Università di Firenze, Università di Perugia, INdAM consorziate nel CIAFM

DOTTORATO DI RICERCA
IN MATEMATICA, INFORMATICA, STATISTICA

CURRICULUM IN MATEMATICA
CICLO XXXIV

Sede amministrativa Università degli Studi di Firenze

Elementwise accurate algorithms for nonsymmetric algebraic Riccati equations associated with M-matrices

Settore Scientifico Disciplinare MAT/08

Dottorando
Elena Addis

Tutor
Bruno Iannazzo
Advisor
Federico Poloni

Coordinatore
Prof. Matteo Focardi

Anni 2018/2021

To Linda,
we wrote it together

Abstract

We consider the nonsymmetric algebraic Riccati equation

$$XBX - XA - DX - C = 0, \tag{1}$$

where A, B, C, D are real matrices of sizes $n \times n$, $n \times m$, $m \times n$, $m \times m$, respectively. We focus on the case in which the matrix

$$M = \begin{bmatrix} A & -B \\ C & D \end{bmatrix},$$

is an M-matrix. The problem of finding the minimal nonnegative solution of such Riccati equations arises in applied probability, transport theory, fluid queues.

A *structure-preserving doubling algorithm (SDA)* for computing the minimal solution of (1) has been proposed in [17]. This iterative method relies on the fact that the problem of solving (1) can be reduced to the computation of certain invariant subspaces of the matrix $\mathcal{H} = \begin{bmatrix} I_n & 0 \\ 0 & -I_m \end{bmatrix} M$, and its convergence properties are connected with a quotient involving the eigenvalues of $\mathcal{H}$.

In [15] Guo et al. studied the doubling algorithm in the case where M is an irreducible singular M-matrix and, in order to speed up the convergence, proposed a shift technique to move one zero eigenvalue of $\mathcal{H}$ to a positive real number. This approach modifies the equation (1) by introducing a rank-one correction of $\mathcal{H}$ that leads to a shifted equation sharing with (1) the solution. This modification induces a reduction of the quotient that controls the convergence so as to produce in certain cases a dramatic decrease of the number of steps of the algorithm.

When M is a nonsingular M-matrix or an irreducible singular M-matrix, algorithms computing the minimal nonnegative solution of (1) with high elementwise relative accuracy have been proposed in [39], [23], [38]. The general approach is based on the idea that a nonsingular M-matrix can be inverted by the GTH-like algorithm [1] that consists in a modification of the standard Gaussian elimination in a cancellation-free fashion, when a *triplet representation* of the matrix is known. A triplet representation of A is a triple (u, v, w) such that $u = -\text{offdiag}(A) \geq 0$, $v > 0$, $w \geq 0$ and $Av = w$.

Unfortunately, the shifted matrix $\widehat{M}$ constructed in [15] in general is no longer an M-matrix, so the known elementwise accurate algorithms can not be applied directly together with the shift technique in order to improve the accuracy and also accelerate the convergence.

We present an elementwise accurate algorithm which incorporates the shift technique for the computation of the minimal nonnegative solution of (1), when M is an irreducible singular M-matrix.

We propose the idea of *delayed shift* and some results that guarantee the applicability and the convergence of structured doubling algorithm based only on the properties of the matrix of the initial setup of doubling algorithm instead of matrix M or $\widehat{M}$. We provide a componentwise error analysis for the algorithm and we also show some numerical experiments that illustrate the advantage in terms of accuracy and convergence speed.

Contents

List of symbols and notations

$A \in \mathbb{R}^{n\times m}$	$A = (a_{ij})$, where $a_{ij} \in \mathbb{R}$, $(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,m\}$
e	column vector whose components are equal to 1
I_n	identity matrix of size n
$\mathrm{ind}(A)$	index of A, i.e. the minimum nonnegative integer k such that $\mathrm{rank}(A^{k+1}) = \mathrm{rank}(A^k)$
$\sigma(A)$	spectrum of A, i.e. set of the eigenvalues of A
$\rho(A)$	spectral radius of matrix A, i.e. the maximum of the moduli of the eigenvalues of A
$A \geq B$	$a_{ij} \geq b_{ij}$, for all $(i,j) \in \{1,\ldots,n\} \times \{1,\ldots,m\}$
$\mathrm{offdiag}(A)$	operator from $\mathbb{R}^{n\times n}$ to $\mathbb{R}^{n^2-n}$ that stacks the off-diagonal entries of A by column into one long vector
$P[X = a \mid Y = b]$	conditional probability that the random variable X takes the value a given that the random variable Y takes the value b
u	unit roundoff or machine precision in double precision floating point arithmetic (in MATLAB $\mathsf{u} \approx 2.2 \times 10^{-16}$)
$A \dot{\leq} B$	$A \leq B + O(\mathsf{u}^2)$

Chapter 1

Preliminary notions

This chapter contains some preliminary notions about M-matrices, algebraic Riccati equations and elementwise accuracy, with basic definitions, fundamental properties and some applications as Markov chains and fluid queues.

1.1 M-matrices

In a wide variety of problems of the real world (for instance, in the biological, physical, computer and social sciences) arise matrices that, due to the constraints of the specific problem to solve, have the special structure of Z-matrices or M-matrices.

Definition 1.1 ([26]). A matrix $A \in \mathbb{R}^{n\times n}$ is called a Z-matrix if all its offdiagonal entries are nonpositive (i.e. $\text{offdiag}(A) \leq 0$). Any Z-matrix can be expressed in the form

$$A = sI_n - B,$$

where $s \in \mathbb{R}$ and $B \in \mathbb{R}^{n\times n}$, $B \geq 0$. A is called an M-matrix if $s \geq \rho(B)$; it is called a singular M-matrix if $s = \rho(B)$ and a nonsingular M-matrix if $s > \rho(B)$.

In [5], Chapter 6, the authors present a wide characterization of nonsingular M-matrices in the style of Plemmons's survey paper [26]. We report below a result that will be really useful hereafter:

Lemma 1.2 ([5, Theorem 6.2.3]). *For a Z-matrix A, these properties are equivalent:*

1. *A is a nonsingular M-matrix.*
2. *$A^{-1} \geq 0$.*
3. *$Av > 0$ for some vector $v > 0$.*
4. *All eigenvalues of A have positive real parts.*

Another essential notion for the following is that of *reducibility*:

Definition 1.3 ([5, Definition 2.1.2]). A matrix $A \in \mathbb{R}^{n\times n}$ is called reducible if there is a permutation matrix $\Pi \in \mathbb{R}^{n\times n}$ such that

$$\Pi^T A \Pi = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{1.1}$$

where $A_{11} \in \mathbb{R}^{k\times k}$ and $A_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$, with $0 < k < n$; it is irreducible if it is not reducible.

For the singular M-matrices, under the additional assumption of irreducibility, next lemma can be stated :

Lemma 1.4 ([5, Theorem 6.4.16]). *If A is an irreducible singular M-matrix, the following properties hold:*

1. *rank$(A) = n - 1$;*
2. *there exists a vector $v > 0$ such that $Av = 0$;*
3. *each principal submatrix of A other than A itself is a nonsingular M-matrix.*

It is interesting to observe that the two results above provide sufficient conditions for the existence of the LU factorization of an M-matrix:

Corollary 1.4.1 ([5, Theorems 6.4.6-17-18]). *If A is a nonsingular M-matrix or an irreducible singular M-matrix, then there exists an LU factorization of A, i.e. there exists a pair (L, U) with L is and U lower and upper triangular, respectively, and every diagonal element of L equal to 1, such that:*

$$A = LU.$$

If A is an M-matrix, then there exists a permutation matrix Π, such that

$$\Pi A \Pi^T = LU.$$

From Lemmas 1.2-1.4 it can be derived an idea that is fundamental for the following: the *triplet representation* of M-matrices. We start with a definition.

Definition 1.5 ([1], [23], [40]). A triple $(u, v, w) \in \mathbb{R}^{n^2-n} \times \mathbb{R}^n \times \mathbb{R}^n$ with $u \geq 0$, $v > 0$, $w \geq 0$ is called a (right) triplet representation for the matrix A if $u = -\text{offdiag}(A)$ and $Av = w$. We write $A = (u, v, w)$.

We observe that the triplet representation (u, v, w) as defined above uniquely identifies the matrix A, because the offdiagonal entries of A are described by the vector u whereas for the diagonal components of A we have:

$$a_{ii} = \frac{w_i + \sum_{j\neq i} p_{ij} v_j}{v_i}, \tag{1.2}$$

where $P = (p_{ij})$ is a matrix having the same offdiagonal entries of $-A$, but the diagonal entries equal to zero, so offdiag$(P) = u$.

Moreover, we can state the following lemma on the *triplet representation* of M-matrices deriving from Lemmas 1.2-1.4 and more generally from Perron-Frobenius theory:

Theorem 1.6. *If there exists a triplet representation (u, v, w) of A, then A is an M-matrix. In particular, if $A = (u, v, w)$ with $w > 0$, then A is a nonsingular M-matrix. Conversely, if A is a nonsingular M-matrix or an irreducible singular M-matrix then there exists a triplet representation (u, v, w) of A. In particular*

i. if A is a nonsingular M-matrix, then $A = (u, v, w)$ with $w > 0$,

ii. if A is an irreducible singular M-matrix, then $A = (u, v, w)$ with $w = 0$.

Proof. If $A = (u, v, w)$, then A is a Z-matrix because $u \geq 0$, thus $A = sI_n - B$ with $B \geq 0$, and in addition there exists $v > 0$ such that $Av = sv - Bv = w \geq 0$. Theorem 2.1.11 in [5] states that if $C \geq 0$ and there exists $x > 0$ such that $cx \geq Cx$ then $c \geq \rho(C)$, thus we have that $s \geq \rho(B)$. We conclude that A is an M-matrix. The remainder follows directly from Lemma 1.2 and Lemma 1.4. □

In [14] Guo called *regular* the M-matrices having a triplet representation. It is important to note that the class of the regular M-matrices includes not only nonsingular M-matrices or irreducible singular M-matrices, but not every reducible singular M-matrix can be expressed in triplet form. The following lemma, implied by a result in [5], determines a stronger necessary condition for the existence of a triplet representation of an M-matrix A:

Lemma 1.7. *If there exists a triplet representation (u, v, w) of A, then A is an M-matrix with ind$(A) \leq 1$.*

Example 1.8. The matrix

$$A = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}$$

is a reducible singular M-matrix for which a vector $w \geq 0$ such that $Av = w$ cannot be found, if $v > 0$. On the other hand, it is easy to show some reducible singular M-matrix admitting triplet representation: for instance, the matrix

$$B = \begin{bmatrix} b & -1 \\ 0 & 0 \end{bmatrix}$$

is such that $B = (u, v, w)$ where

$$u = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \quad v = \begin{bmatrix} 1 \\ b/2 \end{bmatrix}, \quad w = \begin{bmatrix} b/2 \\ 0 \end{bmatrix},$$

with $w \geq 0$ if $b > 0$. We observe that ind$(A) = 2$ and ind$(B) = 1$, according to Lemma 1.7.

In view of Lemma 1.7 we can state a result that will be useful in the following and that is a slightly different version of Lemma 4 of [16]:

Theorem 1.9. *Let $A \in \mathbb{R}^{n\times n}$ be a matrix having a triplet representation. In particular*

i. A is nonsingular if and only if $ind(A) = 0$ and, equivalently, A does not have any zero eigenvalues;

ii. A is singular if and only if $ind(A) = 1$ and, equivalently, 0 is an eigenvalue of A of multiplicity $r \geq 1$ with r linearly independent corresponding eigenvectors. Moreover, if A is an irreducible singular M-matrix with a triplet representation, then $ind(A) = 1$ and 0 is a simple eigenvalue of A with only one corresponding eigenvector, up to scalar multiples.

Proof. The first statement follows from the fact that A is nonsingular if and only if $\text{rank}(A^k) = n$ for all $k \geq 0$, that is $\text{ind}(A) = 0$ by the definition of index. Thus, for the second item, by Lemma 1.7 we have that if A is a singular matrix having a triplet representation, then $\text{ind}(A) = 1$. But it happens only when all the blocks corresponding to 0 in the Jordan canonical form of A have order 1. Moreover, if A is an irreducible singular M-matrix, we know that $\text{rank}(A) = n - 1$ from Lemma 1.4, then $r = 1$ in this case. □

It is an interesting fact that the class of the matrices having a triplet representation is strictly included in the class of the M-matrices with index less or equal to 1.

Example 1.10. We consider the reducible singular matrix

$$A = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}.$$

It is easily seen that A is an M-matrix and $\text{ind}(A) = 1$. But, for every positive vector $v = [a, b, c]^T$,

$$Av = A \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} a \\ -a \\ -a \end{bmatrix},$$

then it is impossible to find a vector $w \geq 0$ such that $Av = w$.

1.2 Markov chains

The M-matrices are strictly connected to the theory of Markov chains that arise in many areas of application including statistics, economics, queuing theory, and mobile networks ([5],[7],[32]). In this section we recall some basic information about the elementary concepts on the Markov chains. For a more specific presentation of the theoretical and applicative aspects of the stochastic processes we refer the reader, for instance, to [9].

A Markov chain is a stochastic process given by the sequence of random variables $\{X_t\}_{t\in T}$ characterized by the fundamental *Markov property* that the probability of transition of the system to the next state only depends on the present state (memorylessness). The variables X_t take values in the countable set S, the *state space* of the process. Each element of S represents the status of the system at a certain time t. The set T of the index, or of the *times*, using a more applicatively evocative language, may be discrete, usually $T = \mathbb{N}$, or continuous, usually $T = \mathbb{R}^+$: in the first case the Markov chain is said discrete-time (DTMC), in the second case is said continuous-time (CTMC).
If a discrete-time Markov chain has the additional property that the probability $p_{ij}^{(t)}$ of transition from state i to state j is independent of the time t, i.e.

$$p_{ij}^{(t)} = P[X_{t+1} = j \mid X_t = i] = P[X_t = j \mid X_{t-1} = i] = p_{ij}^{(t-1)},$$

for all $t \geq 1$, the Markov chain is called *homogeneous*. Thus, for every homogeneous Markov chain with a finite space state S can be defined a *transition probability matrix* $P = (p_{ij})$ of size $n \times n$, where i, j represent states of S and $n = \text{card}(S)$. Since the probability of transition from a state i to all other must be equal to 1, we observe that P is stochastic, that is, $0 \leq p_{ij} \leq 1$ and

$$\sum_{j=1}^{n} p_{ij} = 1.$$

We can rewrite the equation above in matrix form, so we get

$$Pe = e.$$

The evolution of the system is described by the equation

$$(\pi^{(t+1)})^T = (\pi^{(t)})^T P, \tag{1.3}$$

where $\pi^{(t)}$ is called *probability state vector* and is defined as

$$\pi_i^{(t)} = P[X_t = i \mid X_0].$$

We observe that $0 \leq \pi_i^{(t)} \leq 1$ and $\sum_{i\in S} \pi_i^{(t)} = 1$, from the probability laws.

In the case where the system has an asymptotic behaviour, the limit is defined

$$\pi := \lim_t \pi^{(t)}.$$

If there exists, the vector π is called *invariant probability vector* or *steady state vector* or *stationary vector* and, from (1.3), we have

$$\pi^T = \pi^T P$$

that is, π^T is a nonnegative left eigenvector of P corresponding to the eigenvalue 1 normalized so that $\pi^T e = 1$.

Thus, setting $x = \pi$ and $A = (I_n - P)^T = I_n - P^T$ (sometime we refer to A as the *generator* of the discrete Markov process), the computation of the invariant probability vector of the Markov chain having P as transition matrix can be reduced to the finding of a nontrivial solution of the linear system $Ax = 0$. If the matrix P is stochastic and irreducible, the matrix A is an irreducible singular M-matrix and Lemma 1.4 guarantees the existence of a solution $x > 0$ that is unique up to positive scalar multiples.

The computation of the stationary vector can be performed with a large variety of methods, both direct and iterative ([7], [33]). The direct solution algorithms for Markov chains are in general based on the Gaussian elimination, that has a variant known as the Grassmann-Taksar-Heyman (GTH) algorithm introduced in [12], which is guaranteed to compute the solution with low relative error in each component, removing possible cancellation: this is the starting point of this work.

1.3 Fluid queues and Riccati equation

In this section we examine the relation between the stochastic processes called *fluid queues* and the minimal nonnegative solution of a special type of matrix equations known as *Riccati equations*: for more details we refer to [6], [13], [28] and [29].

A fluid queue, also called *stochastic fluid model*, is a mathematical model used to describe every real phenomenon similar to the flow of fluid in a reservoir subject to randomly determined periods of filling and emptying. For instance in the literature many authors use the fluid queues for modelling data communication channels and networks. In particular, we are interested in a process where the input rates are controlled by a continuous time Markov chain (CTMC) on a finite state space $\mathcal{S}$, with irreducible generator Q. For a continuous time Markov chain the *generator* is a matrix with order equal to $\text{card}(\mathcal{S})$, describing the instantaneous rate of the process transitions from state i to state j and having the following two properties:

i. $Qe = 0$,

ii. $q_{ij} \geq 0$ for $i \neq j$ and $i, j \in \mathcal{S}$.

It is easily seen that Q is a special case of Q-matrix (a Q-matrix has nonnegative offdiagonal elements and nonpositive row sums) and every $-Q$-matrix is an M-matrix.

If $\{X_t\}_{t\geq 0}$ is a CTMC with a finite space state $\mathcal{S}$, some quantities of interest for the buffer content process ξ (for instance, the invariant law for the buffer content process, i.e. the limit

$$\lim_{t\to\infty} P[X_t = j, \xi_t > x], \quad \text{for} j \in \mathcal{S}, x \geq 0)$$

can be expressed in term of Wiener-Hopf factorization of the matrix Q. For more details, see [28] and [29].

If Q has the form

$$Q = \begin{bmatrix} -A & B \\ -C & -D \end{bmatrix}$$

the Wiener-Hopf factorization is a quadruple (Π_1, Q_1, Π_2, Q_2) such that

$$\begin{bmatrix} -A & B \\ C & D \end{bmatrix} \begin{bmatrix} I & \Pi_2 \\ \Pi_1 & I \end{bmatrix} = \begin{bmatrix} I & \Pi_2 \\ \Pi_1 & I \end{bmatrix} \begin{bmatrix} Q_1 & 0 \\ 0 & -Q_2 \end{bmatrix} \tag{1.4}$$

where A, D are square matrices and Q_1 and Q_2 are Q-matrices.
Since the matrix $M = -Q$ defined by

$$M = \begin{bmatrix} A & -B \\ C & D \end{bmatrix} \tag{1.5}$$

is an M-matrix, the definition of Wiener-Hopf factorization can be easily extended for M-matrices (see [13]): in this case the problem is to find a quadruple (Φ, R, Ψ, S) such that

$$\begin{bmatrix} A & -B \\ -C & -D \end{bmatrix} \begin{bmatrix} I_n & \Psi \\ \Phi & I_m \end{bmatrix} = \begin{bmatrix} I_n & \Psi \\ \Phi & I_m \end{bmatrix} \begin{bmatrix} R & 0 \\ 0 & -S \end{bmatrix}. \tag{1.6}$$

where Ψ, Φ are nonnegative matrices and R and S are M-matrices.

By extending the previous results involving nonsingular M-matrices or irreducible singular M-matrices (see [13]), in Corollary 3 of [14] has been proved that a factorization of the form (1.6) exists in the case where M is a regular M-matrix and it can be obtained from the solution of the certain matrix equations:

Theorem 1.11 ([14, Corollary 3]). *If M is a regular M-matrix, a quadruple (Φ, R, Ψ, S) that verifies the relation (1.6) exists and:*

i. Φ is the minimal nonnegative solution of the equation

$$XBX - XA - DX - C = 0; \tag{1.7}$$

ii. $R = A - B\Phi$ is a regular M-matrix;

iii. Ψ is the minimal nonnegative solution of the equation

$$B - AY - YD - YCY = 0; \tag{1.8}$$

iv. $S = D + C\Psi$ is a regular M-matrix.

An equation of the form (1.7) is called *nonsymmetric algebraic Riccati equation* (NARE) and the equation (1.8) is its *dual equation*. In the case of our interest the nonsymmetric algebraic Riccati equation is also called MARE or M-NARE, because its coefficients A, B, C, D are the blocks of the M-matrix M, so the Riccati equation is associated to a regular M-matrix. It is important to notice here that for this class of Riccati equations the minimal nonnegative solution always exists:

Theorem 1.12 ([14, Theorem 2-Proposition 4], [16, Theorem 1]). *If the matrix M in (1.5) is a regular M-matrix then (1.7) and (1.8) have minimal nonnegative solutions Φ and Ψ, respectively. In addition $R = A - B\Phi$ and $S = D + C\Psi$ are regular M-matrices. Moreover, $I_n - \Psi\Phi$ and $I_m - \Phi\Psi$ are both regular M-matrices.*

The accurate computation of the minimal nonnegative solution of M-NARE is the topic of Chapter 3.

1.4 Normwise vs elementwise accuracy

In this section we focus on the difference between normwise and elementwise accuracy of a computed solution $\tilde{X}$ with respect to the exact solution X, because in the following we discuss the design of elementwise accurate algorithms for M-matrices.

In general, if we want to measure the distance between two matrices A and B in $\mathbb{R}^{n\times m}$, we can choose a *normwise* approach based on the distance

$$d_{\mathrm{nw}} := ||A - B||,$$

where $||\cdot||$ is a suitable matrix norm on $\mathbb{R}^{n\times m}$, or an *elementwise* (or *componentwise*, or *entrywise*) approach based on the distance

$$d_{\mathrm{ew}} := \max_{(i,j)} |a_{ij} - b_{ij}|.$$

Thus, we can define two different types of relative error, ϵ_{nw} and ϵ_{ew}, in this way:

$$\epsilon_{\mathrm{nw}} := \frac{||\tilde{X} - X||}{||X||}, \quad \epsilon_{\mathrm{ew}} := \max_{(i,j)} \frac{|\tilde{x}_{ij} - x_{ij}|}{|x_{ij}|},$$

which are undefined, respectively, if $X = 0$ and $x_{i,j} = 0$ for some pair (i, j). We observe that latter relation is equivalent to

$$|\tilde{X} - X| \leq \epsilon_{\mathrm{ew}}|X|.$$

We say that a solution is *elementwise accurate* if the elementwise relative error ϵ_{ew} is small, i.e. such that

$$\epsilon_{\mathrm{ew}} \leq K\varepsilon + O(\varepsilon^2),$$

where K is a constant in general depending on the size of X and ε is a value "acceptably small". In the following, we consider $\varepsilon := \mathsf{u}$, the unit roundoff (or machine precision) in

floating-point arithmetic (cfr. [18]). The definition of *normwise accuracy* is analogous, mutatis mutandis.

We observe that in general a small entrywise relative error implies that the normwise relative error is also small, but the converse is not true. Indeed, a huge difference of magnitude on the entries of X may provide that the computed solution $\tilde{X}$ has a tiny normwise relative error, despite an extremely large elementwise relative error: we illustrate this behaviour in the example below.

Example 1.13. Let $\tilde{X}$ be a solution of a certain problem with X as exact solution. We suppose that, for some reason, the only component $\tilde{x}_{11}$ is affected by an arbitrarily large error $\epsilon > 0$, for example

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 10^k \end{bmatrix}, \quad \tilde{X} = \begin{bmatrix} 1+\epsilon & 1 \\ 1 & 10^k \end{bmatrix}.$$

If we compute ϵ_{nw} by using the norm $||\cdot||_2$, we have

$$\epsilon_{\text{nw}} = \frac{||\tilde{X} - X||_2}{||X||_2} \approx \epsilon \cdot 10^{-k},$$

whereas the elementwise error is provided by the relation

$$\epsilon_{\text{ew}} = \max_{(i,j)} \frac{|\tilde{x}_{ij} - x_{ij}|}{|x_{ij}|} = \max_{(i,j)} \begin{bmatrix} \epsilon & 0 \\ 0 & 0 \end{bmatrix} = \epsilon.$$

Thus, if the value $k > 0$ is reasonably large, the normwise relative error, depending on ϵ and k, may result small although the elementwise relative error, depending only on ϵ, is large.

It is easy to see that, in a case like this, a small normwise relative error produces the undesirable effect of "hiding" a huge relative error on the single components of the solution: this is the fundamental reason for developing the elementwise accurate algorithms which we will describe in the following. We note that this error model is useful in particular for the applications in the field of probability, in which typically small quantities are present that however have to be computed with high accuracy.

Remark 1.14. Hereafter, where there is no risk of confusion, we will write simply *accurate* instead of *elementwise accurate*.

Chapter 2

Basic accurate algorithms for M-matrices

Grassmann, Taksar, and Heyman in [12] proposed a variant of the Gaussian elimination method for computing the steady state vector of a Markov chain, called GTH algorithm. The authors suggested that the accuracy of their algorithm is higher than the "standard" Gaussian elimination because it involves no subtractions. A formal error analysis supporting this intuition has been provided by O'Cinneide in [24] and improved by Xue in [36]. Later, Alfa et al. in [1, 2] extended the GTH algorithm to the computation of quantities of interest for diagonally dominant M-matrices: they called this algorithm GTH-like. Moreover, in [1, 2, 40], some results concerning the perturbation theory have been estabilished: it is proved that if an M-matrix is affected by small relative errors, then its inverse and other quantities of interest can be computed with high entrywise accuracy by using methods based on GTH-like. This property justifies the interest in studying this algorithm. In this section we present the fundamental methods for elementwise accurate computing involving M-matrices with triplet representation and we report the more important results about their elementwise accuracy.

2.1 GTH algorithm

Grassmann, Taksar, and Heyman [12] have shown how to modify the standard algorithm of Gaussian elimination in order to construct an elegant algorithm providing accurate computation of the stationary vector of a Markov chain. We refer the reader interested in other methods direct and iterative for solving Markov chains to the monograph of Stewart [33].

As we have seen previously, if the transition matrix P connected to the Markov process is an irreducible and stochastic (i.e. $Pe = e$) $n \times n$ matrix, the stationary vector exists and it is the unique vector such that

$$\pi^T = \pi^T P$$

with $0 < \pi_i \leq 1$ and $\pi^T e = 1$. For the sake of simplicity, letting $A = I_n - P^T$ and $x = \pi$ we reduce the problem to finding a non trivial solution of the linear system $Ax = 0$.

Remark 2.1. Different implementations of the algorithm are possible working with $A = I_n - P$ rather than $A = I_n - P^T$ (see [33]).

By applying the steps of the Gaussian elimination, the GTH algorithm produces a sequence of matrices of decreasing order $\{A^{(k)}\}_{1 \leq k \leq n}$, starting from $A = A^{(1)}$, where $A^{(k)}$ denotes the matrix of order $n - k + 1$ to the southeast of the k-th pivot entry (and including that pivot entry), just before the k-th Gaussian elimination is performed. Thus we have

$$A^{(k)} = \begin{bmatrix} \alpha_k & -s_k^T \\ -z_k & B^{(k)} \end{bmatrix} \tag{2.1}$$

where α_k is the k-th pivot and the order of $B^{(k)}$ is $n - k$. The relation between $A^{(k)}$ and the next matrix of the sequence is:

$$A^{(k+1)} = B^{(k)} - \frac{z_k s_k^T}{\alpha_k}. \tag{2.2}$$

In the "standard" Gaussian elimination the pivotal elements are given explicitily by:

$$\alpha_1 = a_{11}^{(1)} = a_{11}; \quad \alpha_k = a_{11}^{(k)} = a_{22}^{(k-1)} - \frac{a_{2,k-1}^{(k-1)} a_{k-1,2}^{(k-1)}}{\alpha_{k-1}}. \tag{2.3}$$

Since A is an irreducible singular M-matrix, from Lemma 1.4 follows that the Gaussian elimination method is applicable (without pivoting) because all the pivots are nonzero.

A crucial observation is that, under the assumptions above on A, the k-th pivot α_k is positive for all k and during its computation no subtractions occur. This is easily seen at the first step, since from the relations $A = I_n - P^T$ and $Pe = e$ we have

$$\alpha_1 = a_{11} = 1 - p_{11} = \sum_{j=2}^{n} p_{1j}, \tag{2.4}$$

thus the first pivot can be computed with no subtractions.
Moreover, from (2.2) the elements of the matrix $A^{(2)}$ can be computed by the relation

$$a_{ij}^{(2)} = a_{i+1,j+1}^{(1)} - \frac{a_{i+1,1}^{(1)} a_{1,j+1}^{(1)}}{\alpha_1}, \quad \text{for } i, j \geq 1. \tag{2.5}$$

We observe that all the quantities involved in the computation are nonpositive (except than α_1 when $i \neq j$), so the offdiagonal elements can be computed by negating a sum:

$$a_{ij}^{(2)} = -\left(-a_{i+1,j+1}^{(1)} + \frac{a_{i+1,1}^{(1)} a_{1,j+1}^{(1)}}{\alpha_1} \right), \quad \text{for } i, j \geq 1, i \neq j. \tag{2.6}$$

The fundamental point is that in GTH algorithm the second pivot α_2 is not obtained from the relation (2.3), but by observing that the properties that characterize the matrix A here (i.e. $A = I_n - P^T$ with P irreducible and stochastic) are invariant under the elementary operations performed during the Gaussian elimination process.

Remark 2.2. For the second and the successive steps, in their original paper Grassmann, Taksar, and Heyman show these invariant properties by using a sophisticated probabilistic argument and in [31] and [4] are present proofs based on the idea of *stochastic complementation* (see also [22]). In the next section we give a general proof of this property based on the triplet representation of M-matrices.

Thus, assuming that $e^T A^{(2)} = 0^T$, the current pivot is given by the negated sum of the new offdiagonal elements in the corresponding column:

$$\alpha_2 = a_{11}^{(2)} = -\sum_{i=2}^{n-1} a_{i1}^{(2)}. \tag{2.7}$$

This procedure can be extended to cover each step of the Gaussian elimination process.

Remark 2.3. We refer the reader to [12], [25], [31], [33] for different versions of GTH algorithm. We observe that in [4] has been introduced a variant of a direct projection method based on the GTH idea that computes the stationary probabilities accurately.

It is easily seen that the GTH algorithm requires more numerical operations than the standard implementation of the Gaussian elimination, but the extra additions are not very costly when compared with the overall cost of the elimination procedure. In addition, the advantage of the higher accuracy of the solution leads to the conclusion that the GTH algorithm should be exploited where possible in elimination procedures. Stewart in [33] also noticed that difficulties arise in implementing GTH when the structure of the matrix A is sparse, since the construction of L and U produces fill-in phenomena that destroy the sparsity of A (this is a usual problem in the context of direct methods applied to sparse matrices [30]).

2.2 GTH-like algorithm

In [1] it has been introduced an extension of the GTH algorithm to diagonally dominant M-matrices, the so-called GTH-like algorithm, with the purpose of computing the smallest eigenvalue of A besides the inverse of A. Later, in [2], the same authors developed a entrywise perturbation theory for diagonally dominant M-matrices. In this section we extend the approach of [1, 2] based on the triplet representation of a diagonally dominant M-matrix to the matrices having a triplet representation.

Firstly, we need the following lemma regarding the well-definition of the sequence of matrices of decreasing order $\{A^{(k)}\}_{1 \le k \le n}$, generated by the Gaussian elimination, where $A^{(k)}$ denotes the matrix of order $n-k+1$ to the southeast of the k-th pivot entry (and including that pivot entry), just before the k-th Gaussian elimination is performed.

Lemma 2.4. *Let $A \in \mathbb{R}^{n\times n}$ and let $\{A^{(k)}\}_{1\le k\le n}$ be the sequence starting from $A^{(1)} = A$ of matrices of the form:*

$$A^{(k)} = \begin{bmatrix} \alpha_k & -s_k^T \\ -z_k & B^{(k)} \end{bmatrix} \tag{2.8}$$

with $B^{(k)}$ of order $n-k$, defined by

$$A^{(k+1)} = B^{(k)} - \frac{z_k s_k^T}{\alpha_k}. \tag{2.9}$$

If A is a nonsingular M-matrix or an irreducible singular M-matrix, then the sequence $\{A^{(k)}\}_{1\le k\le n}$ is well-defined, that is $\alpha_k \neq 0$ for $k = 1, ..., n-1$.

Proof. The well-definition of the sequence of the $A^{(k)}$ is equivalent to the nonsingularity of each of the first $n-1$ principal submatrices of A. If A is a nonsingular M-matrix or an irreducible singular M-matrix this condition is satisfied (Lemma 1.4). □

We are now ready to describe a fundamental property about the Gaussian elimination applied to a matrix A having a triplet representation: at each step of the Gaussian elimination, the active matrix $A^{(k)}$ inherits from A the representability in triplet form.

Theorem 2.5. *With the notation above, if $A^{(1)} = A$ has a triplet representation $A = (u, v, w)$ and the sequence $\{A^{(k)}\}_{1\le k\le n}$ is well-defined, then $A^{(k)}$ has a triplet representation for all $k \ge 1$:*

$$A^{(k)} = (u^{(k)}, v^{(k)}, w^{(k)}).$$

Proof. We will prove the theorem by induction on k. The base case $k = 1$ is true for the assumption of existence of a triplet representation of $A = A^{(1)}$: we have $u^{(1)} = u$, $v^{(1)} = v$, $w^{(1)} = w$. For the inductive step we show that, if there exists a triplet representation $(u^{(k)}, v^{(k)}, w^{(k)})$ of $A^{(k)}$, then $A^{(k+1)}$ also has a triplet representation $(u^{(k+1)}, v^{(k+1)}, w^{(k+1)})$. Firstly, we note that if $A^{(k)}$ has a triplet representation, then it holds that $\mathrm{offdiag}(B_k) \le 0$, $z_k, s_k \ge 0$, and $A^{(k)}v^{(k)} = w^{(k)}$ with $v^{(k)} > 0$, $w^{(k)} \ge 0$. Thus, partitioning $v^{(k)} = \begin{bmatrix} v_1^{(k)} \\ v_2^{(k)} \end{bmatrix}$ and $w^{(k)} = \begin{bmatrix} w_1^{(k)} \\ w_2^{(k)} \end{bmatrix}$ according to $A^{(k)}$, we get

$$\begin{cases} \alpha_k v_1^{(k)} - s_k^T v_2^{(k)} &= w_1^{(k)} \\ -v_1^{(k)} z_k + B^{(k)} v_2^{(k)} &= w_2^{(k)} \end{cases} \tag{2.10}$$

From the first equation it follows that $\alpha_k > 0$, since

$$\alpha_k = \frac{s_k^T v_2^{(k)} + w_1^{(k)}}{v_1^{(k)}}$$

and $\alpha_k \neq 0$ for assumption of well-definition of the sequence $\{A^{(k)}\}_{1\le k\le n}$. We can now derive a triplet representation of $A^{(k+1)}$. Indeed, by defining $u^{(k+1)} = \mathrm{offdiag}(A^{(k+1)})$, we have that $u^{(k+1)} \le 0$, by definition of $A^{(k+1)}$ as in (2.9) and by the fact that $\mathrm{offdiag}(B_k) \le$

0, $z_k, s_k \geq 0$ and $\alpha_k > 0$. In addition, if we define $v^{(k+1)} = v_2^{(k)}$, then $v^{(k+1)} > 0$ for the assumption of the positivity of $v^{(k)}$. Now we can conclude if $w^{(k+1)} \geq 0$, where $w^{(k+1)}$ is defined by $w^{(k+1)} = A^{(k+1)}v^{(k+1)}$. From (2.9) and (2.10):

$$\begin{aligned} w^{(k+1)} = A^{(k+1)}v^{(k+1)} &= B^{(k)}v_2^{(k)} - \frac{z_k s_k^T}{\alpha_k} v_2^{(k)} \\ &= w_2^{(k)} + v_1^{(k)} z_k - \frac{z_k s_k^T}{\alpha_k} v_2^{(k)} \\ &= w_2^{(k)} + \frac{w_1^{(k)}}{\alpha_k} z_k. \end{aligned} \tag{2.11}$$

Since all the addenda in last relation are nonnegative, the proof is completed. □

Next corollary specifies the definition of the triplet $(u^{(k)}, v^{(k)}, w^{(k)})$ for the matrix $A^{(k)}$:

Corollary 2.5.1. *Under the assumptions of Theorem 2.5 the sequences* $\{u^{(k)}\}_{1\leq k\leq n}$, $\{v^{(k)}\}_{1\leq k\leq n}$, $\{w^{(k)}\}_{1\leq k\leq n}$, *starting with* $u^{(1)} = u$, $v^{(1)} = v$, $w^{(1)} = w$, *are defined by*

$$u^{(k)} = -\mathit{offdiag}(A^{(k)}), \tag{2.12}$$

$$v^{(k+1)} = v_2^{(k)}, \tag{2.13}$$

$$w^{(k+1)} = w_2^{(k)} + \frac{w_1^{(k)}}{\alpha_k} z_k, \tag{2.14}$$

and $v^{(k)}$, $w^{(k)}$ *are partitioned according to* $A^{(k)}$ *and they have the form*

$$v^{(k)} = \begin{bmatrix} v_1^{(k)} \\ v_2^{(k)} \end{bmatrix}, \; w^{(k)} = \begin{bmatrix} w_1^{(k)} \\ w_2^{(k)} \end{bmatrix},$$

with $v_1^{(k)}, w_1^{(k)} \in \mathbb{R}$ *and* $v_2^{(k)}, w_2^{(k)} \in \mathbb{R}^{n-k}$.

Proof. The statement follows from Theorem 2.5. □

Remark 2.6. We know from Corollary 1.4.1 that if A is a reducible singular M-matrix having a triplet representation, it is possible to perform a Gaussian elimination with pivoting [11] that produces an LU factorization of $\Pi A \Pi^T$, where Π is a suitable permutation matrix. It is easily seen that it is not sufficient to guarantee that $A^{(k)}$ has a triplet representation for the necessary number of steps: indeed, the permutation of rows or columns does not preserve the nonpositivity of the offdiagonal elements. Nevertheless, if $\alpha_k = 0$ for any $1 \leq k < n$, and $z_k = 0$, the Gaussian process can be carried out and all the matrices of the sequence $\{A^{(k)}\}_{1\leq k\leq n}$ have a triplet representation (permutation is not required).

Remark 2.7. The process of GTH algorithm described in previous section is based on the fact that each matrix of the sequence $\{A^{(k)}\}_{1\leq k\leq n}$ has a (left) triplet representation:

$$A^{(k)} = (u^{(k)}, v^{(k)}, w^{(k)}), \text{ for all } k \geq 0,$$

where $u^{(k)} = \text{offdiag}(-A^{(k)})$, ${v^{(k)}}^T A^{(k)} = {w^{(k)}}^T$, with $v^{(k)} = e$ and $w^{(k)} = 0$ of length $n-k+1$.

The crucial importance of the existence of a triplet representation for all the matrices $A^{(k)}$ relies in the fact that in this case all the process of Gaussian elimination can be performed with no subtractions. Indeed, we can construct the matrix $A^{(k)}$ by using the relation (2.9) and by noticing that:

- the offdiagonal entries of $A^{(k)}$ are the elements of the matrix $-P^{(k)}$, where $P^{(k)}$ is the nonnegative matrix containing all the offdiagonal elements of $-A^{(k)}$ and zeros on the diagonal (so $A^{(k)} = D^{(k)} - P^{(k)}$, where $D^{(k)}$ is a diagonal matrix with $d_{ii}^{(k)} = a_{ii}^{(k)}$);

- the pivotal element is:

$$\alpha_k = \frac{s_k^T v_2^{(k)} + w_1^{(k)}}{v_1^{(k)}}, \tag{2.15}$$

  where the j-th element of s_k^T is the $(j+1)$-th element of the first row of $P^{(k)}$.

We report here the resulting Algorithm 1 for the computation of the LU factorization of A in the case where A is a nonsingular or an irreducible singular M-matrix (see [1]). For easier comprehension of the pseudocode, we observe that the i-th element of $v^{(k)}$ is equal to the $(k+i-1)$-th element of $v^{(1)}$ and in (2.14) the vector z_k contains the last $n-k$ elements of the first column of $P^{(k)}$. Note that the diagonal elements are computed just before they are needed.

One of the most important applications of the GTH-algorithm in the case in which A is a nonsingular M-matrix is the accurate computation of the inverse A^{-1}, by solving one column at a time the matrix equation $AX = I$.

Thus, we conclude this section with the solution method for linear system $Ax = b$, where A is a nonsingular M-matrix and b is a nonnegative vector (see Algorithm 2). The pseudocode corresponds to that of the Algorithm 1 with the addition of the backward and forward substitution for the computation of the solution of the linear system. We observe that these substitutions too are subtraction-free.

Algorithm 1: GTH-like algorithm for the computation of the LU factorization of A, where A is a nonsingular or an irreducible singular M-matrix

Input: $P \in \mathbb{R}^{n\times n}$, $v = (v(1), \ldots, v(n))^T$, $w = (w(1), \ldots, w(n))^T$, where offdiag$(P) = u$ and (u, v, w) is a triplet representation of A, satisfying $P \geq 0$ (with null diagonal entries), $v > 0$, $w \geq 0$.

Output: L, U

```
1  L ← I_n; U ← 0_n;
   // Perform the LU factorization
2  for k = 1 : n − 1 do
3      α(k) ← (w(k) + P(k, k + 1 : n)v(k + 1 : n))/v(k); // see equation (2.15)
4      for i = k + 1 : n do
5          w(i) ← w(i) + w(k)P(i, k)/α(k);// see equation (2.14)
6          for j = k + 1 : n do
7              if i ≠ j then
8                  P(i, j) ← P(i, j) + (P(i, k)P(k, j))/α(k);// see equation (2.9)
9  α(n) = w(n)/v(n);
   // Restore L and U
10 for i = 1 : n do
11     for j = 1 : n do
12         if i == j then
13             U(i, i) ← α(i);
14         else if i < j then
15             U(i, j) ← −P(i, j);
16         else
17             L(i, j) ← −P(i, j)/α(j);
18 end;
```

Algorithm 2: GTH-like algorithm for solving $Ax = b$, where A is a nonsingular M-matrix

Input: $b \in \mathbb{R}^n$, $b \geq 0$ and $P \in \mathbb{R}^{n \times n}$, $v = (v(1), \ldots, v(n))^T$, $w = (w(1), \ldots, w(n))^T$, where offdiag$(P) = u$ and (u, v, w) is a triplet representation of A, satisfying $P \geq 0$ (with null diagonal entries), $v > 0$, $w \geq 0$.

Output: x

```
1  α ← 0_n; y ← 0_n; x ← 0_n;
   // Perform the LU factorization
2  for k = 1 : n − 1 do
3      α(k) ← (w(k) + P(k, k + 1 : n)v(k + 1 : n))/v(k); // see equation (2.15)
4      for i = k + 1 : n do
5          w(i) ← w(i) + w(k)P(i, k)/α(k);// see equation (2.14)
6          for j = k + 1 : n do
7              if i ≠ j then
8                  P(i, j) ← P(i, j) + (P(i, k)P(k, j))/α(k);// see equation (2.9)
9  α(n) = w(n)/v(n);
   // Compute the solution x by solving Ly = b and Ux = y
   // Forward substitution
10 y(1) ← b(1)/α(1);
11 for k = 2 : n do
12     y(k) ← b(k) + P(k, 1 : (k − 1))y(1 : (k − 1));
13     y(k) ← y(k)/α(k);
   // Backward substitution
14 x(n) ← y(n);
15 for k = (n − 1) : −1 : 1 do
16     x(k) ← y(k) + (P(k, k + 1 : n)x(k + 1 : n))/α(k);
17 end;
```

2.3 Error analysis

This section is devoted to reporting the most important results about error analysis concerning GTH and GTH-like algorithm: it justifies the interest in the study of computational problems related to M-matrices having a triplet representation.

In [24] O'Cinneide originally developed an entrywise perturbation theorem for Markov chains. The error bound he obtained has been improved by Xue in [36]. Later, Alfa et al. ([2]) showed a more general theorem concerning diagonally dominant M-matrices: the next lemma is a slightly improvement of both O'Cinneide and Xue results.

Lemma 2.8 ([2, Corollary 2.4]). *Let P and $\tilde{P}$ be two stochastic matrices with respectively*

stationary distributions $\pi = [\pi_1, \ldots, \pi_n]$ *and* $\tilde{\pi} = [\tilde{\pi_1}, \ldots, \tilde{\pi_n}]$. *If* $|P - \tilde{P}| \leq \epsilon P$, *then*

$$\left(\frac{1-\epsilon}{1+\epsilon}\right)^{n-1} \pi_i \leq \tilde{\pi}_i \leq \pi_i \left(\frac{1+\epsilon}{1-\epsilon}\right)^{n-1}, \quad 1 \leq i \leq n$$

The lemma above ensures us that the solution of the linear system $Ax = 0$ where $A = I - P$ (under the usual assumptions on P: P nonnegative, irreducible, stochastic) is well determined by the floating point representation of the data. We have included this result in our discussion because it is a justification and a significant starting point for the development of next research on variant of GTH algorithm.

An analogue of lemma 2.8 in the case of the solution of the linear system $Ax = b$ with $b \geq 0$ is a minor modification of Theorem 2.5 in [1] (that worked with $v = e$):

Theorem 2.9 ([40, Theorem 2.2]). *Let* $A, \tilde{A} \in \mathbb{R}^{n \times n}$ *be a nonsingular M-matrix and a perturbation of it, respectively. If there exist* $\epsilon \in \mathbb{R}$ *and* $v \in \mathbb{R}^n$ *such that* $0 \leq \epsilon < 1$, $v > 0$, *and*

$$|a_{ij} - \tilde{a}_{ij}| \leq \epsilon |a_{ij}| \ \text{ for } i \neq j \ \text{ and } \ |Av - \tilde{A}v| \leq \epsilon Av,$$

then $\tilde{A}$ *is a nonsingular M-matrix, and*

$$\frac{(1-\epsilon)^{n-1}}{(1+\epsilon)^n} A^{-1} \leq \tilde{A}^{-1} \leq \frac{(1+\epsilon)^{n-1}}{(1-\epsilon)^n} A^{-1}$$

.

The crucial consequence of this is that, if a triplet representation of $A = (u, v, w)$ is available then performing the computation of A^{-1} by using (u, v, w) is numerically advantageous because, if the entrywise perturbation of the triplet is small, then the entries of the inverse can be computed with high entrywise accuracy.

Remark 2.10. We observe that in general w numerically is not exactly Av, but an approximation of it. However it is not a real trouble, because it is possible to show that this approximation always can be computed with high elementwise accuracy. We refer to [40] for an exhaustive treatment of this aspect of the problem.

The possibility of computing an elementwise accurate solution of $Ax = b$ suggested by Theorem 2.9 is realized by GTH-like algorithm:

Theorem 2.11 ([1, Theorem 3.1]). *Suppose Algorithm 2 is carried out in floating point arithmetic with machine precision* u. *Then, the computed approximation* $\tilde{x}$ *of the exact solution* $x \geq 0$ *satisfies the componentwise inequality*

$$|\tilde{x} - x| \leq (\phi(n)\mathsf{u} + O(\mathsf{u}^2))x,$$

where $\phi(n) := \frac{2}{3}(2n+5)(n+2)(n+3)$.

Remark 2.12. We point out the difference between the bound in the theorem above and the usual error bound for the standard Gaussian elimination: it is a well-known fact that latter depends on the condition number of the matrix A (see, for instance, [10]).

Chapter 3

Accurate doubling algorithms for M-NARE

This chapter is devoted to analysing the accurate doubling algorithms for M-NARE. The doubling methods for the computation of the minimal nonnegative solution of M-NARE have been extensively studied over the last two decades. The first structure-preserving doubling algorithm (SDA) was proposed by X. Guo et al. in 2006 [17]. Later, SDA has been improved by two variants known as SDA-ss [8] and ADDA [34] that differ only for the initial setup: ADDA is most recent, general and fast among all doubling algorithms derivable from bilinear transformations. Accurate implementations of ADDA have been given in [23] for an M-NARE defined by an irreducible singular M-matrix and in [38] for a nonsingular M-matrix.

We recall here the fundamental notions about doubling algorithms and its accurate implementations, starting by a presentation of the most important spectral features of the matrix $\mathcal{H}$ on which it depends the rate of convergence of the doubling algorithms.

3.1 Spectral properties of $\mathcal{H}$

In this section we examine some interesting spectral properties of the matrix $\mathcal{H}$ under suitable assumptions on M. Firstly, we need a preliminary result:

Lemma 3.1 ([14]). *If M is a singular M-matrix with a simple zero eigenvalue and having a triplet representation, then there are two nonnegative nonzero vectors $\nu^T = [\nu_1^T, \nu_2^T]$, and $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, such that*

$$\nu^T M = 0, \quad M\mu = 0,$$

with $\nu_1, \mu_1 \in \mathbb{R}^n$ and $\nu_2, \mu_2 \in \mathbb{R}^m$. The vectors ν, μ are unique up to scalar multiplication.

We notice that, if M is an irreducible singular M-matrix, then M satisfies the hypothesis of the lemma above (cfr. Theorem 1.9).

Remark 3.2. Similarly to what has been done in [21], the matrix M can be defined *strongly regular* if is a regular M-matrix and if $\text{rank}(M) \geq m+n-1$, and *super-regular* if is a nonsingular or an irreducible singular M-matrix.

We recall that under the hypothesis that the matrix M of (1.5) is a regular M-matrix, there exist minimal solutions Φ and Ψ to the M-NARE equation (1.7) and its dual, respectively (cfr. Theorem 1.12).

Since the convergence of the doubling algorithms applied to the M-NARE (1.7) can be related to the spectral radius of the two matrices $R = A - B\Phi$ and $S = D + C\Psi$, we recap some properties of $\mathcal{H}$ and then give a more precise relationship between the eigenvalues of R and S and the ones of $\mathcal{H}$.

Theorem 3.3 (Theorems 6-10, Lemma 5, Lemma 8, [14])**.** *Let M be an M-matrix having a triplet representation and let $\lambda_1, \ldots, \lambda_{m+n}$ be the eigenvalues of $\mathcal{H}$ arranged by a nondecreasing order by their real parts. Then the spectra of R and S are $\sigma(R) = \{\lambda_{m+1}, \ldots, \lambda_{m+n}\}$ and $\sigma(S) = \{-\lambda_1, \ldots, -\lambda_m\}$, respectively.*
In particular, λ_m and λ_{m+1} are real numbers and $\lambda_m \leq 0 \leq \lambda_{m+1}$, with strict inequalities if M is a nonsingular M-matrix.
In addition, if M is a regular singular M-matrix and 0 is a simple eigenvalue of M, then:

- *if $\nu_1^T \mu_1 = \nu_2^T \mu_2$, 0 is a simple eigenvalue of R and S, thus $\lambda_m = \lambda_{m+1} = 0$;*
- *if $\nu_1^T \mu_1 \neq \nu_2^T \mu_2$, 0 is a simple eigenvalue of $\mathcal{H}$. In particular, if $\nu_1^T \mu_1 > \nu_2^T \mu_2$ then $\lambda_{m+1} = 0$ and $\lambda_m \neq 0$; $\nu_1^T \mu_1 < \nu_2^T \mu_2$ then $\lambda_m = 0$ and $\lambda_{m+1} \neq 0$.*

We give here a different version of Lemma 7 of [16]:

Corollary 3.3.1. *With the notations of Theorem 3.3 and assuming that M is a regular M-matrix, we have that*

1. *$R = A - B\Phi$ is a regular M-matrix, and its spectrum is such that*

$$\sigma(R) \subset \{|z - \beta| \leq \beta\} \subset \{\text{Re}(z) \geq 0\},$$

   where β is a scalar such that $\beta \geq \max_i a_{ii}$;

2. *$S = D + C\Psi$ is a regular M-matrix, and the spectrum of $-S$ is such that*

$$\sigma(-S) \subset \{|z + \alpha| \leq \alpha\} \subset \{\text{Re}(z) \leq 0\},$$

   where α is a scalar such that $\alpha \geq \max_i d_{ii}$.

In particular,

- *if M is a nonsingular M-matrix, then $\sigma(R) \subset \{|z-\beta| < \beta\}$ and $\sigma(-S) \subset \{|z-\alpha| < \alpha\}$;*

- *if M is a regular singular M-matrix and 0 is a simple eigenvalue of M, then $\sigma(R) \subset \{|z-\beta| < \beta\}$ in the case $\nu_1^T \mu_1 < \nu_2^T \mu_2$, and $\sigma(-S) \subset \{|z-\alpha| < \alpha\}$ in the case $\nu_1^T \mu_1 > \nu_2^T \mu_2$.*

Proof. We need to prove that $\sigma(R) \subset \{|z-\beta| \leq \beta\}$ and $\sigma(-S) \subset \{|z+\alpha| \leq \alpha\}$. We prove the statement for R, the proof for $-S$ is analogous: since β is such that $\beta \geq \max_i a_{ii}$, we can write $R = \beta I - P$ for a nonnegative matrix P. As R is an M-matrix, $\beta \geq \rho(P)$, and hence the result follows. In the case in which R is nonsingular, $\beta > \rho(P)$, thus the strict inequality $|z-\beta| < \beta$ holds. □

Hence, the imaginary axis splits the eigenvalues of $\mathcal{H}$ in the two sets of m and n eigenvalues, respectively. We prove that the circle $\{|z| = \frac{\beta}{\alpha}\}$ separates their images under f. These separation properties are called *splittings* in [6].

Lemma 3.4. *Under the same assumptions as Corollary 3.3.1, and for $f(z) = (z-\beta)(z+\alpha)^{-1}$, we have*

1. $\sigma(f(R)) = \{f(\lambda_{m+1}), f(\lambda_{m+2}), \ldots, f(\lambda_{m+n})\} \subset \{|z| \leq \frac{\beta}{\alpha}\}$.
2. $\sigma(f(-S)) = \{f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_m)\} \subset \{|z| \geq \frac{\beta}{\alpha}\}$.

In addition, if M is a regular nonsingular M-matrix, or a regular singular M-matrix and 0 is a simple eigenvalue of M and $\nu_1^T \mu_1 \neq \nu_2^T \mu_2$ (i.e. non critical case), then

$$r = \rho(f(R))\rho(f(-S)^{-1}) = \left| \frac{f(\lambda_{m+1})}{f(\lambda_m)} \right| < 1;$$

if M is a regular singular M-matrix and 0 is a simple eigenvalue of M and $\nu_1^T \mu_1 = \nu_2^T \mu_2$ (i.e. critical case) then

$$r = \rho(f(R))\rho(f(-S)^{-1}) = \left| \frac{f(\lambda_{m+1})}{f(\lambda_m)} \right| = 1.$$

Proof. As the map f is a linear fractional (Möbius) transformation, it maps circles and lines into circles and lines; moreover, as its coefficients are real, circles that have their center on the real line (and hence are symmetric with respect to the real axis) are mapped into circles with the same property, with vertical lines as degenerate case.

We use the fact that for any matrix Q, with eigenvalues $\tau_1, \ldots, \tau_\ell$, counted with multiplicity, the eigenvalues of $f(Q)$ are $f(\tau_1), \ldots, f(\tau_\ell)$.

By Lemma 3.3.1, $\sigma(R)$ belongs to the closed disk with diameter $[0, 2\beta]$; the latter region is mapped by f inside the disk with diameter $[f(0), f(2\beta)] = [-\frac{\beta}{\alpha}, \frac{\beta}{2\beta+\alpha}]$, which is included in $\{|z| \leq \frac{\beta}{\alpha}\}$ as $\frac{\beta}{2\beta+\alpha} < \frac{\beta}{\alpha}$. Analogously, $\sigma(-S)$ belongs to the closed disk with diameter $[-2\alpha, 0]$, which is mapped into the *exterior* (as the region includes the pole $-\alpha$ of f) of the circle with diameter $[0, f(-2\alpha)] = [-\frac{\beta}{\alpha}, \frac{2\alpha+\beta}{\alpha}]$; this exterior is included in $\{|z| \geq \frac{\beta}{\alpha}\}$ as $\frac{2\alpha+\beta}{\alpha} > \frac{\beta}{\alpha}$.

We know that $f(R) = (R - \beta I)(R + \alpha I)^{-1} \leq 0$, since $R - \beta I \leq 0$ by definition of β and $R + \alpha I$ is an M-matrix. If v is the Perron vector of R with eigenvalue λ_{m+1},

then it is also the Perron vector of $f(R)$ with eigenvalue $f(\lambda_{m+1})$, and hence $\rho(f(R)) = f(\lambda_{m+1}) \leq \beta/\alpha$. Analogously, $f(-S)^{-1} = (-S+\alpha I)(-S-\beta I)^{-1} \leq 0$, since $-S+\alpha I \geq 0$ by definition of α and $S+\beta I$ is an M-matrix. Then $\rho(f(-S)^{-1})$ is the Perron eigenvalue of this matrix, that is $f(\lambda_m)^{-1} \leq \alpha/\beta$.

We have that $\rho(f(R))\rho(f(-S)^{-1}) \leq 1$ and the equality holds if and only if $\rho(f(R)) = \rho(f(-S)) = \beta/\alpha$ and this happens only when $\lambda_m = \lambda_{m+1} = 0$. □

This splitting property of the spectrum of $f(\mathcal{H})$ appears really useful for better understanding the results about convergence theory presented in next section.

3.2 Doubling algorithms for M-NARE

In this section we outline some information about the doubling algorithm for M-NARE with a special focus on the convergence theory for M-matrices having triplet representation (also called regular) described by Guo [14] and Guo and Lu [16].

Doubling algorithms can be seen as a way to implicitly construct, through an iteration, the factorization

$$f(\mathcal{H})^{2^k} = \begin{bmatrix} I_m & -G_k \\ 0 & F_k \end{bmatrix}^{-1} \begin{bmatrix} E_k & 0 \\ -H_k & I_n \end{bmatrix}$$

for $k = 0, 1, 2, \ldots$ and $f(\mathcal{H}) = (a\mathcal{H} + b)(c\mathcal{H} + d)^{-1}$, with $a, b, c, d \in \mathbb{R}$.

In order to describe the general doubling iteration, we define as in [23] the partition of the matrix $P \in \mathbb{R}^{N\times N}$, with $N = n + m$, and the *doubling map* $\mathcal{F} : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$

$$P = \begin{bmatrix} E & G \\ H & F \end{bmatrix}, \quad \mathcal{F}(P) = \begin{bmatrix} \widetilde{E} & \widetilde{G} \\ \widetilde{H} & \widetilde{F} \end{bmatrix},$$

where $E, \widetilde{E} \in \mathbb{R}^{n\times n}$, $F, \widetilde{F} \in \mathbb{R}^{m\times m}$, with $N = n + m$, and

$$\widetilde{E} = E(I - GH)^{-1}E, \tag{3.1}$$
$$\widetilde{F} = F(I - HG)^{-1}F, \tag{3.2}$$
$$\widetilde{G} = G + E(I - GH)^{-1}GF, \tag{3.3}$$
$$\widetilde{H} = H + F(I - HG)^{-1}HE. \tag{3.4}$$

We note that $\mathcal{F}$ is well-defined if $I - GH$ (or equivalently $I - HG$) is nonsingular. Applying the doubling map iteratively with a given matrix P_0, one obtains the sequence $P_k = \mathcal{F}^{\circ k}(P_0)$, for $k = 1, 2, \ldots$

The first appearance of the doubling algorithm is the structure-preserving doubling algorithm (SDA) for nonsymmetric algebraic Riccati equations (1.7), originally developed by Guo et al. [17]. Later, two variants have been proposed by Bini et al. [8] and Wang et al. [34], named SDA-ss and ADDA respectively. These algorithms share the iterative

process but use different functions f and hence differ in the initialization step. If we define

$$\alpha_{\text{opt}} = \max_i d_{ii}, \qquad \beta_{\text{opt}} = \max_i a_{ii}, \tag{3.5}$$

then SDA corresponds to choosing $f(z) = (z-\alpha)(z+\alpha)^{-1}$ with $\alpha = \max(\alpha_{\text{opt}}, \beta_{\text{opt}})$, SDA-ss uses $f(z) = z - \beta_{\text{opt}}$, while in ADDA, $f(z) = (z-\beta)(z+\alpha)^{-1}$ where α and β are two nonnegative reals not both being zero, such that

$$\alpha \geq \alpha_{\text{opt}}, \qquad \beta \geq \beta_{\text{opt}}. \tag{3.6}$$

The best results in terms of convergence speed are obtained with ADDA, that we will consider from now on. One typically chooses $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$; indeed, this choice gives the fastest convergence, as shown in [34].

Now we discuss on how to apply the doubling algorithm, in the ADDA variant, to the M-NARE (1.7). For the moment we do not make any assumption on the matrix M of (1.5).

The initial values, that correspond to the ADDA algorithm, can be computed with the following formula that is analogous to (3.5) of [38] (derived from [27], or [23]):

$$P_0 = \begin{bmatrix} E_0 & G_0 \\ H_0 & F_0 \end{bmatrix} = -M_{\alpha,\beta}^{-1} M_{-\beta,-\alpha}, \tag{3.7}$$

where

$$M_{\alpha,\beta} = \begin{bmatrix} A+\alpha I & -B \\ C & D+\beta I \end{bmatrix}, \quad M_{-\beta,-\alpha} = \begin{bmatrix} A-\beta I & -B \\ C & D-\alpha I \end{bmatrix}. \tag{3.8}$$

When M is a nonsingular M-matrix or an irreducible singular M-matrix, the convergence theory of SDA and ADDA has been widely studied: for instance we refer to [34], [17], [15], [6], [20].

The following theorem on the convergence of the doubling algorithm for the non critical case with initial setup (3.7) holds (see [20], [34], [15]):

Theorem 3.5. *Let M be a nonsingular or an irreducible singular M-matrix, and let $\{E_k\}_{k\geq 0}$, $\{F_k\}_{k\geq 0}$, $\{G_k\}_{k\geq 0}$ and $\{H_k\}_{k\geq 0}$ be the sequences generated by the iteration (3.1)-(3.4) with initial value (3.7). Then, in the non critical case, the ADDA is well defined with $I - G_k H_k$ and $I - H_k G_k$ being nonsingular M-matrices for each $k \geq 0$. Moreover $E_k \geq 0$, $F_k \geq 0$, $0 \leq H_k \leq H_{k+1} \leq \Phi$, $0 \leq G_k \leq G_{k+1} \leq \Psi$ for all $k \geq 0$, and*

$$\limsup_{k\to\infty} \|\Psi - G_k\|^{1/2^k} \leq r, \quad \limsup_{k\to\infty} \|\Phi - H_k\|^{1/2^k} \leq r, \tag{3.9}$$

where $r = \rho(f(R)) \cdot \rho(f(-S)^{-1}) < 1$.

More recently, the case M regular has been studied in [14] and [16]. We report here the principal results for (irreducible and reducible) regular singular M-matrices, under suitable assumptions. For the non critical case we have:

Theorem 3.6 ([16, Theorem 3]). *Let M be a regular singular M-matrix with a simple zero eigenvalue and $\nu_1^T \mu_1 \neq \nu_2^T \mu_2$. Assume that $\alpha \geq \max_i d_{ii} > 0$ and $\beta \geq \max a_{ii} > 0$. Then the ADDA is well defined with $I - G_k H_k$ and $I - H_k G_k$ being nonsingular M-matrices for each $k \geq 0$. Moreover $E_k \geq 0$, $F_k \geq 0$, $0 \leq H_k \leq H_{k+1} \leq \Phi$, $0 \leq G_k \leq G_{k+1} \leq \Psi$ for all $k \geq 0$, and*

$$\limsup_{k\to\infty} \|\Psi - G_k\|^{1/2^k} \leq r, \quad \limsup_{k\to\infty} \|\Phi - H_k\|^{1/2^k} \leq r,$$

where $r = \rho(f(R)) \cdot \rho(f(-S)^{-1}) < 1$.

Next theorem is the first result with a rigorous convergence analysis of ADDA in the critical case.

Theorem 3.7 ([16, Theorem 8]). *Let M be a regular singular M-matrix with a simple zero eigenvalue and $\nu_1^T \mu_1 = \nu_2^T \mu_2$. Assume that $\alpha > \max_i d_{ii}$ and $\beta > \max a_{ii}$. Then, by using the notation of Theorem 3.5,*

$$H_k - \Phi = O(2^{-k}), \quad G_k - \Psi = O(2^{-k}).$$

Thus, in the critical case ADDA always converges at least linearly with a linear convergence rate $1/2$, regardless of the choices of the parameters α and β. The authors point out that in the critical case the "optimal" choice $\alpha = \alpha_{\text{opt}}$ and $\beta = \beta_{\text{opt}}$ does not guarantee the well-definedness of ADDA.

In order to improve the convergence rate r a shift technique has been proposed in [15]: the description of this method, and the study of an accurate algorithm that realizes it, will be the topic of the Chapter 4.

3.3 Accurate doubling algorithm

We report in this section the accurate implementations of doubling algorithm for M-NARE, proposed by Nguyen and Poloni [23] and Xue and Li [38]. Before describing the so-called accADDA, we examine the triplet representation of all the matrices to be inverted in the iterative process of ADDA and, at last, we point out a technique suggested in [38] in order to make safe the unavoidable subtractions.

3.3.1 Triplet representations

In [23] Nguyen et al. provided triplet representations for all the matrices to be inverted in ADDA in the case where the matrix M in (1.5) is an irreducible singular M-matrix such that $Mv = 0$, with $v = e$. Later, Xue et al. [38] generalized the result for M

nonsingular or irreducible singular M-matrix such that $Mv = w$, where $v > 0$, $w \geq 0$ (with $w > 0$ in the nonsingular case and $w = 0$ in the irreducible singular case). We state here similar results with our notation and under the assumption that M is a regular M-matrix, nonsingular or singular (irreducible or not) having a simple zero eigenvalue.

Next theorem provides a triplet representation of $M_{\alpha,\beta}$:

Theorem 3.8. *Let M be a regular singular M-matrix having a simple zero eigenvalue or a nonsingular M-matrix and let $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ be a positive vector such that $Mv = t \geq 0$. Let $M_{\alpha,\beta}$ be the matrix to be inverted in the initial step (3.7). Assume that:*

- *$\alpha \geq \max_i d_{ii}$ and $\beta \geq \max_i a_{ii}$, if M is singular and we are in the noncritical case with $m, n \geq 2$ or if M is nonsingular;*
- *$\alpha > \max_i d_{ii}$ and $\beta > \max_i a_{ii}$ if M is singular and we are in the critical case or we are in the noncritical case with $m = 1$ or $n = 1$.*

Then a triplet representation for $M_{\alpha,\beta}$ is

$$(offdiag(-M_{\alpha,\beta}), v, t + \begin{bmatrix} \alpha v_1 \\ \beta v_2 \end{bmatrix}). \tag{3.10}$$

In particular $M_{\alpha,\beta}$ is a nonsingular M-matrix.

Proof. We notice that a positive vector v such that $Mv = t \geq 0$ exists from the definition of regular matrix. In particular, $t > 0$ when M is nonsingular. So we have:

$$M_{\alpha,\beta}v = t + \begin{bmatrix} \alpha v_1 \\ \beta v_2 \end{bmatrix} \geq 0.$$

The relation above ensures that $M_{\alpha,\beta}$ has a triplet representation. In order to show the nonsingularity of $M_{\alpha,\beta}$ it is sufficient that

$$w := t + \begin{bmatrix} \alpha v_1 \\ \beta v_2 \end{bmatrix} > 0.$$

If M is a regular nonsingular M-matrix, w is positive since $t > 0$. If M is a regular singular M-matrix having a simple zero eigenvalue, we prove that $\alpha, \beta > 0$. This is true for assumption in the non critical case when $m = 1$ or $n = 1$ and in the critical case. In the non critical case with $m, n \geq 2$, if M has a single zero eigenvalue, then M has at most one zero diagonal entry, thus $\alpha, \beta > 0$. □

We state here a lemma that points out a useful relation between P_k and P_0, without specific assumptions on M:

Lemma 3.9. *With the notations of Section 3.2, let*

$$P_0 = \begin{bmatrix} E_0 & G_0 \\ H_0 & F_0 \end{bmatrix}$$

be the initial value of a sequence $P_k = \mathcal{F}^{\circ k}(P_0) = \left[\begin{smallmatrix} E_k & G_k \\ H_k & F_k \end{smallmatrix}\right]$, *such that* $\mathcal{F}$ *is well-defined. If there exist* $v, w \in \mathbb{R}^N$ *such that* $(I - P_0)v = w$ *then*

$$(I - P_k)v = w^{(k)}, \quad \text{for all } k > 0 \tag{3.11}$$

where the sequence $w^{(k)}$ *is defined by*

$$\begin{aligned} w^{(0)} &= w, && (3.12)\\ w_1^{(k+1)} &= w_1^{(k)} + E_k(I - G_kH_k)^{-1}(w_1^{(k)} + G_kw_2^{(k)}), && (3.13)\\ w_2^{(k+1)} &= w_2^{(k)} + F_k(I - H_kG_k)^{-1}(w_2^{(k)} + H_kw_1^{(k)}). && (3.14) \end{aligned}$$

In addition, for all $k \geq 0$, *the matrices* $I - G_kH_k$ *and* $I - H_kG_k$ *are such that*

$$\begin{aligned} (I - G_kH_k)v_1 &= E_kv_1 + G_kF_kv_2 + w_1^{(k)} + G_kw_2^{(k)}, && (3.15)\\ (I - H_kG_k)v_2 &= F_kv_2 + H_kE_kv_1 + w_2^{(k)} + H_kw_1^{(k)}. && (3.16) \end{aligned}$$

Proof. We prove the result by induction on k. For $k = 0$, the first statement is true by the hypothesis, posing $w^{(0)} = w$. In order to prove the relation

$$(I - G_0H_0)v_1 = E_0v_1 + G_0F_0v_2 + w_1 + G_0w_2, \tag{3.17}$$

and the analogous

$$(I - H_0G_0)v_2 = F_0v_2 + H_0E_0v_1 + w_2 + H_0w_1 \tag{3.18}$$

we observe that $(I_N - P_0)v = w$ can be rewritten as

$$\begin{bmatrix} I_n & -G_0 \\ -H_0 & I_m \end{bmatrix} v = \begin{bmatrix} E_0 & 0 \\ 0 & F_0 \end{bmatrix} v + w. \tag{3.19}$$

Premultiplying by $[I_n, G_0]$ (and respectively by $[H_0, I_m]$) both sides of last equation, we obtain (3.17) (and respectively (3.18)). Now, we assume that the result holds for a certain k and prove it for $k + 1$. We can rewrite the inductive assumption in the form:

$$\begin{aligned} E_kv_1 + G_kv_2 &= v_1 - w_1^{(k)}\\ H_kv_1 + F_kv_2 &= v_2 - w_2^{(k)} \end{aligned}$$

By using the definition of the sequence P_k and the inductive hypothesis, we have

$$\begin{aligned} E_{k+1}v_1 + G_{k+1}v_2 &= E_k(I - G_kH_k)^{-1}(E_kv_1 + G_kF_kv_2) + G_kv_2\\ &= E_kv_1 + G_kv_2 - E_k(I - G_kH_k)^{-1}\left(w_1^{(k)} + G_kw_2^{(k)}\right)\\ &= v_1 - w_1^{(k+1)}. \end{aligned}$$

The proof of

$$H_{k+1}v_1 + F_{k+1}v_2 = v_2 - w_2^{(k+1)}$$

is analogous. □

Next lemma provides a similar relation for a slightly different initial setup of ADDA:

Lemma 3.10. *Under the assumptions of Theorem 3.8 and with the notations of Section 3.2, there exists a nonnegative vector $w \in \mathbb{R}^N$ such that $(I - \tilde{P}_0)v = w$, where*

$$\tilde{P}_0 = \begin{bmatrix} \frac{\alpha}{\beta}E_0 & G_0 \\ H_0 & \frac{\beta}{\alpha}F_0 \end{bmatrix}.$$

Proof. We notice that

$$\begin{aligned} \tilde{P}_0 &= \begin{bmatrix} \tilde{E}_0 & \tilde{G}_0 \\ \tilde{H}_0 & \tilde{F}_0 \end{bmatrix} && (3.20) \\ &= \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix} P_0 \begin{bmatrix} \frac{1}{\beta} & 0 \\ 0 & \frac{1}{\alpha} \end{bmatrix} && (3.21) \\ &= \begin{bmatrix} \frac{\alpha}{\beta}E_0 & G_0 \\ H_0 & \frac{\beta}{\alpha}F_0 \end{bmatrix}. && (3.22) \end{aligned}$$

In addition, from (3.21), we have that

$$\tilde{P}_0 = -\tilde{M}_{\alpha,\beta}^{-1}\tilde{M}_{-\beta,-\alpha}, \tag{3.23}$$

where

$$\tilde{M}_{\alpha,\beta} = \begin{bmatrix} \frac{1}{\alpha}A + I_n & -\frac{1}{\beta}B \\ \frac{1}{\alpha}C & \frac{1}{\beta}D + I_m \end{bmatrix}, \quad \tilde{M}_{-\beta,-\alpha} = \begin{bmatrix} \frac{1}{\beta}A - I_n & -\frac{1}{\alpha}B \\ \frac{1}{\beta}C & \frac{1}{\alpha}D - I_m \end{bmatrix}. \tag{3.24}$$

The matrix $\tilde{M}_{\alpha,\beta}^{-1}$ is a nonsingular M-matrix: indeed, we can provide its triplet representation with positive third vector by observing that

$$\tilde{M}_{\alpha,\beta} = \begin{bmatrix} \frac{1}{\alpha}A + I_n & -\frac{1}{\beta}B \\ \frac{1}{\alpha}C & \frac{1}{\beta}D + I_m \end{bmatrix} = \begin{bmatrix} \frac{1}{\alpha}A & -\frac{1}{\beta}B \\ \frac{1}{\alpha}C & \frac{1}{\beta}D \end{bmatrix} + I_N \tag{3.25}$$

thus, from $Mv = t$, we have

$$\tilde{M}_{\alpha,\beta}\begin{bmatrix} \alpha v_1 \\ \beta v_2 \end{bmatrix} = \begin{bmatrix} t_1 + \alpha v_1 \\ t_2 + \beta v_2 \end{bmatrix} > 0.$$

From (3.23) and (3.24) we get

$$\begin{aligned} I - \tilde{P}_0 &= I + \tilde{M}_{\alpha,\beta}^{-1}\tilde{M}_{-\beta,-\alpha} && (3.26) \\ &= \tilde{M}_{\alpha,\beta}^{-1}(\tilde{M}_{\alpha,\beta} + \tilde{M}_{-\beta,-\alpha}) && (3.27) \\ &= \tilde{M}_{\alpha,\beta}^{-1}\left(\frac{1}{\alpha} + \frac{1}{\beta}\right)M. && (3.28) \end{aligned}$$

From last equivalence and from $Mv = t$ it follows that

$$(I - \tilde{P}_0)v = \left(\frac{1}{\alpha} + \frac{1}{\beta}\right) \tilde{M}_{\alpha,\beta}^{-1} t,$$

where the vector $w := \left(\frac{1}{\alpha} + \frac{1}{\beta}\right) \tilde{M}_{\alpha,\beta}^{-1} t$ is nonnegative due to the nonsingularity of $\tilde{M}_{\alpha,\beta}^{-1}$ (see Lemma 1.2). □

Remark 3.11. Inductively it can be proved that the sequence $\tilde{P}_k = \mathcal{F}^{\circ k}(\tilde{P}_0) = \left[\begin{smallmatrix} \tilde{E}_k & \tilde{G}_k \\ \tilde{H}_k & \tilde{F}_k \end{smallmatrix}\right]$, under the assumption that $\mathcal{F}$ is well-defined, is given by:

$$\tilde{E}_k = \left(\frac{\alpha}{\beta}\right)^{2^k} E_k, \quad \tilde{F}_k = \left(\frac{\beta}{\alpha}\right)^{2^k} F_k, \quad \tilde{G}_k = G_k, \quad \tilde{H}_k = H_k, \text{ for } k \geq 0.$$

Remark 3.12. Analogously to$\{P_k\}_{k\geq 0}$, the sequence $\{\tilde{P}_k\}_{k\geq 0}$ can be obtained by considering a suitable function f in the definition of ADDA process. In this case we have the scaled function $\tilde{f} := \frac{\alpha}{\beta} f$, for which $\tilde{f}(0) = -1$. We observe that such a choice of $\tilde{f}$ corresponds to a splitting of the spectrum of $\tilde{f}(\mathcal{H})$ with respect to the unit disk in the complex plane.

So, we are ready to provide a triplet representation for $I - G_k H_k$ and $I - H_k G_k$:

Theorem 3.13 (based on [38, Section 3])**.** *Under the assumptions of Theorem 3.8 and Lemma 3.10, there exist triplet representations for $I - G_k H_k$ and $I - H_k G_k$, for all $k \geq 0$:*

$$(\mathit{offdiag}(G_k H_k), v_1, \left(\frac{\alpha}{\beta}\right)^{2^k} E_k v_1 + \left(\frac{\beta}{\alpha}\right)^{2^k} G_k F_k v_2 + \tilde{w}_1^{(k)} + G_k \tilde{w}_2^{(k)}) \tag{3.29}$$

and

$$(\mathit{offdiag}(H_k G_k), v_2, \left(\frac{\beta}{\alpha}\right)^{2^k} F_k v_2 + \left(\frac{\alpha}{\beta}\right)^{2^k} H_k E_k v_1 + \tilde{w}_2^{(k)} + H_k \tilde{w}_1^{(k)}), \tag{3.30}$$

where the sequence $\tilde{w}^{(k)}$ is defined by

$$\tilde{w}^{(0)} = \left(\frac{1}{\alpha} + \frac{1}{\beta}\right) \tilde{M}_{\alpha,\beta}^{-1} t \tag{3.31}$$

$$\tilde{w}_1^{(k+1)} = \tilde{w}_1^{(k)} + \left(\frac{\alpha}{\beta}\right)^{2^k} E_k (I - G_k H_k)^{-1} (\tilde{w}_1^{(k)} + G_k \tilde{w}_2^{(k)}), \tag{3.32}$$

$$\tilde{w}_2^{(k+1)} = \tilde{w}_2^{(k)} + \left(\frac{\beta}{\alpha}\right)^{2^k} F_k (I - H_k G_k)^{-1} (\tilde{w}_2^{(k)} + H_k \tilde{w}_1^{(k)}). \tag{3.33}$$

Proof. We observe that $v > 0$ by assumption and $E_k \geq 0$, $F_k \geq 0$, $G_k \geq 0$, $H_k \geq 0$, so offdiag$(G_k H_k) \geq 0$, offdiag$(H_k G_k) \geq 0$ for all $k \geq 0$. Thus, the expressions (3.29) and (3.30) are triplet representation for $I - G_k H_k$ and $I - H_k G_k$, respectively, if $w^{(k)} \geq 0$ for all $k \geq 0$. This is true since the vector $w^{(0)}$ is nonnegative and $w^{(k+1)} \geq w^{(k)}$. □

As a corollary, we have a triplet representation for the matrices to be inverted in accADDA, in the case M irreducible singular M-matrix:

Corollary 3.13.1 ([23, Theorem 4.2], [38, Section 3]). *Let M be an irriducible singular M-matrix s.t. $Mv = 0$ with $v > 0$, let $M_{\alpha,\beta}$ be the matrix to be inverted in the initial step (3.7) and let E_k, F_k, G_k, H_k defined as in (3.1)-(3.4).*

1. *A triplet representation for $M_{\alpha,\beta}$ is*

$$(\mathit{offdiag}(-M_{\alpha,\beta}), v, \begin{bmatrix} \alpha v_1 \\ \beta v_2 \end{bmatrix}) \tag{3.34}$$

2. *A triplet representation for $I - G_k H_k$ is*

$$(\mathit{offdiag}(G_k H_k), v_1, \left(\frac{\alpha}{\beta}\right)^{2^k} E_k v_1 + \left(\frac{\beta}{\alpha}\right)^{2^k} G_k F_k v_2). \tag{3.35}$$

3. *A triplet representation for $I - H_k G_k$ is*

$$(\mathit{offdiag}(H_k G_k), v_2, \left(\frac{\beta}{\alpha}\right)^{2^k} F_k v_2 + \left(\frac{\alpha}{\beta}\right)^{2^k} H_k E_k v_1). \tag{3.36}$$

3.3.2 Algorithm

The following algorithm proposed in [23] computes accurate solutions of (1.7)-(1.8) by using the GTH-like algorithm (Algorithm 2) for the inversion of $M_{\alpha,\beta}$ and $I - G_k H_k$, $I - H_k G_k$ in the case where M is an irreducible singular M-matrix, by using the triplet representations proposed in Corollary 3.13.1. We present here this version of the accurate ADDA implementation for sake of simplicity, but we refer to [21] for a variant for strongly regular M-matrices.

Algorithm 3: Elementwise accurate ADDA for M irreducible singular M-matrix

Input: M, v and ε, where M is an irreducible singular M-matrix as in (1.5), $v > 0$ s.t. $Mv = 0$ and ε is the convergence tolerance

Output: Φ and Ψ, minimal nonnegative solutions of (1.7) and (1.8)

1 set $\alpha \leftarrow \theta\alpha_{\text{opt}}$ and $\beta \leftarrow \theta\beta_{\text{opt}}$ for a moderate θ, e.g., $\theta = 1.1$, where α_{opt} and β_{opt} are defined in (3.5);

2 compute initial values E_0, F_0, G_0, H_0 according to (3.7), performing the matrix inversion with Algorithm 2 and triplet representation (3.34);

3 $k \leftarrow 0$;

4 **do**

5 compute $E_{k+1}, F_{k+1}, G_{k+1}, H_{k+1}$ according to (3.1)-(3.4), using Algorithm 2 for inversions and triplet representations (3.35)-(3.36);

6 $k \leftarrow k + 1$;

7 **while** *a suitable stopping criterion is not verified*

8 $\Psi \leftarrow G_k$; $\Phi \leftarrow H_k$

A perturbation analysis of the doubling algorithm is presented in [23, Section 7], for the case when $v = e$ (the vector of all ones) and $w = 0$. Their main result is the following, converted to our notation and keeping only the hypotheses used in the proof.

Theorem 3.14 ([23, Theorem 7.7]). *Let $P_0 \in \mathbb{R}^{(n+m)\times(n+m)}$ be stochastic (i.e., $P_0 \geq 0$ and $P_0 v = v$, where v is the vector of all ones). Suppose that the doubling algorithm with initial value P_0 is applicable and converges quadratically to $\lim H_k = \Phi$. In particular, let $\widehat{K}_0, \delta$ be constants such that $\Phi - H_k \leq \widehat{K}_0 \delta^{2^k} \Phi$ for each $k \geq 0$. Let $\tilde{P}_0$ be another stochastic matrix that satisfies the elementwise inequality $|\tilde{P}_0 - P_0| \leq \varepsilon P_0$, and let $\tilde{H}_k$ be the sequence generated by the doubling algorithm with initial value $\tilde{P}_0$. Then, for each $k \geq 0$ we have*

$$|\tilde{H}_k - H_k| \dot{\leq} \left(1 + \frac{4\widehat{K}_0(n+m)}{1-\delta}\right) \varepsilon \Phi.$$

A related result [23, Theorem 4.4] shows that the quantity H_k can be computed with small componentwise error using machine arithmetic, provided k is moderate. Together, these two results can be used to bound the componentwise forward error of a computed approximation $\tilde{H}_k \approx \Phi$, under the condition that the initial value $\tilde{P}_0$ is computed with a small componentwise forward error. While these results are only proved for v equal to the vector of all ones, it seems plausible that a small modification can yield a similar result for other choices of v.

3.3.3 Ensuring safe subtractions

Theorems 3.8-3.13 ensure that we can perform the steps of a doubling algorithm in a subtraction-free way. However, the initialization phase still requires subtractions in the computation of the diagonal of $M_{-\beta,-\alpha}$ in (3.8). When one takes $\alpha = \alpha_{\text{opt}}$, $\beta = \beta_{\text{opt}}$ one of the diagonal entries becomes exactly zero; however, there can still be catastrophic cancellation if other entries of $\text{diag}(A)$ or $\text{diag}(D)$ are close to the maximum one. Indeed, we start our computation from the machine representation $\tilde{A} = fl(A)$ of A, which we use to compute $\tilde{\beta}_{\text{opt}} = \max \text{diag}(\tilde{A})$ and then $\tilde{a}_{ii} \ominus \tilde{\beta}_{\text{opt}}$; the combined effect of these approximations can cause this quantity to have a large relative error over $a_{ii} - \tilde{\beta}_{\text{opt}}$, by the classical error analysis of subtractions.

A remedy to this type of problem is obtained by adapting a technique in [40]. In the initialization step of the doubling algorithm we will use $\tilde{\beta} = fl((1+\theta)\tilde{\beta}_{\text{opt}})$ instead of $\tilde{\beta} = \tilde{\beta}_{\text{opt}}$, for a sufficiently large $\theta > 0$, e.g., $\theta \approx 0.1$. This modification slightly degrades the convergence speed of the iteration, but at the same time ensures that there cannot be catastrophic cancellation.

We sketch the following forward error bound. Assume that θ is chosen so that $\tilde{\beta}$ is a floating point number, and set $\beta = (1+\theta)\beta_{\text{opt}}$. We have

$$\beta - a_{ii} = \theta\beta_{\text{opt}} + \beta_{\text{opt}} - a_{ii} \geq \theta\beta_{\text{opt}} = \frac{\theta}{1+\theta}\beta \geq \theta a_{ii},$$

from which it follows that

$$\frac{a_{ii}}{\beta - a_{ii}} \leq \frac{1}{\theta}, \text{ and } \frac{\beta}{\beta - a_{ii}} \leq \frac{\theta + 1}{\theta}$$

Hence, if we start from an approximation $\tilde{A}$ such that $|\tilde{a}_{ii} - a_{ii}| \leq \mathsf{u}\mathsf{a}a_{ii}$ holds for all $i = 1, 2, \ldots, n$ for a moderate constant $\mathsf{a} > 0$ we obtain (see also Lemma 4.15 for more details)

$$\begin{aligned} |\tilde{\beta} \ominus \tilde{a}_{ii} - (\beta - a_{ii})| &\leq |\tilde{\beta} \ominus \tilde{a}_{ii} - (\tilde{\beta} - \tilde{a}_{ii})| + |\tilde{\beta} - \beta| + |\tilde{a}_{ii} - a_{ii}| \\ &\leq \mathsf{u}(\tilde{\beta} - \tilde{a}_{ii}) + \mathsf{u}\mathsf{a}\beta + \mathsf{u}\mathsf{a}a_{ii} \\ &\leq \mathsf{u}\left(1 + \mathsf{a}\frac{2 + \theta}{\theta}\right)(\beta - a_{ii}) + O(\mathsf{u}^2). \end{aligned}$$

Thus $\beta - a_{ii}$ is computed with high relative accuracy. The same arguments apply to D and α, *mutatis mutandis*.

Chapter 4

Accurate doubling algorithms for shifted M-NARE

The aim of this chapter is to introduce an elementwise accurate doubling algorithm for shifted M-NARE. In [15] a shift technique has been proposed that improves the speed of convergence of the doubling algorithm by modifying the coefficients of the original Riccati equation. This approach is useful when the M-matrix defining the M-NARE is irreducible singular and it is especially advantageous when the corresponding matrix $\mathcal{H}$ has a zero eigenvalue with (algebraic) multiplicity 2: in the latter case, if no shift technique is performed, the order of convergence of the doubling algorithm is indeed linear instead of quadratic as usual.

We recall the main results presented in [15], [14], [16] and later we propose an elementwise accurate doubling algorithm for shifted M-NARE. It is important to point out here that it is not merely an application of the known accurate algorithms for M-NARE to the shifted case, but we needed developing some new ideas in order to guarantee the effectiveness of our approach.

4.1 Shift technique for M-NARE

In order to accelerate the convergence of the structure preserving doubling algorithm, Guo et al. in [15] proposed a shift technique based on the rank-one correction of $\mathcal{H}$ in the case where M is an irreducible singular M-matrix.

We define the shifted matrix $\widehat{\mathcal{H}}$ as

$$\widehat{\mathcal{H}} = \mathcal{H} + \eta v p^T, \tag{4.1}$$

where v is a nonnegative vector in the kernel of $\mathcal{H}$, $\eta > 0$ is a properly chosen scalar and $p \geq 0$ is a vector with $p^T v = 1$.

Partitioning p, v and $\widehat{\mathcal{H}}$ according to the structure of matrix M, we obtain

$$p^T = \begin{bmatrix} p_1^T & p_2^T \end{bmatrix}, \quad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix},$$

and we write

$$\widehat{\mathcal{H}} = \begin{bmatrix} \widehat{A} & -\widehat{B} \\ -\widehat{C} & -\widehat{D} \end{bmatrix}, \quad \widehat{M} = \begin{bmatrix} \widehat{A} & -\widehat{B} \\ \widehat{C} & \widehat{D} \end{bmatrix} \tag{4.2}$$

where

$$\begin{aligned} \widehat{A} &= A + \eta v_1 p_1^T, & \widehat{B} &= B - \eta v_1 p_2^T, \\ \widehat{C} &= C - \eta v_2 p_1^T, & \widehat{D} &= D - \eta v_2 p_2^T. \end{aligned} \tag{4.3}$$

Since $\widehat{M} = \mathcal{J}\widehat{\mathcal{H}}$ and $M = \mathcal{J}\mathcal{H}$, we observe that $\widehat{M} = M + \eta \mathcal{J} v p^T$.

An important fact is that the matrices $\widehat{\mathcal{H}}$ and $\mathcal{H}$ have the same spectrum except for one zero eigenvalue of $\mathcal{H}$ which is shifted to η in $\mathcal{H}$ (see Lemma 5.3 of [15]).

Assuming the notations of Section 3.1, according to Theorem 3.3, if M is a regular singular M-matrix having a simple zero eigenvalue (so, in particular, in the case where M is an irreducible singular M-matrix), we have three cases about the zero eigenvalues of $\mathcal{H}$:

- if $\nu_1^T \mu_1 = \nu_2^T \mu_2$, then $\lambda_m = \lambda_{m+1} = 0$;
- if $\nu_1^T \mu_1 > \nu_2^T \mu_2$ then $\lambda_{m+1} = 0$ and $\lambda_m \neq 0$;
- if $\nu_1^T \mu_1 < \nu_2^T \mu_2$ then $\lambda_m = 0$ and $\lambda_{m+1} \neq 0$.

For M irreducible singular M-matrix, Lemma 5.1 and Theorem 2.1 in [15] show that the case $\nu_1^T \mu_1 < \nu_2^T \mu_2$ can be reduced to the case $\nu_1^T \mu_1 > \nu_2^T \mu_2$. Thus in the following we consider only the case $\nu_1^T \mu_1 \geq \nu_2^T \mu_2$ and $\eta > 0$ without loss in generality.

The shifted NARE corresponding to $\widehat{M}$ and $\widehat{\mathcal{H}}$ is

$$X\widehat{B}X - X\widehat{A} - \widehat{D}X - \widehat{C} = 0 \tag{4.4}$$

and its dual equation is

$$\widehat{B} - \widehat{A}Y - Y\widehat{D} - Y\widehat{C}Y = 0. \tag{4.5}$$

A fundamental property holds:

Theorem 4.1 ([15, Theorem 5.4])**.** *Let M be an irreducible singular M-matrix. If $\nu_1^T \mu_1 \geq \nu_2^T \mu_2$, then $\widehat{X} = \Phi$ is a solution of the shifted equation* (4.4) *where Φ is the minimal nonnegative solution of the original M-NARE* (1.7)*.*
Moreover,

$$\sigma(\widehat{A} - \widehat{B}\Phi) = \{\eta, \lambda_{m+2}, \ldots, \lambda_{m+n}\}.$$

In addition, there exists $\widehat{Y}$ solution of (4.5) with interesting spectral properties:

Theorem 4.2 ([15, Theorems 5.8, 5.9]). *Let M be an irreducible singular M-matrix. If $\nu_1^T\mu_1 > \nu_2^T\mu_2$, or $\nu_1^T\mu_1 = \nu_2^T\mu_2$ and $p_1 > 0$, then the equation (4.5) has a solution $\widehat{Y}$ of such that*

$$\sigma(\widehat{D} + \widehat{C}\widehat{Y}) = \{-\lambda_1, \ldots, -\lambda_m\}.$$

The results above ensure that the spectrum of $\widehat{\mathcal{H}}$ consists in the union of the spectrum of $\widehat{R}$ and $-\widehat{S}$, where we have defined:

$$\widehat{R} := \widehat{A} - \widehat{B}\Phi, \quad \widehat{S} := \widehat{D} + \widehat{C}\widehat{Y}.$$

Analogously to what we have seen in Lemma 3.4 and in Section 3.2, the rate of convergence of the doubling algorithm applied to (4.4) is related to the product

$$\hat{r} = \rho(f(\widehat{R}))\rho(f(-\widehat{S})^{-1})$$

If the latter product $\hat{r}$ is smaller than $r = \rho(f(R))\rho(-f(S)^{-1})$ then there is an acceleration. This fact has been proved for the SDA, with $f(z) = (z-\gamma)(z+\gamma)^{-1}$ in [15], but a similar argument can be used also for the ADDA (see Corollary 4.8.1 below).

If $\nu_1^T\mu_1 > \nu_2^T\mu_2$, or $\nu_1^T\mu_1 = \nu_2^T\mu_2$ and $p_1 > 0$, then, in view of Lemma 3.4, we have that

$$\sigma(f(\widehat{R})) = \{f(\eta), f(\lambda_{m+2}), \ldots, f(\lambda_{m+n})\}$$

and analogously for $\sigma(f(-\widehat{S})^{-1})$. Therefore $\rho(f(-\widehat{S})^{-1}) = \rho(f(-S)^{-1})$ and

$$\rho(f(\widehat{R})) = \max\{|f(\eta)|, |f(\lambda_{m+2})|, \ldots, |f(\lambda_{m+n})|\}.$$

In order to improve the rate of convergence one should choose η such that

$$\rho(f(\widehat{R})) < |f(\lambda_{m+1})| = \rho(f(R)). \tag{4.6}$$

This is possible if $\lambda_{m+2} \neq 0$ (and Theorem 3.3 ensures this), with $\eta > 0$ for $\alpha \leq \beta$ or $0 < \eta < 2\alpha\beta/(\alpha-\beta)$ for $\alpha > \beta$; in both cases $|f(\eta)| < \beta/\alpha = \rho(f(R))$.

We observe that all the choices of η such that $\rho(f(\widehat{R})) = \max_{i=2,\ldots,n}\{|f(\lambda_{m+i})|\}$, or equivalently

$$|f(\eta)| \leq \max_{i=2,\ldots,n}\{|f(\lambda_{m+i})|\},$$

are optimal in relation to the convergence ratio. In particular, the choice $\eta = \beta$ is optimal (but it is not always possible for our purpose: see Section 4.2.1 for more details) because in this case $f(\eta) = 0$.

In addition, we note that, in the case $\nu_1^T\mu_1 = \nu_2^T\mu_2$, the condition (4.6) guarantees that

$$\rho(f(\widehat{R}))\rho(f(-\widehat{S})^{-1}) < \rho(f(R))\rho(f(-S)^{-1}) = 1,$$

that implies the change of ADDA order of convergence from linear to quadratic.

Despite in [35] Wang at al. outline that the shifting idea of Guo et al. [15] should be combined with ADDA, at the best of our knowledge a complete proof of convergence of ADDA applied to the shifted equation (4.4) has not been reported in the literature: in the following we state a convergence result of ADDA for shifted M-NARE under suitable assumptions on the matrix P_0.

4.2 Accurate doubling algorithm for shifted M-NARE

Before describing the elementwise accurate doubling algorithm for shifted M-NARE we need some preliminary results: we introduce the idea of delayed shift and we exhibit the triplet representations for all the nonsingular M-matrices to be inverted during the iterative doubling process.

4.2.1 Delayed shift

The elementwise accurate doubling algorithm of the previous chapter cannot be applied directly to the shifted equation because the rank-one correction on $\mathcal{H}$ in general does not preserve the M-matrix structure of $\widehat{M}$ for any η and p. In this section we will discuss a remedy to this problem by a technique that we called *delayed shift*.

Applying the ADDA to the shifted equation (4.4) we get the initial values

$$\widehat{P}_0 = \begin{bmatrix} \widehat{E}_0 & \widehat{G}_0 \\ \widehat{H}_0 & \widehat{F}_0 \end{bmatrix} = -\begin{bmatrix} \widehat{A}+\alpha I & -\widehat{B} \\ \widehat{C} & \widehat{D}+\beta I \end{bmatrix}^{-1} \begin{bmatrix} \widehat{A}-\beta I & -\widehat{B} \\ \widehat{C} & \widehat{D}-\alpha I \end{bmatrix}. \tag{4.7}$$

Using the elementwise accurate doubling algorithm requires that the matrix to be inverted

$$\widehat{M}_{\alpha,\beta} = \begin{bmatrix} \widehat{A}+\alpha I & -\widehat{B} \\ \widehat{C} & \widehat{D}+\beta I \end{bmatrix} = \begin{bmatrix} A+\eta v_1 p_1^T+\alpha I & -B+\eta v_1 p_2^T \\ C-\eta v_2 p_1^T & D-\eta v_2 p_2^T+\beta I \end{bmatrix} \tag{4.8}$$

has a triplet representation, but in general $\widehat{M}_{\alpha,\beta}$ is not even a Z-matrix: for example, if an off-diagonal entry of B is close to 0, we should impose very strong constraints on η in order to guarantee this property. This could be a serious obstacle for the effectiveness of the shift technique because, as we have seen in the previous section, η has to be sufficiently large in order to provide a significant acceleration to the algorithm: indeed, η has to be such that:

$$\max\{|f(\eta)|, |f(\lambda_{m+2})|, \ldots, |f(\lambda_{m+n}|\} = \rho(f(\widehat{R})) < \rho(f(R)) = |f(\lambda_{m+1})|\,.$$

A way around this difficulty is the remark that the shift can be seen as a rank-one modification of the matrix P_0 and then one can invert $M_{\alpha,\beta}$ and apply the correction to

P_0 that is computed with elementwise accuracy. We will call this technique *delayed shift* since in some sense the effect of the shift is imposed at a later stage than the usual one.

By manipulating (3.7) and from the relations (4.3), it is easy to obtain

$$\widehat{P}_0 = -I_N + (\alpha+\beta)\begin{bmatrix} A+\eta v_1 p_1^T + \alpha I & -B+\eta v_1 p_2^T \\ C-\eta v_2 p_1^T & D-\eta v_2 p_2^T + \beta I \end{bmatrix}^{-1}.$$

Recalling (3.7)-(3.8) and using the Shermann-Morrison-Woodbury formula, we get

$$\begin{aligned} \widehat{P}_0 &= -I_N + (\alpha+\beta)(M_{\alpha,\beta} + \eta J v p^T)^{-1} \\ &= P_0 - (\alpha+\beta)\frac{{M_{\alpha,\beta}}^{-1} J v \eta p^T {M_{\alpha,\beta}}^{-1}}{1+\eta p^T {M_{\alpha,\beta}}^{-1} J v} \\ &= P_0 - \Sigma \end{aligned} \tag{4.9}$$

with

$$\Sigma = (\alpha+\beta)\frac{{M_{\alpha,\beta}}^{-1} J v q^T}{1+q^T J v}, \quad q^T = \eta p^T {M_{\alpha,\beta}}^{-1}. \tag{4.10}$$

Hence, one can construct $\widehat{P}_0$ not only by applying the analogues of (3.7)–(3.8) to $\widehat{M}$, but also by constructing P_0 and Σ separately and then applying the formula (4.9): this second option is the delayed shift.

We will see in the next sections how this approach, skipping the construction of $\widehat{M}$, allows one to keep the positivity structure by imposing on η much milder constraints than the ones necessary to let $\widehat{M}_{\alpha,\beta}$ be an M-matrix. In this way we can keep the advantages of the shift technique, in terms of acceleration of the convergence, and, on the other hand, we can guarantee an accurate computation.

Remark 4.3. We note that the choice of η and p in the construction phase of Σ has been made with the aim of avoiding the cancellation in the subtraction $\widehat{P}_0 = P_0 - \Sigma$. Indeed, if we choose $\theta \approx 0.1$ analogously to the Section 3.3.3, we obtain $\Sigma \leq (1-\theta)P_0$ by construction.

4.2.2 Triplet representations

In order to obtain an elementwise accurate version of the doubling algorithm applied to the shifted equation we will give a condition on $\widehat{P}_0$ under which there exists a triplet representation for the matrices to be inverted during the iteration, but first we need a lemma about the doubling algorithm that is a slightly different version of [38, Theorem 3.3]. The key difference between the original version and our theorem is that we impose conditions on the matrix P_0, without assuming that it is obtained from an M-matrix, and thus it can be extended to the shifted case.

Lemma 4.4. *With the notation of Section 3.2 and of Lemma 3.9, let*

$$P_0 = \begin{bmatrix} E_0 & G_0 \\ H_0 & F_0 \end{bmatrix}$$

be the initial value of a sequence $P_k = \mathcal{F}^{\circ k}(P_0) = \left[\begin{smallmatrix} E_k & G_k \\ H_k & F_k \end{smallmatrix}\right]$. *If there exist* $v, w \in \mathbb{R}^N$ *such that* $(I - P_0)v = w$, *with* $P_0 > 0$, $v > 0$ *and* $w \geq 0$, *then* $P_k > 0$, $w^{(k)} \geq 0$ *for all* $k > 0$ *and there exist triplet representations for* $I - G_k H_k$ *and* $I - H_k G_k$, *for all* $k \geq 0$:

$$(\mathit{offdiag}(G_k H_k), v_1, E_k v_1 + G_k F_k v_2 + w_1^{(k)} + G_k w_2^{(k)}) \tag{4.11}$$

and

$$(\mathit{offdiag}(H_k G_k), v_2, F_k v_2 + H_k E_k v_1 + w_2^{(k)} + H_k w_1^{(k)}). \tag{4.12}$$

Proof. If $v > 0$ and $P_k > 0$, $w^{(k)} \geq 0$, the positivity of P_{k+1} and the non-negativity of $w^{(k+1)}$ follows from the absence of subtractions in the definition of the sequences $\{P_k\}_{k\geq 0}$ and $\{w^{(k)}\}_{k\geq 0}$ and by the equivalence of the second and third items of Lemma 1.2 applied to $I - G_k H_k$ and $I - H_k G_k$, that are Z-matrices under the condition $P_k > 0$.

We observe that, if $P_0 > 0$, $v > 0$ and $w \geq 0$ with $(I - P_0)v = w$, the doubling map $\mathcal{F}$ is automatically well-defined. Indeed, by Lemma 1.2, the existence of a positive triplet representations for $I - G_k H_k$ and $I - H_k G_k$ ensures that these matrices are nonsingular M-matrices and this is a sufficient condition for the applicability of a doubling algorithm ([6]). □

Remark 4.5. We observe that the condition $P_0 > 0$ is naturally satisfied under suitable assumptions on M. However it is crucial to notice that for our purpose we need not to make any assumption on M as we have seen in Section 4.2.1.

Now we derive conditions under which there exists a triplet representation for all the matrices to be inverted in the doubling algorithm with the delayed shift technique when applied to equation (1.7).

Corollary 4.5.1. *Let* $\widehat{P}_k$ *be the* k*-th iteration of doubling map* $\mathcal{F}$ *starting with* $\widehat{P}_0$ *as in* (4.9)

$$\widehat{P}_k = \mathcal{F}^{\circ k}(\widehat{P}_0), \quad \widehat{P}_k = \begin{bmatrix} \widehat{E}_k & \widehat{G}_k \\ \widehat{H}_k & \widehat{F}_k \end{bmatrix}, \quad k \geq 0.$$

Assume that η *is a scalar,* $\eta \neq -\alpha$, $\eta \neq \beta$. *The matrix* $\widehat{P}_k$ *satisfies the following property, for all* $k \geq 0$:

$$\begin{bmatrix} \xi^{2^k} \widehat{E}_k & \widehat{G}_k \\ \widehat{H}_k & \xi^{-2^k} \widehat{F}_k \end{bmatrix} v = v, \tag{4.13}$$

where $\xi = \xi_1/\xi_2$ *and* $\xi_1 = \alpha + \eta$, $\xi_2 = \beta - \eta$.

In addition, for all $k \geq 0$, *the matrices* $I - \widehat{G}_k \widehat{H}_k$ *and* $I - \widehat{H}_k \widehat{G}_k$ *are such that*

$$(I - \widehat{G}_k \widehat{H}_k) v_1 = \xi^{2^k} \widehat{E}_k v_1 + \xi^{-2^k} \widehat{G}_k \widehat{F}_k v_2, \tag{4.14}$$

$$(I - \widehat{H}_k \widehat{G}_k) v_2 = \xi^{-2^k} \widehat{F}_k v_2 + \xi^{2^k} \widehat{H}_k \widehat{E}_k v_1. \tag{4.15}$$

In particular, if $\widehat{P}_0 > 0$ *and* $0 < \eta < \beta$, $\widehat{P}_k$ *is positive for all* $k > 0$ *and there exist triplet representations for* $I - \widehat{G}_k\widehat{H}_k$ *and* $I - \widehat{H}_k\widehat{G}_k$, *for all* $k \geq 0$ *:*

$$(\mathit{offdiag}(\widehat{G}_k\widehat{H}_k), v_1, \xi^{2^k}\widehat{E}_k v_1 + \xi^{-2^k}\widehat{G}_k\widehat{F}_k v_2) \tag{4.16}$$

and

$$(\mathit{offdiag}(\widehat{H}_k\widehat{G}_k), v_2, \xi^{-2^k}\widehat{F}_k v_2 + \xi^{2^k}\widehat{H}_k\widehat{E}_k v_1). \tag{4.17}$$

Proof. For $k = 0$, by the definition of $\widehat{P}_0$ in (4.7), we have

$$(I_N + \widehat{P}_0)\begin{bmatrix} (\alpha+\eta)v_1 \\ (\beta-\eta)v_2 \end{bmatrix} = (\alpha+\beta)v$$

and, equivalently,

$$\begin{bmatrix} \widehat{E}_0 & \widehat{G}_0 \\ \widehat{H}_0 & \widehat{F}_0 \end{bmatrix}\begin{bmatrix} \xi_1 v_1 \\ \xi_2 v_2 \end{bmatrix} = \begin{bmatrix} \xi_2 v_1 \\ \xi_1 v_2 \end{bmatrix}.$$

Since $\beta \neq \eta$ and $\alpha \neq -\eta$ the equation above can be rewritten as

$$\begin{bmatrix} \xi\widehat{E}_0 & \widehat{G}_0 \\ \widehat{H}_0 & \xi^{-1}\widehat{F}_0 \end{bmatrix} v = v.$$

So, we are under the assumptions of Lemma 4.4 with $w = 0$, if we set $E_0 = \xi\widehat{E}_0$, $F_0 = \xi^{-1}\widehat{F}_0$, $G_0 = \widehat{G}_0$ and $H_0 = \widehat{H}_0$. Inductively can be shown that, by applying (3.1)-(3.4) from this starting point, the following sequences are generated:

$$E_k = \xi^{2^k}\widehat{E}_k, \;\; F_k = \xi^{-2^k}\widehat{F}_k, \;\; G_k = \widehat{G}_k, \;\; H_k = \widehat{H}_k, \text{ for } k \geq 0.$$

□

Remark 4.6. The assumption $0 < \eta < \beta$ is necessary in order to guarantee both the positivity of $\widehat{P}_k$ and the effectiveness of the shift in improving the rate of convergence. Since ξ and its inverse have to be positive and

$$\xi^{-1} = \frac{\beta - \eta}{\alpha + \eta} = -f(\eta),$$

we have that $\widehat{P}_k > 0$ if $f(\eta) < 0$. It easy to see that $f(\eta) < 0$ when $-\alpha < \eta < \beta$, but in the case $-\alpha < \eta \leq 0$ we have $|f(\eta)| \geq \beta/\alpha$. Last condition is contrary to (4.6) that guarantees an improvement of the convergence.

Remark 4.7. One of the conditions for the existence of the triplet is that $\eta < \beta$ and in order to get the triplets we need to compute $\xi_2 = \beta - \eta$ that is prone to cancellation. Indeed, analogously to the Section 3.3.3, we can choose a sufficiently large $\theta > 0$, e.g., $\theta \approx 0.1$ and impose $\eta \leq (1-\theta)\beta$ in order to make the operation safe, slightly degrading, in some cases, the convergence speed.

4.2.3 Algorithm

We are ready to describe the algorithm for the shifted elementwise accurate ADDA that uses the new idea of delayed shift and the techniques of choice of η (see Section 4.4) in combination with the known elementwise accurate doubling algorithms ([38], [23]) and with the shift technique proposed in [15].

Algorithm 4: Shifted elementwise accurate ADDA

Input: M, v and ε, where M is the M-matrix defined in (1.5), v is s.t. $Mv = 0$ and ε is the convergence tolerance

Output: Φ and $\widehat{Y}$, extremal solutions of (1.7) and of the dual equation of (4.4)

1 set $\alpha \leftarrow \theta\alpha_{\rm opt}$ and $\beta \leftarrow \theta\beta_{\rm opt}$ for a moderate θ, e.g., $\theta = 1.1$;
2 compute initial values E_0, F_0, G_0, H_0 according to (3.7), performing the matrix inversion with Algorithm 2 and triplet representation (3.34);
3 choose η and the corresponding p by using one of the techniques described in Section 4.4;
4 compute initial values $\widehat{E}_0$, $\widehat{F}_0$, $\widehat{G}_0$, $\widehat{H}_0$ according to (4.9);
5 $k \leftarrow 0$;
6 **do**
7 | compute $\widehat{E}_{k+1}, \widehat{F}_{k+1}, \widehat{G}_{k+1}, \widehat{H}_{k+1}$ according to (3.1)-(3.4), using Algorithm 2 for inversions and triplet representations (4.16)-(4.17);
8 | $k \leftarrow k + 1$;
9 **while** *a suitable stopping criterion is not verified*
10 $\widehat{Y} \leftarrow \widehat{G}_k$; $\Phi \leftarrow \widehat{H}_k$;
11 end;

The result below provides sufficient conditions for the applicability and the convergence of Algorithm 4, but it can be more generally applied to a sequence $\{P_k\}_{k\geq 0}$ generated by any doubling algorithm, without specific hypotheses on the matrix M.

Theorem 4.8. *Assume that there exist* $\left[\begin{smallmatrix} I_n \\ X \end{smallmatrix}\right]$ *and* $\left[\begin{smallmatrix} Y \\ I_m \end{smallmatrix}\right]$ *such that*

$$\begin{bmatrix} E_0 & 0 \\ -H_0 & I_m \end{bmatrix} \begin{bmatrix} I_n \\ X \end{bmatrix} = \begin{bmatrix} I_n & -G_0 \\ 0 & F_0 \end{bmatrix} \begin{bmatrix} I_n \\ X \end{bmatrix} \Gamma \tag{4.18}$$

$$\begin{bmatrix} E_0 & 0 \\ -H_0 & I_m \end{bmatrix} \begin{bmatrix} Y \\ I_m \end{bmatrix} \Lambda = \begin{bmatrix} I_n & -G_0 \\ 0 & F_0 \end{bmatrix} \begin{bmatrix} Y \\ I_m \end{bmatrix}, \tag{4.19}$$

with $r := \rho(\Gamma)\rho(\Lambda) < 1$, *where* $P_0 = \left[\begin{smallmatrix} E_0 & G_0 \\ H_0 & F_0 \end{smallmatrix}\right]$ *is the initial matrix of the doubling algorithm.*

If $P_0 > 0$, $v > 0$ *and* $w \geq 0$, *with* $(I - P_0)v = w$, *then the sequence* $\{P_k\}_{k\geq 0}$ *starting from* P_0 *can be computed with no breakdown. Moreover,* $X, Y > 0$, $E_k, F_k > 0$ *for* $k > 0$, *and* $0 < G_k \leq G_{k+1} \leq Y$, $0 < H_k \leq H_{k+1} \leq X$, *so that the sequences* $\{G_k\}_{k\geq 0}$ *and* $\{H_k\}_{k\geq 0}$ *monotonically converge quadratically to* Y *and* X, *respectively.*

Proof. The assumptions on P_0 ensure that the map $\mathcal{F}$ is well-defined, i.e. the sequence $\{P_k\}_{k\geq 0}$ can be computed with no breakdown (see the proof of Lemma 4.4). Moreover, by Lemma 4.4 and by the definition of $\{P_k\}_{k\geq 0}$ (see equations (3.1)-(3.4)), we get that E_k, $F_k > 0$ for $k > 0$ and $G_{k+1} \geq G_k > 0$, $H_{k+1} \geq H_k > 0$. In addition the existence of X and Y and of subspaces (4.18)-(4.19) guarantees that

$$E_k = (I - G_k X)\Gamma^{2^k}, \tag{4.20}$$
$$X - H_k = F_k X \Gamma^{2^k}, \tag{4.21}$$
$$Y - G_k = E_k Y \Lambda^{2^k}, \tag{4.22}$$
$$F_k = (I - H_k Y)\Lambda^{2^k}. \tag{4.23}$$

Thus the quadratical convergence of the method can be derived from (4.20)-(4.23), that implies

$$\limsup_{k\to\infty} \|Y - G_k\|^{1/2^k} \leq r, \qquad \limsup_{k\to\infty} \|X - H_k\|^{1/2^k} \leq r.$$

Moreover, the sequence $\{G_k\}_{k\geq 0}$ is bounded: since $v = P_k + w^{(k)}$, we have $P_k v \leq v$, thus $E_k v_1 + G_k v_2 \leq v_1$. From last relation and the positivity of v follows that G_k is bounded for all $k \leq 0$. An analogous argument holds for $\{H_k\}_{k\geq 0}$.

Finally, since the convergence of $\{G_k\}_{k\geq 0}$ and $\{H_k\}_{k\geq 0}$ is monotonic, we have that X and Y are positive. □

We can apply the theorem above to the doubling algorithm for the shifted equation (Algorithm 4):

Corollary 4.8.1. *Let Φ and $\widehat{Y}$ be the solutions of the shifted NARE and its dual equations as in Section 4.1 and let $\nu_1^T\mu_1 > \nu_2^T\mu_2$, or $\nu_1^T\mu_1 = \nu_2^T\mu_2$ and $p_1 > 0$, with M irreducible singular M-matrix. If $\widehat{P}_0 > 0$ and $0 < \eta < \beta$ then the sequence $\{\widehat{P}_k\}_{k\geq 0}$ starting from $\widehat{P}_0$ can be computed with no breakdown. Moreover, $\widehat{Y} > 0$ and $\widehat{E}_k$, $\widehat{F}_k > 0$ for $k > 0$, and $0 < \widehat{G}_k \leq \widehat{G}_{k+1} \leq \widehat{Y}$, $0 < \widehat{H}_k \leq \widehat{H}_{k+1} \leq \Phi$, so that the sequences $\{\widehat{H}_k\}_{k\geq 0}$ and $\{\widehat{G}_k\}_{k\geq 0}$ monotonically converge to Φ and $\widehat{Y}$ respectively. In addition, the convergence is quadratic.*

Proof. The existence of Φ and $\widehat{Y}$ implies that

$$\begin{bmatrix} \widehat{E}_0 & 0 \\ -\widehat{H}_0 & I_m \end{bmatrix} \begin{bmatrix} I_n \\ \Phi \end{bmatrix} = \begin{bmatrix} I_n & -\widehat{G}_0 \\ 0 & \widehat{F}_0 \end{bmatrix} \begin{bmatrix} I_n \\ \Phi \end{bmatrix} (-f(\widehat{R})),$$
$$\begin{bmatrix} \widehat{E}_0 & 0 \\ -\widehat{H}_0 & I_m \end{bmatrix} \begin{bmatrix} \widehat{Y} \\ I_m \end{bmatrix} (-f(-\widehat{S})^{-1}) = \begin{bmatrix} I_n & -\widehat{G}_0 \\ 0 & \widehat{F}_0 \end{bmatrix} \begin{bmatrix} \widehat{Y} \\ I_m \end{bmatrix}$$

and moreover $\rho(f(\widehat{R}))\rho(f(-\widehat{S})^{-1}) < 1$. Theorem 4.8 implies the quadratic convergence of the sequences G_k and H_k to Φ and $\widehat{Y}$, respectively. The algorithm is applicable because $\widehat{P}_0 > 0$ and $\eta < \beta$, and we have that $\widehat{Y} > 0$. □

4.3 Elementwise stability

In this section, we present an elementwise forward error bound on the initial values $\widehat{P}_0$ computed with the delayed shift techniques in machine arithmetic in order to apply results similar to [23, Theorem 7.7].

With a slight abuse of notation, we use the notation $fl(\cdot)$ to denote the intermediate quantities computed by performing Algorithm 4 in machine arithmetic; hence, for instance, $fl(\widehat{P}_0)$ is the result of the machine-arithmetic subtraction of $fl(P_0) \ominus fl(\Sigma)$, where $fl(P_0)$ and $fl(\Sigma)$ already contain all the arithmetic errors accumulated in the previous steps. We assume that all the entries of the starting quantities M, v are machine numbers. Hence, here we look for a bound

$$|fl(\widehat{P}_0) - \widehat{P}_0| \leq O(\mathsf{u})\overline{\widehat{P}}_0.$$

Notice on the right-hand side the presence of a matrix $\overline{\widehat{P}}_0$ that differs from $\widehat{P}_0$, which provides a weaker result than those in [23]; nevertheless this computation still provides a useful (and computable) bound, especially if $\overline{\widehat{P}}_0 \approx \widehat{P}_0$.

4.3.1 Elementwise perturbation bound

In this section we provide a perturbation bound of the shift component Σ of $\widehat{P}_0$. In order to achieve this objective, we give some preliminary results.

Lemma 4.9. *Let α be a nonzero real number, let B be a matrix, let $\bar{\alpha}$ and $\bar{B}$ be a nonnegative number and a matrix such that $|\alpha| \leq \bar{\alpha}$ and $|B| \leq \bar{B}$, and let $\tilde{\alpha}$ and $\tilde{B}$ be a number and a matrix such that $|\tilde{\alpha} - \alpha| \dot{\leq} a\varepsilon\bar{\alpha}$ and $|\tilde{B} - B| \dot{\leq} b\varepsilon\bar{B}$, with $0 < \varepsilon < 1$. Then*

$$\left|\frac{\tilde{B}}{\tilde{\alpha}} - \frac{B}{\alpha}\right| \dot{\leq} (a+b)\varepsilon\frac{\bar{\alpha}\bar{B}}{\alpha^2}.$$

Proof.

$$\begin{aligned}
\left|\frac{\tilde{B}}{\tilde{\alpha}} - \frac{B}{\alpha}\right| &= \left|\frac{\tilde{B}}{\tilde{\alpha}} - \frac{\tilde{B}}{\alpha} + \frac{\tilde{B}}{\alpha} - \frac{B}{\alpha}\right| \\
&\leq \left|\frac{\tilde{B}}{\tilde{\alpha}} - \frac{\tilde{B}}{\alpha}\right| + \left|\frac{\tilde{B}}{\alpha} - \frac{B}{\alpha}\right| \\
&= \tilde{B}\left|\frac{1}{\tilde{\alpha}} - \frac{1}{\alpha}\right| + \frac{1}{|\alpha|}|\tilde{B} - B| \\
&\dot{\leq} (\bar{B} + b\varepsilon\bar{B})\frac{a\varepsilon\bar{\alpha}}{|\alpha|\tilde{\alpha}} + \frac{b\varepsilon\bar{B}}{|\alpha|} \\
&\dot{\leq} (a+b)\varepsilon\frac{\bar{\alpha}\bar{B}}{\alpha^2}.
\end{aligned}$$

□

In view of lemma above, we can obtain a perturbation bound for the denominator d of Σ. Since $d = 1 + q^T Jv$, we begin by deriving a bound for q, where $q^T = \eta p^T M_{\alpha\beta}^{-1}$. For simplicity we assume that v, α, β are exact:

Lemma 4.10. *Let $(\eta, \tilde{\eta})$, $(p, \tilde{p})$ and $(M_{\alpha\beta}, \tilde{M}_{\alpha\beta})$ be three pairs of positive scalars, positive $(m+n)$ vectors, $(m+n)\times(m+n)$ M-matrices with triplet representations $(offdiag(M_{\alpha\beta}), v, w)$ and $(offdiag(\tilde{M}_{\alpha\beta}), v, \tilde{w})$, respectively, such that $|\eta - \tilde{\eta}| < \varepsilon\eta$, $|p - \tilde{p}| < \varepsilon p$, $|w - \tilde{w}| < \varepsilon w$, and $|offdiag(M_{\alpha,\beta}) - offdiag(\tilde{M}_{\alpha,\beta})| < \varepsilon |offdiag(M_{\alpha,\beta})|$. If $q^T = \eta p^T M_{\alpha\beta}^{-1}$ and $\tilde{q}^T = \tilde{\eta}\tilde{p}^T \tilde{M}_{\alpha\beta}^{-1}$, then*

$$|\tilde{q}^T - q^T| \dot{\leq} \varepsilon k q^T,$$

where $k = 2(m+n) + 1$.

Proof. From [23, Lemmas 7.1, 7.2], we have

$$|M_{\alpha\beta}^{-1} - \tilde{M}_{\alpha\beta}^{-1}| \dot{\leq} (2(m+n) - 1)\varepsilon M_{\alpha\beta}^{-1}$$

and

$$|p^T M_{\alpha\beta}^{-1} - \tilde{p}^T \tilde{M}_{\alpha\beta}^{-1}| \dot{\leq} 2(m+n)\varepsilon p^T M_{\alpha\beta}^{-1}$$

and, consequently,

$$|\eta p^T M_{\alpha\beta}^{-1} - \tilde{\eta}\tilde{p}^T \tilde{M}_{\alpha\beta}^{-1}| \dot{\leq} (2(m+n)+1)\varepsilon\eta p^T M_{\alpha\beta}^{-1}.$$

□

Lemma 4.11. *Let d and $\tilde{d}$ be two numbers such that $d = 1 + q^T Jv$ is the denominator of the formula for Σ in (4.10) and $\tilde{d}$ is defined by $\tilde{d} = 1 + \tilde{q}^T Jv$, with $(q, \tilde{q})$ and v as defined in lemma above. Then*

$$|d - \tilde{d}| \dot{\leq} k\varepsilon\bar{d}$$

with $\bar{d} = 1 + q^T v$ and $k = 2(m+n) + 1$.

Proof. From [23, Lemma 7.2], we have

$$\begin{aligned} |d - \tilde{d}| &= |1 + q^T Jv - (1 + \tilde{q}^T Jv)| \\ &\leq |(1 + q_1^T v_1) - (1 + \tilde{q}_1^T v_1)| + |q_2^T v_2 - \tilde{q}_2^T v_2| \\ &\dot{\leq} k\varepsilon(1 + q_1^T v_1) + k\varepsilon q_2^T v_2 \\ &= k\varepsilon(1 + q^T v). \end{aligned}$$

We conclude that $|d - \tilde{d}| \dot{\leq} k\varepsilon\bar{d}$. □

An analogous bound can be derived for the numerator N of Σ:

Lemma 4.12. *Let N and $\tilde{N}$ be two matrices such that $N = (\alpha + \beta)M_{\alpha\beta}^{-1} Jvq^T$ is the numerator of Σ and $\tilde{N}$ is defined by $\tilde{N} = (\alpha + \beta)\tilde{M}_{\alpha\beta}^{-1} Jv\tilde{q}^T$, with the notation of the lemmas above. Then*

$$|N - \tilde{N}| \dot{\leq} \nu\varepsilon\bar{N}$$

with $\bar{N} = (\alpha + \beta)M_{\alpha\beta}^{-1} vq^T$ and $\nu = 4(m+n)$.

Proof. By denoting

$$N_1 := (\alpha+\beta)M_{\alpha\beta}^{-1}\begin{bmatrix} v_1 \\ 0 \end{bmatrix} q^T, \quad N_2 := (\alpha+\beta)M_{\alpha\beta}^{-1}\begin{bmatrix} 0 \\ v_2 \end{bmatrix} q^T$$

and by using [23, Lemma 7.2] and Lemma 4.10, we have that

$$\begin{aligned} |N_1 - \tilde{N}_1| &\dot{\leq} 4(m+n)\varepsilon N_1, \\ |N_2 - \tilde{N}_2| &\dot{\leq} 4(m+n)\varepsilon N_2 \end{aligned}$$

And finally,

$$\begin{aligned} |N - \tilde{N}| &= |N_1 - \tilde{N}_1| + |N_2 - \tilde{N}_2| \\ &\dot{\leq} 4(m+n)\varepsilon N_1 + 4(m+n)\varepsilon N_2 \\ &= 4(m+n)\varepsilon \bar{N}. \end{aligned} \tag{4.24}$$

□

To conclude, we provide a perturbation bound for Σ:

Lemma 4.13. *Let S and $\tilde{\Sigma}$ be two matrices such that $\Sigma = N/d$ and $\tilde{\Sigma} = \tilde{N}/\tilde{d}$, with the notation of the lemmas above. Then*

$$|\Sigma - \tilde{\Sigma}| \dot{\leq} \sigma\varepsilon\frac{\bar{d}\bar{N}}{d^2}$$

where $\sigma = 6(m+n)+1$.

Proof. The statement follows by Lemmas 4.9, 4.11, and 4.12. □

4.3.2 Numerical stability

In order to obtain a componentwise error bound for the computed approximation of $\widehat{P}_0$, we state some results for $fl(\Sigma)$, the computed approximation of Σ. We denote by u the machine precision and ignore the second-order terms in u by hiding them in the syntax $A \dot{\leq} B$. Here, and in the following, we use the common floating point arithmetic model in which $fl(x \operatorname{op} y) = (x \operatorname{op} y)(1+\varepsilon)$, $|\varepsilon| \leq \mathsf{u}$, for $\operatorname{op} \in \{+, -, \times, \div\}$ and machine precision u.

We start from three lemmas providing error bounds for the sum and subtraction of two matrices and the division matrix by scalar performed in floating point arithmetic.

Lemma 4.14. *Let A and B be matrices of the same size, let $\bar{A}$ and $\bar{B}$ be nonnegative matrices such that $|A| \leq \bar{A}$ and $|B| \leq \bar{B}$, and let $fl(A)$ and $fl(B)$ be matrices of floating point numbers such that $|fl(A) - A| \dot{\leq} a\mathsf{u}\bar{A}$ and $|fl(B) - B| \dot{\leq} b\mathsf{u}\bar{B}$. Then*

$$|fl(fl(A) + fl(B)) - (A+B)| \dot{\leq} (\max(a,b)+1)\mathsf{u}(\bar{A}+\bar{B}).$$

Proof. We define $\epsilon_{A+B} := |fl(fl(A) + fl(B)) - (A + B)|$:

$$\begin{aligned}
\epsilon_{A+B} &= |fl(fl(A) + fl(B)) - (A + B) - fl(A) + fl(A) - fl(B) + fl(B)| \\
&\leq |fl(fl(A) + fl(B)) - (fl(A) + fl(B))| + |fl(A) - A| + |fl(B) - B| \\
&\leq |fl(A) + fl(B)|\mathsf{u} + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&= |fl(A) - A + A + fl(B) - B + B|u + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&\leq (|fl(A) - A| + |A| + |fl(B) - B| + |B|)\mathsf{u} + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&\dot{\leq} (a\mathsf{u}\bar{A} + \bar{A} + b\mathsf{u}\bar{B} + \bar{B})\mathsf{u} + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&\dot{\leq} \mathsf{u}(\bar{A} + \bar{B}) + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&\leq (1 + max(a, b))\mathsf{u}(\bar{A} + \bar{B}).
\end{aligned}$$

□

Lemma 4.15. *Let A and B be matrices of the same size, let $\bar{A}$ and $\bar{B}$ be nonnegative matrices such that $|A| \leq \bar{A}$ and $|B| \leq \bar{B}$, and let $fl(A)$ and $fl(B)$ be matrices of floating point numbers such that $|fl(A) - A| \dot{\leq} a\mathsf{u}\bar{A}$ and $|fl(B) - B| \dot{\leq} b\mathsf{u}\bar{B}$. Then*

$$|fl(fl(A) - fl(B)) - (A - B)| \dot{\leq} \max(a, b)\mathsf{u}(\bar{A} + \bar{B}) + \mathsf{u}|A - B|.$$

Proof. We define $\epsilon_{A-B} := |fl(fl(A) - fl(B)) - (A - B)|$:

$$\begin{aligned}
\epsilon_{A-B} &= |fl(fl(A) - fl(B)) - (A - B) - fl(A) + fl(A) - fl(B) + fl(B)| \\
&\leq |fl(fl(A) - fl(B)) - (fl(A) - fl(B))| + |fl(A) - A| + |fl(B) - B| \\
&\leq |fl(A) - fl(B)|\mathsf{u} + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&= |fl(A) - A + A + fl(B) - B + B|\mathsf{u} + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&\leq (|fl(A) - A| + |fl(B) - B| + |A - B|)\mathsf{u} + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&\dot{\leq} (a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} + |A - B|)\mathsf{u} + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&\dot{\leq} \mathsf{u}|A - B| + a\mathsf{u}\bar{A} + b\mathsf{u}\bar{B} \\
&\leq \max(a, b)\mathsf{u}(\bar{A} + \bar{B}) + \mathsf{u}|A - B|.
\end{aligned}$$

□

Lemma 4.16. *Let α be a nonzero real number, let B be a matrix, let $\bar{\alpha}$ and $\bar{B}$ be a nonnegative number (matrix) such that $|\alpha| \leq \bar{\alpha}$ and $|B| \leq \bar{B}$, and let $fl(\alpha)$ and $fl(B)$ be a floating point number (a matrix of floating point numbers) such that $|fl(\alpha) - \alpha| \dot{\leq} a\mathsf{u}\bar{\alpha}$ and $|fl(B) - B| \dot{\leq} b\mathsf{u}\bar{B}$. Then*

$$\left| fl\left(\frac{fl(B)}{fl(\alpha)}\right) - \frac{B}{\alpha} \right| \dot{\leq} (a + b + 1)\mathsf{u} \left| \frac{\bar{B}}{\alpha^2} \right| \bar{\alpha}.$$

Proof.

$$\begin{aligned}
\left|fl\left(\frac{fl(B)}{fl(\alpha)}\right)-\frac{B}{\alpha}\right| &= \left|fl\left(\frac{fl(B)}{fl(\alpha)}\right)-\frac{B}{\alpha}+\frac{fl(B)}{\alpha}-\frac{fl(B)}{\alpha}+\frac{fl(B)}{fl(\alpha)}-\frac{fl(B)}{fl(\alpha)}\right| \\
&\leq \left|\frac{fl(B)}{\alpha}-\frac{B}{\alpha}\right|+\left|\frac{fl(B)}{fl(\alpha)}-\frac{fl(B)}{\alpha}\right|+\left|fl\left(\frac{fl(B)}{fl(\alpha)}\right)-\frac{fl(B)}{fl(\alpha)}\right| \\
&\dot{\leq} \frac{1}{|\alpha|}b\mathsf{u}\bar{B}+|fl(B)|\left|\frac{1}{fl(\alpha)}-\frac{1}{\alpha}\right|+\left|\frac{fl(B)}{fl(\alpha)}\right|\mathsf{u} \\
&\dot{\leq} \frac{1}{|\alpha|}b\mathsf{u}\bar{B}+\left(b\mathsf{u}\bar{B}+\bar{B}\right)\frac{a\mathsf{u}\bar{\alpha}}{|fl(\alpha)|\cdot|\alpha|}+\left|\frac{fl(B)}{fl(\alpha)}+\frac{B}{fl(\alpha)}-\frac{B}{fl(\alpha)}\right|\mathsf{u} \\
&\dot{\leq} \frac{1}{|\alpha|}b\mathsf{u}\bar{B}+\frac{a\mathsf{u}\bar{\alpha}\bar{B}}{|fl(\alpha)|\cdot|\alpha|}+\frac{1}{|fl(\alpha)|}\mathsf{u}\bar{B} \\
&\leq \frac{1}{|fl(\alpha)|\cdot|\alpha|}\left(b|fl(\alpha)|+a\bar{\alpha}+\bar{\alpha}\right)\mathsf{u}\bar{B} \\
&\dot{\leq} \frac{1}{|\alpha^2|}\left(b(a\mathsf{u}\bar{\alpha}+\bar{\alpha})+a\bar{\alpha}+\bar{\alpha}\right)\mathsf{u}\bar{B} \\
&\dot{\leq} (b+a+1)\,\mathsf{u}\bar{\alpha}\bar{B}\frac{1}{|\alpha^2|}.
\end{aligned}$$

□

Now we can prove a theorem concerning $fl(\Sigma)$ by applying the lemma above, but we need two lemmas on the numerator and the denominator of Σ.

Lemma 4.17. *Let $fl(d)$ be the computed approximation of $d = 1+q^TJv$, the denominator of Σ. Then there exist δ and $\bar{d}$ such that $|d| \leq \bar{d}$ and*

$$|fl(d) - d| \leq \delta\mathsf{u}\bar{d},$$

with $\delta = \psi(m) + 2$ and $\bar{d} = 1 + q^Tv$, where m is the number of columns of T and $\psi(m) = \frac{2}{3}(2m+5)(m+2)(m+3)$.

Proof. We denote by T and $fl(T)$ the product $p^TM_{\alpha\beta}^{-1}$ and its computed approximation, respectively. If we assume p and the triplet for $M_{\alpha\beta}$ are exact, from [23, Lemma 7.9]

$$|fl(T) - T| \dot{\leq} \psi(m)\mathsf{u}T.$$

Since $q^T = \eta T$, the computed approximation $fl(q)$ satifies the bound

$$|fl(q^T) - q^T| \dot{\leq} (\psi(m) + 1)\mathsf{u}q^T,$$

by [23, Lemma 7.8.2] applied to a product of a matrix by a scalar. Because $d = 1+q^TJv = 1 + {q_1}^Tv_1 - {q_2}^Tv_2$, we obtain the result in the statement of the lemma

$$\begin{aligned}
|fl(d) - d| &\leq |fl(1 + {q_1}^Tv_1) - (1 + {q_1}^Tv_1)| + |fl({q_2}^Tv_2) - {q_2}^Tv_2| \\
&\dot{\leq} (\psi(m) + 2)|1 + {q_1}^Tv_1|\mathsf{u} + (\psi(m) + 1)|{q_2}^Tv_2|\mathsf{u} \\
&\dot{\leq} (\psi(m) + 2)\bar{d}\mathsf{u},
\end{aligned}$$

where $\bar{d} = |1 + {q_1}^Tv_1| + |{q_2}^Tv_2| = 1 + q^Tv \geq |d| = |1 + q^TJv|$. □

Lemma 4.18. *Let $fl(N)$ be the computed approximation of $N = (\alpha+\beta)M_{\alpha\beta}^{-1}Jvq^T$, numerator of Σ. Then there exist ν and $\bar{N}$ such that $|N| \leq \bar{N}$ and*

$$|fl(N) - N| \leq \nu \mathsf{u} \bar{N},$$

where $\nu = \psi(m) + 2$ and $\bar{N} = (\alpha+\beta)M_{\alpha\beta}^{-1}vq^T$.

Proof. We write

$$N_1 := (\alpha+\beta)M_{\alpha\beta}^{-1}\begin{bmatrix} v_1 \\ 0 \end{bmatrix} q^T, \quad N_2 := (\alpha+\beta)M_{\alpha\beta}^{-1}\begin{bmatrix} 0 \\ v_2 \end{bmatrix} q^T$$

and let $fl(N_1)$ and $fl(N_2)$ be their computed approximations. Analogously to the previous lemma, we have:

$$|fl(N_i) - N_i| \dot{\leq} (\psi(m)+2)\mathsf{u}N_i, \quad (i = 1, 2)$$

and, since $N = N_1 - N_2$,

$$|fl(N) - N| \dot{\leq} (\psi(m)+2)\mathsf{u}\bar{N},$$

where $\bar{N} := N_1 + N_2 \geq |N|$. □

Now we are ready to derive the error bound for $fl(\Sigma)$, the computed approximation of Σ

Theorem 4.19. *Let $fl(\Sigma)$ be the computed approximation of $\Sigma = N/d$. Then there exist σ and $\bar{\Sigma}$ such that $|\Sigma| \leq \bar{\Sigma}$ and*

$$|fl(\Sigma) - \Sigma| \dot{\leq} \sigma\mathsf{u}\frac{1+q^Tv}{(1+q^TJv)^2}(\alpha+\beta)M_{\alpha\beta}^{-1}vq^T = \sigma\mathsf{u}\frac{\bar{d}}{d^2}\bar{N},$$

where $\sigma = 2\psi(m) + 5$ and $\bar{\Sigma} = \frac{\bar{d}}{d^2}\bar{N}$.

To conclude, we return to the error bound for the computed approximation of $\widehat{P}_0$, the initial value of ADDA applied to the shifted equation. We write, as is done in (4.9),

$$\widehat{P}_0 = P_0 - \Sigma.$$

Theorem 4.20. *Let $fl(\widehat{P}_0)$ be the computed approximation of $\widehat{P}_0 = P_0 - \Sigma$, with $P_0 \geq 0$, $|fl(P_0) - P_0| \dot{\leq} k\mathsf{u}P_0$, and $|\Sigma| \leq \bar{\Sigma}$. Then there exist σ and $\bar{\widehat{P}}_0$ such that $\widehat{P}_0 \leq \bar{\widehat{P}}_0$ and*

$$|fl(\widehat{P}_0) - \widehat{P}_0| \dot{\leq} (max(k, \sigma) + 1)\mathsf{u}\bar{\widehat{P}}_0,$$

where $\sigma = 2\psi(m) + 5$ and $\bar{\widehat{P}}_0 = P_0 + \bar{\Sigma}$. In addition, if $\bar{\Sigma} \leq \theta P_0$, then

$$|fl(\widehat{P}_0) - \widehat{P}_0| \dot{\leq} (max(k, \sigma) + 1)\mathsf{u}(1+\theta)P_0.$$

Proof. Lemma 4.15 and

$$\widehat{P}_0 = P_0 - \Sigma \leq P_0 + |\Sigma| \leq P_0 + \bar{\Sigma}$$

and

$$|fl(\widehat{P}_0) - \widehat{P}_0| \dot{\leq} \max(k, \sigma)\mathsf{u}(P_0 + \bar{\Sigma}) + \mathsf{u}\widehat{P}_0.$$

□

4.4 Choice of η

In this section we present two alternative ways to choose the value η that we used in the numerical experiments.

Case 1:
In order to choose a value of η satisfying (4.6) and the conditions of Corollary 4.5.1, and to maximize the speed of convergence one looks for the maximum nonnegative $\eta < \beta$ such that $\widehat{P}_0 > 0$, or equivalently $\Sigma < P_0$. As discussed in Remark 4.3 it may be convenient to slightly reduce η in order to guarantee $\Sigma \leq (1-\theta)P_0$, with $\theta \approx 0.1$.

We find that, given p, the parameter η should satisfy the $(n+m)^2$ inequalities of the form:

$$g(\eta) = \frac{a\eta}{1+b\eta} \leq k \tag{4.25}$$

where

$$\begin{aligned} k &= ((1-\theta)P_0)_{ij} \\ a &= (T)_{ij} \\ T &= (\alpha+\beta){M_{\alpha,\beta}}^{-1} Jvp^T {M_{\alpha,\beta}}^{-1} \\ b &= q^T Jv/\eta = p^T {M_{\alpha,\beta}}^{-1} Jv \end{aligned} \tag{4.26}$$

with $i,j = 1,\ldots,n+m$. The functions $g(\eta)$ describe $(n+m)^2$ hyperbolas with shapes depending on the sign of a and b. Since $k > 0$ by the definition of P_0, we get that the set of solutions of the system of the inequalities (4.25) is always non empty: in the worst case, $\eta = 0$, and this means that our algorithm fails in the attempt of accelerate the convergence of the elementwise accurate doubling algorithm because in this case the zero eigenvalue is not shifted.

Case 2:
We might try to construct the matrix $\bar{\Sigma} = \bar{d}\bar{N}/d^2$ (as in previous section) such that $\bar{\Sigma} \leq \theta P_0$. This choice guarantees that the error bound of Theorem 4.20 is satisfied. In order to satisfy the inequality $\bar{\Sigma} \leq \theta P_0$, for fixed p we choose η such that

$$h(\eta) = \frac{(1+c\eta)\eta}{(1+b\eta)^2} \leq k = \min_{i,j} K_{ij} \tag{4.27}$$

where

$$\begin{aligned} K_{ij} &= (\theta P_0)_{ij}/[(\alpha+\beta)M_{\alpha\beta}^{-1} vp^T M_{\alpha\beta}^{-1}]_{ij} \\ c &= q^T v/\eta = p^T M_{\alpha\beta}^{-1} v \\ b &= q^T Jv/\eta = p^T M_{\alpha\beta}^{-1} Jv \end{aligned}$$

By a simple manipulation of (4.27) we get

$$(c - kb^2)\eta^2 + (1-2kb)\eta - k \leq 0, \tag{4.28}$$

that is a system of inequalities described by $(n+m)^2$ parabolas. As in the previous case, the set of the solutions can not be empty because it contains always the value $\eta = 0$: it ensures that the algorithm does not fail but in the worst case returns the same result of the non shifted version.

We expect that the bound (4.27) in general is harder to satisfy than the "usual" bound $\Sigma \leq \theta P_0$, with a loss of the practical effectiveness of the shift technique: we give some examples of this behavior in the next section.

About the choice of p, in the numerical experiments we tested as $p = e_i/v_i$, where e_i is the i-th vector of the canonical basis of $\mathbb{R}^{n+m}$, and $p = v/\|v\|^2$ and we picked the value of p maximizing η.

Chapter 5

Numerical experiments

We present several examples to compare the behaviour of the algorithm presented in this work with the non-shifted elementwise accurate ADDA ([23]), the shift technique proposed in [15] and other algorithms solving (1.7) in the case where M is an irreducible singular M-matrix.

The numerical experiments are performed by using Matlab; the stopping criteria are:

- $\min\{||E_k||_1, ||F_k||_1\} < 10^{-15}$, for the normwise convergence;
- $(G_{k+1} == G_k)$ or $(H_{k+1} == H_k)$, where "==" indicates the symbol of equality in floating point arithmetic, for the elementwise convergence.

We computed the exact solution Φ with a high number of significant digits (in our tests we used 60 significant digits) using variable precision arithmetic, and compared the normwise and elementwise errors defined as:

$$\epsilon_{\mathrm{nw}} = \frac{||\tilde{\Phi} - \Phi||}{||\Phi||}, \quad \epsilon_{\mathrm{ew}} = \max_{i,j} \frac{\left|(\tilde{\Phi} - \Phi)_{ij}\right|}{(\Phi)_{ij}}$$

We compare the behaviour of the algorithms:

- ADDA: introduced in [34], as described in this work at Section 3.2;
- ewaADDA: elementwise accurate ADDA, in the case where M is an irreducible singular M-matrix ([23], [38]) as in Algorithm 3;
- shifted ADDA: ADDA applied to the shifted Riccati equation, with the shift technique originally proposed in [15] for SDA, with choices of $\eta = \beta$ and $p = v/||v||^2$;
- shifted ewaADDA: the subject of this thesis, see Algorithm 4;
- dADDA: deflated ADDA, i.e. ADDA with a deflation technique presented in [35] in order to improve the rate of convergence of a doubling algorithm (we performed the version by orthogonal transformation [35, Section 5.2]).

Algorithm	Choice of η	Iterations nw	Iterations ew
ADDA	$\eta = 0$	15	16
ewaADDA	$\eta = 0$	15	16
shifted ADDA	$\eta_{\text{GIM}} = 14.9850$	7	8
shifted ewaADDA	$\eta_{\text{ewa}} = 1.1208$	8	10
dADDA		7	8

Table 5.1: Number of iterations to reach the convergence. *Note that the choice of η_{ewa} has been performed as in Section 4.4 (case 1), and the corresponding p is $p = e_5$.*

5.1 Example 1

In [23, Example 5.1] a fluid model has been described by the pair (T, C) of the transition and rate matrices. We present the same example by writing directly the corresponding irreducible singular M-matrix M such that $Mv = 0$ (with $v = e$ and $n = m = 3$):

$$M = \begin{bmatrix} 15/1.001 & -5/1.001 & 0 & 0 & -5/1.001 & -5/1.001 \\ -5/1.001 & 15/1.001 & 0 & 0 & -5/1.001 & -5/1.001 \\ 0 & 0 & 5/1.001 & -4/1.001 & -1/1.001 & 0 \\ 0 & 0 & -4 & 4 & 0 & 0 \\ -5 & -5 & -10^{-8} & 0 & 15 + 10^{-8} & -5 \\ -5 & -5 & 0 & 0 & -5 & 15 \end{bmatrix}.$$

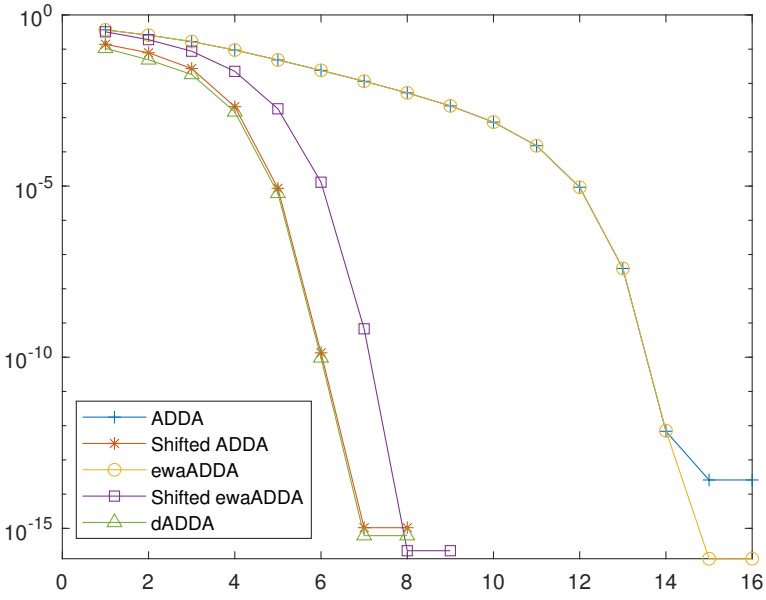
In Table 5.1 we compare the results of the algorithms in terms of number of iterations to reach the convergence.

We note that the shifted ewaADDA converges in a number of steps lower than ADDA and ewaADDA (so we obtain the improvement of convergence speed that we expected), but shifted ADDA and dADDA have apparently better performances.

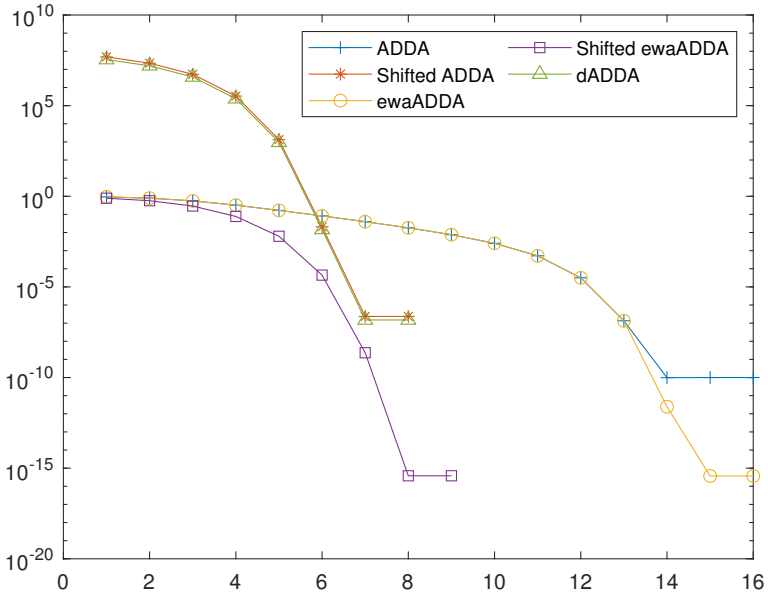
Indeed, if we examine the relative error of the five algorithms (see Table 5.2), we can understand that the relative elementwise error for shifted ADDA and dADDA is dramatically larger than for the shifted ewaADDA, specifically designed with the purpose of reducing the relative elementwise error. We also observe that a seemingly perfect normwise convergence actually hides the fact that the elementwise error is large.

The figures 5.1(a) and 5.1(b) illustrate the trend of the relative error normwise and elementwise during the execution of the algorithms, step by step.

In addition, it is interesting to test the behaviour of the algorithm for two different choices of η (η_1 and η_2) that correspond to cases 1 and 2 of Section 4.4. We observe that, as we argued in Section 4.4, the choice of η is in general such that $\eta_2 < \eta_1$ and it can lead to an increase of the number of iterations necessary for the convergence (see Figure 5.2).



(a) Relative error normwise



(b) Relative error elementwise

Figure 5.1: Relative errors for Example 1, step by step

Algorithm	ϵ_{nw}	ϵ_{ew}
ADDA	2.6e−14	1.0e−10
ewaADDA	1.0e−16	3.0e−16
shifted ADDA	1.0e−15	2.3e−07
shifted ewaADDA	1.3e−16	3.7e−16
dADDA	6.1e−16	1.5e−07

Table 5.2: Relative error achieved in the final step, normwise (ϵ_{nw}) and elementwise (ϵ_{ew})

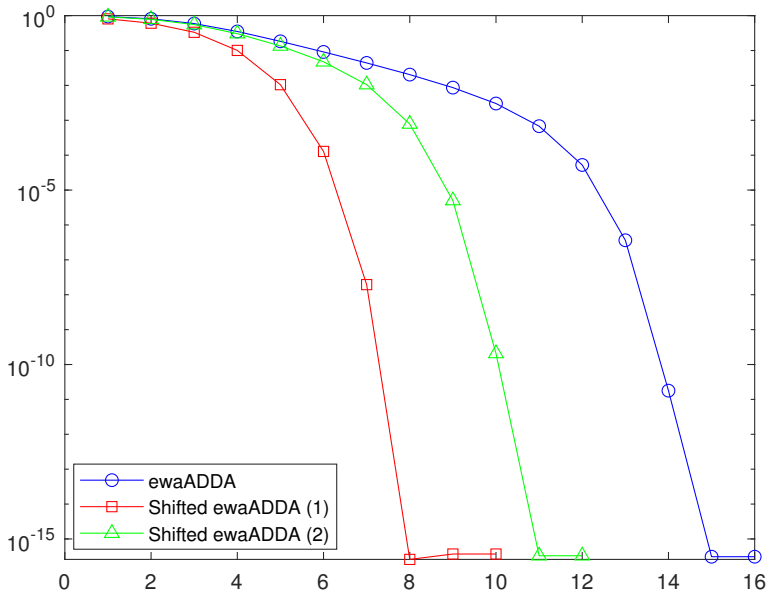


Figure 5.2: Relative errors for Example 1, step by step

5.2 Example 2

We examine the example [15, Example 7.2] and [3, Example 1]: this is a null recurrent (or critical) case.

$$M = \begin{bmatrix} 0.003 & -0.001 & -0.001 & -0.001 \\ -0.001 & 0.003 & -0.001 & -0.001 \\ -0.001 & -0.001 & 0.003 & -0.001 \\ -0.001 & -0.001 & -0.001 & 0.003 \end{bmatrix}$$

As in Section 5.1, we experimented two choices of η (η_1 and η_2) that correspond to cases 1 and 2 of Section 4.4. In this case we can compare the computed solutions with the known exact solution that is:

$$\Phi = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

As we can see in the Figure 5.3, the convergence orders of ewaADDA and shifted ewaADDA are linear and quadratic, respectively.

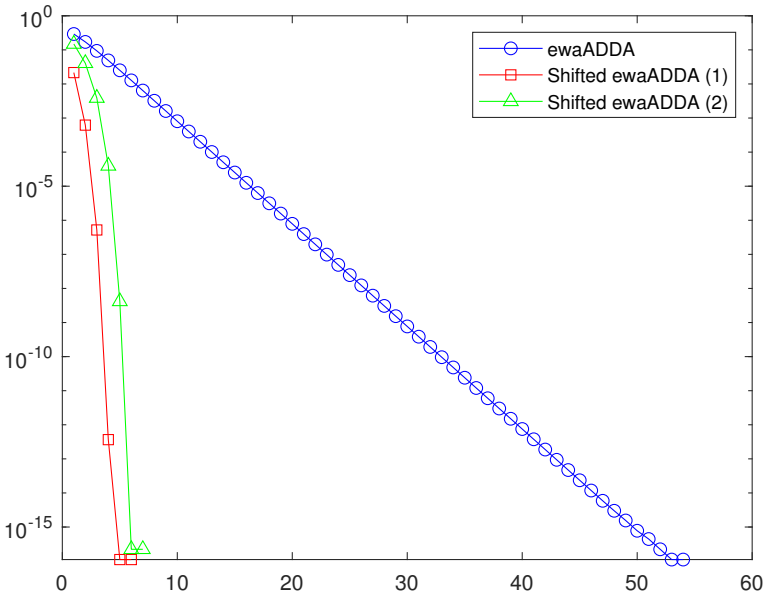


Figure 5.3: Relative error elementwise for Example 2, step by step

The difference of the number of iterations is significant (see Table 5.3). We observe that, in this case, the choice $\eta_2 < \eta_1$ does not lead to a huge deterioration of the speed of convergence.

Algorithm	Choice of η	Iterations nw	Iterations ew
ewaADDA	$\eta = 0$	50	54
shifted ewaADDA (1)	$\eta_1 = 0.0023$	5	6
shifted ewaADDA (2)	$\eta_2 = 0.0009$	6	7

Table 5.3: Number of iterations to reach convergence for Example 2

5.3 Example 3

This example is a slight modification of Example 4 in [3]. In the original work is presented the generator of a Markov process, but according to our exposition we write directly the matrix M describing the Riccati equation:

$$M = \begin{bmatrix} 0.0030 & -0.0001 & -0.0019 & -0.0010 \\ -0.0001 & 0.0030 & -0.0019 & -0.0010 \\ -0.0015 & -0.0015 & 0.0030 & 0 \\ -0.0029 & -0.0001 & 0 & 0.0030 \end{bmatrix}.$$

We observe that M is an irreducible singular M-matrix such that $Mv = 0$, $u^T M = 0$, with drift $\mu = u_2^T v_2 - u_1^T v_1 < 0$ (with $v = e$ and $n = m = 2$).

This is an "unfortunate" case where the use of the elementwise accurate shifted algorithm that we proposed (shifted ewaADDA) is not a real improvement in terms of accuracy over the "classical" (non-elementwise accurate) shifted ADDA performed as in [15], with $\eta_{\text{GIM}} = \beta = 0.0030$ and $p = v/||v||^2$. Note that the solution computed is the exact solution of the M-NARE for all algorithms; the values appearing in Table 5.4 reflect error in the computation of our reference solution with 60 significant digits.

Algorithm	ϵ_{nw}	ϵ_{ew}
ADDA	3.42e−39	6.70e−39
ewaADDA	3.42e−39	6.70e−39
shifted ADDA	3.42e−39	6.70e−39
shifted ewaADDA	3.42e−39	6.70e−39

Table 5.4: Relative error normwise (ϵ_{nw}) and elementwise (ϵ_{ew})

From the point of view of the number of iterations, as we have seen in Section 4.1, the choice of $\eta = \beta$ is always optimal, but, in accord to properties of the spectrum of $\widehat{\mathcal{H}}$,

$$\Lambda(\widehat{\mathcal{H}}) = \{-0.003, -0.0001, \eta, 0.0031\},$$

all the choices of η such that $|f(\eta)| \leq |f(0.0031)|$ produce the same effects on the convergence speed. But if we choose η in accord to Section 4.4 (case 1), we obtain that $\eta_{\text{ewa}} = 0.0021$ (and $p = e_3$): this choice of η leads to

$$|f(\eta_{\text{ewa}})| = 0.1765 > 0.0164 = |f(0.0031)|.$$

Remark 5.1. As we have seen in Section 4.2.1, one may ask why we designed an algorithm to "make accurate" the shifted ADDA when, if it is possible a choice of η such that $\widehat{M}$ is a nonsingular or an irreducible singular M-matrix, we can apply the known accurate algorithms directly to this matrix $\widehat{M}$. This example provides us the opportunity to show that the choice of η as we have described it in Section 4.4 enlarges the range of choice of η. Fixed $p = v/||v||^2$, the maximum value of η for which $\widehat{M} = M + J\eta vp^T$ is a Z-matrix is $\eta_Z = 0.0004$. This choice of η is worse than η_{ewa} (and obviously also than η_{GIM}): in this case we have $|f(\eta_Z)| = 0.7647 > |f(\eta_{\text{ewa}})|$. So, if we perform the ADDA shifted algorithm with this choice of η we obtain a higher number of iterations necessary for the convergence.

We summarize the results of the experiments with Table 5.5 that shows the number of iterations necessary for the convergence (normwise and elementwise) of the different versions of doubling algorithm and for different choices of η (and p).

Algorithm	Choice of η	Iterations nw	Iterations ew
shifted ADDA	$\eta_{\text{GIM}} = 0.0030$	4	4
shifted ewaADDA	$\eta_{\text{ewa}} = 0.0021$	5	6
shifted ADDA	$\eta_Z = 0.0004$	7	8
ADDA	$\eta = 0$	9	10
ewaADDA	$\eta = 0$	10	11

Table 5.5: Number of iterations to reach convergence for Example 3

5.4 Example 4

This example is [23, Example 5.2], with $\kappa = 10^6$. In Figures 5.4(a) and 5.4(b) we plot the trend of the relative error normwise and elementwise during the execution of the algorithms, step by step, for the same algorithms as in Example 1.

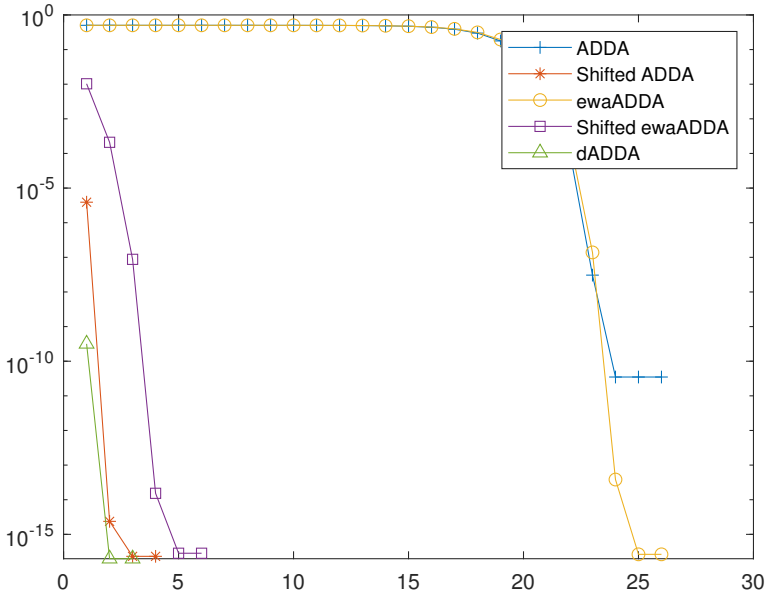
We notice that, by varying the parameter κ across several orders of magnitude, the elementwise error of the ewa algorithms remains small, but the number of iteration of ewaADDA considerably increases, while for shifted ewaADDA this value is stable. Figures 5.5(a) and 5.5(b) illustrate this behaviour.

5.5 Example 5

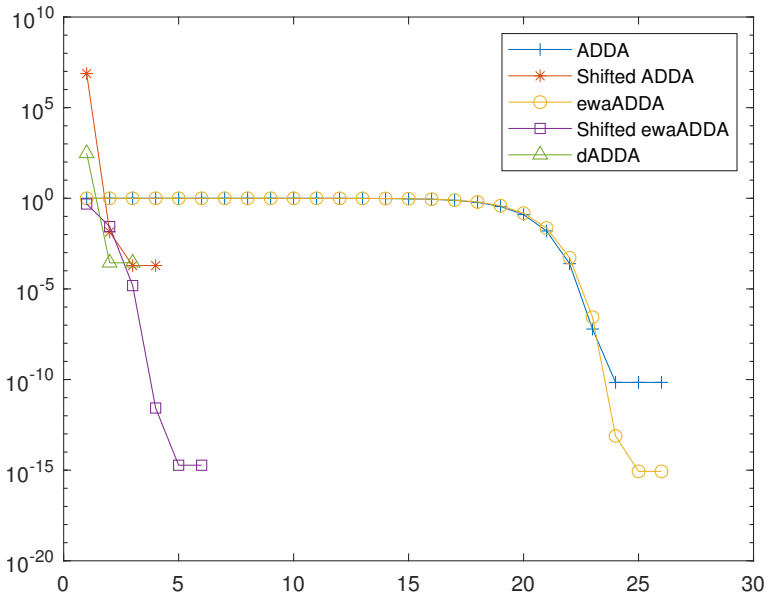
In this example we construct a random matrix R with $n = m = 50$ by using the command

$$R = \text{rand}(n + m).^* \exp(s^*\text{randn}(n + m)),$$

later we multiply the offdiagonal blocks of R by a parameter κ and finally we define $M = \text{diag}(Re) - R$. We generate five different matrices M in this way, with $s = 3$ and $\kappa = 10^{-6}$.

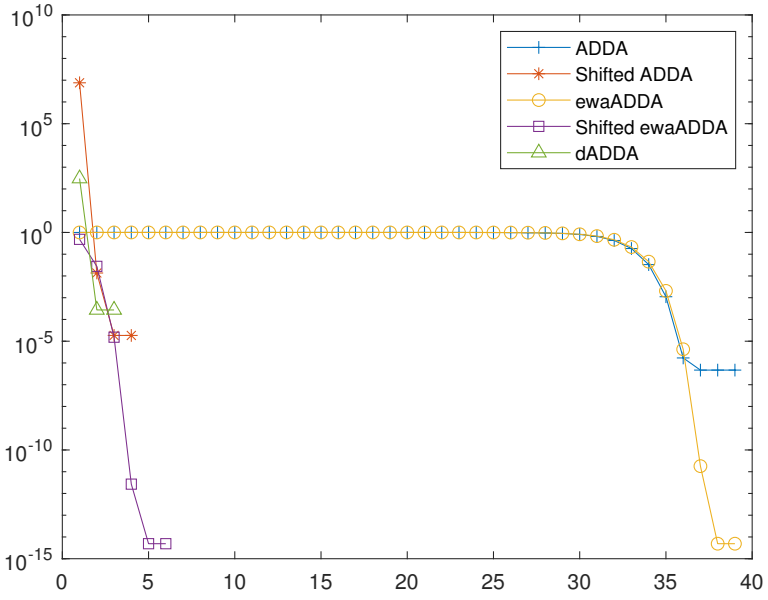


(a) Relative error normwise for $\kappa = 10^6$

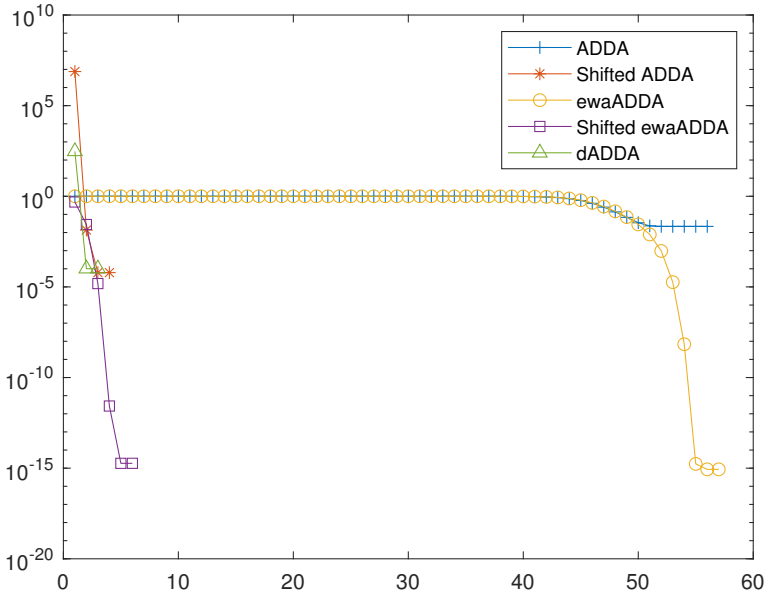


(b) Relative error elementwise for $\kappa = 10^6$

Figure 5.4: Relative errors for Example 4, step by step



(a) Relative error elementwise for $\kappa = 10^{10}$



(b) Relative error elementwise for $\kappa = 10^{14}$

Figure 5.5: Relative errors for Example 4, step by step, for different value of κ

Figure 5.6 shows the typical behaviour of the elementwise relative error in this situation.

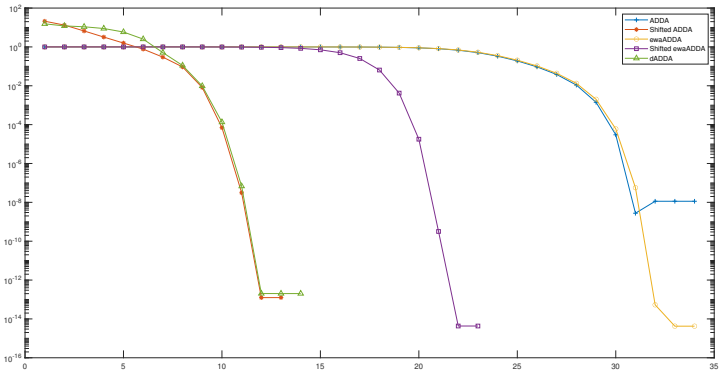


Figure 5.6: Relative elementwise errors for Example 5, step by step

We report the elementwise relative error and the number of iterations for the five tests in Tables 5.6-5.7.

Algorithm	Test 1	Test 2	Test 3	Test 4	Test 5
shifted ADDA	11	12	12	13	12
shifted ewaADDA	21	22	22	23	23
deflated ADDA	12	12	12	14	13
ADDA	29	28	35	34	29
ewaADDA	29	28	35	34	29

Table 5.6: Number of iterations to reach convergence elementwise for Example 5

Algorithm	Test 1	Test 2	Test 3	Test 4	Test 5
shifted ADDA	3.00e−14	9.65e−14	9.84e−14	1.25e−13	6.21e−14
shifted ewaADDA	2.74e−15	4.40e−15	3.96e−15	4.31e−15	3.28e−15
deflated ADDA	2.15e−14	2.96e−14	8.93e−14	2.01e−13	5.06e−14
ADDA	5.16e−10	5.23e−10	3.16e−08	1.14e−08	9.80e−10
ewaADDA	4.15e−15	3.74e−15	4.51e−15	4.23e−15	4.87e−15

Table 5.7: Relative error elementwise for Example 5

Finally, we show the behaviour of the algorithms when $s = 5$ and $\kappa = 10^{-10}$: see Figure 5.7.

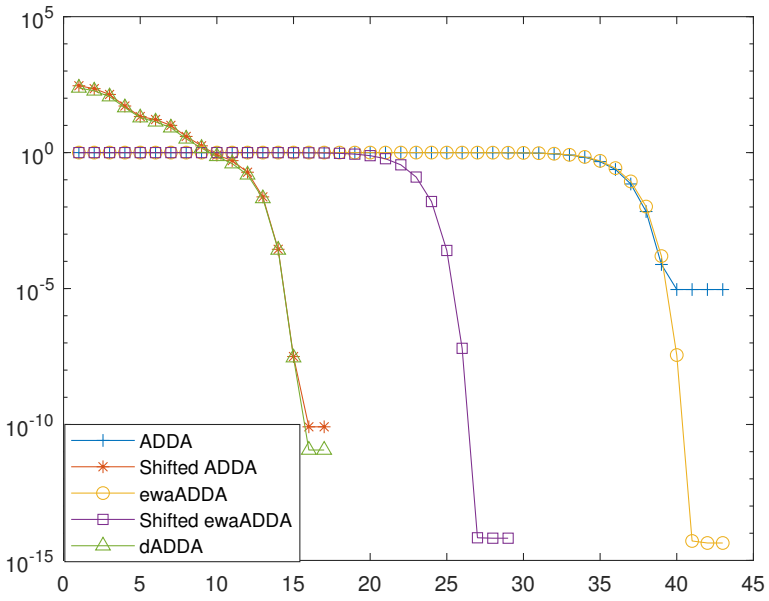


Figure 5.7: Relative elementwise errors for Example 5, step by step ($s = 5$, $\kappa = 10^{-10}$)

Chapter 6

Conclusions

In this chapter we wish to give some final remarks on this work.

It is well known that the existence of a triplet representation for nonsingular and for certain singular M-matrices allows one to perform an elementwise accurate version of the Gaussian elimination for these matrices, the so-called GTH-like algorithm. In the first chapters of this thesis, we have recalled the fundamental notions about the triplet representation of M-matrices and GTH-like algorithm and we have focused on the explicit construction of a triplet representation for the active matrix at k-th step of the Gauss process.

We have tried to summarise, as much as possible, the state of the art in elementwise accurate doubling algorithms, because in the last decade many authors studied how to use the triplet representation of M-matrices in order to compute the minimal nonnegative solution of an M-NARE with high relative elementwise accuracy. We have found an interesting property of splitting with respect to a suitable disk in the complex plan for the spectrum of the matrix $f(\mathcal{H})$ that is connected to the speed of convergence of the ADDA algorithm for M-NARE. We have noticed that this spectral splitting can be reduced to a splitting with respect to the unit disk by a suitable rescaling of the function f in the definition of ADDA iteration. We have pointed out the existence of a triplet representation for all the matrices to be inverted in performing ADDA iterative process, when the matrix M defining the M-NARE is a nonsingular or a regular singular M-matrix with a simple zero eigenvalue.

The last part of the work consists in the description of an elementwise accurate algorithm for shifted M-NARE. The shift technique for M-NARE has been proposed by Guo et al. in 2007 and produces an acceleration of the convergence of SDA (and consequently of ADDA) in the case where M is an irreducible singular M-matrix. In order to apply this approach in an elementwise accurate fashion, we have proposed the delayed shift idea, that allows one to accurately compute the inverse of the matrix in the initial setup of ADDA by using the triplet representation of the M-matrix M instead of the shifted

matrix $\widehat{M}$, that in general is no longer an M-matrix. In addition, we have found sufficient conditions on P_0 (the starting point of the ADDA iteration process) for which all the matrices to be inverted during the ADDA process have a triplet representation. We have provided some results about the convergence properties and about the elementwise stability of the algorithm that we proposed.

Finally, we have shown the effectiveness of our algorithm by testing it on examples known in literature. In particular, we have observed that, when a suitable choice of the parameter of shift η is possible, our approach speeds up the convergence of elementwise accurate ADDA (in the critical case the order of convergence increases from linear to quadratic) and the elementwise relative error in certain cases is considerably smaller than that of the non-elementwise accurate shifted ADDA.

Acknowledgements

I am extremely grateful to my tutor Prof. Bruno Iannazzo for his continuous help and advice during the course of my PhD degree. My gratitude extends to Prof. Federico Poloni for his tresaured support, for his insightful comments and suggestions throughout the research project.

Finally, I would like to offer my special thanks to my parents, my brother, my husband and my children. Without their encouragement and support in the past few years, it would be impossible for me to complete my study.

Bibliography

[1] Alfa, A.S., Xue, J., Ye, Q.: Accurate computation of the smallest eigenvalue of a diagonally dominant M-matrix. Math. Comput. **71**(237), 217-236 (2002).

[2] Alfa, A.S., Xue, J., Ye, Q.: Entrywise perturbation theory for diagonally dominant M-matrices with applications. Numer. Math. **90**(3), 401-414 (2002).

[3] Bean, N.G., O'Reilly, M.M., Taylor, P.G.: Algorithms for return probabilities for stochastic fluid flows. Stochastic Models **21**(1), 149-184 (2005).

[4] Benzi, M.: A direct projection method for Markov chains. Linear Algebra Appl. **486**, 27-49 (2004).

[5] Berman, A., Plemmons, R. J.: Nonnegative matrices in the mathematical sciences. Revised reprint of the 1979 original. Classics in Applied Mathematics, 9. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1994).

[6] Bini, D.A., Iannazzo, B., Meini, B.: Numerical solution of algebraic Riccati equations. Fundamentals of Algorithms, 9. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2012).

[7] Bini, D. A., Latouche, G., Meini, B.: Numerical methods for structured Markov chains. Numerical Mathematics and Scientific Computation. Oxford University Press, New York (2005).

[8] Bini, D.A., Meini, B., Poloni, F.: Transforming algebraic Riccati equations into unilateral quadratic matrix equations. Numer. Math. **116**(4), 553-578 (2010).

[9] Cinlar, E.: Introduction to stochastic processes. Prentice-Hall, Englewood Cliffs (1975).

[10] Demmel, J.: Applied numerical linear algebra. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1997).

[11] Golub, G., Van Loan, C.: Matrix computations, 3rd edition. The Johns Hopkins University Press, Baltimore (1996).

[12] Grassmann, W.K., Taksar, M.T., Heyman, D.P.: Regenerative analysis and steady-state distributions for Markov chains. Oper. Res. **33**(5), 1107-1116 (1985).

[13] Guo, C.H.: Nonsymmetric algebraic Riccati equations and Wiener-Hopf factorization for M-matrices. SIAM J. Matrix Anal. Appl. **23**, 225-242 (2001).

[14] Guo, C.H.: On algebraic Riccati equations associated with M-matrices. Linear Algebra Appl. **439**, 2800-2814 (2013).

[15] Guo, C.H., Iannazzo, B., Meini, B.: On the doubling algorithm for a (shifted) nonsymmetric algebraic Riccati equation. SIAM J. Matrix Anal. Appl. **29**(4), 1083-1100 (2007).

[16] Guo, C.H., Lu, D.: On algebraic Riccati equations associated with regular singular M-matrices. Linear Algebra Appl. **493**, 108-119 (2016).

[17] Guo, X.-X., Lin,W.-W., Xu, S.-F.: A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. Numer. Math. **103**, 393-412 (2006).

[18] Higham, N. J.: Accuracy and stability of numerical algorithms, second edition. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2002).

[19] Huang, T.M., Huang, W.Q., Li, R.-C., Lin, W.W.: A new two-phase structure-preserving doubling algorithm for critically singular M-matrix algebraic Riccati equations. Numer. Linear Algebra Appl. **23**, 291-313 (2016).

[20] Huang, T.-M., Li, R.-C., Lin, W.-W.: Structure-preserving doubling algorithms for nonlinear matrix equations. Fundamentals of algorithms, 14. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2018).

[21] Liu, C., Wang, W.-G., Xue, J., Li, R.-C.: Accurate numerical solution for structured M-matrix algebraic Riccati equations. Journal of Computational and Applied Mathematics **396**, art. no. 113614, (2021).

[22] Meyer, C.D.: Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. SIAM Rev. **31**, 240-272 (1989).

[23] Nguyen, G.T., Poloni, F.: Componentwise accurate fluid queue computations using doubling algorithms. Numer. Math. **130**(4), 763-792 (2015).

[24] O'Cinneide, C.A.: Entrywise perturbation theory and error analysis for Markov chains. Numer. Math. **65**, 109-120 (1993).

[25] O'Cinneide, C.A.: Relative-error bounds for the LU decomposition via the GTH algorithm. Numer. Math. **73**, 507-519 (1996).

[26] Plemmons, R.J.: M-Matrix Characterizations. I – Nonsingular M-Matrices. Linear Algebra Appl. **18**(2), 175-188 (1977).

[27] Poloni, F., Reis, T.: The SDA Method for Numerical Solution of Lur'e Equations. Numer. Linear Algebra Appl. **23**, 169-186 (2016).

[28] Ramaswami, V.: Matrix analytic methods for stochastic fluid flows. Smith, D., Hey, P. (eds.) Proceedings of the 16th International Teletraffic Congress. Teletraffic Engineering in a Competitive World, pp. 1019-1030. Elsevier Science B.V, Edinburgh (1999).

[29] Rogers, L.C.G.: Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. Ann. Appl. Probab. **4**, 390-413 (1994).

[30] Saad, Y.: Iterative methods for sparse linear systems, second edition. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (2003).

[31] Seneta, E.: Complementation in stochastic matrices and GTH algorithm. SIAM J. Matrix Anal. Appl. **19**, 556-563 (1998).

[32] Seneta, E.: Non-negative matrices and Markov chains. Springer, New York (1981).

[33] Stewart, W. J.: Introduction to the numerical solution of Markov chains. Princeton University Press, Princeton (1994).

[34] Wang, W.G., Wang, W.C., Li, R.-C.: Alternating-directional doubling algorithm for M-matrix algebraic Riccati equations. SIAM J. Matrix Anal. Appl. **33**(1), 170-194 (2012).

[35] Wang, W.G., Wang, W.C., Li, R.-C.: Deflating irreducible singular M-matrix algebraic Riccati equations. Numer. Algebra Control Optim. **3**, 491-518 (2013).

[36] Xue, J.: A note on entrywise perturbation theory for Markov chains. Linear Alg. Appl. **260**, 209-213 (1997).

[37] Xue, J., Jiang, E.: Entrywise relative perturbation theory for nonsingular M-matrices and applications. BIT **35**, 417-427 (1995).

[38] Xue, J., Li, R.-C.: Highly accurate doubling algorithms for M-matrix algebraic Riccati equations. Numer. Math. **135**(3), 733-767 (2017).

[39] Xue, J., Xu, S., Li, R.-C.: Accurate solutions of M-matrix algebraic Riccati equations. Numer. Math. **120**(4), 671-700 (2012).

[40] Xue, J., Xu, S., Li, R.-C.: Accurate solutions of M-matrix Sylvester equations. Numer. Math. **120**(4), 639-670 (2012).