ELSEVIER Linear Algebra and its Applications 285 (1998) 7–31

LINEAR ALGEBRA
AND ITS
APPLICATIONS

Perturbation bounds and characterisation of the solution of the associated algebraic Riccati equation

M.M. Konstantinov [a,*], M.O. Stanislavova [b,1], P.Hr. Petkov [c,2]

[a] *University of Architecture and Civil Engineering, 1 Hr. Smirnenski Blv, 1421 Sofia, Bulgaria*
[b] *202 Math. Sci. Bldg., University of Missouri, Columbia, MO 65211, USA*
[c] *Department of Automatics, Technical University of Sofia, 1756 Sofia, Bulgaria*

Received 27 February 1995; received in revised form 26 April 1998; accepted 12 May 1998

Submitted by R.A. Brualdi

Abstract

The paper deals with the associated algebraic matrix Riccati equation (AAMRE), closely related to the standard algebraic matrix Riccati equation arising in the theory of linear-quadratic optimisation and filtering. The sensitivity of the AAMRE relative to perturbations in its coefficients is studied. Both linear local (norm-wise and component-wise) and non-linear non-local perturbation bounds are obtained. The conditioning of the AAMRE is determined in particular. A full characterisation of the solution of AAMRE in terms of neutral subspaces of certain Hermitian matrix is given which is a counterpart of the characterisation of the solutions to the standard Riccati equation in terms of the invariant subspaces of the corresponding Hamiltonian matrix. A reliable method to obtain all solutions to AAMRE is briefly outlined. © 1998 Published by Elsevier Science Inc. All rights reserved.

AMS classification: 15A24

Keywords: Matrix quadratic equations; Associated algebraic matrix Riccati equations; Perturbation analysis; Condition estimates

[*] Corresponding author. E-mail: mmk_fte@uacg.acad.bg.

[1] E-mail: mstanis@sindy5.math.missouri.edu.

[2] E-mail: php@mbox.digsys.bg.

PII: S0024-3795(98)10094-0

1. Introduction

Recently there is a permanent interest in the sensitivity analysis of the matrix Riccati equations arising in the solution of linear-quadratic optimisation and estimation problems in control theory. This interest is motivated by the fact that these equations are usually subject to perturbations in the data reflecting either parameter uncertainties or rounding errors, accompanying the numerical solution of the problem. Thus we have to deal with a family of Riccati equations rather than with a single equation. Also, if a backwardly stable numerical method is implemented for the solution of the equation, then the computed solution will be close to the exact solution of an equation with slightly perturbed coefficients. If we have a quantitative measure for the sensitivity of the Riccati equation we may derive an accuracy estimate for the computed solution. Without such accuracy estimate the corresponding computational algorithm will not meet the modern standards of reliability.

In this paper we study the sensitivity of the solutions of the complex associated algebraic matrix Riccati equation (AAMRE) relative to perturbations in its coefficients. The AAMRE is closely related to the standard algebraic Riccati equation, arising in the theory of linear continuous time-invariant systems. The sensitivity of the standard Riccati equation is studied in [2,4,10,13–15,20]. We also give a full description and a parametrisation of the set of all solution to AAMRE. Similar results for the real AAMRE are outlined in [11].

In Section 2 we give the statement of the problem. In Section 3 we consider special cases of AAMRE. General properties and a parametrisation of the solutions of AAMRE are given in Section 4. Here we characterise the solution set by the neutral subspaces of a Hermitian matrix, related to the Hamiltonian matrix of the standard Riccati equation. A method for reliable computation of all solutions of AAMRE is presented in Section 5. In Section 6 local linear (norm-wise and component-wise) and non-local non-linear perturbation analysis of the AAMRE is presented. In the first case we suppose that the perturbations in the data are asymptotically small and the corresponding bounds contain first order terms only. In this way the conditioning of the equation is determined as well. We also give a local perturbation bound (first order homogeneous but not additive), which is better or equal to the bound, based on condition numbers. In the second case an upper bound for the norm of the perturbation in the solution is obtained without the assumption that the coefficient perturbations are asymptotically small. This bound is a non-linear function of the perturbations in the data. Illustrative examples are presented in Section 7.

We use the following abbreviations: $\mathscr{F}^{m\times n}$ is the linear space of $m \times n$ matrices over the field $\mathscr{F}$ of real ($\mathscr{F} = \mathbb{R}$) or complex ($\mathscr{F} = \mathbb{C}$) numbers; $\jmath := \sqrt{-1}$; $\mathbb{R}_+ = [0, \infty)$; $\mathbb{C}^m = \mathbb{C}^{m\times 1}$; $\|.\|$ a norm in $\mathbb{C}^m$ or the corresponding induced norm in $\mathbb{C}^{m\times n}$ (if necessary we use the subscript 2 or F to denote the spectral or Frobenius norm); I_n the unit $n \times n$ matrix; $\overline{A} \in \mathbb{C}^{m\times n}$, $A^{\mathrm{T}} \in \mathbb{C}^{n\times m}$ and

$A^{\mathrm{H}} = \overline{A}^{\mathrm{T}} \in \mathbb{C}^{n\times m}$ the complex conjugate, transpose and complex conjugate transpose of $A \in \mathbb{C}^{m\times n}$; $A^{\dagger} \in \mathbb{C}^{n\times m}$ the pseudo-inverse of the matrix $A \in \mathbb{C}^{m\times n}$; $|A| = [|a_{ij}|] \in \mathbb{R}_{+}^{m\times n}$ the matrix module of $A \in \mathbb{C}^{m\times n}$; $\det(A)$ the determinant of $A \in \mathbb{C}^{n\times n}$; $\operatorname{rank}(A)$ the rank of $A \in \mathbb{C}^{m\times n}$; $\preceq$ the partial component-wise order relation in $\mathbb{R}^{m\times n}$, i.e. $A \preceq B$ if $a_{ij} \leqslant b_{ij}$, where $A = [a_{ij}], B = [b_{ij}] \in \mathbb{R}^{m\times n}$; $\mathscr{L}(n)$ the space of linear operators $L : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ (each $L \in \mathscr{L}(n)$ may be defined from $L(Z) = \sum_i A_i Z B_i$, where $Z, A_i, B_i \in \mathbb{C}^{n\times n}$); $\mathscr{I}_{n^2} \in \mathscr{L}(n)$ the identity operator, i.e. $\mathscr{I}_{n^2}(Z) = Z$ for $Z \in \mathbb{C}^{n\times n}$; $\mathscr{PL}(n)$ the set of pseudo-linear operators $\mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$, i.e. $L \in \mathscr{PL}(n)$ if $L(Z) = L_1(Z) + L_2(Z^{\mathrm{H}})$, where $L_1, L_2 \in \mathscr{L}(n)$ (pseudo-linear operators are continuous but not in general differentiable over $\mathbb{C}$, while the realification of a pseudo-linear operator is a linear operator over $\mathbb{R}$); $\operatorname{vec}(Z) \in \mathbb{C}^{mn}$ the vector column-wise representation of the matrix $Z \in \mathbb{C}^{m\times n}$; $\operatorname{vec}_{\mathbb{R}}(Z) = [\operatorname{vec}(Z_1)^{\mathrm{T}}, \operatorname{vec}(Z_2)^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{2mn}$ the real vector column-wise representation of the matrix $Z = Z_1 + \jmath Z_2 \in \mathbb{C}^{m\times n}$, where $Z_j \in \mathbb{R}^{m\times n}$; $\mathscr{U}(n) \subset \mathscr{GL}(n)$ and $\mathscr{GL}(n) \subset \mathbb{C}^{n\times n}$ the groups of unitary ($A^{\mathrm{H}}A = I_n$) and non-singular matrices respectively; $\mathscr{H}(n) \subset \mathbb{C}^{n\times n}$ the set of Hermitian matrices ($A^{\mathrm{H}} = A$); $\mathscr{H}_{+}(n) \subset \mathscr{H}(n)$ the set of non-negative definite Hermitian matrices (we write $A \geqslant 0$ if A is non-negative and $A > 0$ if A is positive definite); $\operatorname{sign}(A) = (n_+, n_-)$ the signature of the matrix $A \in \mathscr{H}(n)$, where n_+ and n_- are the numbers of positive and negative eigenvalues of A respectively; $\operatorname{Rg}(A) \subset \mathbb{C}^m$ and $\operatorname{Ker}(A) \subset \mathbb{C}^n$ the range and kernel of the matrix $A \in \mathbb{C}^{m\times n}$; $\dim(\mathscr{T})$ and $\operatorname{codim}(\mathscr{T})$ the (complex) dimension and codimension of the (complex) variety $\mathscr{T}$; $\dim_{\mathbb{R}}(\mathscr{T})$ and $\operatorname{codim}_{\mathbb{R}}(\mathscr{T})$ the real dimension and codimension of the complex variety $\mathscr{T}$. The notation ":=" stands for "equal by definition". The end of proofs is marked by □.

2. Problem statement

Consider the complex algebraic matrix quadratic equation

$$Q + A^{\mathrm{H}}X + X^{\mathrm{H}}A - X^{\mathrm{H}}MX = 0, \tag{1}$$

where $Q, M \in \mathscr{H}_{+}(n)$ and $A \in \mathbb{C}^{n\times n}$ are given matrices, such that the triple $\Sigma := (Q, A, M)$ is regular (i.e. the pair $(Q, A]$ is detectable and the pair $[A, M)$ is stabilizable) and $X \in \mathbb{C}^{n\times n}$ is the unknown matrix. As shown below, Eq. (1) is closely related to the famous algebraic matrix Riccati equation

$$Q + A^{\mathrm{H}}X + XA - XMX = 0 \tag{2}$$

arising in the theory of optimisation and filtering of linear continuous time-invariant systems. For this reason, Eq. (1) is further referred to as the *associated algebraic matrix Riccati equation (AAMRE). Setting*

$$R(X, Y, \Sigma) := Q + A^{\mathrm{H}}X + YA - YMX,$$

we may rewrite Eqs. (1) and (2) as $R(X,X^{\mathrm{H}},\Sigma)=0$ and $R(X,X,\Sigma)=0$.

The connection between Eqs. (1) and (2) is revealed as follows. Denote by $\Xi\subset\mathbb{C}^{n\times n}$ and $\mathsf{P}\subset\mathbb{C}^{n\times n}$ the sets of all solutions of Eqs. (1) and (2) respectively. Note that while for Eq. (2) we have $X\in\mathsf{P}$ if and only if $X^{\mathrm{H}}\in\mathsf{P}$, the inclusion $X\in\Xi$ does not imply $X^{\mathrm{H}}\in\Xi$ and vice versa.

The solution sets Ξ and P may have very complicated structure. However, the realification of the set Ξ is a closed algebraic variety of real dimension n^2 and, unlike P, contains no isolated points. The examples, presented in Section 7, give an idea how the set Ξ may look for $n=1$ and $n=2$.

As it is well known [7,8] the solution set P of Eq. (2) may be characterised by the invariant n-dimensional subspaces of the Hamiltonian matrix

$$H:=\begin{bmatrix} A & -M \\ -Q & -A^{\mathrm{H}} \end{bmatrix}\in\mathscr{GL}(2n). \tag{3}$$

In turn, the solution set Ξ of (1) may be characterised by the neutral n-dimensional subspaces of the related Hermitian matrix

$$S:=\begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix}H=\begin{bmatrix} Q & A^{\mathrm{H}} \\ A & -M \end{bmatrix}\in\mathscr{H}(2n)\cap\mathscr{GL}(2n). \tag{4}$$

The solutions of Eqs. (1) and (2) may be Hermitian and non-Hermitian. Let $\Xi^*:=\Xi\cap\mathscr{H}(n)$ and $\mathsf{P}^*:=\mathsf{P}\cap\mathscr{H}(n)$ be the sets of Hermitian solutions of Eqs. (1) and (2) respectively. Since every Hermitian solution of Eq. (1) satisfies Eq. (2) and vice versa then $\Xi^*=\mathsf{P}^*$ and the solution sets Ξ and P may be represented as disjoint unions $\Xi=\Xi^*\cup\Xi^0$, $\mathsf{P}=\mathsf{P}^*\cup\mathsf{P}^0$, where $\Xi^0:=\Xi\setminus\Xi^*$, $\mathsf{P}^0:=\mathsf{P}\setminus\Xi^*$.

Let

$$L(Z):=\begin{bmatrix} I_n & 0 \\ Z & I_n \end{bmatrix}\in\mathscr{GL}(2n),$$

where $Z\in\mathbb{C}^{n\times n}$. Then

$$L(-Y)HL(X)=\begin{bmatrix} A-MX & -M \\ -R(X,Y,\Sigma) & -(A-MY^{\mathrm{H}})^{\mathrm{H}} \end{bmatrix}. \tag{5}$$

Setting $Y=X$ for $X\in\mathsf{P}$ and $X\in\mathsf{P}^*$ in Eq. (5) we get

$$\det(H)=(-1)^n d(X)\overline{d(X^{\mathrm{H}})},\quad X\in\mathsf{P},$$

and

$$\det(H)=(-1)^n|d(X)|^2,\quad X\in\mathsf{P}^*$$

where $d(X):=\det(A-MX)\neq 0$. Similarly, setting $Y=X^{\mathrm{H}}$ for $X\in\Xi$ it follows from Eq. (5)

$$\det(H) = (-1)^n |d(X)|^2, \quad X \in \Xi.$$

Hence $A - MX \in \mathscr{GL}(n)$.

Our next observation is that $\Xi \cap \mathrm{P} = \Xi^*$ or, equivalently, $\Xi^0 \cap \mathrm{P}^0 = \emptyset$. Indeed, if $X \in \Xi \cap \mathrm{P}$ then $(X - X^{\mathrm{H}})(A - MX) = 0$. Since $A - MX \in \mathscr{GL}(n)$ it follows $X = X^{\mathrm{H}}$ and the assertion is proved.

The general properties of the solution of the real AAMRE

$$Q + A^{\mathrm{T}}X + X^{\mathrm{T}}A - X^{\mathrm{T}}MX = 0, \quad Q, A, M, X \in \mathbb{R}^{n\times n},$$

have been considered in [11]. The extension of these results to the complex case, however is presented in Section 3. As may be expected, the structure of the solution set of the real AAMRE is more involved than that of the complex AAMRE due to the fact that the field $\mathbb{R}$ is not algebraically closed.

Consider now the perturbation analysis of Eq. (1), rewritten as $F(X, \Sigma) = F(X, Q, A, M) = 0$, where $F(X, \Sigma) := R(X, X^{\mathrm{H}}, \Sigma)$. We note that the matrix functions $F(., \Sigma) : \mathbb{C}^{n\times n} \to \mathscr{H}(n)$ and $F(X, Q, ., M) : \mathbb{C}^{n\times n} \to \mathscr{H}(n)$ are pseudo-polynomials, i.e. they are continuous but not differentiable over $\mathbb{C}$, see Section A.1 of Appendix A. However, in the framework of the realifications $\mathbb{C}^{n\times n} \simeq \mathbb{R}^{2n^2}$ and $\mathscr{H}(n) \simeq \mathbb{R}^{n^2}$, the realifications of these functions are real analytic.

We shall refer to Eq. (1) as the *unperturbed equation* and to a fixed solution $X = X_0$ of (1) as the *unperturbed solution.* We note that the solution set Ξ of Eq. (1) is a closed algebraic variety in $\mathbb{C}^{n\times n} \simeq \mathbb{R}^{2n^2}$ and the perturbation analysis presented below is relative to a fixed solution $X_0 \in \Xi$ rather than to the whole variety Ξ.

Let $\Delta Q, \Delta A, \Delta M \in \mathbb{C}^{n\times n}$ be perturbations of the matrix coefficients Q, A, M in (1) with $\Delta Q, \Delta M \in \mathscr{H}(n)$. Consider the perturbed equation

$$Q + \Delta Q + (A + \Delta A)^{\mathrm{H}} Y + Y^{\mathrm{H}}(A + \Delta A) - Y^{\mathrm{H}}(M + \Delta M)Y = 0 \tag{6}$$

and denote $\Delta := [\Delta_Q, \Delta_A, \Delta_M]^{\mathrm{T}} \in \mathbb{R}^3_+$, where $\Delta_Z = \|\Delta Z\|$ and $\|.\|$ is the Frobenius (F-) or spectral (2-) norm in $\mathbb{C}^{n\times n}$.

As it was shown above, $A_0 := A - MX_0 \in \mathscr{GL}(n)$ for $X_0 \in \mathrm{P}$. But then the partial Fréchet pseudo-derivative $F_X(X_0, \Sigma) \in \mathscr{PL}(n)$,

$$F_X(X_0, \Sigma)(Z) := A_0^{\mathrm{H}} Z + Z^{\mathrm{H}} A_0$$

of the left-hand side $F(X, \Sigma)$ of (1) in X at $X = X_0$ is surjective (operators $P(., B) \in \mathscr{PL}(n)$, acting according to the rule $P(Z, B) = B^{\mathrm{H}}Z + ZB$, are called *associated Lyapunov operators,* see Section A.2 of Appendix A). Then according to the implicit function theorem [9] we get the following assertion.

Theorem 1. *The perturbed equation* (6) *has a solution*

$$Y = Y(\Delta\Sigma) = X_0 + \Delta X, \quad \Delta\Sigma := (\Delta Q, \Delta A, \Delta M),$$

in a neighbourhood of X_0, such that $Y(0) = X_0$. Moreover, the realification of $Y(.)$ is a real analytic function of the realifications of the perturbations $\Delta Q, \Delta A, \Delta M$ in certain neighbourhood of the origin, e.g. for $\|\Delta\|$ sufficiently small.

The perturbation problem solved in this paper is formulated as:

(i) Find a local linear norm-wise estimate of the type

$$\Delta_X \leqslant K_Q \Delta_Q + K_A \Delta_A + K_M \Delta_M + \mathrm{O}(\|\Delta\|^2) \tag{7}$$

for the norm $\Delta_X = \|\Delta X\|$ of the perturbation ΔX as a function of $\Delta_Q, \Delta_A, \Delta_M$, where $K_Z \in \mathbb{R}_+$ are the *condition numbers* of AAMRE relative to $Z \in \{Q, A, M\}$, which is valid for $\|\Delta\|$ asymptotically small. Find a local linear component-wise estimate

$$|\mathrm{vec}_{\mathbb{R}}(\Delta X)| \preceq L_Q |\mathrm{vec}_{\mathbb{R}}(\Delta Q)| + L_A |\mathrm{vec}_{\mathbb{R}}(\Delta A)| + L_M |\mathrm{vec}_{\mathbb{R}}(\Delta M)| + \mathrm{O}(\|\Delta\|^2),$$
$$\|\Delta\| \to 0,$$

for the matrix module $|\Delta X|$ of ΔX as a function of $|\Delta Q|, |\Delta A|, |\Delta M|$, where $L_Z \in \mathbb{R}_+^{2n^2 \times 2n^2}$ are the *condition matrices* of AAMRE.

(ii) Find a domain $\mathscr{D} \subset \mathbb{R}_+^3$, $0 \in \mathscr{D}$, such that for each $\Delta \in \mathscr{D}$ Eq. (6) has a solution $Y = Y(\Delta\Sigma) = X_0 + \Delta X$ in the neighbourhood of X_0, whose realification is a real analytic function of the realification of $\Delta\Sigma$, and $Y(0) = X_0$. Find an estimate

$$\Delta_X \leqslant f(\Delta), \quad \Delta \in \mathscr{D}, \tag{8}$$

where the function $f : \mathscr{D} \to \mathbb{R}_+$ is continuous, non-decreasing in each component of Δ and $f(0) = 0$.

Note that Eq. (8) is a non-local estimate since it holds for all (possibly small but finite) perturbation vectors $\Delta \in \mathscr{D}$, i.e. $\|\Delta\|$ needs not to be asymptotically small.

The above perturbation bounds are understood in the sense that there exists a solution $Y = X_0 + \Delta X$ of the perturbed equation (6), for which the estimates Eq. (7) or Eq. (8) hold. At the same time the perturbed equation may have solutions Y, for which the perturbation $\Delta X = Y - X_0$ does not satisfy these estimates (in particular, Eq. (6) may have solutions of arbitrary large norm).

3. Special cases of AAMRE

In this section we consider special cases of Eq. (1) in which either the solution is obtained in explicit form or the actual order of the equation may be reduced.

3.1. The case $M = 0$

In this "completely uncontrollable" case AAMRE reduces to the *associated Lyapunov equation*

$$Q + A^{\mathrm{H}}X + X^{\mathrm{H}}A = 0. \tag{9}$$

In view of the regularity of Σ the matrix A is stable and hence invertible. Thus the solution of Eq. (9) is $X = -A^{-\mathrm{H}}(Q + Z - Z^{\mathrm{H}})/2$, where $Z \in \mathbb{C}^{n\times n}$ is an arbitrary matrix, and

$$\Xi = \left\{-A^{-\mathrm{H}}(Q + Z - Z^{\mathrm{H}})/2 \colon Z \in \mathbb{C}^{n\times n}\right\}.$$

We see that here Ξ is isomorphic to the set of $n \times n$ complex screw-Hermitian matrices, i.e. $\Xi \simeq \mathbb{R}^{n^2}$ and $\dim_{\mathbb{R}}(\Xi) = n^2$ (for comparison, in the real case $\dim(\Xi) = n(n-1)/2$). At the same time the only member of Ξ^* is the positive definite solution X^* of the Lyapunov equation $Q + A^HX + XA = 0$.

3.2. The case $M > 0$

In this "most controllable" case, Eq. (1) may be rewritten as

$$\left(M^{1/2}X - M^{-1/2}A\right)^{\mathrm{H}}\left(M^{1/2}X - M^{-1/2}A\right) = Q + A^{\mathrm{H}}M^{-1}A, \tag{10}$$

where $M^{1/2}$ is the positive definite square root of M. It follows from Eq. (10) that

$$\Xi = \left\{M^{-1}A + M^{-1/2}U\left(Q + A^{\mathrm{H}}M^{-1}A\right)^{1/2} \colon\ U \in \mathcal{U}(n)\right\} \tag{11}$$

see also [6,12]. The detectability of $(Q, A]$ yields $Q + A^{\mathrm{H}}M^{-1}A \in \mathcal{GL}(n)$. Hence, according to (10), the set Ξ is isomorphic to $\mathcal{U}(n)$. As in the previous case $\dim_{\mathbb{R}}(\Xi) = n^2$ but here Ξ is a compact. As it is shown below, the real dimension of Ξ is equal to n^2 not only in the special cases $M = 0$ and $M > 0$ but also in the general case $M \geqslant 0$.

3.3. The effective order of AAMRE

Let $r := \operatorname{rank}[M, AM, \ldots, A^{n-1}M]$ be the dimension of the controllable subspace of the pair $[A, M)$. The integer r may be considered as the *effective order* of AAMRE in the following sense. If $1 \leqslant r < n$ then the AAMRE reduces to three matrix equations, only one of which is in fact quadratic and is in the form (1), while the other two are linear and are solved explicitly. Indeed, if $1 \leqslant r < n$ then there exists a matrix $D \in \mathcal{GL}(n)$ such that

$$\hat{A} := D^{-1}AD = \begin{bmatrix} \hat{A}_{11} & \hat{A}_{12} \\ 0 & \hat{A}_{22} \end{bmatrix}, \qquad \hat{M} := D^{-1}MD^{-\mathrm{H}} = \begin{bmatrix} \hat{M}_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

where $\hat{A}_{11} \in \mathbb{C}^{r\times r}$, $\hat{M}_{11} \in \mathscr{H}_+(r)$, the pair $[\hat{A}_{11}, \hat{M}_{11})$ is controllable and the matrix $\hat{A}_{22} \in \mathbb{C}^{(n-r)\times(n-r)}$ is stable. Setting

$$\hat{X} := D^{\mathrm{H}} X D = \begin{bmatrix} \hat{X}_{11} & \hat{X}_{12} \\ \hat{X}_{21} & \hat{X}_{22} \end{bmatrix}, \qquad \hat{Q} := D^{\mathrm{H}} Q D = \begin{bmatrix} \hat{Q}_{11} & \hat{Q}_{21}^{\mathrm{H}} \\ \hat{Q}_{21} & \hat{Q}_{22} \end{bmatrix},$$

where $\hat{X}_{11} \in \mathbb{C}^{r\times r}$, $\hat{Q}_{11} \in \mathscr{H}_+(r)$, we get the equations

$$\hat{X}_{11}^{\mathrm{H}} \hat{M}_{11} \hat{X}_{11} - \hat{A}_{11}^{\mathrm{H}} \hat{X}_{11} - \hat{X}_{11}^{\mathrm{H}} \hat{A}_{11} - \hat{Q}_{11} = 0, \tag{12}$$

$$\hat{X}_{12}^{\mathrm{H}} \left(\hat{M}_{11} \hat{X}_{11} - \hat{A}_{11} \right) - \hat{A}_{12}^{\mathrm{H}} \hat{X}_{11} - \hat{Q}_{21} - \hat{A}_{22}^{\mathrm{H}} \hat{X}_{21} = 0, \tag{13}$$

$$\hat{X}_{12}^{\mathrm{H}} \hat{M}_{11} \hat{X}_{12} - \hat{A}_{12}^{\mathrm{H}} \hat{X}_{12} - \hat{X}_{12}^{\mathrm{H}} \hat{A}_{12} - \hat{Q}_{22} - \hat{A}_{22}^{\mathrm{H}} \hat{X}_{22} - \hat{X}_{22}^{\mathrm{H}} \hat{A}_{22} = 0. \tag{14}$$

We see that only Eq. (12) is quadratic, while equations Eqs. (13) and (14) are linear and are explicitly solved in $\hat{X}_{21}$ and $\hat{X}_{22}$ by

$$\hat{X}_{21} = \hat{A}_{22}^{-\mathrm{H}} \left(\hat{X}_{12}^{\mathrm{H}} (\hat{M}_{11} \hat{X}_{11} - \hat{A}_{11}) - \hat{A}_{12}^{\mathrm{H}} \hat{X}_{11} - \hat{Q}_{21} \right)$$

and

$$\hat{X}_{22} = \hat{A}_{22}^{-\mathrm{H}} \left(Z - Z^{\mathrm{H}} + \hat{X}_{12}^{\mathrm{H}} \hat{M}_{11} \hat{X}_{12} - \hat{A}_{12}^{\mathrm{H}} \hat{X}_{12} - \hat{X}_{12}^{\mathrm{H}} \hat{A}_{12} - \hat{Q}_{22} \right),$$

where the matrices $\hat{X}_{12} \in \mathbb{C}^{r\times(n-r)}$ and $Z \in \mathbb{C}^{(n-r)\times(n-r)}$ are arbitrary. According to Section 3.2, the solution $\hat{X}_{11}$ of Eq. (12) depends on r^2 free real parameters. Hence the solution $\hat{X}$ depends on $2r(n-r) + (n-r)^2 + r^2 = n^2$ real parameters. In this case the solution set Ξ has a compact component, homeomorphic to $\mathscr{U}(r)$ and of real dimension r^2, and a non-compact component, homeomorphic to $\mathbb{R}^{n^2-r^2}$ and of real dimension $n^2 - r^2$.

In view of the above considerations only the case when the pair $[A, M)$ is controllable (i.e. $r = n$) is of interest. That is why in the rest of the paper the controllability of $[A, M)$ is assumed.

4. Properties and parametrisation of the solution set

4.1. Properties of the solution

The set Ξ is a quadric – a closed algebraic variety in the Zariski topology of the realification of $\mathbb{C}^{n\times n} \simeq \mathbb{C}^{n^2} \simeq \mathbb{R}^{2n^2}$. The tangent set $\mathscr{T}_{X_0} \subset \mathbb{C}^{n\times n}$ of Ξ at $X_0 \in \Xi$ is $\mathscr{T}_{X_0} = X_0 + \mathrm{Ker}(F_X(X_0, \Sigma))$, where

$$\mathrm{Ker}(F_X(X_0, \Sigma)) = \left\{ A_0^{-\mathrm{H}} (Z - Z^{\mathrm{H}}) \colon Z \in \mathbb{C}^{n\times n} \right\}.$$

Hence the real dimension $\dim_{\mathbb{R}}(\Xi)$ of Ξ at X_0 satisfies

$$\dim_{\mathbb{R}}(\Xi) \leqslant \dim_{\mathbb{R}}(\mathscr{T}_{X_0}) = \dim_{\mathbb{R}}(F_X(X_0, \Sigma)) = n^2. \tag{15}$$

Since the field $\mathbb{C}$ is algebraically closed, then $\dim_{\mathbb{R}}(\Xi)$ is not less than $2n^2 (= \dim_{\mathbb{R}}(\mathbb{C}^{n\times n}))$ minus n^2 (= the number of scalar real equations in Eq. (1)), i.e. $\dim_{\mathbb{R}}(\Xi) \geqslant n^2$. Combining with Eq. (15) we get $\dim_{\mathbb{R}}(\Xi) = n^2$. For comparison, the dimension of Ξ in the real case is $n(n-1)/2$, see [11].

As it is well known the characterisation of the solution set P of the standard Riccati equation (2) as well as the direct methods to solve this equation are based on the $\mathscr{GL}(n)$-similarity invariant eigenstructure [17,7,8], or the $\mathscr{U}(n)$-similarity invariant Schur structure [16] of the Hamiltonian matrix (3). Similar role for the associated equation (1) play the $\mathscr{GL}(n)$-congruent invariant structure, or the $\mathscr{U}(n)$-congruent invariant structure of the Hermitian matrix (4).

Denote by $\mathscr{S} \subset \mathbb{C}^{n\times n}$ the set of all subspaces of $\mathbb{C}^{2n}$ of complex dimension n, which are simultaneously S-neutral and complementary to the subspace

$$\mathrm{Rg}\begin{bmatrix} 0 \\ I_n \end{bmatrix},$$

i.e. $\mathscr{W} \subset \mathscr{S}$ if and only if there exist matrices $W_1 \in \mathscr{GL}(n)$ and $W_2 \in \mathbb{C}^{n\times n}$ such that $\mathscr{W} = \mathrm{Rg}(W)$ and $W^{\mathrm{H}}SW = 0$, where

$$W := \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}.$$

Let $\mathscr{S}^* \subset \mathscr{W}$ be the set of all $\mathscr{W} = \mathrm{Rg}(W)$ such that $W_1^{\mathrm{H}}W_2 = W_2^{\mathrm{H}}W_1$. Then the following characterization of Ξ and Ξ^* may be given.

Theorem 2. *There is a bijection λ between the sets Ξ and $\mathscr{S}$ and a bijection λ^* between the sets Ξ^* and $\mathscr{S}^*$.*

Proof. If $\mathscr{W} = \mathrm{Rg}(W) \in \mathscr{S}$ is defined as above, then

$$\mathscr{W} = \mathrm{Rg}\begin{bmatrix} I_n \\ W_2W_1^{-1} \end{bmatrix}$$

and

$$W_1^{\mathrm{H}}QW_1 + W_1^{\mathrm{H}}A^{\mathrm{H}}W_2 + W_2^{\mathrm{H}}AW_1 - W_2^{\mathrm{H}}MW_2 = 0,$$

i.e. $X := W_2W_1^{-1}$ is a solution to Eq. (1). Now the mapping $\lambda : \Xi \to \mathscr{W}$ defined from

$$\lambda(X) := \mathrm{Rg}\begin{bmatrix} I_n \\ X \end{bmatrix}, \quad X \in \Xi;$$

$$\lambda^{-1}\left(\mathrm{Rg}\begin{bmatrix} W_1 \\ W_2 \end{bmatrix}\right) := W_2W_1^{-1}, \quad \mathrm{Rg}\begin{bmatrix} W_1 \\ W_2 \end{bmatrix} \in \mathscr{S},$$

is the desired bijection.

To prove the existence of λ^* note that for $W_1 \in \mathscr{GL}(n)$ we have $W_2W_1^{-1} \in \mathscr{H}(n)$ if and only if $W_1^{\mathrm{H}}W_2 = W_2^{\mathrm{H}}W_1$. Hence the restriction $\lambda^* = \lambda|_{\Xi^*} : \Xi^* \to \mathscr{S}^*$ of $\lambda : \Xi \to \mathscr{S}$ on $\Xi^* \subset \Xi$ is a bijection between Ξ^* and $\mathscr{S}^*$. □

We note that Theorem 2 gives a characterisation of Ξ and Ξ^* by $\mathscr{S}$ and $\mathscr{S}^*$, similar to that of P by the n-dimensional H-invariant subspaces of $\mathbb{C}^{2n}$ as described in [18,7,8].

4.2. Parametrisation of the solution

Having in mind Theorem 2, it is possible in principle to construct the solution set Ξ using the subspaces $\mathscr{W} \in \mathscr{S}$ and their representations as images of the matrices $W \in \mathbb{C}^{2n\times n}$. Instead, we shall adopt another approach constructing matrices

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \in \mathscr{GL}(2n)$$

with $R_{11} \in \mathscr{GL}(n)$, such that

$$\hat{S} := R^{\mathrm{H}}SR = \begin{bmatrix} 0 & S_1 \\ S_1^{\mathrm{H}} & S_2 \end{bmatrix} \in \mathscr{GL}(2n),$$

where $S_1 \in \mathscr{GL}(n)$, $S_2 \in \mathscr{H}(n)$. In this way the solution of Eq. (1) is obtained in the form $X = R_{21}R_{11}^{-1}$.

First we shall show that $\operatorname{sign}(S) = (n, n)$, i.e. that the matrix S is congruent to $\operatorname{diag}(I_n, -I_n)$. For this purpose we shall construct a matrix

$$T = \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} \in \mathscr{GL}(2n).$$

such that $T^{\mathrm{H}}ST = \operatorname{diag}(I_n, -I_n)$.

Let $X^* \in \Xi^*$ be the (unique) Hermitian non-negative solution to Eqs. (1) and (2) such that the matrix $A^* := A - MX^*$ is stable. Let, in addition, $Z^* > 0$ be the unique solution of the Lyapunov equation $A^*Z + Z(A^*)^{\mathrm{H}} + M = 0$. Then one of the matrices T, which congruently transform S into its diagonal form diag $(I_n, -I_n)$, is [3]

$$T := \begin{bmatrix} I_n & 0 \\ X^* & I_n \end{bmatrix}\begin{bmatrix} I_n & -Z^* \\ 0 & I_n \end{bmatrix}\begin{bmatrix} I_n & 0 \\ (A^*)^{-\mathrm{H}}/2 & -I_n \end{bmatrix}\begin{bmatrix} I_n & I_n \\ 0 & (A^*)^{-\mathrm{H}} \end{bmatrix}. \tag{16}$$

Hence we have $\det(T) = (-1)^n/\det(A^*) \neq 0$. It follows from Eq. (16) that

$$T_{11} = I_n - V, \qquad T_{21} = X^*(I_n - V) + (A^*)^{-1}/2,$$
$$T_{12} = I_n + V, \qquad T_{22} = X^*(I_n + V) - (A^*)^{-1}/2,$$

[3] To construct the real counterpart of the matrix T was the trickiest point when writing the paper [11].

where $V := Z^*(A^*)^{-\mathrm{H}}/2$. Note that $T_{11} \neq 0$. Indeed, if $T_{11} = 0$ then $V = I_n$ and $A^* = Z^*/2$, which is in a contradiction with the stability of A^*.

Representing the matrix R as $R = TK$, where $K \in \mathscr{GL}(2n)$, we get

$$K^{\mathrm{H}} \operatorname{diag}(I_n, -I_n)K = \hat{S}.$$

The general solution of this equation in $K \in \mathscr{GL}(2n)$ is

$$\mathscr{K} = \left\{ \begin{bmatrix} UK_{21} & L + UK_{22} \\ K_{21} & K_{22} \end{bmatrix} : U \in \mathscr{U}(n), K_{22} \in \mathbb{C}^{n\times n}; L, K_{21} \in \mathscr{GL}(n) \right\}.$$

Hence for $K \in \mathscr{K}$ we have

$$\begin{aligned} R_{11} &= R_{11}(U) = (T_{11}U + T_{12})K_{21}, \\ R_{21} &= R_{21}(U) = (T_{21}U + T_{22})K_{21}. \end{aligned}$$

Denote by Ω the set of all unitary matrices U, for which the matrix $R_{11}(U)$ is singular, i.e.

$$\Omega := \{U \in \mathscr{U}(n) \colon \det(T_{11}U + T_{12}) = 0\}. \tag{17}$$

It may be shown that the set Ω is either empty, or is a hyper surface in $\mathscr{U}(n)$ with $\operatorname{codim}(\Omega) = 1$. Indeed, if neither of these options is valid, then Ω should coincide with $\mathscr{U}(n)$. But for $U = I_n$ the matrix $T_{11}U + T_{12} = 2I_n$ is non-singular, i.e. $I_n \notin \Omega$. Hence $\Omega \neq \mathscr{U}(n)$.

Denote $\Omega^\# = \mathscr{U}(n) \setminus \Omega$ and define the function $G \colon \Omega^\# \to \Xi$ from

$$G(U) := R_{21}(U)R_{11}^{-1}(U) = (T_{21}U + T_{22})(T_{11}U + T_{12})^{-1}, \quad U \in \Omega^\#. \tag{18}$$

The function G is continuous and in view of Theorem 2 we have $G(\Omega^\#) = \Xi$. We shall show that the inverse function G^{-1} is defined and continuous on Ξ. Indeed, let $X \in \Xi$ be arbitrary. In view of the surjectivity of G there exists $U \in \Omega^\#$ such that $X = G(U)$. Hence $(XT_{11} - T_{21})U = -(XT_{12} - T_{22})$. Since

$$[X, -I_n]T = [XT_{11} - T_{21}, XT_{12} - T_{22}] = (XT_{11} - T_{21})[I_n, -U]$$

and rank $([X, -I_n]T) = n$ we see that both matrices $XT_{11} - T_{21}$ and $XT_{12} - T_{22}$ are non-singular. Thus the inverse function G^{-1} is well defined from

$$G^{-1}(X) = -(XT_{11} - T_{21})^{-1}(XT_{12} - T_{22}), \ X \in \Xi.$$

Hence we have proved

Theorem 3. *The mapping G, defined by* (18), *is a homeomorphism between the sets Ξ and $\Omega^\#$.*

Thus an efficient parametrisation of the solution set

$$\Xi = \{G(U) \colon U \in \Omega^\#\} \tag{19}$$

is obtained, based on Eq. (18). For this purpose one has first to find the stabilising solution X^* of the standard Riccati equation (2) and then to solve the Lyapunov equation $A^*Z + Z(A^*)^{\mathrm{H}} + M$, $A^* := A - MX^*$. In Section 5 we shall show that even these computations may be avoided using a numerically stable $\mathcal{U}(2n)$-reduction of the Hermitian matrix S into its diagonal form.

The fact that the sets Ξ and $\Omega^{\#}$ are homeomorhpic suggests the following assertion.

Theorem 4. *The set Ξ is compact (homeomorphic to $\mathcal{U}(n)$) if and only if $\Omega = \emptyset$, or, equivalently, the set Ξ is non-compact (closed but unbounded) if and only if $\Omega \neq \emptyset$.*

Proof. If $\Omega = \emptyset$ then $\Omega^{\#} = \mathcal{U}(n)$ and the "if" statement is obvious. Suppose now that Ξ is compact. Then $\Omega^{\#}$ is also compact and the function $p : \Omega^{\#} \to \mathbb{R}_+$, defined from $p(U) := |\det(T_{11}U + T_{12})|^2$, reaches its exact lower bound $p_0 \geqslant 0$ for some $U = U_0$, i.e. $p_0 = p(U_0)$. We shall show that $p_0 > 0$, which means that the set Ω, defined from (17), is empty. Indeed, if $p_0 = 0$ then for $X_0 := G(U_0)$ we have $X_0(T_{11}U_0 + T_{12}) = T_{21}U_0 + T_{22}$. Since $\sqrt{p_0} = |\det(T_{11}U_0 + T_{12})| = 0$, this contradicts to

$$n = \operatorname{rank}\left(T\begin{bmatrix} U_0 \\ I_n \end{bmatrix}\right) = \operatorname{rank}\begin{bmatrix} T_{11}U_0 + T_{12} \\ T_{21}U_0 + T_{22} \end{bmatrix} = \operatorname{rank}\left(\begin{bmatrix} I_n \\ X_0 \end{bmatrix}(T_{21}U_0 + T_{12})\right).$$

Hence $p_0 > 0$ and $\Omega = \emptyset$. □

The topological characterisation of Ξ, provided by Theorem 4, seems not to be convenient for practical purpose since one has to check whether $\Omega = \emptyset$ or $\Omega \neq \emptyset$, which may not be an easy task. Fortunately, the following easily verifiable assertion is valid.

Theorem 5. *The set Ξ is compact and homeomorphic to $\mathcal{U}(n)$ (respectively, Ξ is not compact – closed but unbounded) if and only if* rank $(M) = n$ *(respectively, if and only if* rank $(M) < n$*).*

Proof. We already know that if $M > 0$ then Ξ, as given by (11), is homeomorphic to $\mathcal{U}(n)$. To complete the proof we shall show that if $\operatorname{rank}(M) < n$ then $\Omega \neq \emptyset$ and hence Ξ is non-compact in view of Theorem 4. For this purpose we shall construct a particular member U of Ω thus showing that $\Omega \neq \emptyset$.

Let the polar decompositions of the matrices T_{1j} be $T_{1j} := T_{1j}^* V_j$, where $T_{1j}^* := (T_{1j}T_{1j}^{\mathrm{H}})^{1/2} \in \mathcal{H}_+(n)$ and $V_i \in \mathcal{U}(n)$; $j = 1, 2$. Let $U := -V_1^{\mathrm{H}}V_2$. Then $T_{11}U + T_{12} = (T_{11}^* - T_{12}^*)(-V_2)$. Recalling that $V = Z^*(A^*)^{-\mathrm{H}}/2$, where $A^*Z^* + Z^*(A^*)^{\mathrm{H}} + M = 0$, set

$$V^* := -(V + V^{\mathrm{H}}) = \tfrac{1}{2}(A^*)^{-1}M(A^*)^{-\mathrm{H}} \geqslant 0. \tag{20}$$

Since $V^* \leqslant I_n + WW^{\mathrm{H}}$ then

$$\begin{aligned}(I_n + V^* + WW^{\mathrm{H}})^{1/2} &\leqslant (I_n + WW^{\mathrm{H}})^{1/2} + (V^*)^{1/2},\\ -(I_n - V^* + WW^{\mathrm{H}})^{1/2} &\leqslant -(I_n + WW^{\mathrm{H}})^{1/2} + (V^*)^{1/2}.\end{aligned}$$

Hence

$$0 \leqslant T_{11}^* - T_{12}^* = (I_n + V^* + WW^{\mathrm{H}})^{1/2} - (I_n - V^* + WW^{\mathrm{H}})^{1/2} \leqslant 2(V^*)^{1/2}.$$

Since according to Eq. (20) the matrix V^* is singular, we obtain that $T_{11}^* - T_{12}^*$ is also singular. Thus $U \in \Omega$ and $\Omega \neq \emptyset$, which completes the proof. □

5. Reliable computation of Σ

The use of general $\mathscr{GL}(2n)$-congruent transformations, as described in Section 4, may lead to numerical difficulties. To avoid this we shall consider the use of numerically more reliable $\mathscr{U}(2n)$-congruent transformations on S in order to construct the members of Ξ and eventually Ξ^*.

Denote by $\mathscr{N}$ the set of all matrices

$$N = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} \in \mathscr{U}(2n)$$

such that

$$\mathrm{Rg}\begin{bmatrix} N_{11} \\ N_{21} \end{bmatrix} \subset \mathscr{W},$$

and by $\mathscr{N}^*$ the set of all $N \in \mathscr{N}$ such that $N_{11}^{\mathrm{H}}N_{21} = N_{21}^{\mathrm{H}}N_{11}$. The sets $\mathscr{N}$ and $\mathscr{N}^*$ may be divided into disjoint orbits relative to the equivalence relation $\sim$, where $N \sim P$ if and only if $N_{21}N_{11}^{-1} = P_{21}P_{11}^{-1}$. Denote by $\mathscr{N}/\sim$ and $\mathscr{N}^*/\sim$ the corresponding orbit spaces. Then the following statement, similar to Theorem 2, may be proved.

Theorem 6. *There is a bijection μ between the sets Ξ and $\mathscr{N}/\sim$ and a bijection μ^* between the sets Ξ^* and $\mathscr{N}^*/\sim$.*

Proof. The mapping $\gamma : \mathscr{N} \to \Xi$, defined from $\gamma(N) = N_{21}N_{11}^{-1}$, is surjective since for each $X \in \Xi$ the matrix

$$N = N(X) := \begin{bmatrix} J(X) & -X^{\mathrm{H}}J(X^{\mathrm{H}}) \\ XJ(X) & J(X^{\mathrm{H}}) \end{bmatrix}, \quad J(X) := (I_n + X^{\mathrm{H}}X)^{-1/2},$$

belongs to $\mathcal{N}$ and $\gamma(N) = X$. Hence γ may be decomposed as $\gamma = \pi \circ \mu$, where $\pi : \mathcal{N} \to \mathcal{N}/\sim$ is the canonical projection and $\mu : \mathcal{N}/\sim \to \Xi$ is the desired bijection. The existence of the bijection μ^* is proved in a similar way. □

Note that each $X \in \Xi$ defines the orbit

$$\mu^{-1}(X) = \left\{ \begin{bmatrix} U_1 J(X) & -X^{\mathrm{H}} U_2 J(X^{\mathrm{H}}) \\ X U_1 J(X) & U_2 J(X^{\mathrm{H}}) \end{bmatrix} : U_1, U_2 \in \mathcal{U}(n) \right\}$$

which is homeomorphic to $\mathcal{U}(n) \times \mathcal{U}(n)$, i.e. the members of the orbit spaces are of real dimension $2n^2$.

Let $N = [N_{ij}] \in \mathcal{U}(2n)$, where $N_{ij} \in \mathbb{C}^{n\times n}$, be the matrix which congruently reduces S into diagonal form, i.e. $N^{\mathrm{H}} S N = \mathrm{diag}(\Lambda_1, -\Lambda_2)$, $\Lambda_j := \mathrm{diag}(\lambda_{j1}, \ldots, \lambda_{jn})$, $\lambda_{jk} > 0$, and $\lambda_{1k}, -\lambda_{2k}$ are the eigenvalues of S. Note that this transformation may be accomplished in a numerically reliable way using the corresponding software from EISPACK, LINPACK or LAPACK [19,5,3,1].

For $U \in \mathcal{U}(n)$ define the matrices $\Theta_{j1}(U) := N_{j1} \Lambda_1^{-1/2} U + N_{j2} \Lambda_2^{-1/2}$; $j = 1, 2$, and let

$$\hat{\Omega} := \{U \in \mathcal{U}(n) : \det(\Theta(U)) \neq 0\}.$$

Now the set Ξ may be effectively parametrized from

$$\Xi = \{\Gamma(U) : U \in \hat{\Omega}\}, \tag{21}$$

where $\Gamma(U) := \Theta_{21}(U) \Theta_{11}^{-1}(U)$.

As in Section 4 it may be shown that Γ is a homeomorphism between Ξ and Ω. The advantage of the parametrisation (21) in comparison to (19) is that it is based on numerically more reliable unitary transformations only (note that obtaining $\Lambda_j^{-1/2}$ does not require matrix computations).

6. Perturbation bounds for AAMRE

6.1. Local linear bounds

Let $\|\cdot\|_{\mathcal{L}(n)}$ and $\|\cdot\|_{\mathcal{PL}(n)}$ be the induced norms in $\mathcal{L}(n)$ and $\mathcal{PL}(n)$, i.e.

$$\|L(Z)\|_{\mathcal{L}(n)} := \max\{\|L(Z)\| : Z \in \mathbb{C}^{n\times n}, \|Z\| = 1\}$$

for $L \in \mathcal{L}$, $\mathcal{L} \in \{\mathcal{L}(n), \mathcal{PL}(n)\}$.

Setting $Y = X_0 + \Delta X$ the perturbed equation (6) may be written as

$$F(X_0 + \Delta X, \Sigma + \Delta\Sigma) = F_1(\Delta X, \Delta\Sigma) + F_2(\Delta X, \Delta\Sigma) = 0, \tag{22}$$

where

$$\begin{aligned} F_1(\Delta X,\Delta\Sigma) &:= \sum_{Z\in\{X,Q,A,M\}} F_Z(X_0,\Sigma)(\Delta Z), \\ F_2(\Delta X,\Delta\Sigma) &:= -\Delta X^{\mathrm{H}}(M+\Delta M)\Delta X \\ &\quad + (\Delta A^{\mathrm{H}} - X_0^{\mathrm{H}}\Delta M)\Delta X + \Delta X^{\mathrm{H}}(\Delta A - \Delta M X_0). \end{aligned} \tag{23}$$

Here $F_Z(X_0,\Sigma)$ is the partial Fréchet (pseudo) derivative of $F(X,\Sigma)$ in Z at $X = X_0$, which is a member of $\mathscr{PL}(n)$ for $Z = X, A$ and a member of $\mathscr{L}(n)$ for $Z = Q, M$. A straightforward calculation gives

$$\begin{aligned} F_X(X_0,\Sigma)(Z) &= A_0^{\mathrm{H}}Z + Z^{\mathrm{H}}A_0, \qquad & F_Q(X_0,\Sigma)(Z) &= Z \\ F_A(X_0,\Sigma)(Z) &= X_0^{\mathrm{H}}Z + Z^{\mathrm{H}}X_0, \qquad & F_M(X_0,\Sigma)(Z) &= -X_0^{\mathrm{H}}ZX_0. \end{aligned} \tag{24}$$

Having in mind Eqs. (23) and (24), it follows from Eq. (22)

$$F_X(X_0,\Sigma)(\Delta X) = -\,\Delta Q - X_0^{\mathrm{H}}\Delta A - \Delta A^{\mathrm{H}}X_0 + X_0\Delta M X_0 - F_2(\Delta X,\Delta\Sigma). \tag{25}$$

Since $F_X(X_0,\Sigma)(Z) \in \mathscr{H}(n)$, the operator $F_X(X_0,\Sigma)$ maps $\mathbb{C}^{n\times n} \simeq \mathbb{R}^{2n^2}$ into $\mathscr{H}(n) \simeq \mathbb{R}^{n^2}$. Hence $F_X(X_0,\Sigma)$ is not invertible. It may be shown (see Section A.2 of Appendix A) that the operator $F_X(X_0,\Sigma)$ is surjective if and only if the matrix $A_0 = A - MX_0$ is non-singular (which fortunately is the case).

Denote by $F_X^{\dagger}(X_0,\Sigma)$ the right inverse operator of $F_X(X_0,\Sigma)$, which is of minimum induced norm $\|.\|_{\mathscr{PL}(n)}$, i.e. $F_X(X_0,\Sigma)\circ F_X^{\dagger}(X_0,\Sigma) = \mathscr{I}_{n^2}$ and let

$$\begin{aligned} \varphi(X_0,\Sigma) &:= \left\|F_X^{\dagger}(X_0,\Sigma)\right\|_{\mathscr{PL}(n)} \\ &= \min\left\{\|L\|_{\mathscr{PL}(n)} : L \in \mathscr{PL}(n), F_X(X_0,\Sigma)\circ L = \mathscr{I}_{n^2}\right\} \end{aligned} \tag{26}$$

(an explicit expression for $\varphi(X_0,\Sigma)$ is given later on). Then Eq. (25) yields

$$\Delta X = F_X^{\dagger}(X_0,\Sigma)(-\Delta Q - X_0^{\mathrm{H}}\Delta A - \Delta A^{\mathrm{H}}X_0 + X_0^{\mathrm{H}}\Delta M X_0 - \ F_2(\Delta X,\Delta\Sigma)). \tag{27}$$

Eq. (27) makes it possible to obtain estimates in terms of absolute or relative condition numbers

$$\begin{aligned} \Delta_X &\leqslant K_Q\Delta_Q + K_A\Delta_A + K_M\Delta_M + \mathrm{O}(\|\Delta\|^2), \quad \Delta \to 0 \\ \delta_X &\leqslant k_Q\delta_Q + k_A\delta_A + k_M\delta_M + \mathrm{O}(\|\delta\|^2), \quad \delta \to 0, \end{aligned} \tag{28}$$

where $\delta_Z := \Delta_Z/\|P\|$ and $\delta := [\delta_Q,\delta_A,\delta_M]^{\mathrm{T}} \in \mathbb{R}_+^3$.

The quantities K_Z are the *absolute condition numbers of AAMRE* (1) relative to perturbations in the matrix coefficients $Z \in \{Q, A, M\}$, while $k_Z := K_Z\|Z\|/\|X_0\|$ are the corresponding *relative condition numbers.*

If the relative perturbations in the data satisfy $\delta_Z \leqslant \delta_0$ for some $\delta_0 > 0$ then

$$\delta_X \leqslant k\delta_0 + \mathrm{O}(\delta_0^2), \quad \delta_0 \to 0; \ k := k_Q + k_A + k_M.$$

Hence the quantity k may be considered as an *overall estimate* of the relative conditioning of AAMRE. However, this number will not be a relevant measure of the real perturbation if some of the quantities k_Q, k_A, k_M is much larger than the others while the corresponding perturbation is small or zero.

Later on we show that local estimates not based on condition numbers may give better results.

It follows from Eqs. (27) and (28) that

$$K_Z = K_Z(X_0, \Sigma) = \left\|F_X^{\dagger}(X_0, \Sigma) \circ F_Z(X_0, \Sigma)\right\|_{\mathscr{PL}(n)} \tag{29}$$

and in particular $K_Q = \varphi(X_0, \Sigma)$. The expressions (29) for K_Z may not be convenient for evaluation of the condition numbers for large values of n. In this case the following estimates may be used.

In both F- and 2-norms

$$K_A \leqslant 2\varphi(X_0, \Sigma)x_0, \quad K_M \leqslant \varphi(X_0, \Sigma)x_0^2; \qquad x_0 := \|X_0\|. \tag{30}$$

If $\Delta_0 := \max\{\Delta_Q, \Delta_A, \Delta_M\}$ then

$$\Delta_X \leqslant \varphi(X_0, \Sigma)(1 + x_0)^2\Delta_0 + \mathrm{O}(\Delta_0^2). \tag{31}$$

It must be stressed that the above results are valid without the assumption that the matrices $Q + \Delta Q$ or $M + \Delta M$ are non-negative definite. Hence we have proved the following theorem.

Theorem 7. *For small* $\|\Delta\|$ *the estimates* (28) *and* (31) *are valid, where the condition numbers* K_Z *are determined or estimated from* (24), (26), (29) *and* (30).

Theorem 7 gives local norm-wise bounds for the perturbation in the solution as a function of the perturbations in the data. However, norm-wise estimates may not be relevant if the modules of the elements of the perturbations ΔQ, ΔA, ΔM vary significantly in magnitude. In this case the implementation of component-wise perturbation estimates seems more adequate. Linear local component-wise perturbation bounds are directly available from Eq. (25) neglecting the term $-F_2$. We have

$$P(\Delta X, A_0) = -\Delta Q - P(\Delta A, X_0) + X_0\Delta M X_0 + \mathrm{O}(\|\Delta\|^2), \tag{32}$$

where

$$P(Z, B) := B^{\mathrm{H}}Z + Z^{\mathrm{H}}B.$$

Taking the $\mathrm{vec}_{\mathbb{R}}$ operation from both sides of Eq. (32) we get

$$\begin{aligned}\Pi(A_0)\,\mathrm{vec}_{\mathbb{R}}(\Delta X) = &- \mathrm{vec}_{\mathbb{R}}(\Delta Q) - \Pi(X_0)\,\mathrm{vec}_{\mathbb{R}}(\Delta A)\\ &+ \Psi(X_0)\,\mathrm{vec}_{\mathbb{R}}(\Delta M) + \mathrm{O}(\|\Delta\|^2).\end{aligned} \tag{33}$$

Here $\Pi(B)$ is the matrix of the realification of the operator $P(.,B)$, $B = B_1 + \jmath B_2 \in \mathbb{C}^{n\times n}$, $B_j \in \mathbb{R}^{n\times n}$, i.e.

$$\Pi(B) := \begin{bmatrix} \Pi_1(B_1) & \Pi_1(B_2) \\ -\Pi_2(B_2) & \Pi_2(B_1) \end{bmatrix} \in \mathbb{R}^{2n^2\times 2n^2},$$

$$\Pi_1(R) := I_n \otimes R^{\mathrm{T}} + (R^{\mathrm{T}} \otimes I_n)P_n, \qquad \Pi_2(R) := I_n \otimes R^{\mathrm{T}} - (R^{\mathrm{T}} \otimes I_n)P_n$$

and $\Psi(B)$ is the matrix of the realification of the operator $Z \mapsto BZB$,

$$\Psi(B) := \begin{bmatrix} \Psi_1(B) & -\Psi_2(B) \\ \Psi_2(B) & \Psi_1(B) \end{bmatrix} \in \mathbb{R}^{2n^2\times 2n^2},$$

$$\Psi_1(B) := B_1^{\mathrm{T}} \otimes B_1 - B_2^{\mathrm{T}} \otimes B_2, \qquad \Psi_2(R) := B_1^{\mathrm{T}} \otimes B_2 + B_2^{\mathrm{T}} \otimes B_1,$$

where $P_n \in \mathbb{R}^{n^2\times n^2}$ is the vec-permutation matrix, $\mathrm{vec}(Z^{\mathrm{T}}) = P_n\,\mathrm{vec}(Z)$.

If the induced norm in $\mathscr{PL}(n)$ is based on the Frobenius matrix norm then the quantity $\varphi(X_0,\Sigma)$ may be calculated using the matrix representation of $P(.,A_0)$, namely $\varphi(X_0,\Sigma) = \|\Pi^{\dagger}\|_2$.

It follows from Eq. (33) that

$$\begin{aligned}|\mathrm{vec}_{\mathbb{R}}(\Delta X)| \preceq\ & |\Pi^{\dagger}(A_0)||\mathrm{vec}_{\mathbb{R}}(\Delta Q)|\\ &+ |\Pi^{\dagger}(A_0)\Pi(X_0)||\mathrm{vec}_{\mathbb{R}}(\Delta A)|\\ &+ |\Pi^{\dagger}(A_0)\Psi(X_0)||\mathrm{vec}_{\mathbb{R}}(\Delta M)| + \mathrm{O}(\|\Delta\|^2),\end{aligned} \tag{34}$$

where $\preceq$ is the component-wise partial order relation in $\mathbb{R}^{2n^2}$.

Using Eq. (33) we may derive at least two more local norm-wise bounds, which are alternative to the estimate (28), based on condition numbers (one of them will always be at least as good as (28)). Set $\xi = \Delta_X$, $\Delta_1 := \Delta_Q$, $\Delta_2 := \Delta_A$, $\Delta_3 := \Delta_M$, $x := \mathrm{vec}_{\mathbb{R}}(\Delta X)$, $a_1 := \mathrm{vec}_{\mathbb{R}}(\Delta Q)$, $a_2 := \mathrm{vec}_{\mathbb{R}}(\Delta A)$, $a_3 := \mathrm{vec}_{\mathbb{R}}(\Delta M)$ and

$$\begin{aligned}&A_1 := -\Pi^{\dagger}(A_0), \qquad A_2 := -\Pi^{\dagger}(A_0)\Pi^{\dagger}(X_0),\\ &A_3 := \Pi^{\dagger}(A_0)\Psi(X_0).\end{aligned}$$

Then we may rewrite Eq. (33) as

$$x = A_1a_1 + A_2a_2 + A_3a_3 + \mathrm{O}(\|a\|^2) = \mathscr{A}a + \mathrm{O}(\|a\|^2), \tag{35}$$

where

$$\mathscr{A} := [A_1, A_2, A_3] \in \mathbb{C}^{n^2\times 3n^2}, \qquad a := \left[a_1^{\mathrm{T}}, a_2^{\mathrm{T}}, a_3^{\mathrm{T}}\right]^{\mathrm{T}}.$$

When using the F-norm for the perturbations in the data and in the solution, the problem is to estimate $\xi = \|x\|_2$ subject to the constraints $\|a_i\|_2 \leqslant \Delta_i$.

The condition based norm-wise estimate (28)

$$\xi \leqslant E_1(\Delta) + \mathrm{O}(\|\Delta\|^2) := K\Delta + \mathrm{O}(\|\Delta\|^2),$$

$$K := [\|A_1\|_2, \|A_2\|_2, \|A_3\|_2],$$

is one of the possibilities. Another estimate is based on the second equality in Eq. (35),

$$\xi \leqslant E_2(\Delta) + \mathrm{O}(\|\Delta\|^2) := \|\mathscr{A}\|_2\|\Delta\| + \mathrm{O}(\|\Delta\|^2).$$

In general $E_1(\Delta)$ and $E_2(\Delta)$ are alternative, i.e. which one is better depends on the particular problem.

We also have

$$\xi^2 = \sum_{i,j=1}^{3} a_i^{\mathrm{H}} A_i^{\mathrm{H}} A_j a_j + \mathrm{O}(\|\Delta\|^3) \leqslant \sum_{i,j=1}^{3} \|A_i^{\mathrm{H}} A_j\|_2 \Delta_i \Delta_j + \mathrm{O}(\|\Delta\|^3).$$

Hence

$$\xi \leqslant E_3(\Delta) + \mathrm{O}(\|\Delta\|^2) := \sqrt{\Delta^{\mathrm{T}} \mathscr{Q} \Delta} + \mathrm{O}(\|\Delta\|^2),$$

where $\mathscr{Q} = \left[\|A_i^{\mathrm{H}} A_j\|_2\right] \in \mathscr{H}(3)$. Since $E_3(\Delta) \leqslant E_1(\Delta)$ we see that the bound E_3 is better (or eventually equal) to the condition based estimate E_1. Finally we get

$$\Delta_X \leqslant E(\Delta) + \mathrm{O}(\|\Delta\|^2) := \min\left\{\|\mathscr{A}\|_2\|\Delta\|, \sqrt{\Delta^{\mathrm{T}} \mathscr{Q} \Delta}\right\} + \mathrm{O}(\|\Delta\|^2). \tag{36}$$

Thus we have proved the following statement.

Theorem 8. *For small* $\|\Delta\|$ *the local component-wise estimate* (34) *and the improved local norm-wise estimate* (36) *for AAMRE are valid.*

6.2. Non-local non-linear bounds

The local bounds of type $\Delta_X \leqslant E(\Delta)$, derived in Section 6.1, have a very serious drawback – they are valid only asymptotically, for $\|\Delta\| \to 0$. But in practice one always has finite perturbations. Even if the latter seem to be small in the sense that $\|\delta\| \ll 1$, the neglected $\mathrm{O}(\|\Delta\|^2)$ terms may be large enough in order to turn the local bound into a NaB (NaB is something that is Not a Bound in the rigorous sense). To overcome this difficulty in Section 7 we derive a non-local non-linear bound for Δ_X, which is valid in a (possibly small but) finite domain for the perturbation vector Δ. For this purpose we use the technique of Lyapunov majorants and the Schauder fixed point principle as proposed in [13,14].

Eq. (27) may be used to obtain non-local perturbation bounds for the solution. For this purpose rewrite the perturbed equation (6) as an operator equation $Z = \Lambda(Z)$ for $Z := \Delta X$, where Λ is the right-hand side of Eq. (27).

We shall show that under some conditions on Δ there exists $\rho = f(\Delta)$ such that the continuous operator Λ maps the closed ball

$$B_\rho := \left\{Z \in \mathbb{C}^{n\times n}\colon \|Z\|_{\mathrm{F}} \leqslant \rho\right\}$$

into itself. Let $Z \in B_\rho$. Then according to the Schauder fixed point principle there exists a solution $Z \in B_\rho$ of the operator equation $Z = \Lambda(Z)$, i.e. $\Delta_X = \|Z\|_{\mathrm{F}} \leqslant \rho$.

Applying the $\mathrm{vec}_{\mathbb{R}}$ operation to $\Lambda(Z)$ we get

$$\|\Lambda(Z)\|_{\mathrm{F}} \leqslant h(\rho, \Delta) := a_0(\Delta) + a_1(\Delta)\rho + a_2(\Delta)\rho^2, \tag{37}$$

where

$$\begin{aligned} a_0(\Delta) &:= E(\Delta), \\ a_1(\Delta) &:= 2\left\|\Pi^\dagger(A_0)\right\|_2\Delta_A \\ &\qquad \left(\left\|\Pi^\dagger(A_0)\left(X_0^{\mathrm{T}} \otimes I_n\right)\right\|_2 + \left\|\Pi^\dagger(A_0)\left(I_n \otimes X_0^{\mathrm{H}}\right)\right\|_2\right)\Delta_M, \\ a_3(\Delta) &:= \left\|\Pi^\dagger(A_0)\right\|_2(\|M\|_2 + \Delta_M). \end{aligned} \tag{38}$$

Due to Eq. (37) the operator Λ will map the closed convex set B_ρ into itself if there exist $\rho > 0$ such that $h(\rho, \Delta) \leqslant \rho$. The last inequality holds true if and only if

$$\Delta \in \mathscr{D} := \left\{\Delta\colon a_1(\Delta) + 2\sqrt{a_0(\Delta)a_2(\Delta)} \leqslant 1\right\}. \tag{39}$$

In this case we may choose $\rho = f(\Delta)$ as the smaller root of the quadratic equation

$$a_2(\Delta)\rho^2 - (1 - a_1(\Delta))\rho + a_0(\Delta) = 0,$$

i.e.

$$f(\Delta) = \frac{2a_0(\Delta)}{1 - a_1(\Delta) + \sqrt{d(\Delta)}}, \tag{40}$$

where

$$d(\Delta) := (1 - a_1(\Delta))^2 - 4a_0(\Delta)a_2(\Delta). \tag{41}$$

Thus we have proved the following theorem.

Theorem 9. *Let the condition* (39) *be fulfilled. Then the perturbed equation* (6) *has a solution* $Y = X_0 + \Delta X$ *in the neighbourhood of* X_0 *such that the estimate* $\Delta_X \leqslant f(\Delta)$, $\Delta \in \mathscr{D}$, *holds, where the function* f *is defined via* (40), (41) *and* (38).

7. Examples

Example 1. Consider the first order AAMRE

$$M|X|^2 - \overline{A}X - A\overline{X} - Q = 0$$

together with the standard Riccati equation

$$MX^2 - (A + \overline{A})X - Q = 0,$$

where $M, Q \geqslant 0$ and $A \in \mathbb{C}$. If $M > 0$ (see Section 3.2) the solution set Ξ is the circle

$$\Xi = \left\{ X \in \mathbb{C} \colon \left| X - \frac{A}{M} \right|^2 = Q + \frac{|A|^2}{M^2} \right\}$$

centred at $A/M \in \mathbb{C}$ and with radius $\sqrt{Q + |A|^2/M^2}$. At the same time the set P contains two members,

$$\mathsf{P} = \left\{ \frac{A + \overline{A}}{2M} \pm \sqrt{\frac{(A + \overline{A})^2}{4M^2} + \frac{Q}{M}} \right\},$$

which are the intersection points of Ξ with the real axis.

If $M = 0$ (see Section 3.1) then, by the regularity of Σ, we have $A + \overline{A} < 0$. Hence the solution of $\overline{A}X + A\overline{X} + Q = 0$ is the straight line

$$\Xi = \left\{ \jmath At - \frac{QA}{2|A|^2} \colon t \in \mathbb{R} \right\}.$$

The set P here has a single member $X = -Q/(A + \overline{A})$, which is the intersection of Ξ with the real axis.

We see that in both cases Ξ is a closed algebraic variety in $\mathbb{C} \simeq \mathbb{R}^2$: a compact isomorphic to $\mathscr{U}(1)$ if $M > 0$ and a straight line if $M = 0$.

Example 2. Consider the second order AAMRE with matrices

$$Q = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad M = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

Let

$$X = \begin{bmatrix} x_1 & x_3 \\ x_2 & x_4 \end{bmatrix}$$

be a solution of the unperturbed equation (1). Then

$$F(X, \Sigma) = \begin{bmatrix} x_2 + \overline{x}_2 - |x_1|^2 & x_4 - \overline{x}_1 x_3 \\ \overline{x}_4 - x_1 \overline{x}_3 & 1 - |x_3|^2 \end{bmatrix}$$

and the general solution, given by

$$\Xi = \left\{ \begin{bmatrix} z & \exp(\jmath\varphi) \\ \jmath t + |z|^2/2 & \bar{z}\exp(\jmath\varphi) \end{bmatrix};\ z \in \mathbb{C};\ \varphi, t \in \mathbb{R} \right\},$$

depends on four real parameters. Further on we have

$$A - MX = \begin{bmatrix} -z & -\exp(\jmath\varphi) \\ 1 & 0 \end{bmatrix}; \quad \det(A) = \exp(\jmath\varphi) \neq 0.$$

There are two Hermitian solutions in Ξ, which in the given case are real and correspond to $\varphi = t = 0$ and $z = \pm\sqrt{2}$:

$$X_+ = \begin{bmatrix} \sqrt{2} & 1 \\ 1 & \sqrt{2} \end{bmatrix}, \qquad X_- = \begin{bmatrix} -\sqrt{2} & 1 \\ 1 & -\sqrt{2} \end{bmatrix}$$

(they are also the Hermitian members of P). The set P has two symmetric non-Hermitian, as well as two anti-Hermitian members, listed below

$$\begin{bmatrix} \jmath\sqrt{2} & -1 \\ -1 & -\jmath\sqrt{2} \end{bmatrix}, \quad \begin{bmatrix} -\jmath\sqrt{2} & -1 \\ -1 & \jmath\sqrt{2} \end{bmatrix}, \quad \begin{bmatrix} 0 & \jmath \\ -\jmath & \end{bmatrix}, \quad \begin{bmatrix} & -\jmath \\ \jmath & 0 \end{bmatrix}.$$

If we choose a particular solution $X_0 \in \Xi$ from $\varphi = t = 0$ and $z = \jmath$ then

$$X_0 = \begin{bmatrix} \jmath & 1 \\ 0.5 & -\jmath \end{bmatrix}, \qquad A_0 = \begin{bmatrix} -\jmath & -1 \\ 1 & 0 \end{bmatrix}.$$

In this case the matrices of the realification of $F_X(X_0, \Sigma)$ and its pseudoinverse are

$$\Pi(A_0) = \begin{bmatrix} 0 & 2 & 0 & 0 & -2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & -1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$\|\Pi(A_0)\|_2 = 1 + \sqrt{5},$$

$$\Pi^{\dagger}(A_0) = \frac{1}{6}\begin{bmatrix} 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -3 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 0 & -2 & 2 & 0 \\ & & & & & & & \end{bmatrix},$$

$$\|\Pi^{\dagger}(A_0)\|_2 = \frac{1+\sqrt{5}}{4}.$$

Acknowledgements

The authors thank the referee for his helpful comments.

Appendix A

A.1. Complex matrix pseudo-polynomials

Denote by $\Psi(n)$ the set of matrix polynomial functions $\mathbb{C}^{n\times n} \times \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ of two matrix variables, i.e. $R \in \Psi(n)$ if $R(X, Y)$ is a polynomial in the matrices $X, Y \in \mathbb{C}^{n\times n}$. Let $\Theta(n)$ be the set of all functions $G: \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$, which may be represented as $G(X) = R(X, X^{\mathrm{H}})$ for some $R \in \Psi(n)$. The members of $\Theta(n)$ are called *pseudo-polynomials*. The pseudo-polynomials are continuous but not, in general, differentiable over $\mathbb{C}$ as a consequence of the non-differentiability of the function $z \mapsto \bar{z}$. However, the realification of a pseudo-polynomial is a real analytic function. Let $\mathscr{PL}(n) \subset \Theta(n)$ be the set of *pseudo-linear* operators, i.e. $P \in \mathscr{PL}(n)$ if $P(Z) = P_1(Z) + P_2(Z^{\mathrm{H}})$, where $P_k \in \mathscr{L}(n)$. The pseudo-linear operators are additive, i.e. $P(X + Y) = P(X) + P(Y)$, but not homogeneous over $\mathbb{C}$, i.e. $P(\lambda Z) \neq \lambda P(Z)$, $\lambda \in \mathbb{C} \setminus \mathbb{R}$.

For an operator $L \in \mathscr{PL}(n)$ we define its *image* and *kernel* as

$$\mathrm{Rg}(L) := \{L(X): X \in \mathbb{C}^{n\times n}\} \subset \mathbb{C}^{n\times n},$$
$$\mathrm{Ker}(L) := \{X \in \mathbb{C}^{n\times n}: L(X) = 0\} \subset \mathbb{C}^{n\times n}.$$

For a function $G \in \Theta(n)$, where $G(X) = R(X, X^{\mathrm{H}})$ and $R \in \Psi(n)$, we define the Fréchet pseudo-derivative $G_X(X_0) \in \mathscr{PL}(n)$ of G at $X = X_0$ as

$$G_X(X_0)(Z) = R_X(X_0, X_0^{\mathrm{H}})(Z) + R_Y(X_0, X_0^{\mathrm{H}})(Z^{\mathrm{H}}),$$

where $R_X(X_0, Y_0)$ and $R_Y(X_0, Y_0)$ are the partial Fréchet derivatives of R in X and Y respectively at the point $(X, Y) = (X_0, Y_0)$.

For $G \in \Theta(n)$ the solution set $\Gamma \subset \mathbb{C}^{n\times n}$ of the equation $G(X) = 0$ is closed in the standard point-wise topology. For $X_0 \in \Gamma$ the tangent set $\mathscr{T}_{X_0}$ to the set Γ at the point X_0 is defined as $T_{X_0} := X_0 + \mathrm{Ker}(G_X(X_0))$.

A.2. Associated Lyapunov operators

Consider the operator $P(.,B)\colon \mathbb{C}^{n\times n} \simeq \mathbb{R}^{2n^2} \to \mathscr{H}(n) \simeq \mathbb{R}^{n^2}$, defined from $P(Z,B) := B^{\mathrm{H}}Z + Z^{\mathrm{H}}B$, where $B \in \mathbb{C}^{n\times n}$ is a given matrix. Obviously $P(.,B) \in \mathscr{PL}(n)$. We shall be concerned with conditions under which the operator $P(.,B)$ is surjective, i.e. $\{P(Z,B)\colon Z \in \mathbb{C}^{n\times n}\} = \mathscr{H}(n)$, or, equivalently, the equation $P(Z,B) = C$ is solvable in Z for each $C \in \mathscr{H}(n)$. The above problem is solved by the following assertion.

Theorem 10. *The operator $P(.,B)$ is surjective if and only if the matrix B is non-singular.*

Proof. We shall prove the theorem by induction on n showing first that the operator $P(.,B)$ is surjective if B is non-singular. For $n = 1$ the equation $\bar{B}Z + \bar{Z}B = C$ is solvable in $Z \in \mathbb{C}$ for each $C \in \mathbb{R}$ if and only if $B \neq 0$, $Z = \jmath Bt + BC/(2|B|^2)$, where $t \in \mathbb{R}$ is arbitrary. Suppose now that every operator $P(.,B_1) : \mathbb{C}^{(n-1)\times(n-1)} \to \mathscr{H}(n-1)$ is surjective if the underlying matrix $B_1 \in \mathbb{C}^{(n-1)\times(n-1)}$ is non-singular. We shall show that the operator $P(..B) : \mathbb{C}^{n\times n} \to \mathscr{H}(n)$ is also surjective provided $B \in \mathbb{C}^{n\times n}$ is non-singular.

The matrix $B \in \mathscr{GL}(n)$ can always be taken in the form

$$B = \begin{bmatrix} B_1 & 0 \\ b^{\mathrm{H}} & \beta \end{bmatrix},$$

where $B_1 \in \mathscr{GL}(n-1)$, $b \in \mathbb{C}^{n-1}$ and $0 \neq \beta \in \mathbb{C}$ (one may use a preliminary transformation $B \to U^{\mathrm{H}}BU$, $U \in \mathscr{U}(n)$, to reduce B in lower Schur form if necessary). Partitioning the matrices

$$Z = \begin{bmatrix} Z_1 & y \\ z^{\mathrm{H}} & \zeta \end{bmatrix} \quad Z_1 \in \mathbb{C}^{(n-1)\times(n-1)};\ y, z \in \mathbb{C}^{n-1};\ \zeta \in \mathbb{C}$$

and $P(Z,B)$ in accordance with the partitioning of B we have

$$P(Z,B) = \begin{bmatrix} B_1^{\mathrm{H}}Z_1 + Z_1^{\mathrm{H}}B_1 + bz^{\mathrm{H}} + zb^{\mathrm{H}} & B_1^{\mathrm{H}}y + \zeta b + \beta z \\ y^{\mathrm{H}}B_1 + \bar{\zeta}b^{\mathrm{H}} + \bar{\beta}z^{\mathrm{H}} & \bar{\beta}\zeta + \beta\bar{\zeta} \end{bmatrix}. \tag{A.1}$$

Consider an arbitrary Hermitian matrix

$$C = \begin{bmatrix} C_1 & c \\ c^{\mathrm{H}} & \eta \end{bmatrix} \in \mathscr{H}(n), \quad C_1 \in \mathscr{H}(n-1),\ y \in \mathbb{C}^{(n-1)},\ \eta \in \mathbb{R}.$$

By the induction assumption the equation $B_1^{\mathrm{H}}Z_1 + Z_1^{\mathrm{H}}B_1 = C_1$ is solvable in Z_1 since $B_1 \in \mathscr{GL}(n-1)$. Denote by Z_1^0 any solution of the latter equation. Then it follows from Eq. (A.1) that the matrix

$$Z^0 := \begin{bmatrix} Z_1^0 & B_1^{-\mathrm{H}}(c - \zeta b) \\ 0 & \beta\zeta/(2|\beta|^2) \end{bmatrix}$$

solves the equation $B^{\mathrm{H}}Z + Z^{\mathrm{H}}B = C$ for arbitrary $C \in \mathscr{H}(n)$. Hence the operator $P(.,B) \in \mathscr{P}_n$ is surjective provided $B \in \mathscr{GL}(n)$.

To show that the surjectivity of $P(.,B) \in \mathscr{P}_n$ implies $B \in \mathscr{GL}(n)$, suppose that $P(.,B) \in \mathscr{P}_n$ is surjective but the matrix B is singular. Then one of the lower Schur forms of B is

$$\begin{bmatrix} B_1 & 0 \\ b^{\mathrm{H}} & 0 \end{bmatrix}, \quad B_1 \in \mathscr{C}^{(n-1)\times(n-1)},$$

i.e. $\beta = 0$. Then according to Eq. (A.1) the (n,n)-element of $P(Z,B)$ will be zero for each $Z \in \mathbb{C}^{n\times n}$ and hence $P(.,B)$ is not surjective. □

References

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. DuCroz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, D. Sorensen, LAPACK Users' Guide, SIAM, Philadelphia, 1992.

[2] R. Byers, Numerical condition of the algebraic Riccati equation, Contemp. Math. 47 (1985) 35–49.

[3] J. Dongarra, J. Bunch, C. Moler, G. Stewart, LINPACK Users' Guide, SIAM, Philadelphia, 1979.

[4] P. Gahinet, A. Laub, Computable bounds for the sensitivity of the algebraic Riccati equation, SIAM J. Control Optim. 28 (1990) 1461–1480.

[5] B. Garbow, J. Boyle, J. Dongarra, C. Moler, Matrix Eigensystem Routines: EISPACK Guide Extension, Springer, Berlin, 1977.

[6] I. Glazman, Yu. Lyubich, Finite-Dimensional Linear Analysis (in Russian), Nauka, Moscow, 1969.

[7] P. Lankaster, L. Rodman, Existence and uniqueness theorems for the algebraic Riccati equation, Int. J. Control 32 (1980) 285–309.

[8] P. Lancaster, L. Rodman, Algebraic Riccati Equations, Clarendon Press, Oxford, 1995.

[9] L. Kantorovich, G. Akilov, Functional Analysis in Normed Spaces (in Russian), Nauka, Moscow, 1977.

[10] C. Kenney, G. Hewer, The sensitivity of the algebraic and differential Riccati equations, SIAM J. Control Optim. 28 (1990) 50–69.

[11] M. Konstantinov, P. Petkov, N. Christov, The associated algebraic Riccati equation, in: Proceedings of Third International Conference on Systems Engr., Wright State Univ., Dayton, OH, 1984, pp. 530–537.

[12] M. Konstantinov, P. Petkov, N. Christov, Invariance of the quadratic cost on a set of non-optimal control laws, Systems Control Lett. 2 (1982) 169–174.

[13] M. Konstantinov, P. Petkov, N. Christov, Perturbation analysis of the continuous and discrete matrix Riccati equations, in: Proc. 1986 ACC, Seattle, vol. 1, 1986, pp. 636–639.

[14] M. Konstantinov, P. Petkov, N. Christov, Perturbation analysis of matrix quadratic equations, SIAM J. Sci. Statist. Comput. 11 (1990) 1159–1163.

[15] M. Konstantinov, P. Petkov, D. Gu, I. Postlethwaite, Perturbation techniques for linear control problems, LUED Rpt. 95-7, Department of Engineering, Leicester University, Leicester, UK, February 1995.

[16] A. Laub, A Schur method for solving algebraic Riccati equations, IEEE Trans. Automat. Control 24 (1979) 913–921.

[17] J. Potter, Matrix quadratic solutions, SIAM J. Appl. Math. 14 (1966) 496–501.

[18] M. Shayman, Geometry of the algebraic Riccati equation (Parts I and II), SIAM J. Control Optim. 21 (1983) 375–409.

[19] B. Smith, J. Boyle, J. Dongarra, B. Garbow, Y. Ikebe, V. Klema, C. Moler, Matrix Eigensystem Routines: EISPACK Guide, Springer, Berlin, 1976.

[20] Ji-Guang Sun, Perturbation theory for algebraic Riccati equations, SIAM J. Matrix Anal. Appl. 19 (1998) 39–65.