Linear Algebra and its Applications 359 (2003) 197–218
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

Perturbation bounds for coupled matrix Riccati equations

Mihail Konstantinov [a,*], Vera Angelova [b], Petko Petkov [c], Dawei Gu [d], Vassilios Tsachouridis [d]

[a] *University of Architecture and Civil Engineering, 1 Hr. Smirnenski Blvd., 1046 Sofia, Bulgaria*
[b] *Institute of Information Technologies, Akad. G. Bonchev Str., Bl. 2, 1113 Sofia, Bulgaria*
[c] *Department of Automatics, Technical University of Sofia, 1756 Sofia, Bulgaria*
[d] *Department of Engineering, University of Leicester, Leicester LE1 7RH, England, UK*

Received 20 August 2001; accepted 6 May 2002

Submitted by V. Mehrmann

Abstract

Local and non-local perturbation bounds for real continuous-time coupled algebraic matrix Riccati equations are derived using the technique of Lyapunov majorants and fixed point principles. Asymptotic expansions of non-linear non-local bounds are also presented. Equations of this type arise in the $\mathscr{H}_2/\mathscr{H}_\infty$ analysis and design of linear control systems.

AMS classification: 15A24; 93B36; 94C73

Keywords: Perturbation analysis; Coupled algebraic matrix Riccati equations; Condition numbers $\mathscr{H}_2/\mathscr{H}_\infty$ control

1. Introduction and notation

The real Continuous-time Coupled Algebraic matrix Riccati Equations (CCARE), considered below, are related to the $\mathscr{H}_2$ and $\mathscr{H}_\infty$ analysis and design of linear multivariable system, see [1,2,5,12]. The numerical solution of these equations is usually

* Corresponding author.
E-mail addresses: mmk_fte@uacg.bg (M. Konstantinov), vangelova@iit.bas.bg (V. Angelova), php@tu-sofia.acad.bg (P. Petkov), dag@le.ac.uk (D. Gu), vt10@le.ac.uk (V. Tsachouridis).

PII: S0024-3795(02)00416-0

contaminated with rounding and parameters errors. This may lead to significant loss of accuracy and, in particular, to divergence of the numerical procedure, carried out in floating point computing environment. The error in the computed solution depends on the sensitivity of the solution of CCARE to perturbations in their matrix coefficients. Hence obtaining perturbation bounds for CCARE is important from both theoretical and computational point of view.

In this paper we present a complete perturbation analysis of CCARE of the form $F_i(X_1, X_2, P_i) = 0$, $i = 1, 2$, where F_i are matrix quadratic functions in the unknown matrices X_i, and P_i are collections of matrix coefficients (see (1) for more details). Suppose that P_i are subject to perturbations $P_i \to P_i + \delta P_i$ which lead to perturbations $X_i \to X_i + \delta X_i$ in the solution matrices. Then the perturbation analysis problem is to estimate the norms of the perturbations δX_i as functions of the norms of the perturbations δP_i in the coefficient matrices. In practice, the perturbations δP_i may be due to parameter uncertainties as well as to rounding errors when solving the equations in finite precision arithmetics.

As a result of the perturbation analysis, using the technique of Lyapunov majorants [3,7] and fixed point principles [11], local first order homogeneous as well as non-local non-linear perturbation bounds are derived. The non-local bounds are rigorous and they are valid in a certain finite domain in the space of perturbations in the coefficient matrices. The local bounds are asymptotic, valid for $\delta P \to 0$. These local bounds are first order homogeneous non-linear functions and are better than the bounds, based on individual condition numbers.

An experimental analysis is made to compare the performance of the proposed perturbation bounds. It is shown that for some particular example the non-local bounds are slightly more pessimistic than the local ones.

Throughout the paper we use the following notation: $\mathbb{R}^{m\times n}$—the space of $m \times n$ real matrices; $\mathbb{R}^m = \mathbb{R}^{m\times 1}$; $\mathbb{R}_+ = [0, \infty)$; $A^{\mathrm{T}} \in \mathbb{R}^{n\times m}$—the transpose of the matrix $A \in \mathbb{R}^{m\times n}$; $\preceq$—the component-wise order relation on $\mathbb{R}^{m\times n}$; $\mathrm{vec}(A) \in \mathbb{R}^{mn}$—the column-wise vector representation of the matrix $A \in \mathbb{R}^{m\times n}$; $\mathrm{Mat}(\mathbf{L}) \in \mathbb{R}^{pq\times mn}$—the matrix representation of the linear matrix operator $\mathbf{L} : \mathbb{R}^{m\times n} \to \mathbb{R}^{p\times q}$, i.e.,

$$\mathrm{vec}(\mathbf{L}(X)) = \mathrm{Mat}(\mathbf{L})\mathrm{vec}(X)$$

for all $X \in \mathbb{R}^{m\times n}$; I_n—the unit $n \times n$ matrix; Π_{n^2}—the $n^2 \times n^2$ vec-permutation matrix such that $\mathrm{vec}(A^{\mathrm{T}}) = \Pi_{n^2}\mathrm{vec}(A)$ for all $A \in \mathbb{R}^{n\times n}$; $A \otimes B = [a_{pq}B]$—the Kronecker product of the matrices $A = [a_{pq}]$ and B; $\|\cdot\|_2$—the Euclidean norm in $\mathbb{R}^m$ or the spectral (or 2-) norm in $\mathbb{R}^{m\times n}$; $\|\cdot\|_{\mathrm{F}}$—the Frobenius (or F-) norm in $\mathbb{R}^{m\times n}$; $\|\cdot\|$—a replacement of either $\|\cdot\|_2$ or $\|\cdot\|_{\mathrm{F}}$; $\mathrm{rad}(A)$—the spectral radius of the square matrix A; $\det(A)$—the determinant of the square matrix A.

If $P = (E_1, \ldots, E_r)$ is a matrix r-tuple, we denote by

$$|\!|\!| P |\!|\!| = [\|E_1\|, \ldots, \|E_r\|]^{\mathrm{T}} \in \mathbb{R}_+^r$$

its generalized norm. We also set $\mathscr{R} = \mathbb{R}^{n\times n}$ and $\mathscr{S} = \{A \in \mathscr{R} : A = A^{\mathrm{T}}\} \subset \mathscr{R}$. The set of non-negative definite matrices from $\mathscr{S}$ is denoted as $\mathscr{S}_+$.

The space of linear operators $\mathscr{L}_1 \to \mathscr{L}_2$, where $\mathscr{L}_1$, $\mathscr{L}_2$ are linear spaces, is denoted by $\mathbf{Lin}(\mathscr{L}_1, \mathscr{L}_2)$, while **Lin** is an abbreviation for $\mathbf{Lin}(\mathscr{R}, \mathscr{R})$.

We usually identify the Cartesian product $\mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}$, endowed with the structure of a linear space, with any of the spaces $\mathbb{R}^{m\times 2n}$, $\mathbb{R}^{2m\times n}$ and $\mathbb{R}^{2mn}$. In particular, the ordered pair $(A, B) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}$ and the matrix $[A, B] \in \mathbb{R}^{m\times 2n}$ are considered as identical objects. Finally, we use the same notation P for an ordered matrix r-tuple $(E_1, \dots, E_r)$ (considered as an element of a linear space) as well as for the collection $\{E_1, \dots, E_r\}$ (a collection is a set with possibly repeated elements). Thus $Z \in P$ means that Z is some of the matrices E_k of P, or that Z varies over the set P.

The notation ':=' stands for 'equal by definition'.

2. Problem statement

Consider the system of CCARE

$$\begin{aligned} F_1(X_1, X_2, P_1) &:= (A_1 + B_1X_2)^{\mathrm{T}}X_1 + X_1(A_1 + B_1X_2) \\ &\quad + C_1 - X_1D_1X_1 = 0, \\ F_2(X_1, X_2, P_2) &:= (A_2 + X_1B_2)X_2 + X_2(A_2 + X_1B_2)^{\mathrm{T}} \\ &\quad + C_2 - X_2D_2X_2 = 0, \end{aligned} \tag{1}$$

where $X_i \in \mathscr{R}$ are the unknown matrices, $A_i, B_i \in \mathscr{R}$, $C_i, D_i \in \mathscr{S}$, $i = 1, 2$, are given matrix coefficients and $P_i := (A_i, B_i, C_i, D_i) \in \mathscr{R}^4$.

We set

$$\begin{aligned} P &:= (P_1, P_2) = (A_1, B_1, C_1, D_1, A_2, B_2, C_2, D_2) \\ &=: (E_1, E_2, E_3, E_4, E_5, E_6, E_7, E_8) \in \mathscr{R}^8. \end{aligned}$$

The *generalized norm* of the matrix 8-tuple P is the vector

$$|||P||| := [\|E_1\|_{\mathrm{F}}, \dots, \|E_8\|_{\mathrm{F}}]^{\mathrm{T}} \in \mathbb{R}^8_+. \tag{2}$$

Although the matrices C_i, D_i are symmetric, system (1) may have solutions (X_1, X_2) in which some of the matrices X_i is not symmetric. In this work we are interested only in symmetric solutions of system (1), i.e., $(X_1, X_2) \in \mathscr{S}^2$. The non-symmetric case is treated similarly.

An important feature of the solutions of (1) is whether they stabilize the corresponding closed-loop system matrices (we recall that a matrix $A \in \mathscr{R}$ is stable if its eigenvalues have negative real parts).

Definition 2.1. The solution pair $(X_1, X_2) \in \mathscr{S}^2$ is called *stabilizing* if the matrices $G_1 := A_1 + B_1X_2 - D_1X_1$ and $G_2 := A_2 + X_1B_2 - X_2D_2$ are stable.

Note that F_i as defined by (1) are functions from $\mathscr{R} \times \mathscr{R} \times \mathscr{R}^4 = \mathscr{R}^6$ to $\mathscr{R}$. It will be convenient to write the system of CCARE as one matrix equation. For this purpose we denote $X := (X_1, X_2)$, $F := (F_1, F_2)$. Then the system (1) may be written as

$$F(X, P) = 0. \tag{3}$$

Here F is considered as a mapping $\mathscr{R}^{10} \to \mathscr{R}^2$, or equivalently, as a mapping $\mathbb{R}^{n\times 2n} \times \mathscr{R}^8 \to \mathbb{R}^{n\times 2n}$, see the end of Section 1.

The problem of existence of (stabilizing) solutions $(X_1, X_2) \in \mathscr{S}_+^2$ of system (1) is a difficult one and is not considered here.

In what follows we assume the following.

Assumption 2.1. The system (1) has a solution $X = (X_1, X_2) \in \mathscr{S}^2$ such that the partial Fréchet derivative $\mathbf{F}_X(X, P)(\cdot)$ of F in X at the point (X, P) is invertible.

The partial Fréchet derivative of F in X at (X, P) is a linear operator $\mathscr{R}^2 \to \mathscr{R}^2$, calculated as follows. Let $Y = (Y_1, Y_2) \in \mathscr{R}^2$ be arbitrary. We have

$$\mathbf{F}_X(X, P)(Y) = (\mathbf{F}_{1,X}(X, P_1)(Y), \mathbf{F}_{2,X}(X, P_2)(Y))$$

and

$$\mathbf{F}_{i,X}(X, P_i)(Y) = \mathbf{F}_{i,X_1}(X, P_i)(Y_1) + \mathbf{F}_{i,X_2}(X, P_i)(Y_2).$$

A direct calculation gives

$$\begin{aligned}
&\mathbf{F}_{1,X_1}(X, P_1)(Z) = G_1^{\mathrm{T}} Z + Z G_1, \ \mathbf{F}_{1,X_2}(X, P_1)(Z) = X_1 B_1 Z + Z^{\mathrm{T}} B_1^{\mathrm{T}} X_1, \\
&\mathbf{F}_{2,X_1}(X, P_2)(Z) = X_2 B_2^{\mathrm{T}} Z^{\mathrm{T}} + Z B_2 X_2, \ \mathbf{F}_{2,X_2}(X, P_2)(Z) = G_2 Z + Z G_2^{\mathrm{T}}.
\end{aligned}$$

Further on we use the following abbreviations for the partial Fréchet derivatives of F and F_i

$$\begin{aligned}
\mathbf{L}(\cdot) &:= \mathbf{F}_X(X, P)(\cdot) \in \mathbf{Lin}(\mathscr{R}^2, \mathscr{R}^2), \\
\mathbf{L}_i(\cdot) &:= \mathbf{F}_{i,X}(X, P_i)(\cdot) \in \mathbf{Lin}(\mathscr{R}^2, \mathscr{R}), \\
\mathbf{L}_{ij}(\cdot) &:= \mathbf{F}_{i,X_j}(X, P_i)(\cdot) \in \mathbf{Lin}(\mathscr{R}, \mathscr{R}).
\end{aligned}$$

Thus

$$\mathbf{F}_X(X, P)(Y) = (\mathbf{L}_1(Y), \mathbf{L}_2(Y)) = (\mathbf{L}_{11}(Y_1) + \mathbf{L}_{12}(Y_2), \mathbf{L}_{21}(Y_1) + \mathbf{L}_{22}(Y_2)).$$

Note that $\mathbf{L}_{ii}(\cdot)$ are Lyapunov operators [6]. At the same time $\mathbf{L}_{ij}(\cdot)$, $i \neq j$, are associated Lyapunov operators when $X_i \in \mathscr{S}$.

Applying the vec operation to the pair $\mathbf{F}_X(X, P)(Y)$ and using the identity $(A \otimes B)\Pi_{n^2} = \Pi_{n^2}(B \otimes A)$ (see [4]) we find that the matrix representation of the linear operator $\mathbf{L}(\cdot)$ is

$$L := \mathrm{Mat}(\mathbf{L}(\cdot)) = \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix} \in \mathbb{R}^{2n^2 \times 2n^2}, \tag{4}$$

where

$$\begin{aligned}
&L_{11} := I_n \otimes G_1^{\mathrm{T}} + G_1^{\mathrm{T}} \otimes I_n, \ L_{12} := (I_{n^2} + \Pi_{n^2})(I_n \otimes (X_1 B_1)), \\
&L_{21} := (I_{n^2} + \Pi_{n^2})\left((B_2 X_2)^{\mathrm{T}} \otimes I_n\right), \ L_{22} := I_n \otimes G_2 + G_2 \otimes I_n.
\end{aligned} \tag{5}$$

Here $L_{ij} \in \mathbb{R}^{n^2 \times n^2}$ is the matrix representation of the operator $\mathbf{L}_{ij}(\cdot)$.

It follows from Assumption 2.1 and the implicit function theorem [11] that the solution X is isolated, i.e., there exists $\varepsilon > 0$ such that Eq. (3) has no other solution $\widetilde{X}$ with $\|\widetilde{X} - X\| < \varepsilon$.

Hereinafter, with certain abuse of notation, we consider P_i as an ordered pair (and hence as an element of the linear space $\mathscr{R}^4$) as well as a collection, i.e., as a set.

The perturbation problem for CCARE (1) is stated as follows. Let the matrices from P_i be perturbed as $A_i \mapsto A_i + \delta A_i$, $B_i \mapsto B_i + \delta B_i$, $C_i \mapsto C_i + \delta C_i$, $D_i \mapsto D_i + \delta D_i$. We assume that the perturbations δC_i and δD_i are symmetric. This assumption is necessary to ensure that the perturbed equation, considered below, also has a solution in $\mathscr{S}^2$. Symmetric perturbations in C_i and D_i arise naturally in many applications, where these matrices are factorized as $C_i = \Gamma_i \Gamma_i^{\mathrm{T}}$, etc.

Denote by $P_i + \delta P_i$ the perturbed collection P_i, in which each matrix $Z \in P_i$ is replaced by $Z + \delta Z$ and let $\delta P = (\delta P_1, \delta P_2)$. Then the perturbed version of Eq. (3) is

$$F(X + \delta X, P + \delta P) = 0. \tag{6}$$

The invertibility of the operator F_X and the symmetry of the matrices $C_i + \delta C_i$, $D_i + \delta D_i$ implies that Eq. (6) has a unique isolated solution $Y = X + \delta X \in \mathscr{S}^2$ in the neighbourhood of X if the perturbation δP is sufficiently small. Moreover, in this case the elements of δX are analytic functions of the elements of δP, see [9].

Let

$$\delta := \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} \in \mathbb{R}_+^8,$$

where $\delta_i := [\delta_{A_i}, \delta_{B_i}, \delta_{C_i}, \delta_{D_i}]^{\mathrm{T}} \in \mathbb{R}_+^4$, be the vector of absolute Frobenius norm perturbations $\delta_Z := \|\delta Z\|_{\mathrm{F}}$ in the data matrices $Z \in P$.

The perturbation problem for CCARE (1) is to find bounds

$$\delta_{X_i} \leqslant f_i(\delta),\ \delta \in \Omega \subset \mathbb{R}_+^8,\ i = 1, 2, \tag{7}$$

for the perturbations $\delta_{X_i} := \|\delta X_i\|_{\mathrm{F}}$. Here Ω is a certain set and f_i are continuous functions, non-decreasing in each of their arguments and satisfying $f_i(0) = 0$. The inclusion $\delta \in \Omega$ guarantees that the perturbed CCARE (6) has a unique solution $Y = X + \delta X$ in a neighbourhood of the unperturbed solution X such that the elements of $\delta X_1, \delta X_2$ are analytic functions of the elements of the matrices δZ, $Z \in P$, provided δ is in the interior of Ω.

First order local bounds

$$\delta_{X_i} \leqslant \mathrm{est}_i(\delta) + \mathrm{O}(\|\delta\|^2), \quad \delta \to 0,\ i = 1, 2, \tag{8}$$

are first derived with $\mathrm{est}_i(\delta) = \mathrm{O}(\|\delta\|)$, $\delta \to 0$, which are then incorporated in the non-local bounds (7). Here the functions $\mathrm{est}_i : \mathbb{R}_+^8 \to \mathbb{R}_+$ are non-linear first order homogeneous, i.e., $\mathrm{est}_i(\lambda\delta) = \lambda \mathrm{est}_i(\delta)$ for every $\lambda \geqslant 0$.

3. Local perturbation analysis

In this section we present a local perturbation analysis for CCARE (1) which consists in determining the functions est_i in (8).

3.1. Condition numbers

Consider first the conditioning of the CCARE (1).

Having in mind that $F_i(X, P_i) = 0$, the perturbed equations may be written as

$$F_i(X + \delta X_i, P_i + \delta P_i) = \sum_{j=1}^{2} \mathbf{L}_{ij}(\delta X_j) + \sum_{Z \in P_i} \mathbf{F}_{i,Z}(\delta Z) + H_i(\delta X, \delta P_i) = 0,$$

where $\mathbf{F}_{i,Z}(\cdot) := \mathbf{F}_{i,Z}(X, P_i)(.) \in \mathbf{Lin}$, $Z \in P_i$, $i = 1, 2$, are the Fréchet derivatives of $F_i(X, P_i)$ in the matrix argument Z, evaluated at the point (X, P_i). The matrix expression $H_i(\delta X, \delta P_i) = \mathrm{O}(\|[\delta X, \delta P_i]\|^2)$ contains second and higher order terms in $\delta X, \delta P_i$. In fact, for $Y = (Y_1, Y_2) \in \mathscr{S}^2$, we have

$$\begin{aligned} H_1(Y, \delta P_1) = {} & (\delta B_1 Y_2 - \delta D_1 Y_1)^{\mathrm{T}} X_1 + X_1(\delta B_1 Y_2 - \delta D_1 Y_1) \\ & + Y_1 \delta B_1 X_2 + X_2 \delta B_1^{\mathrm{T}} Y_1 - Y_1(D_1 + \delta D_1) Y_1 + Y_1 \delta A_1 \\ & + \delta A_1^{\mathrm{T}} Y_1 + Y_1(B_1 + \delta B_1) Y_2 + Y_2(B_1 + \delta B_1)^{\mathrm{T}} Y_1 \end{aligned} \tag{9}$$

and

$$\begin{aligned} H_2(Y, \delta P_2) = {} & X_2 \,(Y_1 \delta B_2 - Y_2 \delta D_2)^{\mathrm{T}} + (Y_1 \delta B_2 - Y_2 \delta D_2)\, X_2 + X_1 \delta B_2 Y_2 \\ & + Y_2 \delta B_2^{\mathrm{T}} X_1 - Y_2(D_2 + \delta D_2) Y_2 + \delta A_2 Y_2 + Y_2 \delta A_2^{\mathrm{T}} \\ & + Y_2(B_2 + \delta B_2)^{\mathrm{T}} Y_1 + Y_1(B_2 + \delta B_2) Y_2. \end{aligned} \tag{10}$$

We stress that the first four terms in the right-hand sides of (9) and (10) have a structure (an outer non-perturbed multiplier X_1 or X_2) which will be exploited later in the derivation of tighter non-local bounds. Indeed, suppose that we want to bound from above the 2-norms of the vector $A\mathrm{vec}(BZC)$, where A, B and C are given matrices and the only information about the matrix Z is that $\|Z\|_{\mathrm{F}} = \|\mathrm{vec}(Z)\|_2 \leqslant \delta_Z$. Then we have the 'rough' bound

$$\begin{aligned} \|A\ \mathrm{vec}(BZC)\|_2 & \leqslant \|A\|_2 \|\mathrm{vec}(BZC)\|_2 = \|A\|_2 \|BZC\|_{\mathrm{F}} \\ & \leqslant \|A\|_2 \|B\|_2 \|C\|_2 \|Z\|_{\mathrm{F}} = \|A\|_2 \|B\|_2 \|C\|_2 \delta_Z. \end{aligned} \tag{11}$$

But we have also the bound

$$\|A\ \mathrm{vec}(BZC)\|_2 = \|A(C^{\mathrm{T}} \otimes B)\mathrm{vec}(Z)\|_2 \leqslant \|A(C^{\mathrm{T}} \otimes B)\|_2 \delta_Z. \tag{12}$$

Since $\|A(C^{\mathrm{T}} \otimes B)\|_2 \leqslant \|A\|_2 \|B\|_2 \|C\|_2$ and the strict inequality is possible, we see that the bound (12) is tighter than (11).

We recall that the matrix representation of $\mathbf{L}_{ij}(\cdot)$ is denoted by L_{ij}. We also have, for $(X_1, X_2) \in \mathscr{S}^2$,

$$\begin{aligned}
&\mathbf{F}_{1,A_1}(Z) = X_1 Z + Z^{\mathrm{T}} X_1,\ \mathbf{F}_{1,B_1}(Z) = X_1 Z X_2 + X_2 Z^{\mathrm{T}} X_1,\\
&\mathbf{F}_{1,C_1}(Z) = Z,\ \mathbf{F}_{1,D_1}(Z) = -X_1 Z X_1,\\
&\mathbf{F}_{2,A_2}(Z) = Z X_2 + X_2 Z^{\mathrm{T}},\ \mathbf{F}_{2,B_2}(Z) = X_1 Z X_2 + X_2 Z^{\mathrm{T}} X_1,\\
&\mathbf{F}_{2,C_2}(Z) = Z,\ \mathbf{F}_{2,D_2}(Z) = -X_2 Z X_2.
\end{aligned} \tag{13}$$

The inverse $\mathbf{M}(\cdot) := \mathbf{L}(\cdot)^{-1} \in \mathbf{Lin}(\mathscr{R}^2, \mathscr{R}^2)$ of the operator $\mathbf{L} = F_X(X, P)(\cdot)$ may be represented as $\mathbf{L}^{-1}(\cdot) = (\mathbf{M}_1(\cdot), \mathbf{M}_2(\cdot))$, where, for $Z := (Z_1, Z_2) \in \mathscr{R}^2$,

$$\mathbf{M}_i(Z) = \mathbf{M}_{i1}(Z_1) + \mathbf{M}_{i2}(Z_2), \mathbf{M}_{ij}(\cdot) \in \mathbf{Lin}, \quad i = 1, 2.$$

Hence $\delta X = -\mathbf{M}(W_1(\delta X, \delta P_1), W_2(\delta X, \delta P_2))$, where

$$W_i(Y, \delta P_i) := \sum_{Z \in P_i} F_{i,Z}(\delta Z) + H_i(Y, \delta P_i).$$

In this way

$$\delta X_i = -\sum_{j=1}^{2} \mathbf{M}_{ij}(W_j(\delta X, \delta P_j)), \quad i = 1, 2,$$

which gives

$$\delta X_i = -\sum_{j=1}^{2} \sum_{Z \in P_j} \mathbf{M}_{ij} \circ F_{j,Z}(\delta Z) - \sum_{j=1}^{2} \mathbf{M}_{ij}(H_j(\delta X, \delta P_j)), \quad i = 1, 2. \tag{14}$$

Therefore

$$\delta_{X_i} \leqslant \sum_{j=1}^{2} \sum_{Z \in P_j} K_{ij,Z} \delta_Z + \mathrm{O}(\|\delta\|^2),\ \delta \to 0, \tag{15}$$

where the quantity $K_{ij,Z} := \|\mathbf{M}_{ij} \circ F_{j,Z}\|_{\mathbf{Lin}}$ is the *absolute condition number* of the solution component X_i with respect to the matrix coefficient $Z \in P_j$. Here $\|.\|_{\mathbf{Lin}}$ is the induced norm in the space $\mathbf{Lin}$ of linear operators $\mathscr{R} \to \mathscr{R}$.

The calculation of the condition numbers $K_{ij,Z}$ is straightforward when the Frobenius norm is used in $\mathscr{R}$. Indeed, let $U \in \mathbf{Lin}$. Then

$$\|U\|_{\mathbf{Lin}} := \max\{\|U(Z)\|_{\mathrm{F}} : \|Z\|_{\mathrm{F}} = 1\} = \|\mathrm{Mat}(U)\|_2.$$

Let $L_{i,Z} \in \mathbb{R}^{n^2 \times n^2}$ be the matrix of the operator $F_{i,Z} \in \mathbf{Lin}$. A direct calculation in view of (13) yields

$$\begin{aligned}
&L_{1,A_1} = (\Pi_{n^2} + I_{n^2})(I_n \otimes X_1),\ L_{2,A_2} = (\Pi_{n^2} + I_{n^2})(X_2 \otimes I_n),\\
&L_{1,B_1} = (\Pi_{n^2} + I_{n^2})(X_2 \otimes X_1),\ L_{2,B_2} = (\Pi_{n^2} + I_{n^2})(X_2 \otimes X_1),\\
&L_{1,C_1} = I_{n^2},\ L_{2,C_2} = I_{n^2},\ L_{1,D_1} = -X_1 \otimes X_1,\ L_{2,D_2} = -X_2 \otimes X_2.
\end{aligned} \tag{16}$$

Denote the matrix representation of the operator

$$\mathbf{M}(\cdot) = F_X^{-1}(X, P)(\cdot) \in \mathbf{Lin}(\mathscr{R}^2, \mathscr{R}^2)$$

as

$$M := \mathrm{Mat}(\mathbf{M}) = L^{-1} := \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}, \; M_{ij} \in \mathbb{R}^{n^2 \times n^2}. \tag{17}$$

Then we have the following result.

Theorem 3.1. *In the Frobenius norm the absolute condition number of the solution component X_i relative to the matrix coefficient $Z \in P_j$ is $K_{ij,Z} = \|M_{ij} L_{j,Z}\|_2$, $i, j = 1, 2$, where the matrices M_{ij} and $L_{j,Z}$ are defined by* (16) *and* (17) *in view of* (4), (5).

Proof. The proof follows from (15) and the exprerssions for the matrix representations $L_{i,Z}$ of the linear matrix operators $\mathbf{L}_{i,Z}$ and for the blocks M_{ij} of the matrix $M = L^{-1}$ of $\mathbf{M}$. □

3.2. First order homogeneous bounds

Rewrite Eq. (14) in vectorized form as

$$\mathrm{vec}(\delta X_i) = \sum_{j=1}^{2} \sum_{Z \in P_j} N_{i,Z} \mathrm{vec}(\delta Z) - \sum_{j=1}^{2} M_{ij} \mathrm{vec}(H_j(\delta X, \delta P_j)), \quad i = 1, 2, \tag{18}$$

where $N_{i,Z} := -M_{ij} L_{j,Z} \in \mathbb{R}^{n^2 \times n^2}$, $Z \in P_j$.

The condition number based perturbation bounds are an immediate consequence of (18),

$$\delta_{X_i} = \|\delta X_i\|_{\mathrm{F}} = \|\mathrm{vec}(\delta X_i)\|_2 \leqslant \mathrm{est}_i^{(1)}(\delta) + \mathrm{O}(\|\delta\|^2), \quad \delta \to 0, \tag{19}$$

where

$$\mathrm{est}_i^{(1)}(\delta) := \sum_{j=1}^{2} \sum_{Z \in P_j} \|N_{i,Z}\|_2 \delta_Z = \sum_{j=1}^{2} \sum_{Z \in P_j} K_{ij,Z} \delta_Z.$$

The bounds $\mathrm{est}_i^{(1)}(\cdot)$ are linear functions in the perturbation vector $\delta \in \mathbb{R}^8$.

Relations (18) also give another perturbation bound

$$\delta_{X_i} \leqslant \mathrm{est}_i^{(2)}(\delta) + \mathrm{O}(\|\delta\|^2), \quad \delta \to 0, \tag{20}$$

where $\mathrm{est}_i^{(2)}(\delta) := \|N_i\|_2 \|\delta\|_2$ and

$$\begin{aligned} N_i &:= [N_{i,1}, N_{i,2}] \in \mathbb{R}^{n^2 \times 8n^2}, \\ N_{i,j} &:= [N_{i,A_j}, N_{i,B_j}, N_{i,C_j}, N_{i,D_j}] \in \mathbb{R}^{n^2 \times 4n^2}, \quad i = 1, 2. \end{aligned} \tag{21}$$

The bounds $\mathrm{est}_i^{(1)}(\delta)$ and $\mathrm{est}_i^{(2)}(\delta)$ are alternative, i.e., which one is better depends on the particular value of δ.

There is a third bound, which is always less or equal to $\mathrm{est}_1^{(1)}(\delta)$, see also [8]. Indeed, we have $\delta_{X_i}^2 = \mathrm{vec}^{\mathrm{T}}(\delta X_i)\mathrm{vec}(\delta X_i) = \eta^{\mathrm{T}} N_i^{\mathrm{T}} N_i \eta + \mathrm{O}(\|\delta\|^2)$, $\delta \to 0$, where

$$\eta := \left[\mathrm{vec}^{\mathrm{T}}(\delta A_1), \mathrm{vec}^{\mathrm{T}}(\delta B_1), \ldots, \mathrm{vec}^{\mathrm{T}}(\delta D_2)\right]^{\mathrm{T}} \in \mathbb{R}^{8n^2}. \tag{22}$$

We shall represent the matrix $N_i^{\mathrm{T}} N_i \in \mathbb{R}_+^{8n^2 \times 8n^2}$ as a 8×8 block matrix with $n^2 \times n^2$ blocks as follows. Let the $n^2 \times n^2$ blocks of N_i be denoted as $\widehat{N}_{i,k}$, $k = 1, \ldots, 8$, i.e.,

$$N_i = \left[\widehat{N}_{i,1}, \widehat{N}_{i,2}, \ldots, \widehat{N}_{i,8}\right], \; \widehat{N}_{i,k} \in \mathbb{R}^{n^2 \times n^2},$$

where

$$\widehat{N}_{i,1} := N_{i,A_1}, \; \widehat{N}_{i,2} := N_{i,B_1}, \ldots, \widehat{N}_{i,8} := N_{i,D_2}.$$

Then $\eta^{\mathrm{T}} N_i^{\mathrm{T}} N_i \eta \leqslant \delta^{\mathrm{T}} \widehat{N}_i \delta$, where $\widehat{N}_i = [n_{i,pq}] \in \mathbb{R}_+^{8 \times 8}$ is a matrix with elements

$$n_{i,pq} := \left\| \widehat{N}_{i,p}^{\mathrm{T}} \widehat{N}_{i,q} \right\|_2, \; p, q = 1, \ldots, 8$$

(note that the non-negative matrices $\widehat{N}_i$ may be indefinite). Therefore we find a third type perturbation bounds

$$\delta_{X_i} \leqslant \mathrm{est}_i^{(3)}(\delta) + \mathrm{O}(\|\delta\|^2), \quad \delta \to 0, \tag{23}$$

where $\mathrm{est}_i^{(3)}(\delta) := \sqrt{\delta^{\mathrm{T}} \; \widehat{N}_i \delta}$.

The overall estimates are summarized in the next theorem.

Theorem 3.2. *It is fulfilled that*

$$\delta_{X_i} \leqslant \mathrm{est}_i(\delta) + \mathrm{O}(\|\delta\|^2), \; \delta \to 0, \quad i = 1, 2,$$

where

$$\mathrm{est}_i(\delta) := \min\left\{\mathrm{est}_i^{(2)}(\delta), \mathrm{est}_i^{(3)}(\delta)\right\}, \quad i = 1, 2,$$

and $\mathrm{est}_i^{(2)}(\delta)$, $\mathrm{est}_i^{(3)}(\delta)$ *are determined by* (20) *and* (23), *respectively*.

Proof. We have three local first order bounds, defined by (19) and (23). The bounds (19) and (20) are alternative, and the bounds (20) and (23) are also alternative. At the same time we have

$$\left\| \widehat{N}_{i,p}^{\mathrm{T}} \widehat{N}_{i,q} \right\|_2 \leqslant \left\| \widehat{N}_{i,p} \right\|_2 \left\| \widehat{N}_{i,q} \right\|_2,$$

which yields $\mathrm{est}_i^{(3)}(\delta) \leqslant \mathrm{est}_i^{(1)}(\delta)$ and completes the proof. □

We stress that the local bounds, given in Theorem 3.2, may be very accurate for certain collections of data and data perturbations. This will be the case when, for example, the vector η in (22) is (approximately) proportional to the right singular vector of the matrix N_i from (21), corresponding to its maximum singular value $\|N_i\|_2$.

The local bounds considered in this section are continuous, first order homogeneous, non-linear functions in δ. Also, for $\delta \neq 0$ these functions are real analytic.

All the three bounds $\mathrm{est}_i^{(k)}$ are in fact majorants for the solution of a complicated optimization problem, defining the conditioning of the problem as follows. Set $\xi_i := \mathrm{vec}(\delta X_i)$ and $\delta := [\delta_1, \dots, \delta_8]^{\mathrm{T}} := [\delta_{A_1}, \dots, \delta_{D_2}]^{\mathrm{T}} \in \mathbb{R}_+^8$. Then we have

$$\xi_i = \sum_{k=1}^{8} \widehat{N}_{i,k}\eta_k + \mathrm{O}(\|\delta\|^2), \quad \delta \to 0$$

and $\delta_{X_i} = \|\xi_i\|_2 \leqslant K_i(\delta) + \mathrm{O}(\|\delta\|^2),\ \delta \to 0$. Here

$$K_i(\delta) := \max\left\{\left\|\sum_{k=1}^{8} \widehat{N}_{i,k}\eta_k\right\|_2 : \|\eta_k\| \leqslant \delta_k,\ k = 1, \dots, 8\right\}$$

is the exact upper bound for the first order term in the perturbation bound for the solution component X_i (note that $K_i(\delta)$ is well defined, since the minimization in η is carried out over a compact set).

The calculation of $K_i(\delta)$ is a difficult task. Instead, one can use a bound above such as $\mathrm{est}_i(\delta) \geqslant K_i(\delta)$.

Let $\gamma \in \mathbb{R}_+^8$ be a given vector. Then we may define the relative conditioning of the problem as follows.

Definition 3.1. Let $X_i \neq 0$. The quantity $\kappa_i(\gamma) := K_i(\gamma)/\|X_i\|_{\mathrm{F}}$ is the *relative condition number of X_i with respect to γ*. If $|||P|||$ is the generalized norm (2) of P, then $\kappa_i(|||P|||)$ is the *relative norm-wise condition number* of X_i.

Note that if all elements γ_k of γ are zero except one, equal to $\|E_l\|_{\mathrm{F}}$ in the lth position, then the quantity $\kappa_i(\gamma)$ is the individual relative condition number of X_i with respect to perturbations in the matrix coefficient E_l.

4. Non-local perturbation analysis

4.1. Introductory remarks

Local bounds of the type considered in Section 3 are valid only asymptotically, for $\delta \to 0$. But in practice they are usually used simply neglecting terms of order $\mathrm{O}(\|\delta\|^2)$, e.g., $\delta_{X_i} \leqslant \mathrm{est}_i(\delta)$. Unfortunately, such chopped bounds may not be correct either because they underestimate the true perturbed quantity or because the solution of the perturbed problem does not exist. The reason is that it is usually impossible to say, having a small but a finite perturbation δ, whether the neglected terms are indeed negligible. Moreover, for some critical values of the perturbations in the coefficient matrices the solution may not exist (or may go to infinity when these critical values are approached). Nevertheless, even in such cases the local estimates will still produce a 'bound' for a very large or even for a non-existing solution which surely is not desirable.

The disadvantages of the local estimates may be overcome using the techniques of non-linear perturbation analysis. As a result, we get a domain $\Omega \subset \mathscr{R}_+^8$ and two non-linear continuous functions $f_1, f_2 : \Omega \to \mathscr{R}_+$, satisfying $f_1(0) = f_2(0) = 0$, and such that $\delta_{X_i} \leqslant f_i(\delta)$, $\delta \in \Omega$, $i = 1, 2$. The inclusion $\delta \in \Omega$ guarantees that the perturbed equation has an unique solution in a neighbourhood of the unperturbed solution. Furthermore, the last estimate is rigorous, i.e., the inequality holds true for all perturbations with $\delta \in \Omega$.

A disadvantage of the non-local bounds is that they may not exist or may be pessimistic for some collections of perturbations.

4.2. *The perturbed equation*

The perturbed equation $F(X + \delta X, P + \delta P) = 0$ may be rewritten as an operator equation for the perturbation δX

$$\delta X = \Pi(\delta X, \delta P), \ \Pi = (\Pi_1, \Pi_2), \tag{24}$$

where $\Pi(Y, \delta P) := -\mathbf{M}(F_P(X, P)(\delta P) + H(Y, \delta P))$. Here

$$H(Y, \delta P) := (H_1(Y, \delta P_1), H_2(Y, \delta P_2))$$

contains second and third order terms in Y and δP, see (9), (10).

Eq. (24) comprises two equations, namely

$$\delta X_i = \Pi_i(\delta X, \delta P_i), \quad i = 1, 2, \tag{25}$$

where the right-hand side of (25) is defined by relations (14). Setting

$$\xi_i := \mathrm{vec}(\delta X_i) \in \mathbb{R}^{n^2}, \quad i = 1, 2, \ \xi := \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \in \mathbb{R}^{2n^2},$$

we obtain the vector operator equation

$$\xi = \pi(\xi, \eta), \tag{26}$$

in $\mathbb{R}^{2n^2}$, which is reduced to two coupled vector equations $\xi_i = \pi_i(\xi, \eta)$, $i = 1, 2$, in $\mathbb{R}^{n^2}$.

Next we present a brief description of the method of Lyapunov majorants [3,7] for the analysis of operator equations of type (26). We recall that our purpose is to find bounds for $\delta_{X_i} = \|\xi_i\|_2$.

Define generalized norms in $\mathbb{R}^{2n^2}$ and $\mathbb{R}^{8n^2}$ by

$$|||\xi||| := \begin{bmatrix} \|\xi_1\|_2 \\ \|\xi_2\|_2 \end{bmatrix} \in \mathbb{R}_+^2$$

and $|||\eta||| := [\|\eta_1\|_2, \ldots, \|\eta_8\|_2]^{\mathrm{T}} \in \mathbb{R}^8_+$. For all $\rho \in \mathbb{R}^2_+$ let

$$\mathscr{B}_\rho := \left\{\xi \in \mathbb{R}^{2n^2} : |||\xi||| \preceq \rho\right\}$$

be the ball centered at the origin and of generalized radius ρ.

Suppose that we can find a continuous function

$$h = \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} : \mathbb{R}^2_+ \times \mathbb{R}^8_+ \to \mathbb{R}^2_+$$

such that the following assumption takes place.

Assumption 4.1

1. The components h_i are non-decreasing in all of their scalar arguments, for all $\delta \in \mathbb{R}^8_+$ the function $h(\cdot, \delta) : \mathbb{R}^2_+ \to \mathbb{R}^2_+$ is differentiable, and
$$h(0, 0) = 0, \ \mathrm{rad}(h_\rho(0, 0)) < 1.$$
2. For all $\rho \in \mathbb{R}^2_+$; $\xi, \widetilde{\xi} \in \mathscr{B}_\rho$ and $\eta \in \mathscr{B}_\delta$ the inequalities $|||\pi(\xi, \eta)||| \preceq h(\rho, \delta)$ and $|||\pi(\xi, \eta) - \pi(\widetilde{\xi}, \eta)||| \preceq h_\rho(\rho, \delta)|||\xi - \widetilde{\xi}|||$ hold.

Here $h_\rho(\rho, \delta)$ is the Jacobi matrix of the function $\rho \mapsto h(\rho, \delta)$ for a fixed value of δ. In our case the matrix $h_\rho(\rho, \delta)$ is non-negative and according to the Perron–Frobenius theorem [10] its spectral radius is equal to its maximum (non-negative) eigenvalue.

Definition 4.1. The function h, satisfying Assumption 4.1, is called a *Lyapunov majorant* for the operator equation (26).

If h is a Lyapunov majorant, then there exists a domain $\Omega \subset \mathbb{R}^8_+$ such that for $\delta \in \Omega$ the vector majorant equation $\rho = h(\rho, \delta)$ has a solution $\rho = f(\delta) = (f_1(\delta), f_2(\delta))$. Here $f : \Omega \to \mathbb{R}^2_+$ is a continuous function, the components f_i of f are non-decreasing in each of their scalar arguments (i.e., $\delta \preceq \widetilde{\delta}$ implies $f(\delta) \preceq f(\widetilde{\delta})$), and $f(0) = 0$.

For $\delta \in \Omega$ the operator $\pi(\cdot, \eta) : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ maps the closed convex set $\mathscr{B}_{f(\delta)}$ into itself. Hence, according to the Schauder fixed point principle [11], there exists a solution $\xi \in \mathscr{B}_{f(\delta)}$ of the operator equation (26). Now the desired non-local perturbation bounds for the solution are

$$\delta_{X_i} = \|\xi_i\|_2 \leqslant f_i(\delta), \ \delta \in \Omega.$$

We have $\pi_i(\xi, \eta) = N_i\eta_i + \psi_i(\xi, \eta)$, where

$$\psi_i(\xi, \eta) := -\mathrm{vec}\left(\sum_{j=1}^{2} M_{ij}\,\mathrm{vec}\left(H_j\left(\mathrm{vec}^{-1}(\xi), \mathrm{vec}^{-1}(\eta_j)\right)\right)\right).$$

We next apply the theory of Lyapunov majorants and fixed point principles of Banach and Schauder [3,7] to show that the operator $\pi(\cdot, \eta) : \mathbb{R}^{2n^2} \to \mathbb{R}^{2n^2}$ is a contraction on a certain ‘small’ set of diameter vanishing together with η. An estimate of this set in terms of δ will give us the desired non-local perturbation bound.

The vectorizations of the matrices $H_i(Y, \delta P_i)$ are

$$\begin{aligned}\mathrm{vec}(H_1(Y, \delta P_1)) = {} & (I_n \times X_1)\left(I_{n^2} + \Pi_{n^2}\right)\mathrm{vec}(\delta B_1 Y_2 - \delta D_1 Y_1)\\ & + (X_2 \otimes I_n)\left(I_{n^2} + \Pi_{n^2}\right)\mathrm{vec}(Y_1 \delta B_1)\\ & - \mathrm{vec}(Y_1(D_1 + \delta D_1)Y_1) + \mathrm{vec}\left(Y_1 \delta A_1 + \delta A_1^{\mathrm{T}} Y_1\right)\\ & + \mathrm{vec}\left(Y_1(B_1 + \delta B_1)Y_2 + Y_2(B_1 + \delta B_1)^{\mathrm{T}} Y_1\right)\end{aligned} \tag{27}$$

and

$$\begin{aligned}\mathrm{vec}(H_2(Y, \delta P_2)) = {} & (X_2 \otimes I_n)\left(I_{n^2} + \Pi_{n^2}\right)\mathrm{vec}\,(Y_1 \delta B_2 - Y_2 \delta D_2)\\ & + (I_n \otimes X_1)\left(I_{n^2} + \Pi_{n^2}\right)\mathrm{vec}(\delta B_2 Y_2)\\ & - \mathrm{vec}(Y_2(D_2 + \delta D_2)Y_2) + \mathrm{vec}\left(\delta A_2 Y_2 + Y_2 \delta A_2^{\mathrm{T}}\right)\\ & + \mathrm{vec}\left(Y_2(B_2 + \delta B_2)^{\mathrm{T}} Y_1 + Y_1(B_2 + \delta B_2)Y_2\right).\end{aligned} \tag{28}$$

4.3. Implicit non–local bounds

Let $\|Y_i\|_{\mathrm{F}} \leqslant \rho_i$, $i = 1, 2$, where ρ_i are non-negative constants. Then it follows from (27), (28) that

$$\begin{aligned}\|\pi_i(\xi, \eta)\|_2 & \leqslant \mathrm{est}_i(\delta) + \left\|\sum_{j=1}^{2} M_{ij}\,\mathrm{vec}(H_j(Y, \delta P_j))\right\|_2\\ & \leqslant \mathrm{est}_i(\delta) + \sum_{j=1}^{2}\left\|M_{ij}\,\mathrm{vec}(H_j(Y, \delta P_j))\right\|_2 \leqslant h_i(\rho, \delta),\\ & \quad \text{where } \rho = \begin{bmatrix}\rho_1\\ \rho_2\end{bmatrix} \in \mathbb{R}_+^2\end{aligned}$$

and

$$\begin{aligned}h_i(\rho_1, \rho_2, \delta) := {} & \mathrm{est}_i(\delta) + a_{i1}(\delta)\rho_1 + a_{i2}(\delta)\rho_2 + 2b_i(\delta)\rho_1\rho_2\\ & + c_{i1}(\delta)\rho_1^2 + c_{i2}(\delta)\rho_2^2.\end{aligned}$$

Here

$$\begin{aligned}& a_{i1}(\delta) := 2\|M_{i1}\|_2 \delta_{A_1} + \nu_{i1}\delta_{D_1} + \nu_{i2}\delta_{B_2} + \nu_{i3}\delta_{B_1},\\ & a_{i2}(\delta) := 2\|M_{i2}\|_2 \delta_{A_2} + \nu_{i1}\delta_{B_1} + \nu_{i2}\delta_{D_2} + \nu_{i4}\delta_{B_2},\\ & b_i(\delta) := \|M_{i1}\|_2(\|B_1\|_2 + \delta_{B_1}) + \|M_{i2}\|_2(\|B_2\|_2 + \delta_{B_2}),\\ & c_{i1}(\delta) := \|M_{i1}\|_2(\|D_1\|_2 + \delta_{D_1}), \quad c_{i2}(\delta) := \|M_{i2}\|_2(\|D_2\|_2 + \delta_{D_2}),\\ & \quad i = 1, 2,\end{aligned} \tag{29}$$

and

$$\begin{aligned}
\nu_{i1} &:= \left\| M_{i1}(I_n \otimes X_1)\left(I_{n^2} + \Pi_{n^2}\right)\right\|_2, \\
\nu_{i2} &:= \left\| M_{i2}(X_2 \otimes I_n)\left(I_{n^2} + \Pi_{n^2}\right)\right\|_2, \\
\nu_{i3} &:= \left\| M_{i1}(X_2 \otimes I_n)\left(I_{n^2} + \Pi_{n^2}\right)\right\|_2, \\
\nu_{i4} &:= \left\| M_{i2}(I_n \otimes X_1)\left(I_{n^2} + \Pi_{n^2}\right)\right\|_2.
\end{aligned} \tag{30}$$

The function $h : \mathbb{R}_+^2 \times \mathbb{R}_+^8 \to \mathbb{R}_+^2$ is a vector Lyapunov majorant for the operator Eq. (26), see [3,7].

Consider the majorant system of two scalar quadratic equations

$$\rho_i = h_i(\rho_1, \rho_2, \delta), \quad i = 1, 2, \tag{31}$$

which may also be written in vector form as $\rho = h(\rho, \delta)$, where

$$h(\rho, \delta) := \begin{bmatrix} h_1(\rho, \delta) \\ h_2(\rho, \delta) \end{bmatrix}.$$

We have

$$h(0, \delta) = \begin{bmatrix} \mathrm{est}_1(\delta) \\ \mathrm{est}_2(\delta) \end{bmatrix}$$

and

$$h_\rho(\rho, \delta) = \begin{bmatrix} a_{11}(\delta) + 2b_1(\delta)\rho_2 + 2c_{11}(\delta)\rho_1 & a_{12}(\delta) + 2b_1(\delta)\rho_1 + 2c_{12}(\delta)\rho_2 \\ a_{21}(\delta) + 2b_2(\delta)\rho_2 + 2c_{21}(\delta)\rho_1 & a_{22}(\delta) + 2b_2(\delta)\rho_1 + 2c_{22}(\delta)\rho_2 \end{bmatrix}.$$

Hence $h(0, 0) = 0$ and $h_\rho(0, 0) = 0$. Therefore, according to the theory of Lyapunov majorants [3,7], for δ sufficiently small, the system (31) has a solution

$$\rho = f(\delta) = \begin{bmatrix} f_1(\delta) \\ f_2(\delta) \end{bmatrix}, \tag{32}$$

which is continuous, real analytic in $\delta \neq 0$ and satisfies $\rho(0) = 0$. The function $f(\cdot)$ is defined in a domain $\Omega \subset \mathbb{R}_+^8$ whose boundary $\partial\Omega$ may be obtained by excluding ρ from the system of equations

$$\rho = h(\rho, \delta), \ \det(I_2 - h_\rho(\rho, \delta)) = 0. \tag{33}$$

The second equation means that the Jacobi matrix $h_\rho(\rho, \delta)$ of h in ρ has an eigenvalue 1. In fact, in this case the spectral radius of $h_\rho(\rho, \delta)$ is equal to 1.

Relations (33) form a system of three scalar functionally independent equations of 4th degree in 10 unknowns (the elements of ρ and δ). This defines a 7-dimensional algebraic variety $\widehat{\Omega} \subset \mathbb{R}_+^{10}$. In a neighbourhood of the origin the variety $\widehat{\Omega}$ may be parametrized as $\rho = \widehat{\rho}(t)$, $\delta = \widehat{\delta}(t)$, $t \in \mathbb{R}^7$, where $\widehat{\rho}(\cdot) : \mathbb{R}^7 \to \mathbb{R}^2_+$, $\widehat{\delta}(\cdot) : \mathbb{R}^7_+ \to \mathbb{R}^8_+$ are algebraic functions. In turn, the surface (an algebraic variety of co-dimension 1) in $\mathbb{R}^8_+$, defined by $\delta = \widehat{\delta}(t)$, $t \in \mathbb{R}^7$, forms part of the boundary of the set $\Omega \subset \mathbb{R}^8_+$.

The second equation in (33) is equivalent to

$$\begin{aligned}\omega(\rho, \delta) := 1 - \varepsilon(\delta) + \alpha_1(\delta)\rho_1 + \alpha_2(\delta)\rho_2 \\ + 2\beta(\delta)\rho_1\rho_2 + \gamma_1(\delta)\rho_1^2 + \gamma_2(\delta)\rho_2^2 = 0,\end{aligned}$$

where

$$\begin{aligned}
\varepsilon(\delta) &:= a_{11}(\delta) + a_{22}(\delta) - a_{11}(\delta)a_{22}(\delta) + a_{12}(\delta)a_{21}(\delta),\\
\alpha_1(\delta) &:= -2(c_{11}(\delta)(1 - a_{22}(\delta)) + b_2(\delta)(1 - a_{11}(\delta))\\
&\qquad + a_{12}(\delta)c_{21}(\delta) + b_1(\delta)a_{21}(\delta)),\\
\alpha_2(\delta) &:= -2(c_{22}(\delta)(1 - a_{11}(\delta)) + b_1(\delta)(1 - a_{22}(\delta))\\
&\qquad + a_{21}(\delta)c_{12}(\delta) + b_2(\delta)a_{12}(\delta)),\\
\beta(\delta) &:= 4(c_{11}(\delta)c_{22}(\delta) - c_{12}(\delta)c_{21}(\delta)),\\
\gamma_1(\delta) &:= 4(b_2(\delta)c_{11}(\delta) - b_1(\delta)c_{21}(\delta)),\\
\gamma_2(\delta) &:= 4(b_1(\delta)c_{22}(\delta) - b_2(\delta)c_{12}(\delta)).
\end{aligned}$$

Thus for the determination of (part of) the boundary $\partial\Omega$ of the set Ω we have a system of three scalar full 2nd degree equations in ρ_1, ρ_2, whose coefficients are 2nd degree polynomials in δ. For $\delta \in \Omega$ denote by $\rho = f(\delta)$ the smallest non-negative solution of the majorant system (31). Speaking about the smallest solution, some remarks are necessary.

Recall that in $\mathbb{R}^p$ the component-wise order relation $\preceq$ ($x \preceq y$ if $x_i \leqslant y_i$, where x_k and y_k are the components of x and y, respectively) is only a partial one, i.e., there are vectors $x, y \in \mathbb{R}^n$ such that neither $x \preceq y$ nor $y \preceq x$ holds. So, we have to assume that the system (31) has a smallest solution in $\mathbb{R}^2_+$. If this is not the case, we can take any solution $\rho = f(\delta) \in \mathbb{R}^2_+$ such that $\omega(f(\delta), \delta) \geqslant 0$.

As a result, we have the following result.

Theorem 4.1. *The implicit non-local non-linear perturbation bounds*

$$\delta_{X_i} \leqslant f_i(\delta),\ \delta \in \Omega,$$

are valid, where $f(\delta)$, *as defined by* (32), *is the smallest solution of* (31).

Note that if δ is not on the boundary of Ω, in the sense that $\omega(\rho, \delta) > 0$, then $\mathrm{rad}(h_\rho(\rho, \delta)) < 1$. In this case $\pi(\cdot, \delta)$ is a generalized contraction on $\mathscr{B}_\rho$ and, according to the Banach fixed point principle, the solution for δX is locally unique. Moreover, its elements are real analytic functions in the elements of the perturbations in the coefficient matrices.

4.4. Asymptotic bounds

For δ sufficiently small the perturbation bound $\rho = f(\delta)$ which is the solution of the majorant equation $\rho = h(\rho, \delta)$, is analytic in δ and, for every integer $m \geqslant 1$, we have the asymptotic expansions

$$f_i(\delta) = \sum_{k=1}^{m} f_{i,k}(\delta) + \mathrm{O}(\|\delta\|^{m+1}), \quad \delta \to 0, \quad i = 1, 2,$$

where $f_{i,k}(\delta) = \mathrm{O}(\|\delta\|^k)$, $\delta \to 0$. The expressions $f_{i,k}(\delta)$ may be derived as follows. Introduce a ficticious 'small' parameter ε and replace δ by $\varepsilon\delta$. Then $f_{i,k}(\varepsilon\delta) = \varepsilon^k f_{i,k}(\delta)$. Substituting these expressions in the majorant system and equating the coefficients of the corresponding powers of ε we obtain recurrence relations for determining $f_{i,k}(\delta)$. Finally, the parameter ε is set to 1. In particular for $m = 2$ we have the following result.

Theorem 4.2. *The asymptotic estimates*

$$\delta_{X_i} \leqslant \mathrm{est}_i(\delta) + f_{i,2}(\delta) + \mathrm{O}(\|\delta\|^3), \ \delta \to 0, \quad i = 1, 2,$$

are valid, where

$$f_{i,2}(\delta) = a_{i1}(\delta)\mathrm{est}_1(\delta) + a_{i2}(\delta)\mathrm{est}_2(\delta) + 2b_i^0\mathrm{est}_1(\delta)\mathrm{est}_2(\delta) + c_{i1}^0\mathrm{est}_1^2(\delta) + c_{i2}^0\mathrm{est}_2^2(\delta)$$

and

$$b_i^0 := \|M_{i1}\|_2\|B_1\|_2 + \|M_{i2}\|_2\|B_2\|_2, \ c_{ij}^0 := \|M_{ij}\|_2\|D_j\|_2.$$

Proof. The proof is a straightforward calculation and is hence omitted. □

4.5. Explicit non-local bounds

In practice it is not necessary to explicitly determine the domain Ω and the functions f_i. It suffices, for a given δ, to solve numerically the majorant system (31) and then to check the condition $\omega(\widetilde{\rho}, \delta) \geqslant 0$, where $\widetilde{\rho}$ is the computed solution. Then, if it exists, one has to choose the smallest non-negative solution of the system (31).

This 'numerical' approach to the non-local perturbation analysis may still be avoided, obtaining explicit perturbation bounds at the price of certain worthening of the corresponding estimates. The idea is to find a new Lyapunov majorant g, such that $h(\rho, \delta) \preceq g(\rho, \delta)$ and for which the equation

$$\rho = g(\rho, \delta) \tag{34}$$

has an explicit form solution. This can be done in many ways. Three of them are described below.

Let

$$\begin{aligned}\text{est}(\delta) &:= \max\{\text{est}_1(\delta), \text{est}_2(\delta)\},\ a_1(\delta) := \max\{a_{11}(\delta), a_{21}(\delta)\},\\ a_2(\delta) &:= \max\{a_{12}(\delta), a_{22}(\delta)\},\ b(\delta) := \max\{b_1(\delta), b_2(\delta)\},\\ c_1(\delta) &:= \max\{c_{11}(\delta), c_{21}(\delta)\},\ c_2(\delta) := \max\{c_{12}(\delta), c_{22}(\delta)\}.\end{aligned}$$

Hereinafter, in order to simplify the notation, we set $a_{ij} := a_{ij}(\delta)$, $a_i := a_i(\delta)$, $b = b(\delta)$, $c_i := c_i(\delta)$, $e_i := \text{est}_i(\delta)$, $e := \text{est}(\delta)$ thus omitting the explicit dependence of the corresponding quantities on the perturbation vector δ.

We have the following result.

Theorem 4.3. *Let*

$$\delta \in \Omega_g := \left\{\delta \in \mathbb{R}^8_+ : a_1 + a_2 + 2\sqrt{e(2b + c_1 + c_2)} \leqslant 1\right\}.$$

Then the non-linear non-local perturbation bounds

$$\delta_{X_1}, \delta_{X_2} \leqslant \frac{2e}{1 - a_1 - a_2 + \sqrt{(1 - a_1 - a_2)^2 - 4e(2b + c_1 + c_2)}} \tag{35}$$

hold true, *where the quantities in the right-hand side of* (35) *are defined by* (30), (29).

Proof. Consider the function g with components

$$g_1(\rho, \delta) = g_2(\rho, \delta) = e + a_1\rho_1 + a_2\rho_2 + 2b\rho_1\rho_2 + c_1\rho_1^2 + c_2\rho_2^2.$$

Obviously g is a Lyapunov majorant for the operator equation (26). Now the majorant equation (34) has solutions with $\rho_1 = \rho_2$, where

$$e - (1 - a_1 - a_2)\rho_1 + (2b + c_1 + c_2)\rho_1^2 = 0. \tag{36}$$

The smaller root $\rho_1(\delta)$ of (36) is the right-hand side of (35). According to the technique of Lyapunov majorants, described in Sections 4.2 and 4.3, Eq. (26) has a solution ξ with $\|\xi_i\|_2, \|\xi\|_2 \leqslant \rho_1(\delta)$ and the proof is complete. □

In Theorem 4.3 one of the bounds (35) is not asymptotically sharp unless $e_1 = e_2$. We next derive two more explicit bounds that are asymptotically sharp in the sense that their first order terms are equal to $\text{est}_i(\delta)$.

Theorem 4.4. *Suppose that*

$$\delta \in \Omega_k := \left\{\delta \in \mathbb{R}^8_+ : d_k(\delta) \geqslant 0\right\},$$

where

$$\begin{aligned}d_k(\delta) = &(1 - a_1 - a_2)^2 - 4(a_1(b + c_2) + (1 - a_2)(b + c_1))e_1\\ &- 4(a_2(b + c_1) + (1 - a_1)(b + c_2))e_2 + 4(b^2 - c_1c_2)(e_1 - e_2)^2.\end{aligned}$$

Then, *in view of* (30), (29), *we have the bounds*

$$\delta_{X_i} \leqslant \rho_i := \frac{2\left(a_j e_j + (1-a_j)e_i + c_j(e_1-e_2)^2\right)}{1-a_1-a_2+2(b+c_j)(e_i-e_j)+\sqrt{d_k}}, \quad i=1,2, \tag{37}$$

Proof. Consider the function k with components

$$k_i(\delta,\rho) := e_i + a_1\rho_1 + a_2\rho_2 + 2b\rho_1\rho_2 + c_1\rho_1^2 + c_2\rho_2^2.$$

It is easy to see that k is again a Lyapunov majorant for Eq. (26). Since $h(\rho,\delta) \preceq k(\rho,\delta) \preceq g(\rho,\delta)$ the solution of the majorant system $\rho = k(\rho,\delta)$ will majorize the solution of the system $\rho = h(\rho,\delta)$ thus producing less sharp bounds, but will give tighter bounds than those based on the majorant g. To compute this solution we observe that $\rho_1 = \rho_2 + e_1 - e_2$. Substituting this expression in any of the equations $\rho_i = k_i(\rho,\delta)$ we obtain quadratic equations for ρ_i. Choosing the smaller solutions, we obtain the perturbation bounds (37). □

Note that relative to e_1, e_2 the equation $d_k = 0$ is a parabola (if $b^2 \neq c_1c_2$) or a straight line (if $b^2 = c_1c_2$).

The bounds (37) are already asymptotically sharp. However, they can still be slightly improved as the next theorem suggests.

Theorem 4.5. *Let*

$$\delta \in \Omega_l := \left\{\delta \in \mathbb{R}_+^8 : \omega_1 + 2\sqrt{\omega_0\omega_2} \leqslant 1\right\}.$$

Then

$$\delta_{X_2} \leqslant \rho_2 := \frac{2\omega_0}{1-\omega_1+\sqrt{(1-\omega_1)^2-4\omega_2\omega_0}} \tag{38}$$

and

$$\delta_{X_1} \leqslant \rho_1 := \alpha\rho_2 + \beta, \tag{39}$$

where

$$\omega_0 := e_2 + a_{21}\beta + c_1\beta^2,$$
$$\omega_1 := a_{21}\alpha + a_{22} + 2(b + c_1\alpha)\beta,$$
$$\omega_2 := 2b\alpha + c_1\alpha^2 + c_2$$

and

$$\alpha := \frac{1+a_{12}-a_{22}}{1+a_{21}-a_{11}}, \quad \beta := \frac{e_1-e_2}{1+a_{21}-a_{11}}.$$

Proof. Consider the Lyapunov function l for (26) with components

$$l_i(\delta,\rho) := e_i + a_{i1}\rho_1 + a_{i2}\rho_2 + 2b\rho_1\rho_2 + c_1\rho_1^2 + c_2\rho_2^2$$

together with the majorant equations $\rho_i = l_i(\rho, \delta)$, $i = 1, 2$. Substracting both sides of these equations we get

$$\rho_1 - \rho_2 = e_1 - e_2 + a_{11}\rho_1 + a_{12}\rho_2 - a_{21}\rho_1 - a_{22}\rho_2.$$

Supposing that $a_{ii} < 1$ we have

$$\rho_1 = \rho_2\alpha + \beta := \rho_2\frac{1 + a_{12} - a_{22}}{1 + a_{21} - a_{11}} + \frac{e_1 - e_2}{1 + a_{21} - a_{11}}. \tag{40}$$

Substituting this expression in any of the equations $\rho_i = l_i(\rho, \delta)$ we get the quadratic equation

$$\omega_2\rho_2^2 - (1 - \omega_1)\rho_2 + \omega_0 = 0$$

for ρ_2. The smaller root of this equation is the desired bound for δ_{X_2} and this is the right-hand side of (38). The other bound (39) now follows from (40). □

5. Experimental results

Consider a pair of CCARE with matrices

$$A_1 = \begin{bmatrix} -0.4503 & -0.0027 \\ -0.0027 & -0.4648 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & 0 \\ 0.4005 & 0 \end{bmatrix},$$

$$C_1 = \begin{bmatrix} 2.0258 & -0.3951 \\ -0.3951 & 0.9296 \end{bmatrix}, \quad D_1 = \begin{bmatrix} 1.1252 & 0 \\ 0 & 0 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} -0.5664 & 0.0500 \\ 0.0500 & -0.3383 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 \\ -0.7865 & 0 \end{bmatrix},$$

$$C_2 = \begin{bmatrix} 0.9568 & 0.6865 \\ 0.6865 & 0.6766 \end{bmatrix}, \quad D_2 = \begin{bmatrix} -0.1760 & 0 \\ 0 & 0 \end{bmatrix}.$$

The perturbations in the data are taken as

$$\delta A_i = \delta B_i = \delta D_i \ (= \delta X_i) = 10^{(-k)} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

$$\delta C_1 = 10^{(-k+1)} \begin{bmatrix} -0.0645 & -0.1755 \\ -0.1755 & -0.2866 \end{bmatrix},$$

$$\delta C_2 = 10^{(-k)} \begin{bmatrix} -0.7462 & -0.7983 \\ -0.7983 & -0.8504 \end{bmatrix}$$

for $k = 10, 9, \ldots, 1$.

Note that the matrices $X_i = I_n$ solve the unperturbed CCARE.

Table 1
Local bounds

k	$\|\delta X\|_F$	$\mathrm{est}^{(1)}$	$\mathrm{est}^{(2)}$	$\mathrm{est}^{(3)}$
10	2.00×10^{-10}	3.02×10^{-9}	3.35×10^{-9}	3.00×10^{-9}
	2.00×10^{-10}	4.79×10^{-9}	7.25×10^{-9}	4.57×10^{-9}
9	2.00×10^{-9}	3.02×10^{-8}	3.35×10^{-8}	3.00×10^{-8}
	2.00×10^{-9}	4.79×10^{-8}	7.25×10^{-8}	4.57×10^{-8}
8	2.00×10^{-8}	3.02×10^{-7}	3.35×10^{-7}	3.00×10^{-7}
	2.00×10^{-8}	4.79×10^{-7}	7.25×10^{-7}	4.57×10^{-7}
7	2.00×10^{-7}	3.02×10^{-6}	3.35×10^{-6}	3.00×10^{-6}
	2.00×10^{-7}	4.79×10^{-6}	7.25×10^{-6}	4.57×10^{-6}
6	2.00×10^{-6}	3.02×10^{-5}	3.35×10^{-5}	3.00×10^{-5}
	2.00×10^{-6}	4.79×10^{-5}	7.25×10^{-5}	4.57×10^{-5}
5	2.00×10^{-5}	3.02×10^{-4}	3.35×10^{-4}	3.00×10^{-4}
	2.00×10^{-5}	4.79×10^{-4}	7.25×10^{-4}	4.57×10^{-4}
4	2.00×10^{-4}	3.00×10^{-3}	3.30×10^{-3}	3.00×10^{-3}
	2.00×10^{-4}	4.80×10^{-3}	7.20×10^{-3}	4.60×10^{-3}
3	2.00×10^{-3}	3.02×10^{-2}	3.35×10^{-2}	3.00×10^{-2}
	2.00×10^{-3}	4.79×10^{-2}	7.26×10^{-2}	4.58×10^{-2}
2	2.00×10^{-2}	3.06×10^{-1}	3.41×10^{-1}	3.03×10^{-1}
	2.00×10^{-2}	4.84×10^{-1}	7.38×10^{-1}	4.63×10^{-1}
1	2.00×10^{-1}	3.3955	4.0889	3.3715
	2.00×10^{-1}	5.3705	8.8534	5.1612

The perturbation $\|\delta X_1\|_F$, $\|\delta X_2\|_F$ in the solution is estimated by the local bounds $\mathrm{est}_i^{(1)}(\delta)$, $\mathrm{est}_i^{(2)}(\delta)$, $\mathrm{est}_i^{(3)}(\delta)$ from Section 3.2 and the non-local bounds (35), (37) from Section 4.5. The cases when the non-local bounds are not valid (since the existence condition is violated) are denoted by asterisk.

The experiments are made in a MATLAB 5.3 environment. The results obtained for different values of k are shown in Tables 1 and 2. The first quantity in each box corresponds to $i = 1$ (e.g., $\|\delta X_1\|_F, \mathrm{est}_1^{(1)}, \mathrm{est}_1^{(2)}, \mathrm{est}_1^{(3)}$), and the second one – to $i = 2$. When k decreases from 10 to 1 the non-local bounds are only slightly more pessimistic than the local bounds $\mathrm{est}_i^{(1)}(\delta), \mathrm{est}_i^{(2)}(\delta), \mathrm{est}_i^{(3)}(\delta)$. We also see that for this particular example the bound $\mathrm{est}^{(3)}(\delta)$ is superior not only to $\mathrm{est}^{(1)}(\delta)$, which is always the case, but also to $\mathrm{est}^{(2)}(\delta)$.

6. Conclusions

In this paper we have presented a complete local and non-local perturbation analysis of coupled continuous-time matrix Riccati equations, arising in the theory of $\mathscr{H}_\infty$ control. It must be pointed out that the results are not a simple extension of

Table 2
Non-local bounds

k	$\|\delta X\|_F$	(35)	(37)
10	2.00×10^{-10}	4.57×10^{-9}	3.00×10^{-9}
	2.00×10^{-10}		4.57×10^{-9}
9	2.00×10^{-9}	4.57×10^{-8}	3.00×10^{-8}
	2.00×10^{-9}		4.57×10^{-8}
8	2.00×10^{-8}	4.57×10^{-7}	3.00×10^{-7}
	2.00×10^{-8}		4.57×10^{-7}
7	2.00×10^{-7}	4.57×10^{-6}	3.00×10^{-6}
	2.00×10^{-7}		4.57×10^{-6}
6	2.00×10^{-6}	4.57×10^{-5}	3.00×10^{-5}
	2.00×10^{-6}		4.57×10^{-5}
5	2.00×10^{-5}	4.59×10^{-4}	3.01×10^{-4}
	2.00×10^{-5}		4.59×10^{-4}
4	2.00×10^{-4}	4.81×10^{-3}	3.14×10^{-3}
	2.00×10^{-4}		4.81×10^{-3}
3	2.00×10^{-3}	*	*
	2.00×10^{-3}		*
2	2.00×10^{-2}	*	*
	2.00×10^{-2}		*
1	2.00×10^{-1}	*	*
	2.00×10^{-1}		*

corresponding perturbation results for single Riccati equations. An important issue here is the construction of the inverse of the operator $\mathbf{L} = F_X(X, P)(\cdot)$.

The technique for perturbation analysis presented and in particular the perturbation bounds derived above may be extended to other more general systems of matrix quadratic equations.

Acknowldgement

The authors would like to thank the anonymous referee for his very helpful comments.

References

[1] P. Bernstein, W. Haddad, LQC control with H_∞ performance bound: a Riccati equation approach, IEEE Trans. Automat. Control 34 (1989) 293–305.
[2] J. Doyle, K. Zhou, K. Glover, B. Bodenheimer, Mixed H_2 and H_∞ performance objectives: optimal control, IEEE Trans. Automat. Control 39 (1994) 1375–1387.
[3] E.A. Grebenikov, Yu.A. Ryabov, Constructive Methods for Analysis of Nonlinear Systems, Nauka, Moscow, 1979.

[4] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis (in Russian), Cambridge University Press, Cambridge, UK, 1991.
[5] V. Kapila, W. Haddad, A multivariable extension of the Tsypkin criterion using a Lyapunov-function approach, IEEE Trans. Automat. Control 41 (1996) 149–159.
[6] M. Konstantinov, V. Mehrmann, P. Petkov, On properties of general Sylvester and Lyapunov operators, Linear Algebra Appl. 312 (2000) 35–71.
[7] M. Konstantinov, P. Petkov, D.W. Gu, I. Postlethwaite, Perturbation analysis in finite dimensional spaces, Technical Report 96–18, Department of Engineering, Leicester University, Leicester, UK, June 1996.
[8] M. Konstantinov, M. Stanislavova, P. Petkov, Perturbation bounds and characterisation of the solution of the associated algebraic Riccati equation, Linear Algebra Appl. 285 (1998) 7–31.
[9] P. Lancaster, L. Rodman, The Algebraic Riccati Equation, Oxford University Press, Oxford, 1995.
[10] P. Lancaster, M. Tismenetsky, The Theory of Matrices, second ed., Academic Press, Orlando, FL, 1985.
[11] J. Ortega, W. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, 1970.
[12] E. Tyan, P. Bernstein, Anti-windup compensator synthesis for systems with saturation actuators, Int. J. Robust. Nonlinear Control 5 (1995) 321–337.