k-1}^{+} \boldsymbol{u}_n$, where $\boldsymbol{K}^{+}$ stands for the Moore–Penrose generalized inverse of the matrix $\boldsymbol{K}$. Upon substituting this into (2.6), we obtain

$$\boxed{\boldsymbol{s}_{n,k} = \boldsymbol{x}_n - \boldsymbol{U}_{k-1} \boldsymbol{W}_{k-1}^{+} \boldsymbol{u}_n.} \tag{2.9}$$

We will be making use of this representation of $\boldsymbol{s}_{n,k}$ in the sequel. For the above developments, see Sidi [25].

3 An error formula for RRE

3.1 RRE on the linear system $x = s + F(s)(x - s)$

Let us now consider the linear system

$$\boldsymbol{x} = \tilde{\boldsymbol{f}}(\boldsymbol{x}), \quad \tilde{\boldsymbol{f}}(\boldsymbol{x}) = \boldsymbol{s} + \boldsymbol{F}(\boldsymbol{s})(\boldsymbol{x} - \boldsymbol{s}), \tag{3.1}$$

where $\boldsymbol{F}(\boldsymbol{s})$ is the Jacobian matrix of $\boldsymbol{f}$ evaluated at $\boldsymbol{s}$, as given in (1.5). Note that $\tilde{\boldsymbol{f}}(\boldsymbol{x})$ is simply the linear part of the Taylor series of $\boldsymbol{f}(\boldsymbol{x})$ in (1.1) about $\boldsymbol{s}$. Clearly, $\boldsymbol{s}$ is the solution to (3.1) since $\tilde{\boldsymbol{f}}(\boldsymbol{s}) = \boldsymbol{s}$.

With the vectors $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ generated nonlinearly as in (1.4) of the preceding section, let

$$\tilde{\boldsymbol{x}}_n = \boldsymbol{x}_n \quad \text{and} \quad \tilde{\boldsymbol{x}}_{m+1} = \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}}_m), \quad m = n, n+1, \ldots. \tag{3.2}$$

Following this, define

$$\tilde{\boldsymbol{\epsilon}}_m = \tilde{\boldsymbol{x}}_m - \boldsymbol{s}, \quad \tilde{\boldsymbol{u}}_m = \tilde{\boldsymbol{x}}_{m+1} - \tilde{\boldsymbol{x}}_m, \quad \tilde{\boldsymbol{w}}_m = \tilde{\boldsymbol{u}}_{m+1} - \tilde{\boldsymbol{u}}_m, \quad m = n, n+1, \ldots, \tag{3.3}$$

$$\tilde{\boldsymbol{U}}_j = [\, \tilde{\boldsymbol{u}}_n \,|\, \tilde{\boldsymbol{u}}_{n+1} \,|\, \cdots \,|\, \tilde{\boldsymbol{u}}_{n+j} \,], \quad \tilde{\boldsymbol{W}}_j = [\, \tilde{\boldsymbol{w}}_n \,|\, \tilde{\boldsymbol{w}}_{n+1} \,|\, \cdots \,|\, \tilde{\boldsymbol{w}}_{n+j} \,], \quad j = 0, 1, \ldots. \tag{3.4}$$

Then, by (2.9), the vector $\tilde{\boldsymbol{s}}_{n,k}$ produced by applying RRE to the sequence $\{\tilde{\boldsymbol{x}}_m\}$ is

$$\boxed{\tilde{\boldsymbol{s}}_{n,k} = \tilde{\boldsymbol{x}}_n - \tilde{\boldsymbol{U}}_{k-1} \tilde{\boldsymbol{W}}_{k-1}^{+} \tilde{\boldsymbol{u}}_n.} \tag{3.5}$$

Upon subtracting $\boldsymbol{s}$ from both sides of this equality and invoking $\tilde{\boldsymbol{\epsilon}}_n = \tilde{\boldsymbol{x}}_n - \boldsymbol{s}$, we obtain the error formula

$$\boxed{\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s} = \tilde{\boldsymbol{\epsilon}}_n - \tilde{\boldsymbol{U}}_{k-1} \tilde{\boldsymbol{W}}_{k-1}^{+} \tilde{\boldsymbol{u}}_n.} \tag{3.6}$$

The error $\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}$ has been studied in detail in [25], [28], [34], [36]; for a summary, see [33, Chapters 6,7].[8]

The following result from [26, Theorem 4.2] concerning the application of RRE to vector sequences from fixed-point iteration of *linear* systems will be crucial in our analysis of RRE concerning *nonlinear* systems in Section 5.

Theorem 3.1 *Denote $\tilde{\boldsymbol{F}} = \boldsymbol{F}(\boldsymbol{s})$ for short; thus $\boldsymbol{G} = \boldsymbol{I} - \tilde{\boldsymbol{F}}$. Then the vector $\tilde{\boldsymbol{s}}_{n,k}$ is the solution to the optimization problem*

$$\|\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}\| = \|\boldsymbol{G}(\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s})\|_2 = \min_{g \in \tilde{\mathcal{P}}_k} \|g(\tilde{\boldsymbol{F}})\boldsymbol{G}(\tilde{\boldsymbol{x}}_n - \boldsymbol{s})\|_2,$$

$$\tilde{\mathcal{P}}_k = \left\{ g(z) = \sum_{j=0}^{k} \alpha_j z^j : \ g(1) = 1 \right\}. \tag{3.7}$$

[8] See also Sidi and Shapira [37] concerning a modified version of restarted GMRES with prior Richardson iterations, that is very closely related to RRE.

Remarks

1. If k is the degree of the minimal polynomial of $\boldsymbol{F}(\boldsymbol{s})$ with respect to the vector $\tilde{\boldsymbol{\epsilon}}_n$, then $\tilde{\boldsymbol{s}}_{n,k} = \boldsymbol{s}$, the solution to (3.1). See footnote[7].
2. Concerning Theorem 3.1, note that the vector $\boldsymbol{G}(\boldsymbol{y} - \boldsymbol{s})$ is simply the *residual* of the vector $\boldsymbol{y}$ for the linear system $\boldsymbol{x} = \tilde{\boldsymbol{f}}(\boldsymbol{x})$ because

$$\boldsymbol{G}(\boldsymbol{y} - \boldsymbol{s}) = \boldsymbol{y} - \tilde{\boldsymbol{f}}(\boldsymbol{y}),$$

and also

$$\boldsymbol{y} - \tilde{\boldsymbol{f}}(\boldsymbol{y}) = \boldsymbol{0} \quad \Leftrightarrow \quad \boldsymbol{y} = \boldsymbol{s}, \quad \text{since } \boldsymbol{G} \text{ is nonsingular}.$$

Thus, what Theorem 3.1 means is that $\|\boldsymbol{G}(\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s})\|_2$, the l_2 norm of the residual vector of $\tilde{\boldsymbol{s}}_{n,k} = \sum_{i=0}^{k} \tilde{\gamma}_i \tilde{\boldsymbol{x}}_{n+i}$ subject to $\sum_{i=0}^{k} \tilde{\gamma}_i = 1$, is the smallest of all the l_2 norms of the residuals of the vectors $\sum_{i=0}^{k} \alpha_i \tilde{\boldsymbol{x}}_{n+i}$ subject to $\sum_{i=0}^{k} \alpha_i = 1$. Here we also recall that $\tilde{\boldsymbol{x}}_n = \boldsymbol{x}_n$ by (3.2).

3.2 RRE on the nonlinear system $x = f(x)$

In the Introduction, we assumed that $\boldsymbol{f}(\boldsymbol{x})$ is twice continuously differentiable in a neighborhood of the solution $\boldsymbol{s}$. We also assumed that $\rho(\tilde{\boldsymbol{F}}) < 1$, where we recall $\tilde{\boldsymbol{F}} = \boldsymbol{F}(\boldsymbol{s})$, thus ensuring the convergence of the sequence $\{\boldsymbol{x}_m\}$ to $\boldsymbol{s}$. We now assume, in addition, that $\|\tilde{\boldsymbol{F}}\|_2 < 1$ too and define the ball $B(\boldsymbol{s}, \delta)$ containing $\boldsymbol{s}$ in its interior via

$$B(\boldsymbol{s}, \delta) = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{s}\| \equiv \|\boldsymbol{G}(\boldsymbol{x} - \boldsymbol{s})\|_2 \le \delta\}. \tag{3.8}$$

Clearly, $B(\boldsymbol{s}, \delta)$ is a convex set. In addition, we assume $\boldsymbol{f}(\boldsymbol{x})$ is twice continuously differentiable in $B(\boldsymbol{s}, \delta)$.

Lemma 3.2 *For all δ sufficiently small, there exists a positive constant $L < 1$ independent of δ, such that*

$$\|\boldsymbol{x}_{m+1} - \boldsymbol{s}\| \le L\|\boldsymbol{x}_m - \boldsymbol{s}\|, \quad m = 0, 1, \ldots, \quad \textit{provided } \boldsymbol{x}_0 \in B(\boldsymbol{s}, \delta). \tag{3.9}$$

Consequently, the whole sequence $\{\boldsymbol{x}_m\}$ is in $B(\boldsymbol{s}, \delta)$ and converges to $\boldsymbol{s}$.

Proof We begin with the following result that follows from Ortega and Rheinboldt [22, p. 69]:

$$\boldsymbol{f}(x) - \boldsymbol{f}(\boldsymbol{s}) = \int_0^1 \boldsymbol{F}(\boldsymbol{s} + t(\boldsymbol{x} - \boldsymbol{s}))(\boldsymbol{x} - \boldsymbol{s})\, dt \quad \text{provided } \boldsymbol{x} \in B(\boldsymbol{s}, \delta).$$

It is important to note that $\boldsymbol{s} + t(\boldsymbol{x} - \boldsymbol{s})$, with $t \in [0, 1]$, is a convex combination of $\boldsymbol{x}$ and $\boldsymbol{s}$ hence is also in $B(\boldsymbol{s}, \delta)$. Multiplying both sides of this equality on the left by $\boldsymbol{G}$, we obtain

$$\boldsymbol{G}[\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{s})] = \int_0^1 [\boldsymbol{G}\boldsymbol{F}(\boldsymbol{s} + t(\boldsymbol{x} - \boldsymbol{s}))\boldsymbol{G}^{-1}]\,[\boldsymbol{G}(\boldsymbol{x} - \boldsymbol{s})]\, dt,$$

which, upon taking l_2 norms on both sides and invoking the known fact that

$$\left\| \int_a^b \boldsymbol{u}(\xi)\, d\xi \right\|_2 \leq \int_a^b \|\boldsymbol{u}(\xi)\|_2\, d\xi, \quad \boldsymbol{u}(\xi) \in \mathbb{C}^N,$$

gives

$$\|\boldsymbol{G}[\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{s})]\|_2 \leq \int_0^1 \|\boldsymbol{G}\boldsymbol{F}(\boldsymbol{s} + t(\boldsymbol{x} - \boldsymbol{s}))\boldsymbol{G}^{-1}\|_2\, \|\boldsymbol{G}(\boldsymbol{x} - \boldsymbol{s})\|_2\, dt.$$

Finally, invoking in this last inequality $\|\boldsymbol{G}\boldsymbol{z}\|_2 = \|\boldsymbol{z}\|$ and the fact that $\|\boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^{-1}\|_2 = \|\boldsymbol{A}\|$, we obtain

$$\begin{aligned} \|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{s})\| &\leq \left(\int_0^1 \|\boldsymbol{F}(\boldsymbol{s} + t(\boldsymbol{x} - \boldsymbol{s}))\| dt \right) \|\boldsymbol{x} - \boldsymbol{s}\| \\ &\leq \left[\max_{0 \leq t \leq 1} \|\boldsymbol{F}(\boldsymbol{s} + t(\boldsymbol{x} - \boldsymbol{s}))\| \right] \|\boldsymbol{x} - \boldsymbol{s}\| \\ &\leq \left[\max_{z \in B(\boldsymbol{s}, \delta)} \|\boldsymbol{F}(z)\| \right] \|\boldsymbol{x} - \boldsymbol{s}\|. \end{aligned} \tag{3.10}$$

Now, by the fact that $\boldsymbol{f}(\boldsymbol{x})$ is twice differentiable in $B(\boldsymbol{s}, \delta)$, it follows that

$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{s} + (\boldsymbol{x} - \boldsymbol{s})) = \boldsymbol{F}(\boldsymbol{s}) + \boldsymbol{\Delta}(\boldsymbol{x} - \boldsymbol{s}), \tag{3.11}$$

where the matrix $\boldsymbol{\Delta}(\boldsymbol{x} - \boldsymbol{s})$ satisfies

$$\|\boldsymbol{\Delta}(\boldsymbol{x} - \boldsymbol{s})\| \leq \alpha \|\boldsymbol{x} - \boldsymbol{s}\| \quad \text{for some } \alpha > 0 \text{ independent of } \boldsymbol{x} \in B(\boldsymbol{s}, \delta). \tag{3.12}$$

Taking norms on both sides of (3.11), realizing that $\|\boldsymbol{F}(\boldsymbol{s})\| = \|\boldsymbol{F}(\boldsymbol{s})\|_2$ because $\boldsymbol{F}(\boldsymbol{s})$ and $\boldsymbol{G} = \boldsymbol{I} - \boldsymbol{F}(\boldsymbol{s})$ commute, and invoking $\boldsymbol{x} \in B(\boldsymbol{s}, \delta)$, we have

$$\|\boldsymbol{F}(\boldsymbol{x})\| \leq \|\boldsymbol{F}(\boldsymbol{s})\|_2 + \alpha \|\boldsymbol{x} - \boldsymbol{s}\| \leq \|\boldsymbol{F}(\boldsymbol{s})\|_2 + \alpha\delta \quad \forall\, \boldsymbol{x} \in B(\boldsymbol{s}, \delta). \tag{3.13}$$

Since we have assumed that $\|\boldsymbol{F}(\boldsymbol{s})\|_2 < 1$, we can choose δ sufficiently small to cause

$$\max_{\boldsymbol{x} \in B(\boldsymbol{s}, \delta)} \|\boldsymbol{F}(\boldsymbol{x})\| = L < 1. \tag{3.14}$$

With this, (3.10) becomes

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{s})\| \leq L \|\boldsymbol{x} - \boldsymbol{s}\| \quad \forall\, \boldsymbol{x} \in B(\boldsymbol{s}, \delta). \tag{3.15}$$

The proof of (3.9) for the sequence $\{\boldsymbol{x}_m\}$ can now be carried out by letting $\boldsymbol{x} = \boldsymbol{x}_m$ in (3.15), recalling that $\boldsymbol{f}(\boldsymbol{x}_m) = \boldsymbol{x}_{m+1}$ and $\boldsymbol{f}(\boldsymbol{s}) = \boldsymbol{s}$, and then proceeding by induction on m. □

In the sequel, we adopt the shorthand notation

$$\boldsymbol{\epsilon}_m = \boldsymbol{x}_m - \boldsymbol{s}, \quad m = 0, 1, \ldots; \quad \tilde{\boldsymbol{F}} = \boldsymbol{F}(\boldsymbol{s}). \tag{3.16}$$

We also make use of the fact that $\|\tilde{\boldsymbol{F}}\| \leq L < 1$, which follows from (3.14), and, along with (3.9), guarantees that the sequence $\{\|\boldsymbol{\epsilon}_m\|\}$ decreases monotonically and converges to zero.

Expanding $\boldsymbol{f}(\boldsymbol{x})$ in a Taylor series about the solution $\boldsymbol{s}$ and using the fact that $\boldsymbol{f}(\boldsymbol{s}) = \boldsymbol{s}$ and $\boldsymbol{f} \in C^2(B(\boldsymbol{s}, \delta))$, we have

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{s} + \tilde{\boldsymbol{F}} \cdot (\boldsymbol{x} - \boldsymbol{s}) + \boldsymbol{\mu}(\boldsymbol{x} - \boldsymbol{s}), \tag{3.17}$$

where

$$\|\boldsymbol{\mu}(\boldsymbol{x} - \boldsymbol{s})\| \le a \, \|\boldsymbol{x} - \boldsymbol{s}\|^2 \quad \forall \, \boldsymbol{x} \in B(\boldsymbol{s}, \delta), \qquad \text{for some } a > 0. \tag{3.18}$$

Consequently,

$$\boldsymbol{x}_{m+1} = \boldsymbol{f}(\boldsymbol{x}_m) = \boldsymbol{s} + \tilde{\boldsymbol{F}}\boldsymbol{\epsilon}_m + \boldsymbol{\mu}(\boldsymbol{\epsilon}_m) \quad \Rightarrow \quad \boldsymbol{\epsilon}_{m+1} = \tilde{\boldsymbol{F}}\boldsymbol{\epsilon}_m + \boldsymbol{\mu}(\boldsymbol{\epsilon}_m). \tag{3.19}$$

Then, by induction,

$$\boldsymbol{\epsilon}_{n+i} = \tilde{\boldsymbol{F}}^i \boldsymbol{\epsilon}_n + \sum_{j=0}^{i-1} \tilde{\boldsymbol{F}}^{i-j-1} \boldsymbol{\mu}(\boldsymbol{\epsilon}_{n+j}), \quad i = 0, 1, 2, \ldots. \tag{3.20}$$

Lemma 3.3 *The vectors $\boldsymbol{\epsilon}_m$, $\boldsymbol{u}_m$, and $\boldsymbol{w}_m$ satisfy*

$$\boldsymbol{\epsilon}_{n+i} = \tilde{\boldsymbol{F}}^i \boldsymbol{\epsilon}_n + \check{\boldsymbol{\epsilon}}_{n+i}; \quad \|\check{\boldsymbol{\epsilon}}_{n+i}\| \le C_i \|\boldsymbol{\epsilon}_n\|^2, \quad C_i = a \frac{1 - L^i}{1 - L} L^{i-1}, \tag{3.21}$$

$$\boldsymbol{u}_{n+i} = (\tilde{\boldsymbol{F}} - \boldsymbol{I})\tilde{\boldsymbol{F}}^i \boldsymbol{\epsilon}_n + \check{\boldsymbol{u}}_{n+i}; \quad \|\check{\boldsymbol{u}}_{n+i}\| \le D_i \|\boldsymbol{\epsilon}_n\|^2, \quad D_i = C_i + C_{i+1}, \tag{3.22}$$

$$\boldsymbol{w}_{n+i} = (\tilde{\boldsymbol{F}} - \boldsymbol{I})^2 \tilde{\boldsymbol{F}}^i \boldsymbol{\epsilon}_n + \check{\boldsymbol{w}}_{n+i}; \quad \|\check{\boldsymbol{w}}_{n+i}\| \le E_i \|\boldsymbol{\epsilon}_n\|^2, \quad E_i = C_i + 2C_{i+1} + C_{i+2}. \tag{3.23}$$

Remark Note that $C_0 = 0$ and $C_1 = a$ by (3.21). Therefore, $D_0 = a$ by (3.22).

Proof We start by noting that, by (3.20),

$$\check{\boldsymbol{\epsilon}}_{n+i} = \sum_{j=0}^{i-1} \tilde{\boldsymbol{F}}^{i-j-1} \boldsymbol{\mu}(\boldsymbol{\epsilon}_{n+j}),$$

which, upon taking norms and invoking $\|\tilde{\boldsymbol{F}}\| \le L$ and (3.18), gives

$$\begin{aligned} \|\check{\boldsymbol{\epsilon}}_{n+i}\| &\le \sum_{j=0}^{i-1} \|\tilde{\boldsymbol{F}}^{i-j-1}\| \, \|\boldsymbol{\mu}(\boldsymbol{\epsilon}_{n+j})\| \le \sum_{j=0}^{i-1} L^{i-j-1} \, a \, (L^j \|\boldsymbol{\epsilon}_n\|)^2 \\ &= a \left(\sum_{j=0}^{i-1} L^{i+j-1} \right) \|\boldsymbol{\epsilon}_n\|^2, \end{aligned}$$

from which (3.21) follows.

The proofs of (3.22)–(3.23) follow from (3.21) and the observation that

$$\check{\boldsymbol{u}}_m = \check{\boldsymbol{\epsilon}}_{m+1} - \check{\boldsymbol{\epsilon}}_m \quad \text{and} \quad \check{\boldsymbol{w}}_m = \check{\boldsymbol{\epsilon}}_{m+2} - 2\check{\boldsymbol{\epsilon}}_{m+1} + \check{\boldsymbol{\epsilon}}_m.$$

We leave the details to the reader. □

Let us now go back to the linear system $\boldsymbol{x} = \tilde{\boldsymbol{f}}(\boldsymbol{x})$ in (3.1), recalling that $\boldsymbol{F}(\boldsymbol{s}) = \tilde{\boldsymbol{F}}$. As already explained, $\tilde{\boldsymbol{f}}(\boldsymbol{x})$ is simply the linear part of the Taylor series of $\boldsymbol{f}(\boldsymbol{x})$

about $\boldsymbol{s}$, obtained from (3.17) by letting $\boldsymbol{\mu}(\boldsymbol{y}) \equiv \mathbf{0}$ there. In addition, $\tilde{\boldsymbol{f}}(\boldsymbol{s}) = \boldsymbol{s}$, that is, $\boldsymbol{s}$ is the solution to $\boldsymbol{x} = \tilde{\boldsymbol{f}}(\boldsymbol{x})$, as well as $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{x})$. Let us now note that $\boldsymbol{\mu}(\boldsymbol{y}) \equiv \mathbf{0}$ also implies that $\check{\boldsymbol{\epsilon}}_m = \mathbf{0}$, $\check{\boldsymbol{u}}_m = \mathbf{0}$, and $\check{\boldsymbol{w}}_m = \mathbf{0}$ in (3.21), (3.22), and (3.23), respectively. Recalling also that $\tilde{\boldsymbol{\epsilon}}_n = \boldsymbol{\epsilon}_n$, we finally realize that, for $i = 0, 1, \ldots,$

$$\begin{aligned} \tilde{\boldsymbol{\epsilon}}_{n+i} &= \tilde{\boldsymbol{F}}^i \tilde{\boldsymbol{\epsilon}}_n = \tilde{\boldsymbol{F}}^i \boldsymbol{\epsilon}_n, \quad \tilde{\boldsymbol{u}}_{n+i} = \tilde{\boldsymbol{F}}^i \tilde{\boldsymbol{u}}_n = (\tilde{\boldsymbol{F}} - \boldsymbol{I})\tilde{\boldsymbol{F}}^i \boldsymbol{\epsilon}_n, \\ \tilde{\boldsymbol{w}}_{n+i} &= \tilde{\boldsymbol{F}}^i \tilde{\boldsymbol{w}}_n = (\tilde{\boldsymbol{F}} - \boldsymbol{I})^2 \tilde{\boldsymbol{F}}^i \boldsymbol{\epsilon}_n; \end{aligned} \tag{3.24}$$

consequently,

$$\boldsymbol{u}_{n+i} = \tilde{\boldsymbol{u}}_{n+i} + \check{\boldsymbol{u}}_{n+i}, \quad \boldsymbol{w}_{n+i} = \tilde{\boldsymbol{w}}_{n+i} + \check{\boldsymbol{w}}_{n+i}. \tag{3.25}$$

As a result of all this, we have

$$\boldsymbol{U}_{k-1} = \tilde{\boldsymbol{U}}_{k-1} + \check{\boldsymbol{U}}_{k-1}, \quad \check{\boldsymbol{U}}_{k-1} = [\, \check{\boldsymbol{u}}_n \,|\, \check{\boldsymbol{u}}_{n+1} \,|\, \cdots \,|\, \check{\boldsymbol{u}}_{n+k-1} \,] \tag{3.26}$$

and

$$\boldsymbol{W}_{k-1} = \tilde{\boldsymbol{W}}_{k-1} + \check{\boldsymbol{W}}_{k-1}, \quad \check{\boldsymbol{W}}_{k-1} = [\, \check{\boldsymbol{w}}_n \,|\, \check{\boldsymbol{w}}_{n+1} \,|\, \cdots \,|\, \check{\boldsymbol{w}}_{n+k-1} \,], \tag{3.27}$$

with $\boldsymbol{U}_j$ and $\boldsymbol{W}_j$ as in (2.2). For simplicity of notation, in what follows, we drop the subscript $k-1$ from the matrices $\boldsymbol{U}_{k-1}$, $\boldsymbol{W}_{k-1}$, $\tilde{\boldsymbol{U}}_{k-1}$, $\tilde{\boldsymbol{W}}_{k-1}$, etc. With these, (2.9) becomes

$$\begin{aligned} \boldsymbol{s}_{n,k} &= \boldsymbol{x}_n - \boldsymbol{U}\boldsymbol{W}^+ \boldsymbol{u}_n \\ &= \boldsymbol{x}_n - (\tilde{\boldsymbol{U}} + \check{\boldsymbol{U}})(\tilde{\boldsymbol{W}} + \check{\boldsymbol{W}})^+ (\tilde{\boldsymbol{u}}_n + \check{\boldsymbol{u}}_n). \end{aligned} \tag{3.28}$$

Letting also

$$\boldsymbol{H} = \boldsymbol{W}^+ - \tilde{\boldsymbol{W}}^+ = (\tilde{\boldsymbol{W}} + \check{\boldsymbol{W}})^+ - \tilde{\boldsymbol{W}}^+, \tag{3.29}$$

we rewrite (3.28) in the form

$$\boldsymbol{s}_{n,k} = \boldsymbol{x}_n - (\tilde{\boldsymbol{U}} + \check{\boldsymbol{U}})(\tilde{\boldsymbol{W}}^+ + \boldsymbol{H})(\tilde{\boldsymbol{u}}_n + \check{\boldsymbol{u}}_n). \tag{3.30}$$

Next, opening the parentheses in (3.30), we obtain the equality

$$\boldsymbol{s}_{n,k} = \boldsymbol{x}_n - \tilde{\boldsymbol{U}}\tilde{\boldsymbol{W}}^+ \tilde{\boldsymbol{u}}_n - \tilde{\boldsymbol{U}}\tilde{\boldsymbol{W}}^+ \check{\boldsymbol{u}}_n - (\tilde{\boldsymbol{U}}\boldsymbol{H} + \check{\boldsymbol{U}}\tilde{\boldsymbol{W}}^+ + \check{\boldsymbol{U}}\boldsymbol{H})(\tilde{\boldsymbol{u}}_n + \check{\boldsymbol{u}}_n). \tag{3.31}$$

Now, by the fact that $\boldsymbol{x}_n = \tilde{\boldsymbol{x}}_n$ and by (3.5), we have that $\boldsymbol{x}_n - \tilde{\boldsymbol{U}}\tilde{\boldsymbol{W}}^+ \tilde{\boldsymbol{u}}_n = \tilde{\boldsymbol{s}}_{n,k}$ in this equality. Next, we invoke $\boldsymbol{u}_n = \tilde{\boldsymbol{u}}_n + \check{\boldsymbol{u}}_n$ and $\boldsymbol{U} = \tilde{\boldsymbol{U}} + \check{\boldsymbol{U}}$ again, and obtain a convenient representation of $\boldsymbol{s}_{n,k}$ and the error in it. We summarize all this in the following lemma.

Lemma 3.4 *Let*

$$\check{\boldsymbol{s}}_{n,k} = -\tilde{\boldsymbol{U}}\tilde{\boldsymbol{W}}^+ \check{\boldsymbol{u}}_n - (\boldsymbol{U}\boldsymbol{H} + \check{\boldsymbol{U}}\tilde{\boldsymbol{W}}^+)\boldsymbol{u}_n. \tag{3.32}$$

Then, $\boldsymbol{s}_{n,k}$ is given by the equality

$$\boldsymbol{s}_{n,k} = \tilde{\boldsymbol{s}}_{n,k} + \check{\boldsymbol{s}}_{n,k}. \tag{3.33}$$

Subtracting $\boldsymbol{s}$ from both sides of this equality, we also obtain the error formula

$$\boldsymbol{s}_{n,k} - \boldsymbol{s} = (\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}) + \check{\boldsymbol{s}}_{n,k}. \tag{3.34}$$

4 Derivation of upper bounds for $\|s_{n,k} - s\|$

4.1 Preliminaries

We now turn to the study of $\boldsymbol{s}_{n,k} - \boldsymbol{s}$. Multiplying both sides of (3.34) on the left by $\boldsymbol{G}$ and taking l_2 norms, and also invoking $\|\boldsymbol{z}\|_2 \leq \|\boldsymbol{G}^{-1}\|_2 \|\boldsymbol{z}\|$, we obtain

$$\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\| \leq \|\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}\| + \|\check{\boldsymbol{s}}_{n,k}\|,$$
$$\frac{\|\check{\boldsymbol{s}}_{n,k}\|}{\|\boldsymbol{G}^{-1}\|_2} \leq \|\boldsymbol{G}\tilde{\boldsymbol{U}}\|_2 \|\tilde{\boldsymbol{W}}^{+}\|_2 \|\check{\boldsymbol{u}}_n\| + \|\boldsymbol{G}\boldsymbol{U}\|_2 \|\boldsymbol{H}\|_2 \|\boldsymbol{u}_n\| + \|\boldsymbol{G}\check{\boldsymbol{U}}\|_2 \|\tilde{\boldsymbol{W}}^{+}\|_2 \|\boldsymbol{u}_n\|. \quad (4.1)$$

Thus, we need to study the behavior of each one of the terms in this bound. We begin with the following lemma.

Lemma 4.1 *The following are true:*

$$\|\boldsymbol{G}\boldsymbol{U}\|_2 \leq K_1\|\boldsymbol{\epsilon}_n\|, \quad \|\boldsymbol{G}\tilde{\boldsymbol{U}}\|_2 \leq K_2\|\boldsymbol{\epsilon}_n\|, \quad \|\boldsymbol{G}\check{\boldsymbol{U}}\|_2 \leq K_3\|\boldsymbol{\epsilon}_n\|^2, \quad (4.2)$$
$$\|\boldsymbol{W}\|_2 \leq K_1'\|\boldsymbol{\epsilon}_n\|, \quad \|\tilde{\boldsymbol{W}}\|_2 \leq K_2'\|\boldsymbol{\epsilon}_n\|, \quad \|\check{\boldsymbol{W}}\|_2 \leq K_3'\|\boldsymbol{\epsilon}_n\|^2, \quad (4.3)$$

with K_i, K_i', $i = 1, 2, 3$, positive constants independent of k and n.

Proof To achieve the proof, we make use of (3.21)–(3.24) and

$$\begin{aligned} \|\boldsymbol{u}_m\| &\leq (1+L)\|\boldsymbol{\epsilon}_m\| \quad \text{and} \quad \|\boldsymbol{w}_m\| \leq (1+L)^2\|\boldsymbol{\epsilon}_m\|, \\ \|\tilde{\boldsymbol{u}}_m\| &\leq (1+L)\|\boldsymbol{\epsilon}_m\| \quad \text{and} \quad \|\tilde{\boldsymbol{w}}_m\| \leq (1+L)^2\|\boldsymbol{\epsilon}_m\|. \end{aligned} \quad (4.4)$$

We prove the validity of the bound on $\|\boldsymbol{G}\boldsymbol{U}\|_2$ only; the others can be proved in exactly the same way.

We start by analyzing $\|\boldsymbol{G}\boldsymbol{U}\|_F$, the Frobenius norm of $\boldsymbol{G}\boldsymbol{U}$. Noting that

$$\boldsymbol{G}\boldsymbol{U} = [\,\boldsymbol{G}\boldsymbol{u}_n \mid \boldsymbol{G}\boldsymbol{u}_{n+1} \mid \cdots \mid \boldsymbol{G}\boldsymbol{u}_{n+k-1}\,],$$

we have

$$\begin{aligned} \|\boldsymbol{G}\boldsymbol{U}\|_F^2 = \sum_{j=0}^{k-1} \|\boldsymbol{G}\boldsymbol{u}_{n+j}\|_2^2 = \sum_{j=0}^{k-1} \|\boldsymbol{u}_{n+j}\|^2 &\leq \sum_{j=0}^{k-1} [(1+L)\|\boldsymbol{\epsilon}_{n+j}\|]^2 \quad \text{by (4.4)} \\ &\leq (1+L)^2 \sum_{j=0}^{k-1} (L^j\|\boldsymbol{\epsilon}_n\|)^2 \quad \text{by (3.9)} \\ &= \frac{1+L}{1-L}(1-L^{2k})\|\boldsymbol{\epsilon}_n\|^2 \\ &< \frac{1+L}{1-L}\|\boldsymbol{\epsilon}_n\|^2. \end{aligned}$$

The result $\|\boldsymbol{G}\boldsymbol{U}\|_2 \leq K_1\|\boldsymbol{\epsilon}_n\|$, with $K_1 = \sqrt{(1+L)/(1-L)}$, now follows by invoking $\|\boldsymbol{G}\boldsymbol{U}\|_2 \leq \|\boldsymbol{G}\boldsymbol{U}\|_F$.[9] □

[9] Recall that, for any matrix $\boldsymbol{K}$ with $\text{rank}(\boldsymbol{K}) = r$, we have $\|\boldsymbol{K}\|_2 \leq \|\boldsymbol{K}\|_F \leq r\|\boldsymbol{K}\|_2$. See Golub and Van Loan [13].

4.2 Upper bounds for $\|\tilde{W}^+\|_2$ and $\|H\|_2$

Next, by Theorem A.3 in the Appendix, we can bound $\|\boldsymbol{H}\|_2$ as in

$$\|\boldsymbol{H}\|_2 \le \sqrt{2}\frac{\Delta}{1-\Delta}\|\tilde{\boldsymbol{W}}^+\|_2 \quad \text{provided } \Delta = \|\tilde{\boldsymbol{W}}^+\|_2\,\|\check{\boldsymbol{W}}\|_2 < 1. \tag{4.5}$$

We realize that all we need is a suitable upper bound on $\|\tilde{\boldsymbol{W}}^+\|_2$ since we already have an upper bound on $\|\check{\boldsymbol{W}}\|_2$ from (4.13). We turn to this issue next.

Now, by (3.4) and (3.24), we have

$$\tilde{\boldsymbol{W}} = [\,(\tilde{\boldsymbol{F}}-\boldsymbol{I})^2\boldsymbol{\epsilon}_n \,|\, (\tilde{\boldsymbol{F}}-\boldsymbol{I})^2\tilde{\boldsymbol{F}}\boldsymbol{\epsilon}_n \,|\, \cdots \,|\, (\tilde{\boldsymbol{F}}-\boldsymbol{I})^2\tilde{\boldsymbol{F}}^{k-1}\boldsymbol{\epsilon}_n\,], \tag{4.6}$$

which can be written in the form

$$\tilde{\boldsymbol{W}} = \|\boldsymbol{\epsilon}_n\|_2\,\mathring{\boldsymbol{W}}, \quad \mathring{\boldsymbol{W}} = \boldsymbol{R}\boldsymbol{S}(\boldsymbol{e}_n), \tag{4.7}$$

where

$$\boldsymbol{R} = (\tilde{\boldsymbol{F}}-\boldsymbol{I})^2 = \boldsymbol{G}^2, \quad \boldsymbol{S}(\boldsymbol{y}) = [\,\boldsymbol{y} \,|\, \tilde{\boldsymbol{F}}\boldsymbol{y} \,|\, \cdots \,|\, \tilde{\boldsymbol{F}}^{k-1}\boldsymbol{y}\,], \quad \boldsymbol{e}_n = \frac{\boldsymbol{\epsilon}_n}{\|\boldsymbol{\epsilon}_n\|_2}. \tag{4.8}$$

[Note that the columns of $\boldsymbol{S}(\boldsymbol{y})$ span the Krylov subspace $\mathcal{K}_k(\tilde{\boldsymbol{F}};\boldsymbol{y}) = \text{span}\{\boldsymbol{y}, \tilde{\boldsymbol{F}}\boldsymbol{y}, \ldots, \tilde{\boldsymbol{F}}^{k-1}\boldsymbol{y}\}$.] First, $\boldsymbol{R}$ is $N \times N$, constant, and nonsingular since $\boldsymbol{G}$ is. Next, we recall that k is at most the degree of the minimal polynomial of $\tilde{\boldsymbol{F}}$ with respect to the vector $\boldsymbol{\epsilon}_n$, which implies that the vectors $\tilde{\boldsymbol{F}}^i\boldsymbol{\epsilon}_n$, $i = 0, 1, \ldots, k-1$, are linearly independent and, therefore, $\text{rank}(\boldsymbol{S}(\boldsymbol{e}_n)) = k$. As a result, $\text{rank}(\tilde{\boldsymbol{W}}) = k = \text{rank}(\mathring{\boldsymbol{W}})$ since $\boldsymbol{R}$ is nonsingular. By the fact that $(a\boldsymbol{K})^+ = a^{-1}\boldsymbol{K}^+$ for every nonzero scalar $a \in \mathbb{C}$, and by Theorem A.1 in the Appendix, we thus have

$$\tilde{\boldsymbol{W}}^+ = \frac{1}{\|\boldsymbol{\epsilon}_n\|_2}\mathring{\boldsymbol{W}}^+ \quad \Rightarrow \quad \|\tilde{\boldsymbol{W}}^+\|_2 = \frac{1}{\|\boldsymbol{\epsilon}_n\|_2}\|\mathring{\boldsymbol{W}}^+\|_2 \tag{4.9}$$

and

$$\|\mathring{\boldsymbol{W}}^+\|_2 \le \|\boldsymbol{R}^{-1}\|_2\,\|\boldsymbol{S}(\boldsymbol{e}_n)^+\|_2. \tag{4.10}$$

We need to bound only $\|\boldsymbol{S}(\boldsymbol{e}_n)^+\|_2$ *uniformly* (i) for all $n = 1, 2, \ldots$ in the n-Mode, and (ii) for all unit vectors $\boldsymbol{e}_n^{(r)} = \boldsymbol{\epsilon}_n^{(r)}/\|\boldsymbol{\epsilon}_n^{(r)}\|_2$ arising in the different cycles of the C-Mode and the MC-Mode. Unfortunately, we are not able to prove the existence of such uniform bounds. In what follows, concerning the application of RRE in the n-Mode and in two cycling modes, we *assume* that, at each step of the different modes of usage of RRE, $\|\boldsymbol{S}(\boldsymbol{e}_n)^+\|_2$ is bounded uniformly throughout, that is, we assume that, for some constant $\tilde{\eta} > 0$,

$$\|\boldsymbol{S}(\boldsymbol{e}_n)^+\|_2 \le \tilde{\eta}. \tag{4.11}$$

Combining (4.9)–(4.11) and invoking also $\|\boldsymbol{\epsilon}_n\| \le \|\boldsymbol{G}\|_2\,\|\boldsymbol{\epsilon}_n\|_2$, we obtain

$$\boxed{\|\tilde{\boldsymbol{W}}^+\|_2 \le \frac{\eta}{\|\boldsymbol{\epsilon}_n\|}, \quad \eta = \tilde{\eta}\|\boldsymbol{G}\|_2\,\|\boldsymbol{R}^{-1}\|_2.} \tag{4.12}$$

We shall comment on this assumption concerning the uniform upper bound for $\|\boldsymbol{S}(\boldsymbol{e}_n)^+\|_2$ in Section 6.

The first thing to do now is to guarantee that $\Delta = \|\tilde{\boldsymbol{W}}^+\|_2 \, \|\check{\boldsymbol{W}}\|_2 < 1$ in (4.5) is satisfied under the assumption in (4.12) concerning $\|\tilde{\boldsymbol{W}}^+\|_2$. By (4.12) and (4.3) and the fact that $\|\boldsymbol{\epsilon}_0\| \leq \delta$ since $\boldsymbol{x}_0 \in B(\boldsymbol{s}, \delta)$, we have

$$\Delta \leq K_3' \eta \|\boldsymbol{\epsilon}_n\| \leq K_3' \eta L^n \|\boldsymbol{\epsilon}_0\| \leq K_3' \eta L^n \delta. \tag{4.13}$$

Clearly, by making δ sufficiently small, we can make the upper bound on Δ smaller than one. The closer δ is to zero, the closer $\boldsymbol{x}_0$ is to $\boldsymbol{s}$. This is precisely what is needed in order to develop a *local* convergence theory for any extrapolation method.

Next, by (4.5), (4.12), and (4.13),

$$\|\boldsymbol{H}\|_2 \leq \lambda_n, \quad \lambda_n = \sqrt{2} \frac{K_3' \eta^2}{1 - K_3' \eta \|\boldsymbol{\epsilon}_n\|}. \tag{4.14}$$

As we will show later, $\boldsymbol{\epsilon}_n$ is bounded in all three modes (n-Mode, C-Mode, and MC-Mode) we study here, which implies that λ_n is bounded too.

Remark Before proceeding further, we would like to discuss an interesting consequence of the global assumption we have made concerning $\tilde{\boldsymbol{W}}^+$. By (4.12) and (4.14) and also by (3.29), namely, that $\boldsymbol{W}^+ = \tilde{\boldsymbol{W}}^+ + \boldsymbol{H}$, we have

$$\|\boldsymbol{W}^+\|_2 \leq \|\tilde{\boldsymbol{W}}^+\|_2 + \|\boldsymbol{H}\|_2 \leq \frac{\eta}{\|\boldsymbol{\epsilon}_n\|} + \lambda_n.$$

As a result, the vector $\boldsymbol{\xi} = -\boldsymbol{W}^+ \boldsymbol{u}_n$ defined via (2.8), satisfies

$$\|\boldsymbol{\xi}\|_2 \leq \|\boldsymbol{W}^+\|_2 \, \|\boldsymbol{u}_n\|_2 \leq (1 + L)(\eta + \lambda_n \|\boldsymbol{G}^{-1}\|_2 \, \|\boldsymbol{\epsilon}_n\|).$$

Here we have made use of (4.4) too. Since $\|\boldsymbol{\epsilon}_n\|$ and λ_n are bounded, so is $\lambda_n \|\boldsymbol{\epsilon}_n\|$, in all three modes. This implies that $\boldsymbol{\xi}$ is bounded, which causes $\boldsymbol{\gamma}$ in (2.3)–(2.5) to be bounded as well. This can be seen by expressing the γ_i in terms of the ξ_i by employing (2.7) as in

$$\gamma_0 = 1 - \xi_0; \quad \gamma_i = \xi_{i-1} - \xi_i, \quad i = 1, \ldots, k-1; \quad \gamma_k = \xi_{k-1}.$$

Thus, we have globally

$$\boxed{\sum_{i=0}^{k} |\gamma_i| \leq \Gamma \text{ for some } \Gamma > 0 \text{ throughout all three modes.}}$$

Interestingly, this is analogous to the global assumption made by Toth and Kelly [41] in the convergence analysis of the acceleration method of Anderson [1]. Note that, when applied to linear systems, Anderson acceleration is equivalent to GMRES (see Walker and Ni [43]), which is equivalent to RRE applied to linear systems (see Sidi [26]).

4.3 Upper bound for $\|s_{n,k} - s\|$

With the different matrices in (4.1) bounded as above, we turn to $\boldsymbol{s}_{n,k} - \boldsymbol{s}$. By (4.2), (4.3), (4.4), and (4.12), we have

$$\|\boldsymbol{G}\tilde{\boldsymbol{U}}\|_2 \, \|\tilde{\boldsymbol{W}}^+\|_2 \, \|\check{\boldsymbol{u}}_n\| \le K_2 \eta D_0 \|\boldsymbol{\epsilon}_n\|^2, \tag{4.15}$$

$$\|\boldsymbol{G}\boldsymbol{U}\|_2 \, \|\boldsymbol{H}\|_2 \, \|\boldsymbol{u}_n\| \le K_1 \lambda_n (1+L) \, \|\boldsymbol{\epsilon}_n\|^2, \tag{4.16}$$

$$\|\boldsymbol{G}\check{\boldsymbol{U}}\|_2 \, \|\tilde{\boldsymbol{W}}^+\|_2 \, \|\boldsymbol{u}_n\| \le K_3 \eta (1+L) \, \|\boldsymbol{\epsilon}_n\|^2. \tag{4.17}$$

Substituting these into (4.1), we obtain

$$\|\check{\boldsymbol{s}}_{n,k}\| \le \tau_n \, \|\boldsymbol{\epsilon}_n\|^2, \quad \tau_n = [K_2 \eta D_0 + (K_1 \lambda_n + K_3 \eta)(1+L)] \, \|\boldsymbol{G}^{-1}\|_2, \tag{4.18}$$

and this leads to the bound on $\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\|$ in the next lemma:

Lemma 4.2 *The norm of the error vector $\boldsymbol{s}_{n,k} - \boldsymbol{s}$ can be bounded as in*

$$\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\| \le \|\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}\| + \tau_n \, \|\boldsymbol{\epsilon}_n\|^2, \quad \tau_n = [K_2 \eta D_0 + (K_1 \lambda_n + K_3 \eta)(1+L)] \, \|\boldsymbol{G}^{-1}\|_2. \tag{4.19}$$

Remark Note that, by (4.14) and (4.19), $\lim_{n\to\infty} \tau_n$ is finite since $\lim_{n\to\infty} \lambda_n$ is finite. Therefore, $\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\|$ cannot be smaller than $\|\check{\boldsymbol{s}}_{n,k}\| \le \tau_n \|\boldsymbol{\epsilon}_n\|^2$, even though $\|\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}\|$ may be smaller. In other words, the term $\|\check{\boldsymbol{s}}_{n,k}\|$ limits the accuracy of $\boldsymbol{s}_{n,k}$ as an approximation to $\boldsymbol{s}$.

5 Convergence analysis

5.1 Preliminaries

We start by studying the term $\|\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}\|$. We recall that $\tilde{\boldsymbol{s}}_{n,k}$ is the vector obtained by applying RRE to the vectors $\tilde{\boldsymbol{x}}_m$, $m = n, n+1, \ldots, n+k$, with $\tilde{\boldsymbol{x}}_n = \boldsymbol{x}_n$, as described in Section 3.1. Our study will be based on the developments of [26, 36], and [33, Chapters 6,7].

We first have

$$\begin{aligned}
\|\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}\| = \|\boldsymbol{G}(\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s})\|_2 &= \min_{g \in \tilde{\mathcal{P}}_k} \|g(\tilde{\boldsymbol{F}})\boldsymbol{G}(\tilde{\boldsymbol{x}}_n - \boldsymbol{s})\|_2 \quad \text{by Theorem 3.1} \\
&= \min_{g \in \tilde{\mathcal{P}}_k} \|g(\tilde{\boldsymbol{F}})\boldsymbol{G}(\boldsymbol{x}_n - \boldsymbol{s})\|_2 \quad \text{because } \tilde{\boldsymbol{x}}_n = \boldsymbol{x}_n \\
&\le \left[\min_{g \in \tilde{\mathcal{P}}_k} \|g(\tilde{\boldsymbol{F}})\|_2 \right] \|\boldsymbol{G}(\boldsymbol{x}_n - \boldsymbol{s})\|_2 \\
&= \theta_k \|\boldsymbol{\epsilon}_n\|,
\end{aligned} \tag{5.1}$$

recalling that $\boldsymbol{x}_n - \boldsymbol{s} = \boldsymbol{\epsilon}_n$ and defining

$$\theta_k = \min_{g \in \tilde{\mathcal{P}}_k} \|g(\tilde{\boldsymbol{F}})\|_2. \tag{5.2}$$

(Note that θ_k depends only on $\tilde{\boldsymbol{F}}$ and k.) Of course, we also have

$$\theta_k \leq \|g(\tilde{\boldsymbol{F}})\|_2 \quad \forall\, g \in \tilde{\mathcal{P}}_k. \tag{5.3}$$

We now would like to bound θ_k appropriately. Choosing $g(z) = z^k$ in (5.3), we obtain,

$$\theta_k = \min_{g \in \tilde{\mathcal{P}}_k} \|g(\tilde{\boldsymbol{F}})\|_2 \leq \|\tilde{\boldsymbol{F}}^k\|_2 \leq \|\tilde{\boldsymbol{F}}\|_2^k \leq L^k, \quad \text{at worst.} \tag{5.4}$$[10]

With all these developments, (5.1) and (4.18) together give the result in the next lemma:

Lemma 5.1 *The error vector $\boldsymbol{s}_{n,k} - \boldsymbol{s}$ satisfies*

$$\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\| \leq \theta_k \|\boldsymbol{\epsilon}_n\| + \tau_n \|\boldsymbol{\epsilon}_n\|^2. \tag{5.5}$$

Remark By choosing $g(z) \in \tilde{\mathcal{P}}_k$ suitably, upper bounds on θ_k that are smaller than L^k can be given for different cases. We give such bounds for two such cases here. For additional cases involving orthogonal polynomials, such as Jacobi polynomials, we refer the reader to Sidi and Shapira [36].

- If the hermitian part of $\boldsymbol{G} = \boldsymbol{I} - \tilde{\boldsymbol{F}}$, namely, the matrix $\boldsymbol{G}_H = \frac{1}{2}(\boldsymbol{G} + \boldsymbol{G}^*)$, is positive definite, then
$$\theta_k \leq (1 - \nu^2/\sigma^2)^{k/2},$$
where σ is the largest singular value of $\boldsymbol{G}$ and ν is the smallest eigenvalue of $\boldsymbol{G}_H$. Of course, $0 < \nu < \sigma$. See [26].
- If $\tilde{\boldsymbol{F}}$ is hermitian with eigenvalues in the (real) interval $[\alpha, \beta]$, $-1 < \alpha < \beta < 1$, then
$$\theta_k \leq \frac{1}{T_k\left(\frac{2-\alpha-\beta}{\beta-\alpha}\right)} < 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k, \quad \kappa = \frac{1-\alpha}{1-\beta} > 1.$$
Here $T_k(z)$ is the Chebyshev polynomial of the first kind of degree k. (See Varga [42, Chapter 5], for example.) Note that, in this case, $\theta_k < L^k$, with $L = \max(|\alpha|, |\beta|) < 1$.

5.1.1 Main assumptions

Before delving into the local convergence analyses of the different modes of usage of RRE, we would like to summarize the assumptions we have made so far. We will be referring to them in the statements of our (local) convergence theorems below.

A1. $\boldsymbol{f} \in C^2(B(\boldsymbol{s}, \delta))$ for some $\delta > 0$. (We can assume δ to be as small as needed in our proofs.)

A2. $\|\tilde{\boldsymbol{F}}\| \leq \max_{\boldsymbol{x} \in B(\boldsymbol{s},\delta)} \|\boldsymbol{F}(\boldsymbol{x})\| = L < 1$, which also implies that $\rho(\tilde{\boldsymbol{F}}) \leq L$.

[10] Clearly, $g(z) = z^k$ is in $\tilde{\mathcal{P}}_k$ and $\theta_k < 1$ since $L < 1$. Next, in general, the polynomial $g(z)$ that gives the optimum in (5.4) is different from z^k. Thus, generally speaking, $\theta_k < L^k$

A3. The very first vector $\boldsymbol{x}_0$, with which we start any of the modes, is in $B(\boldsymbol{s}, \delta)$. Thus, $\|\boldsymbol{x}_0 - \boldsymbol{s}\| < \delta$.

A4. $\|\tilde{\boldsymbol{W}}^+\| \leq \eta/\|\epsilon_n\|$ for every n in the n-Mode and for every cycle in the C-Mode and the MC-Mode. ($\eta > 0$ is fixed.)

5.2 Convergence in *n*-Mode

We recall that, in the n-Mode, we are applying RRE, with $k \geq 1$ fixed throughout, to the *infinite* sequence $\{\boldsymbol{x}_m\}$ that is generated as in (1.4). (That is, no cycling is involved.)

Theorem 5.2 *Under the assumptions A1–A4, RRE converges in the n-Mode. Actually, we have*

$$\limsup_{n\to\infty} \frac{\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\|}{\|\boldsymbol{\epsilon}_n\|} \leq \theta_k < 1. \tag{5.6}$$

Proof Since $\lim_{n\to\infty} \|\boldsymbol{\epsilon}_n\| = 0$ and $\lim_{n\to\infty} \tau_n < \infty$, it is clear from Lemma 5.1 that $\lim_{n\to\infty} \|\boldsymbol{s}_{n,k} - \boldsymbol{s}\| = 0$, hence $\lim_{n\to\infty} \boldsymbol{s}_{n,k} = \boldsymbol{s}$.

Next, again by Lemma 5.1, we have

$$\frac{\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\|}{\|\boldsymbol{\epsilon}_n\|} \leq \theta_k + \tau_n \|\boldsymbol{\epsilon}_n\|.$$

Taking the limsup as $n \to \infty$ of both sides and recalling again that $\lim_{n\to\infty} \tau_n < \infty$ and $\lim_{n\to\infty} \|\boldsymbol{\epsilon}_n\| = 0$, the result in (5.6) follows. □

Remark Let us also rewrite (4.19) as

$$\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\| = O(\psi_n) \quad \text{as } n \to \infty; \qquad \psi_n = \max\{\|\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}\|, \|\boldsymbol{\epsilon}_n\|^2\}. \tag{5.7}$$

This is possible since $\lim_{n\to\infty} \tau_n$ is finite. It is thus clear that $\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\|$ cannot be less than $O(\|\boldsymbol{\epsilon}_n\|^2) = O(L^{2n}) \approx O(\rho(\tilde{\boldsymbol{F}})^{2n})$ as $n \to \infty$, no matter what $\|\tilde{\boldsymbol{s}}_{n,k} - \boldsymbol{s}\|$ is. [See the remark following (4.19).]

5.3 Convergence in C-Mode cycling

In C-Mode cycling, we keep $n \geq 0$ and $k \geq 1$ fixed throughout, k always being assumed to be less than the degree of the minimal polynomial of $\tilde{\boldsymbol{F}}$ with respect to the vector $\boldsymbol{\epsilon}_n$ in every cycle.

Theorem 5.3 *Under the assumptions A1–A4, RRE converges linearly in the C-Mode. Actually, we have*

$$\limsup_{r\to\infty} \frac{\|\boldsymbol{s}_{n,k}^{(r+1)} - \boldsymbol{s}\|}{\|\boldsymbol{s}_{n,k}^{(r)} - \boldsymbol{s}\|} \leq \theta_k L^n < 1. \tag{5.8}$$

Proof We start by observing that, by Lemma 3.2, there holds $\|\boldsymbol{\epsilon}_n\| \le L^n \|\boldsymbol{\epsilon}_0\|$. With this, (5.5) becomes

$$\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\| \le \left(\theta_k L^n + \tau_n L^{2n} \|\boldsymbol{\epsilon}_0\|\right) \|\boldsymbol{\epsilon}_0\|. \tag{5.9}$$

Let us now denote the vectors $\boldsymbol{x}_m$, $\boldsymbol{\epsilon}_m = \boldsymbol{x}_m - \boldsymbol{s}$, and $\boldsymbol{s}_{n,k}$ used/computed in cycle i by $\boldsymbol{x}_m^{(i)}$, $\boldsymbol{\epsilon}_m^{(i)}$, and $\boldsymbol{s}_{n,k}^{(i)}$, respectively, and rewrite (5.9) that is relevant to the cycle $(r+1)$ as

$$\|\boldsymbol{s}_{n,k}^{(r+1)} - \boldsymbol{s}\| \le \left(\theta_k L^n + \tau_n L^{2n} \|\boldsymbol{\epsilon}_0^{(r+1)}\|\right) \|\boldsymbol{\epsilon}_0^{(r+1)}\|. \tag{5.10}$$

Let us also recall that, in the C-Mode, $\boldsymbol{x}_0^{(r+1)} = \boldsymbol{s}_{n,k}^{(r)}$, and hence $\boldsymbol{\epsilon}_0^{(r+1)} = \boldsymbol{s}_{n,k}^{(r)} - \boldsymbol{s}$. With these, (5.10) becomes

$$\|\boldsymbol{s}_{n,k}^{(r+1)} - \boldsymbol{s}\| \le \mu_r \|\boldsymbol{s}_{n,k}^{(r)} - \boldsymbol{s}\|, \quad \mu_r = \theta_k L^n + \tau_n L^{2n} \|\boldsymbol{s}_{n,k}^{(r)} - \boldsymbol{s}\|. \tag{5.11}$$

We now show by induction that for each r, $\boldsymbol{s}_{n,k}^{(r)}$ is in the ball $B(\boldsymbol{s}, \delta)$ and tends to $\boldsymbol{s}$ as $r \to \infty$, provided $\boldsymbol{x}_0$ in step C0 of C-Mode cycling is sufficiently close to $\boldsymbol{s}$.

For $r = 0$, we have $\boldsymbol{x}_0 = \boldsymbol{s}_{n,k}^{(0)} \in B(\boldsymbol{s}, \delta)$ by choice; therefore,

$$\mu_0 \le \theta_k L^n + \tau_n L^{2n} \delta.$$

Since $\theta_k L^n < 1$, we can force $\mu_0 < 1$ by choosing δ sufficiently small or by choosing $\boldsymbol{s}_{n,k}^{(0)}$ sufficiently close to $\boldsymbol{s}$. This, in turn, forces

$$\|\boldsymbol{s}_{n,k}^{(1)} - \boldsymbol{s}\| \le \mu_0 \|\boldsymbol{s}_{n,k}^{(0)} - \boldsymbol{s}\| \le \mu_0 \delta < \delta \quad \Rightarrow \quad \boldsymbol{s}_{n,k}^{(1)} \in B(\boldsymbol{s}, \delta).$$

Continuing by induction on r, we see that $\mu_{r+1} < \mu_r$, $\|\boldsymbol{s}_{n,k}^{(r+1)} - \boldsymbol{s}\| < \|\boldsymbol{s}_{n,k}^{(r)} - \boldsymbol{s}\|$, hence $\boldsymbol{s}_{n,k}^{(r+1)} \in B(\boldsymbol{s}, \delta)$ since $\boldsymbol{s}_{n,k}^{(r)} \in B(\boldsymbol{s}, \delta)$. We also have $\lim_{r\to\infty} \|\boldsymbol{s}_{n,k}^{(r)} - \boldsymbol{s}\| = 0$, hence $\lim_{r\to\infty} \boldsymbol{s}_{n,k}^{(r)} = \boldsymbol{s}$. With the convergence of $\{\boldsymbol{s}_{n,k}^{(r)}\}_{r=0}^{\infty}$ to $\boldsymbol{s}$ established, let us now rewrite (5.11) as

$$\frac{\|\boldsymbol{s}_{n,k}^{(r+1)} - \boldsymbol{s}\|}{\|\boldsymbol{s}_{n,k}^{(r)} - \boldsymbol{s}\|} \le \mu_r. \tag{5.12}$$

Taking the limsup as $r \to \infty$ on both sides of this inequality, we obtain (5.8). □

5.4 Convergence in MC-Mode cycling

We recall that in MC-Mode cycling, we keep n fixed while $k = k_r$ is the degree of the minimal polynomial of $\tilde{\boldsymbol{F}}$ with respect to $\boldsymbol{\epsilon}_n$ in the rth cycle.

Theorem 5.4 *Under the assumptions A1–A4, RRE converges quadratically in the MC-Mode. Actually, we have*

$$\limsup_{r\to\infty} \frac{\|\boldsymbol{s}_{n,k_{r+1}}^{(r+1)} - \boldsymbol{s}\|}{\|\boldsymbol{s}_{n,k_r}^{(r)} - \boldsymbol{s}\|^2} \le \tau_n L^{2n}. \tag{5.13}$$

Proof We start by noting that $\tilde{\boldsymbol{s}}_{n,k_r} = \boldsymbol{s}$ in each cycle, as mentioned at the end of Section 3.1. Thus, (4.19) becomes

$$\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\| \le \tau_n \|\boldsymbol{\epsilon}_n\|^2. \tag{5.14}$$

Proceeding precisely as in the proof of Theorem 5.3 concerning the C-Mode cycling, we next obtain

$$\|\boldsymbol{s}_{n,k} - \boldsymbol{s}\| \le \tau_n L^{2n} \|\boldsymbol{\epsilon}_0\|^2. \tag{5.15}$$

As in the case of the C-Mode, noting that $\boldsymbol{\epsilon}_0^{(r+1)} = \boldsymbol{s}_{n,k_r}^{(r)} - \boldsymbol{s}$, we write (5.15) in the MC-Mode as

$$\|\boldsymbol{s}_{n,k_{r+1}}^{(r+1)} - \boldsymbol{s}\| \le \tau_n L^{2n} \|\boldsymbol{s}_{n,k_r}^{(r)} - \boldsymbol{s}\|^2 = \phi_r \, \|\boldsymbol{s}_{n,k_r}^{(r)} - \boldsymbol{s}\|, \quad \phi_r = \tau_n L^{2n} \|\boldsymbol{s}_{n,k_r}^{(r)} - \boldsymbol{s}\|. \tag{5.16}$$

We now show, by induction on r, that $\boldsymbol{s}_{n,k_r}^{(r)}$ is in the ball $B(\boldsymbol{s}, \delta)$ and tends to $\boldsymbol{s}$ as $r \to \infty$, provided $\boldsymbol{x}_0$ in step MC0 of MC-Mode is sufficiently close to $\boldsymbol{s}$.

For $r = 0$, we have $\boldsymbol{x}_0 = \boldsymbol{s}_{n,k_0}^{(0)} \in B(\boldsymbol{s}, \delta)$ by choice; therefore,

$$\phi_0 \le \tau_n L^{2n} \delta \quad \Rightarrow \quad \phi_0 < 1 \text{ provided } \delta \text{ sufficiently small.}$$

This implies that $\|\boldsymbol{s}_{n,k_1}^{(1)} - \boldsymbol{s}\| < \|\boldsymbol{s}_{n,k_0}^{(0)} - \boldsymbol{s}\|$; therefore, $\boldsymbol{s}_{n,k_1}^{(1)} \in B(\boldsymbol{s}, \delta)$. In addition, we also have $\phi_1 < \phi_0$. Continuing by induction on r, we see that $\phi_r < \phi_{r-1} < 1$ hence $\|\boldsymbol{s}_{n,k_{r+1}}^{(r+1)} - \boldsymbol{s}\| < \|\boldsymbol{s}_{n,k_r}^{(r)} - \boldsymbol{s}\|$, which implies that $\boldsymbol{s}_{n,k_{r+1}}^{(r+1)} \in B(\boldsymbol{s}, \delta)$ since $\boldsymbol{s}_{n,k_r}^{(r)} \in B(\boldsymbol{s}, \delta)$, and that $\lim_{r\to\infty} \|\boldsymbol{s}_{n,k_r}^{(r)} - \boldsymbol{s}\| = 0$, meaning that $\lim_{r\to\infty} \boldsymbol{s}_{n,k_r}^{(r)} = \boldsymbol{s}$. With the convergence of $\{\boldsymbol{s}_{n,k}^{(r)}\}_{r=0}^{\infty}$ to $\boldsymbol{s}$ established, let us now rewrite (5.16) as

$$\frac{\|\boldsymbol{s}_{n,k_{r+1}}^{(r+1)} - \boldsymbol{s}\|}{\|\boldsymbol{s}_{n,k_r}^{(r)} - \boldsymbol{s}\|^2} \le \tau_n L^{2n}. \tag{5.17}$$

Taking the limsup as $r \to \infty$ on both sides of this inequality, we obtain (5.13). Thus, the convergence of the sequence $\{\boldsymbol{s}_{n,k_r}^{(r)}\}_{r=0}^{\infty}$ is quadratic. □

6 Remarks on $\|S(e_n)^+\|_2$

Let us observe that $\boldsymbol{S}(\boldsymbol{y})$ can be written as the product of two matrices as

$$\boldsymbol{S}(\boldsymbol{y}) = \boldsymbol{P}\,\boldsymbol{Q}(\boldsymbol{y}), \tag{6.1}$$

where $\boldsymbol{P} \in \mathbb{C}^{N\times kN}$ and $\boldsymbol{Q}(\boldsymbol{y}) \in \mathbb{C}^{kN\times k}$ are given as

$$\boldsymbol{P} = [\, \boldsymbol{I} \mid \tilde{\boldsymbol{F}} \mid \cdots \mid \tilde{\boldsymbol{F}}^{k-1} \,]; \quad \boldsymbol{Q}(\boldsymbol{y}) = \begin{bmatrix} \boldsymbol{y} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{y} & \cdots & \boldsymbol{0} \\ \vdots & & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{y} \end{bmatrix}, \quad \boldsymbol{y} \in \mathbb{C}^k. \tag{6.2}$$

Clearly, $\boldsymbol{P}$ is a constant matrix and has full row rank, while $\boldsymbol{Q}(\boldsymbol{y})$ has full column rank for all nonzero $\boldsymbol{y}$, that is,

$$\operatorname{rank}(\boldsymbol{P}) = N, \quad \operatorname{rank}(\boldsymbol{Q}(\boldsymbol{y})) = k \ \ \forall\, \boldsymbol{y} \neq \boldsymbol{0}. \tag{6.3}$$

Before going on, we recall that if $\boldsymbol{K} \in \mathbb{C}^{m\times k}$ with $\text{rank}(\boldsymbol{K}) = k$, then it has k nonzero singular values, which we order such that

$$\sigma_1(\boldsymbol{K}) \geq \sigma_2(\boldsymbol{K}) \geq \cdots \geq \sigma_k(\boldsymbol{K}) > 0,$$

and

$$\sigma_k(\boldsymbol{K}) = \min_{z\in\mathbb{C}^k,\ \|z\|_2=1} \|\boldsymbol{K}\boldsymbol{z}\|_2 \quad \text{and} \quad \|\boldsymbol{K}^+\|_2 = 1/\sigma_k(\boldsymbol{K}).$$

Now, $\boldsymbol{P}$ has N positive singular values, and therefore

$$\|\boldsymbol{P}^+\|_2 = 1/\sigma_N(\boldsymbol{P}).$$

Next, $\boldsymbol{Q}(\boldsymbol{y})$ is unitary when $\|\boldsymbol{y}\| = 1$, in the sense that

$$\boldsymbol{Q}(\boldsymbol{y})^*\boldsymbol{Q}(\boldsymbol{y}) = \boldsymbol{I}_{k\times k} \quad \forall\, \boldsymbol{y} \in \mathbb{C}^k,\ \|\boldsymbol{y}\|_2 = 1, \tag{6.4}$$

hence so is $\boldsymbol{Q}(\boldsymbol{e}_n)$ since $\|\boldsymbol{e}_n\|_2 = 1$. As a result $\boldsymbol{Q}(\boldsymbol{y})^+ = \boldsymbol{Q}(\boldsymbol{y})^*$ and $\boldsymbol{Q}(\boldsymbol{y})$ has k singular values, all equal to one, for all $\boldsymbol{y}$, $\|\boldsymbol{y}\|_2 = 1$. Consequently,

$$\|\boldsymbol{Q}(\boldsymbol{y})^+\|_2 = 1 \quad \forall\, \boldsymbol{y} \in \mathbb{C}^k,\ \|\boldsymbol{y}\|_2 = 1. \tag{6.5}$$

Despite these interesting facts—that $\|\boldsymbol{P}^+\|_2$ is fixed and that $\|\boldsymbol{Q}(\boldsymbol{e}_n^{(r)})^+\|_2 = 1$ throughout the cycling process—we are not able to prove that $\|\boldsymbol{S}(\boldsymbol{e}_n^{(r)})^+\|_2 = \|[\boldsymbol{P}\,\boldsymbol{Q}(\boldsymbol{e}_n^{(r)})]^+\|_2 \leq \alpha$ for some *fixed* $\alpha > 0$, for all $r = 0, 1, \ldots$, where $\boldsymbol{e}_n^{(r)} = \boldsymbol{\epsilon}_n^{(r)}/\|\boldsymbol{\epsilon}_n^{(r)}\|_2$ in the rth cycle.

For example, (A.3) in the appendix, which would be extremely useful if applicable, does not apply to $\boldsymbol{S}(\boldsymbol{y})$. If it did, then we would have $\boldsymbol{S}(\boldsymbol{y})^+ = \boldsymbol{Q}(\boldsymbol{y})^*\boldsymbol{P}^+$ hence $\|\boldsymbol{S}(\boldsymbol{y})^+\|_2 \leq \|\boldsymbol{P}^+\|_2$, very conveniently.

We might think that Theorem A.4 in the Appendix would apply to the n-Mode and C-Mode (it does not necessarily apply to the MC-Mode since the $\text{rank}(\boldsymbol{S}(\boldsymbol{e}_n^{(r)})) = k_r$ may vary with r), but this too is problematic. Theorem A.4 requires the following:

- In the n-Mode, the sequence $\{\boldsymbol{e}_n\}_{n=0}^{\infty}$, where $\boldsymbol{e}_n = \boldsymbol{\epsilon}_n/\|\boldsymbol{\epsilon}_n\|_2$, must have a limit $\boldsymbol{e}_\infty$ such that $\text{rank}(\boldsymbol{S}(\boldsymbol{e}_\infty)) = k$. It is obvious from (3.20)–(3.21) that it is very difficult to determine whether such a vector $\boldsymbol{e}_\infty$ exists when $\boldsymbol{f}(\boldsymbol{x})$ is nonlinear.[11]
- In the C-Mode, the sequence $\{\boldsymbol{e}_n^{(r)}\}_{r=0}^{\infty}$, where $\boldsymbol{e}_n^{(r)} = \boldsymbol{\epsilon}_n^{(r)}/\|\boldsymbol{\epsilon}_n^{(r)}\|_2$, must have a limit $\boldsymbol{e}_n^{(\infty)}$ such that $\text{rank}(\boldsymbol{S}(\boldsymbol{e}_n^{(\infty)})) = k$. It is obvious again from (3.20)–(3.21) that it is very difficult to ascertain whether such a limit exists when $\boldsymbol{f}(\boldsymbol{x})$ is nonlinear.

A different approach to the issue, for the C-Mode, would be as follows: Since $\boldsymbol{S}(\boldsymbol{e}_n^{(r)})$ has full column rank, $\|\boldsymbol{S}(\boldsymbol{e}_n^{(r)})^+\|_2 = 1/\sigma_k(\boldsymbol{S}(\boldsymbol{e}_n^{(r)})) > 0$ for every $r = 1, 2, \ldots$. Defining the vector $\boldsymbol{\zeta}(\boldsymbol{y}) \in \mathbb{C}^k$, $\|\boldsymbol{\zeta}(\boldsymbol{y})\|_2 = 1$, via

$$\min_{z\in\mathbb{C}^k,\ \|z\|_2=1} \|\boldsymbol{S}(\boldsymbol{y})z\|_2 = \|\boldsymbol{S}(\boldsymbol{y})\boldsymbol{\zeta}(\boldsymbol{y})\|_2, \tag{6.6}$$

[11] For the linear system $\boldsymbol{x} = \tilde{\boldsymbol{f}}(\boldsymbol{x})$, we have $\boldsymbol{\epsilon}_{n+1} = \tilde{\boldsymbol{F}}\boldsymbol{\epsilon}_n$, $n = 0, 1, \ldots$, as power iterations. Thus, in some cases, $\boldsymbol{e}_\infty = \lim_{n\to\infty} \boldsymbol{e}_n$ exists and is an eigenvector of $\tilde{\boldsymbol{F}}$, hence causes $\text{rank}(\boldsymbol{S}(\boldsymbol{e}_\infty)) = 1$ at most. Clearly, this is a problem when $\text{rank}(\boldsymbol{S}(\boldsymbol{e}_n)) = k > 1$, for $n = 0, 1, \ldots$.

we thus have

$$\sigma_k(\boldsymbol{S}(\boldsymbol{e}_n^{(r)})) = \min_{z\in\mathbb{C}^k,\ \|z\|_2=1} \|\boldsymbol{S}(\boldsymbol{e}_n^{(r)})z\|_2 = \|\boldsymbol{S}(\boldsymbol{e}_n^{(r)})\boldsymbol{\zeta}(\boldsymbol{e}_n^{(r)})\|_2 > 0 \quad \forall\, r = 1, 2, \ldots, \tag{6.7}$$

from which, we obtain

$$\sigma_k(\boldsymbol{S}(\boldsymbol{e}_n^{(r)})) \geq \liminf_{r\to\infty} \|\boldsymbol{S}(\boldsymbol{e}_n^{(r)})\boldsymbol{\zeta}(\boldsymbol{e}_n^{(r)})\|_2 = \alpha \geq 0. \tag{6.8}$$

Clearly, α is independent of r. Now, if we can show that $\alpha > 0$, we will have shown that $\|\boldsymbol{S}(\boldsymbol{e}_n)^+\|_2 \leq 1/\alpha$, hence that $\|\boldsymbol{S}(\boldsymbol{e}_n)^+\|_2$ is bounded uniformly throughout the cycling process. Unfortunately, this does not seem to be the case in general; the best we can say is that $\alpha \geq 0$.

Thus, even though $\sigma_k(\boldsymbol{S}(\boldsymbol{e}_n^{(r)})) > 0$ for $r = 0, 1, \ldots$, it seems we cannot guarantee the existence of a fixed positive constant $\tilde{\alpha}$ such that, when applying RRE in the cycling mode, $\sigma_k(\boldsymbol{S}(\boldsymbol{e}_n^{(r)})) \geq \tilde{\alpha}$ uniformly in every cycle. Therefore, we can only *assume* that such a constant exists for the C-Mode cycling process being studied, for which k is fixed throughout, namely,

$$\boxed{\|\boldsymbol{S}(\boldsymbol{e}_n^{(r)})^+\|_2 \leq 1/\tilde{\alpha} < \infty \quad \forall\, r,\ \operatorname{rank}(\boldsymbol{S}(\boldsymbol{e}_n^{(r)})) = k \leq k_r, \quad r = 0, 1, \ldots,} \tag{6.9}$$

where k_r is the degree of the minimal polynomial of $\tilde{\boldsymbol{F}}$ with respect to $\boldsymbol{e}_n^{(r)}$.

As for the MC-Mode cycling process, we can, similarly, only *assume* that

$$\boxed{\|\boldsymbol{S}(\boldsymbol{e}_n^{(r)})^+\|_2 \leq 1/\tilde{\alpha} < \infty \quad \forall\, r,\ \operatorname{rank}(\boldsymbol{S}(\boldsymbol{e}_n^{(r)})) = k_r, \quad r = 0, 1, \ldots.} \tag{6.10}$$

(This is reasonable because there are only finitely many k_r as $1 \leq k_r \leq N$.) Precisely (6.9) and (6.10) are what we have assumed in (4.11).

Finally, we note that the global condition in (4.12) we have imposed on the three modes for RRE discussed in this work is formulated in terms of $\tilde{\boldsymbol{F}}$, the Jacobian matrix of $\boldsymbol{f}(x)$ at the solution $\boldsymbol{s}$ only, and it concerns $\boldsymbol{s}_{n,k}$ with arbitrary n. This should be contrasted with the global condition introduced in [15] for the MC-Mode only that is formulated in terms of $\boldsymbol{f}(\boldsymbol{x})$, and concerns $\boldsymbol{s}_{0,k}$. Denoting the $\boldsymbol{x}_i$ and the $\boldsymbol{u}_i = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i$ generated at the rth cycle by $\boldsymbol{x}_i^{(r)}$ and $\boldsymbol{u}_i^{(r)}$, respectively, with $\boldsymbol{x}_0^{(r)} = \boldsymbol{s}_{0,k_{r-1}}^{(r-1)}$, the condition of [15] reads as follows:

$$\sqrt{\det(\boldsymbol{Y}_r^*\boldsymbol{Y}_r)} \geq \alpha > 0 \ \forall\, r; \quad \boldsymbol{Y}_r = [\hat{\boldsymbol{u}}_0^{(r)} \mid \hat{\boldsymbol{u}}_1^{(r)} \mid \cdots \mid \hat{\boldsymbol{u}}_{k_r-1}^{(r)}], \quad \hat{\boldsymbol{u}}_i^{(r)} = \boldsymbol{u}_i^{(r)}/\|\boldsymbol{u}_i^{(r)}\|_2.$$

Acknowledgements The author would like to thank one of the anonymous referees for his/her remarks that helped to improve the presentation and results of this work substantially.

Appendix : Some properties of Moore–Penrose inverses

First , we recall the well-known facts

$$\boldsymbol{A} \in \mathbb{C}^{m\times n}, \quad \text{rank}(\boldsymbol{A}) = n \quad \Rightarrow \quad \boldsymbol{A}^{+} = (\boldsymbol{A}^{*}\boldsymbol{A})^{-1}\boldsymbol{A}^{*} \quad \Rightarrow \quad \boldsymbol{A}^{+}\boldsymbol{A} = \boldsymbol{I}_{n\times n}, \tag{A.1}$$

$$\boldsymbol{A} \in \mathbb{C}^{m\times n}, \quad \text{rank}(\boldsymbol{A}) = m \quad \Rightarrow \quad \boldsymbol{A}^{+} = \boldsymbol{A}^{*}(\boldsymbol{A}\boldsymbol{A}^{*})^{-1} \quad \Rightarrow \quad \boldsymbol{A}\boldsymbol{A}^{+} = \boldsymbol{I}_{m\times m}, \tag{A.2}$$

and

$$\boldsymbol{A} \in \mathbb{C}^{m\times n}, \quad \boldsymbol{B} \in \mathbb{C}^{n\times p}, \quad \text{rank}(\boldsymbol{A}) = \text{rank}(\boldsymbol{B}) = n \quad \Rightarrow \quad (\boldsymbol{A}\boldsymbol{B})^{+} = \boldsymbol{B}^{+}\boldsymbol{A}^{+}. \tag{A.3}$$

The following theorems on Moore–Penrose inverses of perturbed matrices can be found in Ben-Israel and Greville [2], Wedin [44], and Stewart [40]. Here we give independent proofs of two of them.

Remark For convenience of notation, throughout this appendix only, we will use $\|\cdot\|$ to denote the l_2 norm. (Thus, $\|\cdot\|$ here does *not* stand for the $\boldsymbol{G}$ norm we have used in Sections 1–6.)

Theorem A.1 *Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, $rank(\boldsymbol{A}) = n$, and let $\boldsymbol{G} \in \mathbb{C}^{m\times m}$ be nonsingular and define $\boldsymbol{B} = \boldsymbol{G}\boldsymbol{A}$. Then* $\text{rank}(\boldsymbol{B}) = n$ *too, and*

$$\|\boldsymbol{B}^{+}\| \le \|\boldsymbol{G}^{-1}\|\|\boldsymbol{A}^{+}\|.$$

Proof That $\text{rank}(\boldsymbol{B}) = n$ is clear since $\boldsymbol{G}$ is nonsingular. Starting now with $\boldsymbol{A} = \boldsymbol{G}^{-1}\boldsymbol{B}$, we first have

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{G}^{-1}(\boldsymbol{B}\boldsymbol{x}) \quad \Rightarrow \quad \|\boldsymbol{A}\boldsymbol{x}\| \le \|\boldsymbol{G}^{-1}\|\,\|\boldsymbol{B}\boldsymbol{x}\| \quad \forall\, \boldsymbol{x} \in \mathbb{C}^{n}, \quad \|\boldsymbol{x}\| = 1.$$

Let $\boldsymbol{x}'$ and $\boldsymbol{x}''$, with $\|\boldsymbol{x}'\| = 1$ and $\|\boldsymbol{x}''\| = 1$, be such that

$$\sigma_{\min}(\boldsymbol{A}) = \min_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\| = \|\boldsymbol{A}\boldsymbol{x}'\| \quad \text{and} \quad \sigma_{\min}(\boldsymbol{B}) = \min_{\|\boldsymbol{x}\|=1} \|\boldsymbol{B}\boldsymbol{x}\| = \|\boldsymbol{B}\boldsymbol{x}''\|,$$

where $\sigma_{\min}(\boldsymbol{K})$ denotes the smallest singular value of a matrix $\boldsymbol{K}$. Then

$$\sigma_{\min}(\boldsymbol{A}) = \|\boldsymbol{A}\boldsymbol{x}'\| \le \|\boldsymbol{A}\boldsymbol{x}''\| \le \|\boldsymbol{G}^{-1}\|\,\|\boldsymbol{B}\boldsymbol{x}''\| = \|\boldsymbol{G}^{-1}\|\,\sigma_{\min}(\boldsymbol{B}).$$

The result follows by recalling that $\|\boldsymbol{K}^{+}\| = 1/\sigma_{\min}(\boldsymbol{K})$ when $\boldsymbol{K}$ has full column rank, which implies that $\sigma_{\min}(\boldsymbol{K}) > 0$. □

Theorem A.2 *Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ and $(\boldsymbol{A} + \boldsymbol{E}) \in \mathbb{C}^{m\times n}$, $m \ge n$, such that* $\text{rank}(\boldsymbol{A}) = n$ *and $\|\boldsymbol{E}\boldsymbol{A}^{+}\| < 1$. Then*

$$\|(\boldsymbol{A} + \boldsymbol{E})^{+}\| \le \frac{\|\boldsymbol{A}^{+}\|}{1 - \|\boldsymbol{E}\boldsymbol{A}^{+}\|}.$$

If $\Delta = \|\boldsymbol{E}\|\,\|\boldsymbol{A}^+\| < 1$ *in addition, then this result can be expressed as*

$$\|(\boldsymbol{A}+\boldsymbol{E})^+\| \le \frac{1}{1-\Delta}\|\boldsymbol{A}^+\|.$$

Proof First, because $\boldsymbol{A}$ is of full column rank, we have that $\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{I}_{n\times n}$. Consequently,

$$\boldsymbol{A}+\boldsymbol{E} = (\boldsymbol{I}+\boldsymbol{E}\boldsymbol{A}^+)\boldsymbol{A}.$$

Since $\|\boldsymbol{E}\boldsymbol{A}^+\| < 1$ by assumption, the matrix $\boldsymbol{G} = \boldsymbol{I}+\boldsymbol{E}\boldsymbol{A}^+$ is nonsingular. The first result now follows from Theorem A.1 and by the fact that $\|\boldsymbol{G}^{-1}\| \le 1/(1-\|\boldsymbol{E}\boldsymbol{A}^+\|)$. The second result follows by invoking $\|\boldsymbol{E}\boldsymbol{A}^+\| \le \|\boldsymbol{E}\|\,\|\boldsymbol{A}^+\| = \Delta$ and the additional assumption that $\Delta < 1$. □

Theorem A.3 *Let* $\boldsymbol{A}$ *and* $\boldsymbol{E}$ *be as in* Theorem A.2, $\Delta = \|\boldsymbol{E}\|\,\|\boldsymbol{A}^+\| < 1$, *and let* $\boldsymbol{H} = (\boldsymbol{A}+\boldsymbol{E})^+ - \boldsymbol{A}^+$. *Then*

$$\|\boldsymbol{H}\| \le \sqrt{2}\frac{\Delta}{1-\Delta}\|\boldsymbol{A}^+\|.$$

Proof By Wedin [44, Theorem 4.1], there holds

$$\|\boldsymbol{H}\| \le \sqrt{2}\,\|(\boldsymbol{A}+\boldsymbol{E})^+\|\,\|\boldsymbol{A}^+\|\,\|\boldsymbol{E}\|.$$

Invoking now Theorem A.2, the result follows. □

The following theorem is due to Stewart [40].

Theorem A.4 *Let* $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots,$ *and* $\boldsymbol{A}$ *be such that* $\lim_{n\to\infty}\boldsymbol{A}_n = \boldsymbol{A}$. *Then* $\lim_{n\to\infty}\boldsymbol{A}_n^+ = \boldsymbol{A}^+$ *if and only if* $\operatorname{rank}(\boldsymbol{A}_n) = \operatorname{rank}(\boldsymbol{A})$, $n \ge n_0$, *for some integer* n_0.

References

1. Anderson, D.G.: Iterative procedures for nonlinear integral equations. J. ACM **12**, 547–560 (1965)
2. Ben-Israel, A.: On error bounds for generalized inverses. SIAM J. Numer. Anal. **3**, 585–592 (1966)
3. Ben-Israel, A., Greville, T.N.E. Generalized Inverses: Theory and Applications. CMS Books in Mathematics, 2nd edn. Springer, New York (2003)
4. Brezinski, C.: Application de l'ϵ-algorithme à la résolution des systèmes non linéaires. C. R. Acad. Sci. Paris **271 A**, 1174–1177 (1970)
5. Brezinski, C.: Sur un algorithme de résolution des systèmes non linéaires. C. R. Acad. Sci. Paris **272 A**, 145–148 (1971)
6. Brezinski, C.: Généralisations de la transformation de Shanks, de la table de Padé, et de l'ϵ-algorithme. Calcolo **12**, 317–360 (1975)
7. Brezinski, C.: Accélération de la Convergence en Analyse Numérique. Number 584 in Lecture Notes in Mathematics, Springer, Berlin (1977)
8. Brezinski, C., Redivo Zaglia, M.: Extrapolation Methods: Theory and Practice. North-Holland, Amsterdam (1991)
9. Cabay, S., Jackson, L.W.: A polynomial extrapolation method for finding limits and antilimits of vector sequences. SIAM J. Numer. Anal. **13**, 734–752 (1976)
10. Campbell, S.L., Meyer, C.D. Jr..: Generalized Inverses of Linear Transformations. Dover, New York (1991)

11. Eddy, R.P.: Extrapolating to the limit of a vector sequence. In: Wang, P.C.C. (ed.) Information Linkage Between Applied Mathematics and Industry, pp. 387–396. Academic Press, New York (1979)
12. Gekeler, E.: On the solution of systems of equations by the epsilon algorithm of Wynn. Math. Comp. **26**, 427–436 (1972)
13. Golub, G.H., Van Loan, C.F. Matrix Computations, 4th edn. The Johns Hopkins University Press, Baltimore (2013)
14. Graves-Morris, P.R., Saff, E.B.: Row convergence theorems for generalised inverse vector-valued Padé, approximants. J. Comp. Appl. Math. **23**, 63–85 (1988)
15. Jbilou, K., Sadok, H.: Some results about vector extrapolation methods and related fixed-point iterations. J. Comp. Appl. Math. **36**, 385–398 (1991)
16. Jbilou, K., Sadok, H.: Analysis of some vector extrapolation methods for linear systems. Numer. Math. **70**, 73–89 (1995)
17. Jbilou, K., Sadok, H.: LU-implementation of the modified minimal polynomial extrapolation method. IMA J. Numer. Anal. **19**, 549–561 (1999)
18. Kaniel, S., Stein, J.: Least-square acceleration of iterative methods for linear equations. J Optimization Theory Appl. **14**, 431–437 (1974)
19. Laurens, J., Le Ferrand, H.: Fonctions d'itérations vectorielles, itérations rationelles. C. R. Acad. Sci. Paris **321 I**, 631–636 (1995)
20. Le Ferrand, H.: Convergence of the topological ϵ-algorithm for solving systems of nonlinear equations. Numer. Algorithms **3**, 273–283 (1992)
21. Mešina, M.: Convergence acceleration for the iterative solution of the equations $x = AX + f$. Comput. Methods Appl. Mech. Engrg. **10**, 165–173 (1977)
22. Ortega, J., Rheinboldt, W.: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, New York (1970)
23. Pugachev, B.P.: Acceleration of the convergence of iterative processes and a method of solving systems of nonlinear equations. U.S.S.R. Comput. Math. Math. Phys. **17**, 199–207 (1978)
24. Shanks, D.: Nonlinear transformations of divergent and slowly convergent sequences. J. Math. and Phys. **34**, 1–42 (1955)
25. Sidi, A.: Convergence and stability properties of minimal polynomial and reduced rank extrapolation algorithms. SIAM J. Numer. Anal. **23**, 197–209 (1986). Originally appeared as NASA TM-83443 (1983)
26. Sidi, A.: Extrapolation vs. projection methods for linear systems of equations. J. Comp. Appl. Math. **22**, 71–88 (1988)
27. Sidi, A.: Efficient implementation of minimal polynomial and reduced rank extrapolation methods. J. Comp. Appl. Math. **36**, 305–337 (1991). Originally appeared as NASA TM-103240 ICOMP-90-20 (1990)
28. Sidi, A.: Convergence of intermediate rows of minimal polynomial and reduced rank extrapolation tables. Numer. Algorithms **6**, 229–244 (1994)
29. Sidi, A.: Extension and completion of Wynn's theory on convergence of columns of the epsilon table. J Approx. Theory **86**, 21–40 (1996)
30. Sidi, A.: Review of two vector extrapolation methods of polynomial type with applications to large-scale problems. J. Comput. Sci. **3**, 92–101 (2012)
31. Sidi, A.: SVD-MPE: An SVD-based vector extrapolation method of polynomial type. Appl. Math. **7**, 1260–1278 (2016). Special issue on Applied Iterative Methods
32. Sidi, A.: Minimal polynomial and reduced rank extrapolation methods are related. Adv. Comput. Math. **43**, 151–170 (2017)
33. Sidi, A.: Vector Extrapolation Methods with Applications. Number 17 in SIAM Series on Computational Science and Engineering. SIAM, Philadelphia (2017)
34. Sidi, A., Bridger, J.: Convergence and stability analyses for some vector extrapolation methods in the presence of defective iteration matrices. J. Comp. Appl. Math. **22**, 35–61 (1988)
35. Sidi, A., Ford, W.F., Smith, D.A.: Acceleration of convergence of vector sequences. SIAM J. Numer. Anal. **23**, 178–196 (1986). Originally appeared as NASA TP-2193 (1983)
36. Sidi, A., Shapira, Y.: Upper bounds for convergence rates of vector extrapolation methods on linear systems with initial iterations. Technical Report 701, Computer Science Dept., Technion–Israel Institute of Technology, 1991. Appeared also as NASA TM-105608 ICOMP-92-09 (1992)
37. Sidi, A., Shapira, Y.: Upper bounds for convergence rates of acceleration methods with initial iterations. Numer. Algorithms **18**, 113–132 (1998)

38. Skelboe, S.: Computation of the periodic steady-state response of nonlinear networks by extrapolation methods. IEEE Trans. Circuits Syst. **27**, 161–175 (1980)
39. Smith, D.A., Ford, W.F., Sidi, A.: Extrapolation methods for vector sequences. SIAM Rev. **29**, 199–233 (1987). Erratum: SIAM Rev. **30**, 623–634 (1988)
40. Stewart, G.W.: On the continuity of the generalized inverse. **17**, 33–45 (1969)
41. Toth, A., Kelly, C.T.: Convergence analysis for Anderson acceleration. SIAM J. Numer Anal. **53**, 805–819 (2015)
42. Varga, R.S. Matrix Iterative Analysis. Number 27 in Springer Series in Computational Mathematics, 2nd edn. Springer, New York (2000)
43. Walker, H.F., Ni, P.: Anderson acceleration for fixed-point iterations. SIAM J. Numer. Anal. **49**, 1715–1735 (2011)
44. Wedin, P.Å.: Perturbation theory for pseudo-inverses. BIT **13**, 217–232 (1973)
45. Wynn, P.: On a device for computing the $e_m(S_n)$ transformation. Math. Tables Aids to Comput. **10**, 91–96 (1956)
46. Wynn, P.: Acceleration techniques for iterated vector and matrix problems. Math. Comp. **16**, 301–322 (1962)
47. Wynn, P.: On the convergence and stability of the epsilon algorithm. SIAM J. Numer. Anal. **3**, 91–122 (1966)