A Subspace Shift Technique for Nonsymmetric Algebraic Riccati Equations

Bruno Iannazzo[*] Federico Poloni[†]

December 7, 2011

The worst situation in computing the minimal nonnegative solution of a nonsymmetric algebraic Riccati equation associated with an M-matrix occurs when the corresponding linearizing matrix has two very small eigenvalues, one with positive and one with negative real part. When both these eigenvalues are exactly zero, the problem is called critical or null recurrent. While in this case the problem is ill-conditioned and the convergence of the algorithms based on matrix iterations is slow, there exist some techniques to remove the singularity and transform the problem to a well-behaved one. Ill-conditioning and slow convergence appear also in close-to-critical problems, but when none of the eigenvalues is exactly zero the techniques used for the critical case cannot be applied.

In this paper, we introduce a new method to accelerate the convergence properties of the iterations also in close-to-critical cases, by working on the invariant subspace associated with the problematic eigenvalues as a whole. We present a theoretical analysis and several numerical experiments which confirm the efficiency of the new method.

1 Introduction

We consider the nonsymmetric algebraic Riccati equation (or NARE)

$$XCX - AX - XD + B = 0, \tag{1}$$

where $X, B \in \mathbb{C}^{m\times n}$, $A \in \mathbb{C}^{m\times m}$, $C \in \mathbb{C}^{n\times m}$, $D \in \mathbb{C}^{n\times n}$. We write equation (1) briefly as $\mathcal{R}(X) = 0$ where $\mathcal{R}(X) = XCX - AX - XD + B$.

[*] Dipartimento di Matematica e Informatica. Via Vanvitelli 1, 06123 Perugia, Italy bruno.iannazzo@dmi.unipg.it. The work of the first author was partly supported by PRIN 2008 N. 20083KLJEZ and by GNCS of Istituto Nazionale di Alta Matematica.

[†] Technische Universität Berlin, Strasse des 17 Juni 137, 10623 Berlin poloni@math.tu-berlin.de. The work of the second author was supported by a postdoctoral grant of the A. von Humboldt Foundation since May 2011.

In certain applications in queueing models [20] and in the numerical solution of transport equations [18], the coefficients of (1) are such that

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

is an M-matrix, either nonsingular or singular irreducible. In this case, we give Equation (1) the acronym M-NARE. We recall that $M \in \mathbb{C}^{n\times n}$ is an M-matrix if it can be written in the form $M = sI_n - N$, where I_n is the identity matrix of size n (denoted also by I if there is no ambiguity), N is a matrix whose elements are nonnegative, for which we use the notation $N \geqslant 0$, and $s \geqslant \rho(N)$, where $\rho(\cdot)$ is the spectral radius of a square matrix. The M-matrix M is singular if $s = \rho(N)$ and nonsingular if $s > \rho(N)$. It can be proved that the eigenvalues of an M-matrix have nonnegative real part [1].

The solutions of the NARE (1) can be put in correspondence with certain n-dimensional invariant subspaces of the matrix

$$\mathcal{H} = \begin{bmatrix} D & -C \\ B & -A \end{bmatrix}. \tag{2}$$

More precisely, a matrix $X \in \mathbb{C}^{m\times n}$ is a solution of (1) if and only if the columns of $\begin{bmatrix} I_n \\ X \end{bmatrix}$ span an invariant subspace of $\mathcal{H}$. In particular, it holds that

$$\mathcal{H} \begin{bmatrix} I_n \\ X \end{bmatrix} = \begin{bmatrix} I_n \\ X \end{bmatrix} (D - CX), \tag{3}$$

and the eigenvalues of $D - CX$ are a subset of the eigenvalues of $\mathcal{H}$.

We say that the NARE (1) is *associated with* the matrix $\mathcal{H}$ of (2) or that $\mathcal{H}$ is the *linearizing* matrix of the NARE. Observe that any 2×2 block matrix with square diagonal blocks yields a NARE associated with it.

In the case of an M-NARE it can be proved that the eigenvalues of $\mathcal{H}$ can be ordered by non increasing real part such that

$$\Re\lambda_1 \geqslant \cdots \geqslant \Re\lambda_{n-1} \geqslant \lambda_n \geqslant 0 \geqslant \lambda_{n+1} \geqslant \cdots \geqslant \Re\lambda_{m+n}, \tag{4}$$

that is, n eigenvalues belong to the closed right half complex plane and the other eigenvalues to the closed left half plane, and the eigenvalues λ_n and λ_{n+1} are real. If moreover $\mathcal{M}$ is irreducible, then $\Re\lambda_{n-1} > \lambda_n \geqslant 0 \geqslant \lambda_{n+1} > \Re\lambda_{n+2}$. If $\lambda_n = 0 = \lambda_{n+1}$, then the eigenvalue zero is associated to a size-2 Jordan block (see [4] and the references therein).

All these spectral properties implies that the matrix $\mathcal{H}$ associated with an M-NARE has a unique n-dimensional invariant subspace corresponding to the n *rightmost* eigenvalues, namely $\lambda_1, \ldots, \lambda_n$, which we call the n-dimensional *antistable invariant subspace* of $\mathcal{H}$ (the term comes from the theory of the symmetric algebraic Riccati equations in dynamical systems [19]). On the other hand, the matrix $\mathcal{H}$ has a unique m-dimensional invariant subspace corresponding to the m *leftmost* eigenvalues, namely $\lambda_{n+1}, \ldots, \lambda_{n+m}$, which we call *stable invariant subspace.*

In the applications, the required solution of the M-NARE is the one for which the columns of $\begin{bmatrix} I_n \\ X \end{bmatrix}$ span the n-dimensional antistable invariant subspace of $\mathcal{H}$, or, equivalently, such that the eigenvalues of $D - CX$ are the n rightmost eigenvalues of $\mathcal{H}$. This solution has been proved to exist and it turns out to be the minimal element-wise nonnegative solution of (1) (see [10]).

Equation (1) is usually solved either by some matrix iteration, e.g., the Cyclic Reduction (CR) [2] or the Structured Doubling Algorithm (SDA) [8, 13], whose limits yield the required solution or using the ordered Schur form of $\mathcal{H}$ [11].

Both the conditioning of the equation and the convergence speed of most iterations are strictly related to some property of good separation between the stable and anti-stable subspace; for this purpose, several different measures of "nearness" are used in literature; in Section 2 we introduce and discuss them. Nevertheless, all approaches identify two important cases:

1. $\lambda_n = \lambda_{n+1} = 0$; in this case, the minimal nonnegative solution of (1) is ill-conditioned [12] and the convergence of most iterations degrades from quadratic to linear [7]. This case is known as *critical case.* Most of this problems can be circumvented by using the so-called *shift technique* [13, 17]. It consists in making a special rank-one correction of $\mathcal{H}$, obtaining a new Riccati equation with the same minimal solution. The new equation has better conditioning and the convergence of iterations is quadratic again. The shift technique works not only in the critical case, but also when only one of λ_n and λ_{n+1} is zero, yielding minor benefits in this case.

2. equations that are (in some sense) "close" to having $\lambda_n = \lambda_{n+1} = 0$, while neither of the two eigenvalues is exactly zero. This case is known as *close to critical.* It is effectively the worst-case scenario, since the same difficulties as in the critical case appear (ill-conditioning, slow convergence of the numerical methods based on matrix iterations), but the shift technique cannot be applied as it requires an eigenvalue to be exactly zero.

The difficulties in the latter case are the main motivation for this work. We present a new technique, which we call *subspace shift*, that aims to extend the shift technique to this case. A necessary assumption is that we can identify a small subset of the eigenvalues which are "responsible" for the ill-conditioning of the equation and well separated from the rest of the spectrum; we define this property more precisely in the following. We call the associated subspace *central subspace.* The dimension k of this subspace may be known *a priori* from the theoretical properties of the problem (as for instance in [18]), or determined at run-time, which is a more difficult task.

The technique consists in building a rank-k modification of the matrix that alters the eigenvalues associated with the central subspace, but does not modify the invariant subspaces and the minimal solution. In this way, we reduce the problem to a "far from critical" one, for which the solution can be computed with a faster and stabler iterative method.

The paper is organized as follows. In Section 2, we describe and compare the different notions of distance from criticality which exist in literature and how they affect the conditioning of the problem and the convergence speed of the numerical algorithms. In Section 3 we introduce the subspace shift technique and outline some results that give an insight of its behavior in terms of the different criticality metrics. In Section 4 we describe its implementation in more detail and discuss the computational aspects. Section 5 contains some experimental results that show the effectiveness of this technique. Finally, Section 6 draws the conclusions.

In the following, $\sigma(M)$ stands for the set of the eigenvalues of $M \in \mathbb{C}^{n\times n}$, and $\|\cdot\|_F$ denotes the Frobenius norm. We define the *Cayley transform* of parameter $\gamma \in \mathbb{R} \setminus \{0\}$ as the map

$$\mathcal{C}_\gamma : z \mapsto \frac{z-\gamma}{z+\gamma}.$$

Notice that, for $\gamma > 0$, $\mathcal{C}_\gamma$ maps the open (closed) right half-plane onto the open (closed) unit circle, and the open (closed) left half-plane onto the exterior of the open (closed) unit circle.

2 Measures of criticality: gap and sep

2.1 The gap between eigenvalues

The simplest measure of criticality, adopted in most works on the shift [17, 13] is the so-called *gap*, i.e., the distance between the two eigenvalues closer to the imaginary axis $\operatorname{gap}(\mathcal{H}) := |\lambda_n - \lambda_{n+1}|$. In the critical case $\operatorname{gap}(\mathcal{H}) = 0$, and a problem is called close-to-critical when $\operatorname{gap}(\mathcal{H})$ is small with respect to $\|\mathcal{H}\|$. A strictly related quantity, which appears explicitly in the expressions for the convergence speeds of SDA and Cyclic Reduction [4], is its Cayley-transformed version

$$\operatorname{gap}_{\mathcal{C}_\gamma}(\mathcal{H}) := \frac{\max_{i=1,\dots,n} |\mathcal{C}_\gamma(\lambda_i)|}{\min_{j=1,\dots,m} |\mathcal{C}_\gamma(\lambda_{n+j})|}, \tag{5}$$

where γ is chosen according to

$$\gamma \geqslant \gamma_* = \max\left\{ \max_{1\leqslant i\leqslant m} a_{ii}, \max_{1\leqslant i\leqslant n} d_{ii} \right\}. \tag{6}$$

We have $\operatorname{gap}_{\mathcal{C}_\gamma}(\mathcal{H}) \leqslant 1$, with equality in the critical case. Since the Cayley transform alters the position of the eigenvalues, it is not apparent that the minimum and maximum in (5) are attained in λ_n and λ_{n+1}; we give here a proof of this result. Let us first assess the following technical lemma.

Lemma 1 *Let Γ be a closed disc in the complex plane with center $C \in \mathbb{R}$ and radius r. The point in Γ with maximal modulus is one among $C + r$ and $C - r$.*

Proof. Let $C + p \in \mathbb{C}$, $|p| \leqslant r$, be a generic point in the disc. By the triangle inequality, $|C + p| \leqslant |C| + |p| \leqslant |C| + r$, with equality if and only if $|p| = r$ and p has the same argument as C, i.e., either real positive or negative. □

Theorem 2 *Let $\mathcal{H}$ be associated with an M-NARE, and γ be chosen according to* (6). *The minimum and the maximum in* (5) *are attained by $i = n$ and $j = 1$, i.e., we may replace* (5) *with*

$$\operatorname{gap}_{\mathcal{C}_\gamma}(\mathcal{H}) := \frac{|\mathcal{C}_\gamma(\lambda_n)|}{|\mathcal{C}_\gamma(\lambda_{n+1})|}.$$

Proof. From (6) we have $\gamma I - D \geqslant 0$ and thus $P = \gamma I - D + CX_* \geqslant 0$. Hence we may write $D - CX_* = \gamma I - P$; from the Perron–Frobenius theory of M-matrices, it follows that all the eigenvalues of $D - CX_*$ are contained in the closed disc with center γ and radius $r = \gamma - \lambda_n$; and in particular, the eigenvalue λ_n lies on its boundary. We call this disc Γ, and proceed to prove that $\mathcal{C}_\gamma(\lambda_n)$ has the maximal modulus among all points in in $\mathcal{C}_\gamma(\Gamma)$. The image of Γ under the Cayley transform is a closed disc Γ', which must be contained in the unit disc and symmetric with respect to the real axis. This means that its center (which is not in general $\mathcal{C}_\gamma(\gamma)$) is real. This disc Γ' intersects the real axis in the two points $\mathcal{C}_\gamma(\lambda_n)$ and $\mathcal{C}_\gamma(2\gamma - \lambda_n)$. By the lemma, the point of maximal modulus in Γ' is one among them; direct computation (using $\lambda_n \leqslant \gamma$) shows that it is the former.

A similar reasoning starting from $A - X_* C$ yields that $\min_{j=1,\dots,m} |\mathcal{C}_\gamma(\lambda_{n+j})|$ is achieved by $j = 1$; we need some extra care with the signs, as $\lambda_{n+1} \leqslant 0$, and with the fact that this time the image of the enclosing disc under the Cayley transform is the *outside* of a suitable disc. □

Therefore an explicit relation among the two concepts of gap can be established. Observe that while the gap changes by scaling the matrix $\mathcal{H}$ by a real parameter $\alpha \neq 0$, the Cayley-transformed gap of $\alpha\mathcal{H}$ is the same as the one of $\mathcal{H}$, if the same value of γ is chosen according to (6). Thus, the Cayley transformed gap can be seen as a relative inverse gap.

2.2 The subspace separation

A third, more accurate notion of nearness is given by the *subspace separation* [9, 21]. We first define the *separation* between the two square matrices M and N as

$$\operatorname{sep}(M, N) := \min_{X \neq 0} \frac{\|MX - XN\|}{\|X\|}, \tag{7}$$

where $\|\cdot\|$ is a suitable matrix norm (for instance we denote by sep_F and sep_2 the separation in the Frobenius and spectral norm, respectively), and recall the bound

$$\operatorname{sep}(M, N) \leqslant \min_{\mu \in \sigma(M), \nu \in \sigma(N)} |\mu - \nu|. \tag{8}$$

Given an invariant subspace $\mathcal{W}$ for a matrix $A \in \mathbb{C}^{(n+m)\times(n+m)}$, let Q be an unitary matrix such that

$$Q^* A Q = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \qquad A_{11} \in \mathbb{C}^{n\times n}, \qquad A_{22} \in \mathbb{C}^{m\times m}, \tag{9}$$

and the first n columns of Q span $\mathcal{W}$. We define $\operatorname{sep}(\mathcal{W}) := \operatorname{sep}(A_{11}, A_{22})$ and $\operatorname{relsep}(\mathcal{W}) := \frac{\operatorname{sep}(\mathcal{W})}{\|A\|}$.

When $A = \mathcal{H}$ and $\mathcal{W}$ is the anti-stable space, $\operatorname{relsep}(\mathcal{W})$ gives a third measure of the distance of $\mathcal{H}$ from the critical case. The conditioning of the Riccati equation depends essentially on this separation measure, as shown by the following results. Let the *distance* between two subspaces be defined as $\operatorname{dist}(\mathcal{U}, \mathcal{V}) := \|P_{\mathcal{U}} - P_{\mathcal{V}}\|$, with $P_{\mathcal{W}}$ the orthogonal projection on the image of $\mathcal{W}$.

Theorem 3 ([21]) *Let $\mathcal{W}$ be an invariant subspace of a matrix $\mathcal{H}$ and $\widetilde{\mathcal{W}}$ be an invariant subspace of $\widetilde{\mathcal{H}} = \mathcal{H} + E$, where $\|E\| \leqslant \varepsilon \|\mathcal{H}\|$. Then, for all sufficiently small ε we have*

$$\operatorname{dist}(\mathcal{W}, \widetilde{\mathcal{W}}) \leqslant C \frac{\varepsilon}{\operatorname{relsep} \mathcal{W}}, \tag{10}$$

for a suitable constant C of moderate size.

The result is stated in a stronger form in [21], using the norms of E_{12} and $\mathcal{H}_{12}$ in two suitable block partitions of the involved matrices, but here we favor this form for the sake of simplicity.

A bound on the subspace distance can be transformed into a bound on the solutions of the associated Riccati equations.

Lemma 4 *Let $\mathcal{W} = \operatorname{span} \begin{bmatrix} I_n \\ X \end{bmatrix}$, $\widetilde{\mathcal{W}} = \operatorname{span} \begin{bmatrix} I_n \\ \widetilde{X} \end{bmatrix}$. Then, for both spectral and Frobenius norm, we have $\left\| X - \widetilde{X} \right\| \leqslant \left(\|I_n\|^2 + \|X\|^2 \right)^{1/2} \left(\|I_n\|^2 + \left\| \widetilde{X} \right\|^2 \right)^{1/2} \operatorname{dist}(\mathcal{W}, \widetilde{\mathcal{W}})$.*

Proof. We use the characterization of dist given in [9, Theorem 2.6.1]; the proof there refers to the spectral norm, but it can be adapted to the Frobenius norm. We apply the result to the orthonormal matrices

$$W_2 = \begin{bmatrix} -X^* \\ I \end{bmatrix} (I + XX^*)^{-1/2}, \qquad Z_1 = \begin{bmatrix} I \\ \widetilde{X} \end{bmatrix} (I + \widetilde{X}^* \widetilde{X})^{-1/2},$$

then use norm submultiplicativity

$$\left\| X - \widetilde{X} \right\| \leqslant \left\| (I + X^* X)^{1/2} \right\| \left\| (I + X^* X)^{-1/2} [-X \quad I] \begin{bmatrix} I \\ \widetilde{X} \end{bmatrix} (I + \widetilde{X}^* \widetilde{X})^{-1/2} \right\| \left\| (I + \widetilde{X}^* \widetilde{X})^{1/2} \right\|.$$

Finally, the terms $\left\| (I + X^* X)^{-1/2} \right\|$ and $\left\| (I + \widetilde{X}^* \widetilde{X})^{-1/2} \right\|$ can be transformed into the desired form by taking an SVD of X and $\widetilde{X}$, respectively. □

A tighter bound, which still behaves essentially as $O((\operatorname{relsep} \mathcal{W})^{-1})$, can be found in [15].

2.3 Relations between gap and sep

If the Frobenius norm is used, the $\operatorname{gap}(\mathcal{H})$ and the $\operatorname{sep}(\mathcal{W})$ are the smallest eigenvalue and the smallest singular value of the matrix $(I \otimes A_{11} - A_{22}^T \otimes I)$, respectively (compare (7) and (9)). It is well known that the two numbers coincide for normal matrices [21, Exercise 5.2.1], but may differ significantly in the nonnormal case [21, Example 5.2.4]. The same happens for any norm, and in view of (8) we obtain that $\operatorname{sep}(\mathcal{W}) \leqslant \operatorname{gap}(\mathcal{H})$.

For nonnormal matrices the subspace separation is a better tool to gauge the distance from criticality, especially when conditioning properties are in exam. However, as far as we know, all the literature regarding shift methods deals only with the gap as a measure of criticality. The geometrical intuition is clearer in the gap setting and the proofs are easier to carry on. On the other hand, the whole point of shifting strategies is getting rid of the two real eigenvalues λ_n and λ_{n+1} close to the origin, and it is less clear how we should define the gap in when the two eigenvalues have been moved. In view of (8), we extend the definition of gap in the following way. Let $\mathcal{W}$ be an invariant subspace of a matrix A, and A_{11}, A_{22} as in (9). We set

$$\operatorname{gap}(\mathcal{W}) := \min_{\lambda \in \sigma(A_{11}), \mu \in \sigma(A_{22})} |\lambda - \mu| .$$

Similarly, we define the gap between two invariant subspaces of A as

$$\operatorname{gap}(\mathcal{U}, \mathcal{V}) := \min_{\lambda \text{ associated to } \mathcal{U},\ \mu \text{ associated to } \mathcal{V}} |\lambda - \mu| .$$

In the following, we define our subspace shift technique in terms of the gap metric, because only by resorting to eigenvalue location criteria we can select suitable subspaces for its application. When discussing conditioning, moving to the sep setting (and having good separation properties in this setting) is necessary. However, as far as we know, even a complete theory of the basic shift technique in terms of sep does not exist at present; the intuitive assertion that "things get better when we move the eigenvalues more far apart" is difficult to formalize in terms of the sep metric. We are not able to give full proof for many of the conditioning-related assertions, but we provide at least partial ones that show our claims when the separation bounds behave as suggested by the gap metric analogy. This in particular includes the case in which $\mathcal{H}$ is normal or departs only slightly from normality.

3 Theoretical bases

3.1 The shift technique

The shift technique has been applied in [13] to the M-NARE (1) where $\mathcal{M}$ is a singular irreducible M-matrix, that is, when at least one between λ_n and λ_{n+1} is 0. Without loss of generality one can assume that $\lambda_n = 0$: the case $\lambda_n > 0 = \lambda_{n+1}$ can be reduced to the case $\lambda_n = 0$ by a simple trick [13, Lemma 5.1].

The shift technique is rooted in the following results.

Lemma 5 (Brauer's theorem [6]) *Let (λ, v) be an eigenpair for the matrix T. Let u be a vector with $u^*v = 1$ and s be a scalar. The eigenvalues of the matrix $\widehat{T} := T + svu^*$ are the same as those of T, except for one occurrence of λ which is replaced by $\lambda + s$.*

Theorem 6 ([13]) *Let $\mathcal{H}$ be the as in (2) associated with the M-NARE (1) with $\lambda_n = 0$, and let v_n be an eigenvector relative to λ_n; consider the matrix*

$$\widehat{\mathcal{H}} := \mathcal{H} + sv_n u^*,$$

with $u^*v_n = 1$ and $s > 0$. Then, the minimal solution X_* of the M-NARE associated with $\mathcal{H}$ is a solution of the NARE associated with $\widehat{\mathcal{H}}$. Moreover, m eigenvalues of $\widehat{\mathcal{H}}$ lie in the closed left half plane and n in the open right half plane, whose corresponding invariant subspace is spanned by the the columns of $\begin{bmatrix} I \\ X_* \end{bmatrix}$.

The shift technique consists in computing one eigenvector v_n corresponding to the eigenvalue $\lambda_n = 0$, and using it to construct the NARE associated with $\widehat{\mathcal{H}}$, which we call the *shifted NARE*. The matrix $\widehat{\mathcal{H}}$ has eigenvalues $\lambda_1, \dots, \lambda_{n-1}, \widehat{\lambda}_n, \lambda_{n+1}, \dots, \lambda_{n+m}$ where $\widehat{\lambda}_n = s$ (the eigenvalue λ_n has been "shifted" from 0 to s, this justifies the name of the technique). Observe that $\operatorname{gap}(\widehat{\mathcal{H}}) > \operatorname{gap}(\mathcal{H})$; thus, better conditioning and faster convergence are expected, once the Cayley parameter γ is fixed. It has been proved in [13] that SDA applied to the shifted equation, using the same Cayley parameter as the nonshifted case, but with the initial values as in (14), constructed from $\widehat{\mathcal{H}}$, converges quadratically with a better rate of convergence than the nonshifted case. Numerical experiments [13, 3] show that this technique reduces dramatically the number of steps of iterations like SDA and CR. In the critical case, the convergence from linear becomes quadratic.

It can happen that the Riccati equation associated with $\widehat{\mathcal{H}}$ is not an M-NARE; that is, $\widehat{\mathcal{M}} = \left[\begin{smallmatrix} I & 0 \\ 0 & -I \end{smallmatrix}\right] \widehat{\mathcal{H}}$ need not be an M-matrix. Hence, there is no guarantee that the SDA can be carried out without breakdown, even if in practice this method works well and the applicability of SDA is usually assumed [13].

Since $\lambda_n = 0$, the vector v_n can be computed easily as $\ker \mathcal{M}$. In principle, the shift technique could be used also for nonsingular M-matrices, i.e., the hypothesis $\lambda_n = 0$ is not actually needed in Theorem 6. In this case there is no simple relation among the eigenvectors of $\mathcal{M}$ and $\mathcal{H}$, and different techniques are needed for the computation of λ_n and v_n; for instance, the power method. However, in the close-to-critical case the eigenvector v_n is ill-conditioned and therefore it cannot be computed with good accuracy.

To solve this problem, we present in Section 3.3 a new technique which shifts a whole invariant subspace containing λ_n and λ_{n+1}, without attempting to separate the eigenvectors. To this purpose we use an invariant subspace whose associated eigenvalues are well separated from the rest of the spectrum.

3.2 Results on separation

We first provide a couple of simple lemmas that will be used in the following.

Lemma 7 *Let the spectrum of $M \in \mathbb{C}^{n\times n}$ be the union of two disjoint sets, say Λ_1 and Λ_2. Let the columns of U and V (with full column rank) span the left and right invariant subspaces of M corresponding to the eigenvalues in Λ_1, respectively. Then U^*V is nonsingular.*

Moreover, let the columns of W (with full column rank) span a right invariant subspace of M whose corresponding eigenvalues belong all to Λ_2, then $U^*W = 0$.

Proof. Let $L^{-1}ML = J$ be the Jordan canonical form of M where the Jordan blocks are ordered such that the k eigenvalues in Λ_1 come first, then, a basis of the right invariant subspace of M corresponding to Λ_1 is made by the first k columns of L, say L_1. Thus $V = L_1P$ for some nonsingular $P \in \mathbb{C}^{k\times k}$. Similarly, $U^* = QR_1^*$ where R_1^* are the first k rows of L^{-1} and $Q \in \mathbb{C}^{k\times k}$ is nonsingular, thus $U^*V = QR_1^*L_1P = QP$ is nonsingular. By a similar argument, it can be proved that left and right invariant subspaces corresponding to different eigenvalues are orthogonal. □

Lemma 8 *Let*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

where $\sigma(A_{11}) \cap \sigma(A_{22}) = \emptyset$.

1. *There is a matrix*

$$Z_e = \begin{bmatrix} I & Z \\ 0 & I \end{bmatrix}$$

with $\|Z\| \leqslant \frac{\|A_{12}\|}{\mathrm{sep}(A_{11},A_{22})}$ such that $Z_eAZ_e^{-1} = \mathrm{diag}(A_{11}, A_{22})$

2. *For each B, $\mathrm{sep}(A_{11}, B) \geqslant \mathrm{sep}(A, B)$ and $\mathrm{sep}(A_{22}, B) \geqslant \mathrm{sep}(A, B)$.*

3. $\mathrm{sep}(M, N) \leqslant \mathrm{sep}(T_1MT_1^{-1}, T_2NT_2^{-1})\kappa(T_1)\kappa(T_2)$.

4. *If $M = \mathrm{diag}(M_1, M_2)$ and $N = \mathrm{diag}(N_1, N_2)$, then $\mathrm{sep}(M, N) = \min_{i,j=1,2} \mathrm{sep}(M_i, N_j)$.*

Items 1, 3 and 4 are in [21]. Item 2 follows from the definition of sep, by noting that the minimum of $AX - XB$ increases if we restrict to matrices X, partitioned conformably with A as $X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}^*$, where one of the two blocks is zero.

3.3 The subspace shift technique

Let $\mathcal{V}$ and $\bar{\mathcal{V}}$ be two complementary invariant subspaces of $\mathcal{H}$ satisfying the following conditions:

1. $\mathcal{V}$ has small dimension k.

2. $\mathcal{V}$ is well separated from $\bar{\mathcal{V}}$, i.e., $\mathrm{gap}(\mathcal{V}, \bar{\mathcal{V}}) \geqslant \delta_1$, for a $\delta_1 > 0$ not excessively small.

3. $\bar{\mathcal{V}}$ does not contain eigenvalues close to the imaginary axis, i.e., $\mathrm{gap}(\bar{\mathcal{V}}_s, \bar{\mathcal{V}}_u) \geqslant \delta_2$, for a $\delta_2 > 0$ not excessively small, where $\bar{\mathcal{V}}_s$ and $\bar{\mathcal{V}}_u$ are the invariant subspaces associated with the stable and anti-stable part of $\bar{\mathcal{V}}$.

Let $V, U \in \mathbb{C}^{(n+m)\times k}$ be matrices whose orthonormal columns span respectively $\mathcal{V}$ and the left invariant subspace $\mathcal{U}$ associated with the same eigenvalues as $\mathcal{V}$. Notice that $\mathcal{U}$ is well defined, since the subset of $\sigma(\mathcal{H})$ associated to $\mathcal{V}$ and $\bar{\mathcal{V}}$ are disjoint because $\operatorname{gap}(\mathcal{V}, \bar{\mathcal{V}}) > 0$. By Lemma 7, U^*V is nonsingular.

We consider the matrix

$$\widehat{\mathcal{H}} = \mathcal{H}(I + sV(U^*V)^{-1}U^*), \tag{11}$$

which is a rank k modification of $\mathcal{H}$. Its spectral properties are summarized by the following result.

Theorem 9 *Let the spectrum of $\mathcal{H} \in \mathbb{C}^{(n+m)\times(n+m)}$ be the union of two disjoint sets, say $\Lambda_1 = \{\xi_1, \dots, \xi_k\}$ and $\Lambda_2 = \{\xi_{k+1}, \dots, \xi_{n+m}\}$. Let $V, U \in \mathbb{C}^{(n+m)\times k}$ be matrices whose orthonormal columns span the right and left invariant subspaces associated with the eigenvalues of Λ_1, respectively. The matrix $\widehat{\mathcal{H}}$ in* (11) *has the same right invariant subspaces as $\mathcal{H}$ and eigenvalues $\{(1+s)\xi_1, \dots, (1+s)\xi_k, \xi_{k+1}, \dots, \xi_{n+m}\}$.*

Proof. As above, let $\mathcal{V}$ be the right invariant subspace of $\mathcal{H}$ corresponding to the eigenvalues in Λ_1 and let $\bar{\mathcal{V}}$ be the right invariant subspace complementary to $\mathcal{V}$, corresponding to the remaining eigenvalues $\{\xi_{k+1}, \dots, \xi_{m+n}\}$. Let $\mathcal{W}_1 \subset \mathcal{V}$ be a right invariant subspace of $\mathcal{H}$ spanned by the column of the matrix $W_1 \in \mathbb{C}^{(n+m)\times \ell_1}$, then $W_1 = VQ$ for some matrix $Q \in \mathbb{C}^{k\times \ell_1}$, thus

$$\widehat{\mathcal{H}}W_1 = \mathcal{H}(W_1 + sV(U^*V)^{-1}U^*VQ) = \mathcal{H}(W_1 + sVQ) = \mathcal{H}W_1(1+s),$$

and thus $\{(1+s)\xi_1, \dots, (1+s)\xi_k\}$ are eigenvalues of $\widehat{\mathcal{H}}$.

Let $\mathcal{W}_2 \subset \bar{\mathcal{V}}$ be a right invariant subspace of $\mathcal{H}$ spanned by the column of the matrix $W_2 \in \mathbb{C}^{(n+m)\times \ell_2}$, then by Lemma 7 we have $U^*W_2 = 0$ and hence $\widehat{\mathcal{H}}W_2 = \mathcal{H}W_2$. Thus, $\{\xi_{k+1}, \dots, \xi_{m+n}\}$ are eigenvalues of $\widehat{\mathcal{H}}$. Since any right invariant subspace of $\mathcal{H}$ can be written as the sum of two invariant subspaces contained in $\mathcal{V}$ and $\bar{\mathcal{V}}$, respectively, the proof is completed. □

Theorem 9 can be applied to the linearizing matrix $\mathcal{H}$ of a close-to-critical M-NARE (1), where $\mathcal{V}$ is chosen to be a subspace containing both λ_n and λ_{n+1}. Then the antistable invariant subspace $\mathcal{W}$ and thus the Riccati solution X_* of the NARE associated to $\widehat{\mathcal{H}}$ are the same as the ones for (1).

From the point of view of the eigenvalue location and of the gap metric, the behavior of the subspace shift technique is clear: the eigenvalues closer to the imaginary axis, which are responsible for the slow convergence and ill-conditioning, are multiplied by a factor $1+s$, which takes them farther from the imaginary axis and thus improve the gap. If the factor $1+s$ is not too large then the Cayley gap is reduced and the numerical algorithms based on matrix iterations converge faster. In order to give a complete treatment of the stability properties of the technique, we have to resort to the separation metric.

3.4 Conditioning of U^*V

For the subspace shift technique, we need to form $(U^*V)^{-1}$; therefore, it is crucial that the condition of this matrix is not worse than the conditioning of the problem we are solving. Similarly to what happens in the original problem, we can relate its conditioning to the separation between $\mathcal{V}$ and $\bar{\mathcal{V}}$.

Theorem 10 *Let U and V be orthonormal bases for $\mathcal{U}$ and $\mathcal{V}$ as defined above. Then, we have*

$$\left\|(U^*V)^{-1}\right\| \leqslant C\,\mathrm{relsep}(\mathcal{V})^{-1} + D$$

for moderate constants $C, D > 0$.

Proof. Perform an orthonormal change of basis so that $V = \begin{bmatrix} I \\ 0 \end{bmatrix}$ and partition

$$\mathcal{H} = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}.$$

Let Z as in item 1 of Theorem 8; in particular we have $\|Z\| \leqslant \mathrm{relsep}(\mathcal{V})^{-1}$. Notice that $(I + ZZ^*)^{-1/2}\begin{bmatrix} I & -Z \end{bmatrix}$ is another orthonormal basis of $\mathcal{U}$, thus $U^* = Q(I + ZZ^*)^{-1/2}\begin{bmatrix} I & -Z \end{bmatrix}$ for a suitable unitary Q. We have then

$$\left\|(U^*V)^{-1}\right\| = \left\|(I + ZZ^*)^{1/2}\right\| = (\|I\|^2 + \|Z\|^2)^{1/2},$$

where the last equality can be established as in the proof of Lemma 4. □

3.5 Conditioning of the shifted problem

While not a formal proof, the following argument can be used to gauge the conditioning of the shifted problem. Apply an orthogonal change of basis to bring $\mathcal{H}$ in the form

$$\begin{bmatrix} V_a & G_a & * & * \\ & \bar{V}_a & * & * \\ & & V_s & G_s \\ & & & \bar{V}_s \end{bmatrix},$$

where $\sigma(V_s)$ and $\sigma(V_a)$ contain respectively the stable and antistable eigenvalues associated with $\mathcal{V}$, and $\sigma(\bar{V}_s)$, $\sigma(\bar{V}_a)$ the stable and antistable eigenvalues associated with $\bar{\mathcal{V}}$. We may find a matrix R_a such that

$$R_a \begin{bmatrix} V_a & G_a \\ 0 & \bar{V}_a \end{bmatrix} R_a^{-1} = \mathrm{diag}(V_a, \bar{V}_a),$$

and $\kappa(R_a) = O\left(\frac{\|G_a\|^2}{\mathrm{sep}(V_a,\bar{V}_a)^2}\right)$. If $\mathrm{gap}(V_a, \bar{V}_a)$ is a good approximation of the corresponding sep, or if $\mathcal{V}$ and $\bar{\mathcal{V}}$ are well separated also in the sep sense, we expect this condition numbers to be moderate. We argue similarly for the blocks indexed by s, and construct a matrix R_s which annihilates G_s. We let $R := \operatorname{diag}(R_a, R_s)$.

Using the special block structure of $R\mathcal{H}R^{-1}$, we may construct the matrices U and V needed for the shift technique as

$$V = R^{-1}\begin{bmatrix} I & * \\ 0 & * \\ 0 & I \\ 0 & 0 \end{bmatrix}, \qquad U = R\begin{bmatrix} I & 0 \\ 0 & 0 \\ * & I \\ * & 0 \end{bmatrix},$$

with $U^*V = I$. Therefore $\widehat{\mathcal{H}}$ takes the form

$$\widehat{\mathcal{H}} = R^{-1}\begin{bmatrix} (s+1)V_a & 0 & * & * \\ 0 & \bar{V}_a & * & * \\ 0 & 0 & (s+1)V_s & 0 \\ 0 & 0 & 0 & \bar{V}_s \end{bmatrix} R,$$

where the entries marked with an asterisk might be affine functions of s. Thanks to Lemma 8, the separation of the anti-stable subspace in $\widehat{\mathcal{H}}$ is

$$\widehat{\operatorname{sep}\mathcal{W}} \geqslant \frac{\min\bigl(\operatorname{sep}((s+1)V_a,(s+1)V_s),\operatorname{sep}((s+1)V_a,\bar{V}_s),\operatorname{sep}(\bar{V}_a,(s+1)V_s),\operatorname{sep}(\bar{V}_a,\bar{V}_s)\bigr)}{\kappa(R_a)\kappa(R_s)}. \tag{12}$$

If s is moderate and good separation properties hold among $\mathcal{V}$ and $\bar{\mathcal{V}}$, and between the stable and unstable part of $\mathcal{V}$, then we may expect, in analogy with the gap setting, that this minimum is attained by the first element, i.e., $\widehat{\operatorname{sep}\mathcal{W}} \geqslant (s+1)\frac{1}{\kappa(R_a)\kappa(R_s)}\operatorname{sep}\mathcal{W}$. This means that, up to factors that depends on the separation of the central subspace $\operatorname{sep}\mathcal{V}$, the conditioning of the shifted problem is improved by a factor $s+1$.

Unfortunately, a full proof of the fact that the minimum in (12) is attained by the critical subspaces seems elusive. As far as we know, there are no results in literature concerning the behavior of the separation when a scaling is applied to one of the two matrices in a way that takes the eigenvalues “more far apart” (in some suitable setting). The analogy with the gap setting and the experimental results in Section 5 seem to back up our claim.

3.6 Stability under perturbations of the computed central subspace

With a similar reasoning, we may try to estimate the impact of errors in the computed bases for the left and right central subspaces on the accuracy of the solution. We choose a basis as in Theorem 10. If we use perturbed versions of U^* and V as

$$V = \begin{bmatrix} I + E_1 \\ E_2 \end{bmatrix} \qquad U^* = \begin{bmatrix} I + F_1 & -Z + F_2 \end{bmatrix},$$

we obtain instead of $\widehat{\mathcal{H}}$

$$\widetilde{\mathcal{H}} = \begin{bmatrix} (s+1)H_{11}+s(E_1H_{11}+H_{11}F_1+E_1H_{11}F_1) & H_{12}-sH_{11}Z+s(-E_1H_{11}Z+H_{11}F_2+E_1H_{11}F_2) \\ sE_2H_{11}(I+F_1) & H_{22}+sE_2H_{11}(-Z+F_2) \end{bmatrix}.$$

If $\|E_i\|, \|F_i\| \leqslant \varepsilon$, then $\left\|\widehat{\mathcal{H}} - \widetilde{\mathcal{H}}\right\| = Ks\varepsilon \|Z\| \|\mathcal{H}\|$ for a moderate constant K. From the reasoning in the previous section, we expect $\widehat{\operatorname{relsep}\mathcal{W}}^{-1} \leqslant \frac{\kappa(R_a)\kappa(R_s)}{s+1} \operatorname{relsep}\mathcal{W}^{-1}$, thus by (10) the computed subspace $\widetilde{\mathcal{W}}$ satisfies

$$\operatorname{dist}(\mathcal{W}, \widetilde{\mathcal{W}}) \leqslant C \frac{Ks\varepsilon \|Z\|}{\widehat{\operatorname{relsep}\mathcal{W}}} = \frac{s}{s+1} \frac{CK\kappa(R_a)\kappa(R_s) \|Z\| \varepsilon}{\operatorname{relsep}\mathcal{W}}.$$

The factor $\frac{s}{s+1}$ is bounded by 1, and this bounds differs from the perturbation bound (10) for the nonshifted problem only by factors depending on $\operatorname{relsep}\mathcal{V}$.

4 The SuShi (Subspace Shift) algorithm

We assume that the space $\mathcal{V}$ corresponding to the smallest in modulus eigenvalues of $\mathcal{H}$ verifies the assumptions stated at the begin of Section 3.3. We call $\mathcal{V}$ the *central subspace* of $\mathcal{H}$.

In this case, the subspace shift technique of Section 3.3 can be easily translated into a numerical algorithm for solving close-to-critical M-NARE (1). We call it SuShi (Subspace Shift) algorithm.

Algorithm 1 SuShi algorithm for the solution of a close-to-critical M-NARE

```
choose $k$
compute two matrices $U, V$ with orthogonal columns which span the left and right
  invariant subspaces of $\mathcal{H}$ corresponding to its $k$ smallest in modulus eigenvalues,
  $\xi_1, \ldots, \xi_k$, respectively
choose $s > 0$ and compute $\widehat{\mathcal{H}} = \mathcal{H}(I + sV(U^*V)^{-1}U^*)$
solve the NARE $\widehat{\mathcal{R}}(X) = 0$ associated with $\widehat{\mathcal{H}}$, to get the minimal nonnegative
  solution $X_*$ of the original M-NARE $\mathcal{R}(X) = 0$
```

Algorithm 1 is very simple. Nevertheless, in order to get a decent implementation, the details need to be tackled with some care. One could ask how to dynamically determine the value of k, how to compute U and V, how to determine the value of s and how to efficiently solve the "shifted" NARE $\widehat{\mathcal{R}}(X) = 0$. This is the topic of the next sections.

4.1 Computation of the central invariant subspaces

The central left and right invariant subspaces are the ones corresponding to the smallest in modulus eigenvalues of $\mathcal{H}$, then the inverse orthogonal iteration of [9, Section 9.3.2] applied to $\mathcal{H}$ and $\mathcal{H}^*$, respectively, converges to these subspaces. Notice that we assume that $\mathcal{H}$ is nonsingular so that the customary shift technique cannot be used.

The luckiest situation arises when $k = 2$ and the eigenvalues λ_n and λ_{n+1} of $\mathcal{H}$ are the smallest in modulus and well separated from the others, as in the problem of [18].

In the general case, setting $k = 2$ may not yield the desired results since we could have close-to-critical settings such that λ_n and λ_{n-1} are very close to each other and to zero, and λ_{n+1} slightly larger in modulus than both of them. Moreover, even if we shift away λ_n and λ_{n+1}, the remaining eigenvalues of $\mathcal{H}$ (e.g., λ_{n-1} in a setting similar to the case above) could be very close to them, and thus the convergence speed of the Riccati solvers based on matrix iterations is almost unchanged; therefore, the subspace shift with $k = 2$ can still be applied but is not much useful.

For these reasons, we would like to handle cases where k may be larger than two, this means that there is some eigenvalue different from λ_n and λ_{n+1} near to the origin. Our minimal assumption is that there exists a value of $k \geqslant 2$ such that the k smallest eigenvalues of $\mathcal{H}$ are near to each other (close-to-critical case) and well separated from the $(2n - k)$ largest eigenvalues (compare the assumptions of Section 3.3).

If the value of k is not known in advance it can be determined using the following strategy: start the inverse orthogonal iteration with $k = 2$, estimate its convergence speed after some steps. If the iteration shows itself too slow, enlarge k until the convergence becomes (possibly) fast. This approach yields together V and k. The matrix U can be obtained applying a subspace iteration to $\mathcal{H}^*$.

4.2 Magnitude of the shift parameter

Another issue which appears in the practical implementation is the selection of the shift parameter s in Algorithm 1, which should be automatic to deal with different problems.

A major concern is that if the chosen value of s is too small, then the two central eigenvalues do not move significantly and the gap remains small; on the other hand, if the shift parameter is excessively large then $\left\|\widehat{\mathcal{H}}\right\|_F$ grows, and the conditioning of the shifted Riccati equation degrades, according to the results of 2.

Let $\xi_1, \ldots, \xi_{m+n}$ be the eigenvalues of $\mathcal{H}$ ordered by nondecreasing modulus. When the main objective is the acceleration of matrix iterations, like the SDA, the value of s should be chosen such that the eigenvalues corresponding to the central subspaces, say $\xi_1, \ldots, \xi_k$, which are responsible of the slow convergence, gets far from the imaginary axis.

We need to estimate how small they are with respect to the other eigenvalues. We may get an estimate using the convergence speed of the inverse orthogonal iteration. Notice that, with our assumptions on the eigenvalues, the convergence rate is determined by

$$t = \frac{|\xi_k|}{|\xi_{k+1}|},$$

hence a rough estimate for t is given by comparing successive iterates of the subspace iteration. The value ξ_k (or ξ_1), for small k, can be easily computed, since it is the largest in modulus eigenvalue of the $k \times k$ matrix $V^*\mathcal{H}V$, an alternative approach is to approximate it using power methods or some steps of the Arnoldi iteration.

Once an approximation of t has been computed, if we choose s such that $(1 + s)\xi_1 > \xi_{k+1}$, then all the eigenvalues $(1 + s)\xi_1, \ldots (1 + s)\xi_k$ become larger in modulus than δ.

4.3 Solution of the shifted equation

A popular algorithm for computing the minimal solution of an M-NARE is the Structured Doubling Algorithm (SDA), which, in the formulation of [16], is a system of rational matrix iterations defined as

$$\begin{aligned} E_{k+1} =& E_k(I - G_kH_k)^{-1}E_k, \\ F_{k+1} =& F_k(I - H_kG_k)^{-1}F_k, \\ G_{k+1} =& G_k + E_k(I - G_kH_k)^{-1}G_kF_k, \\ H_{k+1} =& H_k + F_k(I - H_kG_k)^{-1}H_kE_k, \end{aligned} \tag{13}$$

with suitable initial values $E_0 \in \mathbb{C}^{n\times n}$, $F_0 \in \mathbb{C}^{m\times m}$, $G_0 \in \mathbb{C}^{n\times m}$, $H_0 \in \mathbb{C}^{m\times n}$. We say that the SDA is applicable (for a set of initial values E_0, F_0, G_0, H_0) if the matrix $I - G_kH_k$ (or equivalently $I - H_kG_k$) is nonsingular for each k otherwise we say that the SDA has a breakdown.

In the case of the M-NARE, choosing the initial values of the SDA as

$$\begin{aligned} E_0 =& I - 2\gamma V_\gamma^{-1}, & F_0 =& I - 2\gamma W_\gamma^{-1}, \\ G_0 =& 2\gamma D_\gamma^{-1}CW_\gamma^{-1}, & H_0 =& 2\gamma W_\gamma^{-1}BD_\gamma^{-1}, \\ A_\gamma =& A + \gamma I, & D_\gamma =& D + \gamma I, \\ W_\gamma =& A_\gamma - BD_\gamma^{-1}C, & V_\gamma =& D_\gamma - CA_\gamma^{-1}B, \end{aligned} \tag{14}$$

for $\gamma \geqslant \gamma_*$, defined in (6), yields well defined sequences such that $G_k \to X_*$ and $H_k \to Y_*$ where X_* is the minimal nonnegative solution of the M-NARE, while Y_* is the minimal nonnegative solution of the *dual* M-NARE: $YBY - YA - DY + C = 0$. In the critical cases the convergence is linear, while in the other cases is quadratic with rate

$$\nu = \operatorname{gap}_{\mathcal{C}_\gamma}(\mathcal{H}) = \frac{|\mathcal{C}_\gamma(\lambda_n)|}{|\mathcal{C}_\gamma(\lambda_{n+1})|}. \tag{15}$$

Moreover, the value of $\gamma \geqslant \gamma_*$ that yields faster convergence is $\gamma = \gamma_*$ (see [4] and the references therein).

The SDA can be applied with minor modifications to the equation associated with the subspace shifted matrix

$$\widehat{\mathcal{H}} =: \begin{bmatrix} \widehat{D} & -\widehat{C} \\ \widehat{B} & -\widehat{A} \end{bmatrix}.$$

It is enough to start the SDA with E_0, F_0, G_0, H_0 obtained using formulae (14) with the new coefficients $\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}$ instead of A, B, C, D and with the same γ.

The new matrix $\widehat{\mathcal{H}}$ might not be an M-matrix so that the applicability should be assumed, in this case one can prove that the method converges to the same limit matrices with a rate which is

$$\widehat{\nu} = \frac{\max_{i=1,\ldots,n} |\mathcal{C}_\gamma(\widehat{\lambda}_i)|}{\min_{j=1,\ldots,m} |\mathcal{C}_\gamma(\widehat{\lambda}_{n+j})|}, \tag{16}$$

where $\widehat{\lambda}_1, \dots, \widehat{\lambda}_{n+m}$ are the eigenvalues of $\widehat{\mathcal{H}}$. The proof of convergence may be obtained with similar manipulations as the ones of [16, 13], so we decided to omit it.

Using the same parameter γ for both standard SDA and SDA applied to the shifted equation gives different convergence rates according to (15) and (16). An acceleration is obtained in the shifted case if $\widehat{\nu} < \nu$. On the assumption that the central eigenvalues are near to the origin, there exists $s_{\max}$ such that $\widehat{\nu} < \nu$ for any value of the shift parameter $0 < s < s_{\max}$.

5 Numerical experiments

We present some numerical examples showing the effectiveness of the subspace shift technique, through the algorithms presented in Section 4, in solving close-to-critical nonsingular M-NARE, when the assumptions of Section 3.3 are fulfilled; that is, when the k eigenvalues corresponding to the central subspaces are nonzero and well separated from the other eigenvalues of $\mathcal{H}$. We recall that these assumptions can be identified dynamically by the algorithm.

We compare the SDA applied to the original equation and to the shifted one. We report the number of steps required by the SDA to converge and the number of steps of the inverse orthogonal iteration in the subspace shift algorithm. All steps of the SDA and the first step of the inverse orthogonal iteration are the most expensive part of the algorithms, since their asymptotic cost is cubic with respect to the size of the matrices; for instance, for $m = n$, the cost of a step of the SDA is $O(n^3)$ elementary arithmetic operations.

We estimate the accuracy of the computed solution $\widetilde{X}_*$ by means of the relative error

$$\mathrm{err} = \frac{\left\| \widetilde{X}_* - X_* \right\|_F}{\|X_*\|_F},$$

where X_*, if available, is the exact solution or an approximation of it obtained using a higher precision. Otherwise, we use the relative residual

$$\mathrm{res} = \frac{\left\| \mathcal{R}(\widetilde{X}_*) \right\|_F}{\left\| \widetilde{X}_* C \widetilde{X}_* + B \right\|_F + \left\| A\widetilde{X}_* + \widetilde{X}_* D \right\|_F}.$$

In our experiments the Frobenius norm is used. In the tests we use the value 10^{-15} as stopping tolerance, a lower tolerance has been used sometimes to monitor the error.

Test 1 As a first test, we consider the close-to-critical cases of the transport problem treated in [14, 18, 5]. It is an M-NARE with square coefficients of size n depending on two parameters $0 \leqslant \alpha < 1$ and $0 < c \leqslant 1$ (for the complete definition and the meaning of the parameters see [18]). The problem is critical for $(\alpha, c) = (0, 1)$, and it is close-to-critical if α and c approach simultaneously 0 and 1, respectively. In this problem $k = 2$ and the two eigenvalues λ_n and λ_{n+1} are well separated from the others.

β	$\mathrm{gap}(\mathcal{H})$	$\mathrm{rsep}(\mathcal{W})$	$\mathrm{gap}_{\mathcal{C}_\gamma}(\mathcal{H})$	$\mathrm{rsep}(\mathcal{U})$	$\mathrm{gap}(\widehat{\mathcal{H}})$	$\mathrm{rsep}(\widehat{\mathcal{W}})$	$\mathrm{gap}_{\mathcal{C}_\gamma}(\widehat{\mathcal{H}})$
10^{-3}	0.11	$4.5 \cdot 10^{-3}$	0.98	$2.9 \cdot 10^{-2}$	2.5	$2.3 \cdot 10^{-2}$	0.69
10^{-6}	$3.5 \cdot 10^{-3}$	$1.4 \cdot 10^{-4}$	0.9995	$2.9 \cdot 10^{-2}$	2.5	$7.1 \cdot 10^{-4}$	0.69
10^{-12}	$3.5 \cdot 10^{-6}$	$1.4 \cdot 10^{-7}$	≈ 1	$2.9 \cdot 10^{-2}$	2.5	$7.1 \cdot 10^{-7}$	0.69

Table 1: Several separation measures for the transport problem with $n = 4$

For $n = 4$ and certain values of β such that $(\alpha, c) = (\beta, 1-\beta)$, we compute the absolute gap of $\mathcal{H}$ ($\mathrm{gap}(\mathcal{H}) = |\lambda_n - \lambda_{n+1}|$), the relative sep of its stable subspace ($\mathrm{rsep}(\mathcal{W})$, defined in Section 2.2), the Cayley-transformed gap ($\mathrm{gap}_{\mathcal{C}_\gamma}(\mathcal{H})$, defined in (5)) and the same quantities for the shifted matrix $\widehat{\mathcal{H}}$, where we have chosen $\gamma = \gamma_*$ of (6) in both cases and $s = \xi_3/\xi_1 - 1$, where ξ_i are the eigenvalues of $\mathcal{H}$ ordered by nondecreasing modulus. We have computed moreover the relative sep of the central subspace, indicated by $\mathrm{rsep}(\mathcal{U})$. The results are reported in Table 1.

Since the conditioning of an invariant subspace is proportional to the reciprocal of the relative sep, we observe that the central invariant subspace is much better conditioned than the stable subspace and that the conditioning of the shifted problem is not worse than the one of the original problem.

Recall that $\mathrm{gap}_{\mathcal{C}_\gamma}(\mathcal{H})$ and $\mathrm{gap}_{\mathcal{C}_\gamma}(\widehat{\mathcal{H}})$ are the parameters of quadratic convergence of the SDA in the nonshifted and the shifted case, respectively. If $\mathrm{gap}_{\mathcal{C}_\gamma}(\cdot)$ is near to 1, we expect a large number of steps of SDA to obtain the desired accuracy. This suggest that the SDA applied to the shifted equation converge much faster, as shown in the next tests.

Test 2 We consider the transport problem of Test 1, to which the subspace shift algorithm is applied, where the Riccati equations are solved with the SDA.

In Table 2 we give the number of SDA iterations needed to get the best relative residual for different matrix sizes n and choices of the parameters β (in a stand-alone implementation the number of iterations may be slightly larger due to a non optimal stopping criterion). We provide in parentheses the number of orthogonal iterations needed to approximate the central invariant subspace in the shifted case, where the shift parameter is chosen with the approximation strategy of Section 4.2.

As β approaches zero, the problem becomes close-to-critical; in fact β is strictly related to the relative gap. The table reports also $\mathrm{gap}(\mathcal{H})$ and the minimum distance δ from the two eigenvalues λ_n and λ_{n+1} to the other eigenvalues of $\mathcal{H}$.

As one can see the problem is well suited to be solved by our algorithms since the central eigenvalues are well separated from the others and δ is always large enough while the gap goes to zero; this shows that this example fits adequately the assumptions of Section 3.3.

Test 3 The second example is taken from [10]. The matrix $\mathcal{M}$ is a random M-matrix of size n and depending on a parameter α. As α tends to 0, the matrix $\mathcal{M}$ tends to

n	β	gap($\mathcal{H}$)	δ	SDA its	SDA res	Alg 1 its	Alg 1 res
32	10^{-3}	0.11	0.96	15	$8.8 \cdot 10^{-15}$	11 (12)	$4.2 \cdot 10^{-16}$
32	10^{-6}	$3.5 \cdot 10^{-3}$	1.0	20	$1.0 \cdot 10^{-14}$	11 (6)	$1.1 \cdot 10^{-16}$
32	10^{-12}	$3.5 \cdot 10^{-6}$	1.0	27	$8.1 \cdot 10^{-15}$	11 (3)	$1.1 \cdot 10^{-16}$
128	10^{-3}	0.11	0.95	17	$1.2 \cdot 10^{-13}$	13 (12)	$7.7 \cdot 10^{-15}$
128	10^{-6}	$3.5 \cdot 10^{-3}$	1.0	21	$8.0 \cdot 10^{-13}$	13 (6)	$3.6 \cdot 10^{-16}$
128	10^{-12}	$3.5 \cdot 10^{-6}$	1.0	30	$1.5 \cdot 10^{-13}$	12 (4)	$2.7 \cdot 10^{-16}$

Table 2: Number of iterations for Algorithm 1 vs. SDA on the transport problem

n	α	gap($\mathcal{H}$)	δ	SDA its	SDA res	Alg 1 its	Alg 1 res
50	10^{-3}	0.43	43	12	$3.9 \cdot 10^{-16}$	4 (7)	$1.1 \cdot 10^{-16}$
100	10^{-3}	0.61	90	13	$6.1 \cdot 10^{-16}$	4 (7)	$1.9 \cdot 10^{-16}$
32	10^{-12}	1.1	185	13	$8.6 \cdot 10^{-16}$	4 (7)	$3.3 \cdot 10^{-16}$

Table 3: Number of iterations for Algorithm 1 vs. SDA on the problem of Test 3

a singular matrix. The problem is not close-to-critical, however, there are two central eigenvalues well separated from the others, so the subspace shift algorithm works fine.

In Table 3 we report the results for this problem. The effectiveness of the subspace shift algorithm suggests the possibility to use it in particular problems where a fistful of small eigenvalues are well separated from the others.

Test 4 The residual is not always a good measure of the accuracy of the solution of a matrix equation. To test the accuracy of the subspace shift algorithm we consider the problems of Test 1 and Test 3 and compute the solution with double precision to get a solution X_* exact up to 8 significant digits. Then we run the customary SDA and the SuShi algorithm with precision 10^{-8} (single precision).

The relative error is essentially the same in both cases. For instance, for Test 1 with $n = 4$ and $\alpha = 10^{-3}$ we get for both errors $1.8 \cdot 10^{-7}$, for Test 3 with $n = 100$ and $\alpha = 10^{-3}$ we get for both errors $1.4 \cdot 10^{-7}$.

6 Conclusions

We have provided a generalization of the shift technique which is aimed to handle close-to-critical nonsymmetric algebraic Riccati equations. The technique consists in computing explicitly the (hopefully moderate-sized) invariant subspace relative to the smallest eigenvalues, which are responsible for the slow convergence of the solution algorithms, and modifying the problem in order to remove them. A theoretical analysis is outlined, not only in terms of eigenvalue location, but also using the more powerful separation metric, which is the one related to the conditioning of the problem; numerical experiments are presented and prove that the application of the shift technique is effective on the analyzed problems.

Acknowledgments

We are thankful to an anonymous referee for his/her observations that helped us direct our efforts during the completion of this article. Some of the results presented here have been obtained while the second author, F. Poloni, was supported by a doctoral grant of Scuola Normale Superiore, Pisa, Italy.

References

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

[2] D. Bini and B. Meini. On the solution of a nonlinear matrix equation arising in queueing problems. *SIAM J. Matrix Anal. Appl.*, 17(4):906–926, 1996.

[3] D. A. Bini, B. Iannazzo, G. Latouche, and B. Meini. On the solution of algebraic Riccati equations arising in fluid queues. *Linear Algebra Appl.*, 413(2-3):474–494, 2006.

[4] D. A. Bini, B. Iannazzo, B. Meini, and F. Poloni. Nonsymmetric Algerbraic Riccati Equations Associated with an M-Matrix: Recent Advances and Algorithms. In V. Olshevsky and E. Tyrtyshnikov, editors, *Matrix Methods: Theory, Algorithms and Applications*, pages 176–209. World Scientific, Singapore, 2010.

[5] D. A. Bini, B. Iannazzo, and F. Poloni. A fast Newton's method for a nonsymmetric algebraic Riccati equation. *SIAM J. Matrix Anal. Appl.*, 30(1):276–290, 2008.

[6] A. Brauer. Limits for the characteristic roots of a matrix. IV. Applications to stochastic matrices. *Duke Math. J.*, 19:75–91, 1952.

[7] C.-Y. Chiang, E. K.-W. Chu, C.-H. Guo, T.-M. Huang, W.-W. Lin, and S.-F. Xu. Convergence analysis of the doubling algorithm for several nonlinear matrix equations in the critical case. *SIAM J. Matrix Anal. Appl.*, 31(2):227–247, 2009.

[8] E. K.-W. Chu, H.-Y. Fan, and W.-W. Lin. A structure-preserving doubling algorithm for continuous-time algebraic Riccati equations. *Linear Algebra Appl.*, 396:55–80, 2005.

[9] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[10] C.-H. Guo. Nonsymmetric algebraic Riccati equations and Wiener-Hopf factorization for M-matrices. *SIAM J. Matrix Anal. Appl.*, 23(1):225–242, 2001.

[11] C.-H. Guo. Efficient methods for solving a nonsymmetric algebraic Riccati equation arising in stochastic fluid models. *J. Comput. Appl. Math.*, 192(2):353–373, 2006.

[12] C.-H. Guo and N. J. Higham. Iterative solution of a nonsymmetric algebraic Riccati equation. *SIAM J. Matrix Anal. Appl.*, 29(2):396–412, 2007.

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

[14] C.-H. Guo and A. J. Laub. On the iterative solution of a class of nonsymmetric algebraic Riccati equations. *SIAM J. Matrix Anal. Appl.*, 22(2):376–391, 2000.

[15] X.-X. Guo and Z.-Z. Bai. On the minimal nonnegative solution of nonsymmetric algebraic Riccati equation. *J. Comput. Math.*, 23(3):305–320, 2005.

[16] X.-X. Guo, W.-W. Lin, and S.-F. Xu. A structure-preserving doubling algorithm for nonsymmetric algebraic Riccati equation. *Numer. Math.*, 103(3):393–412, 2006.

[17] C. He, B. Meini, and N. H. Rhee. A shifted cyclic reduction algorithm for quasi-birth-death problems. *SIAM J. Matrix Anal. Appl.*, 23(3):673–691, 2001/02.

[18] J. Juang and W.-W. Lin. Nonsymmetric algebraic Riccati equations and Hamiltonian-like matrices. *SIAM J. Matrix Anal. Appl.*, 20(1):228–243, 1999.

[19] P. Lancaster and L. Rodman. *Algebraic Riccati equations.* Oxford Science Publications. The Clarendon Press Oxford University Press, New York, 1995.

[20] L. C. G. Rogers. Fluid models in queueing theory and Wiener-Hopf factorization of Markov chains. *Ann. Appl. Probab.*, 4(2):390–413, 1994.

[21] G. W. Stewart and J. G. Sun. *Matrix perturbation theory.* Computer Science and Scientific Computing. Academic Press Inc., Boston, MA, 1990.