Technische Universität Berlin
Institut für Mathematik

A general framework for the perturbation theory of matrix equations

M. Konstantinov, V. Mehrmann, P. Petkov, and D. Gu

Technical Report 760-02

Preprint-Reihe des Instituts für Mathematik
Technische Universität Berlin

Report 760-02 January 2003

A GENERAL FRAMEWORK FOR THE PERTURBATION THEORY OF GENERAL MATRIX EQUATIONS

Mihail Konstantinov* Volker Mehrmann† Petko Petkov‡
Dawei Gu§

11.12.02

Abstract

A general framework is presented for the local and non-local perturbation analysis of general real and complex matrix equations in the form $F(P, X) = 0$, where F is a continuous, matrix valued function, P is a collection of matrix parameters and X is the unknown matrix.

The local perturbation analysis produces condition numbers and improved first order homogeneous perturbation bounds for the norm $\|\delta X\|$ or the absolute value $|\delta X|$ of δX. The non-local perturbation analysis is based on the method of Lyapunov majorants and fixed point principles. It gives rigorous non-local perturbation bounds as well as conditions for solvability of the perturbed equation.

The general framework can be applied to various matrix perturbation problems in science and engineering. We illustrate the procedure with several simple examples. Furhermore, as a model problem for the new framework we derive a new perturbation theory for continuous-time algebraic matrix Riccati equations in descriptor form, $Q + A^H XE + E^H XA - E^H XSXE = 0$. The associated equation $Q + A^H XE + E^H X^H A - E^H X^H SXE = 0$ is also briefly considered.

Keywords: Perturbation analysis, general matrix equations, descriptor Riccati equations.

MSC 2000: 15A24, 93C73.

*Department of Mathematics, University of Architecture and Civil Engineering, 1046 Sofia, Bulgaria, E-mail: mmk_fte@uacg.bg. Supported in part by DFG Research Center FZT86 'Mathematics for Key Technologies' in Berlin.

†Institüt für Mathematik, MA 4–5, Technische Universität Berlin, Strasse des 17. Juni 136, D–10623 Berlin, FRG, E-mail: mehrmann@math.tu-berlin.de. Supported in part by DFG Research Center FZT86 'Mathematics for Key Technologies' in Berlin.

‡Department of Automatics, Technical University of Sofia, 1756 Sofia, Bulgaria, E-mail: php@mbox.digsys.bg. Supported in part by DFG Research Center FZT86 'Mathematics for Key Technologies' in Berlin.

§Department of Engineering, University of Leicester, Leicester LE1 7RH, UK, E-mail: dag@le.ac.uk

Contents

1 Introduction and notation

General matrix equations and in particular algebraic matrix equations are of great theoretic and practical importance. For example, most methods of modern control theory are based on the solution of linear algebraic equations (Lyapunov and Sylvester equations), quadratic or fractional-affine algebraic equations (Riccati equations), or higher order polynomial matrix equations. Finding a solution of a general matrix equation, however, may be a difficult problem due to the possibly large sensitivity of the equation and/or the properties of the numerical algorithm implemented in a finite precision arithmetic. Even if a numerically stable algorithm is used, there may still be large errors in the computed solution if the problem is ill-conditioned in the context of the finite precision machine arithmetic that is used, due to unavoidable the roundoff errors. Besides, the problem itself may come from a mathematical model where parametric and/or measurement uncertainties [58] may occur. In these cases the error in the computed solution may be estimated on the basis of the corresponding perturbation bounds. Obtaining such bounds that are also easily computable is the aim of perturbation analysis, see [52, 36].

If X_0 solves the matrix equation $F(P,X) = 0$ then the perturbed equation $F(P + \delta P, X_0 + \delta X) = 0$ determines the perturbation δX as an implicit function of δP. There are variants of the implicit function theorem [47, 31, 7] which give simple conditions for solvability of the perturbed equation as well as for continuity (or smoothness of given class) of the function $\delta P \mapsto \delta X$. However, a challenging problem here remains to obtain tight quantitative bounds for the norm of the matrix δX or for the absolute values of its elements.

There are many results devoted to the perturbation analysis of matrix equations, see for example [27, 21, 29, 13, 22, 23, 24, 2, 28, 26, 35, 37] for linear matrix equations and [8, 41, 19, 42, 12, 32, 48, 49, 43, 3, 15, 36, 59, 53, 55, 54, 39, 56, 40, 34] for quadratic and fractional-affine matrix equations. Another general approach to the sensitivity analysis of algebraic problems is proposed in [51]. Interesting results relating the conditioning of a problem to its distance to ill-posed (or singular) problems are given in [10].

The main feature of our general framework is to rewrite the perturbed equation as an operator equation for the perturbation in the solution and then to apply the method of Lyapunov majorants which, in combination with topological fixed point principles, allows to estimate the norm or generalized norm of the solution of certain operator equations in finite dimensional and abstract spaces [18, 36].

Apart from simple illustrating examples, we will demonstrate the general framework by deriving the complete perturbation analysis for continuous-time algebraic Riccati equations in descriptor form.

We use the following notation: $\mathbb{N}$ – the set of positive integers; $\mathbb{F}$ – the field of real ($\mathbb{F} = \mathbb{R}$) or complex ($\mathbb{F} = \mathbb{C}$) numbers; $\mathbb{R}_+ = [0, \infty)$; $\imath = \sqrt{-1}$ – the imaginary unit; $\mathbb{F}^{m\times n}$ – the space of $m \times n$ matrices over $\mathbb{F}$; I_n – the $n \times n$ identity matrix; $A^\top$ – the transpose of the matrix A; $\overline{A}$ – the complex conjugate of A; $A^{\mathrm{H}} = \overline{A}^\top$; $A^* = A^{\mathrm{H}}$ if A is complex and $A^* = A^\top$ if A is real; $\lambda_1(A), \ldots, \lambda_n(A)$ – the eigenvalues of $A \in \mathbb{F}^{n\times n}$ counted according to their algebraic multiplicity; $\det(A)$ – the determinant of A; $\mathrm{vec}(A) \in \mathbb{F}^{mn}$ – the column-wise vector representation of the matrix $A \in \mathbb{F}^{m\times n}$; Π_{n^2} – the $n^2 \times n^2$ vec permutation matrix such that $\mathrm{vec}(A^\top) = \Pi_{n^2}\mathrm{vec}(A)$; $J_{n^2} = I_{n^2} + \Pi_{n^2}$; $A \otimes B$ – the Kronecker product of matrices A and B; $\|\cdot\|_2$ – the Euclidean norm in $\mathbb{F}^m$ or the spectral norm in $\mathbb{F}^{m\times n}$; $\|\cdot\|_{\mathrm{F}}$ – the Frobenius norm in $\mathbb{F}^{m\times n}$ (we use $\|\cdot\|$ for any of these norms); $|A| = [|a_{ij}|] \in \mathbb{R}_+^{m\times n}$ – the absolute value of the matrix $A \in \mathbb{F}^{m\times n}$; $z^{\mathbb{R}} \in \mathbb{R}^{2n}$ $\preceq$ – the component-wise order relation in $\mathbb{R}^{m\times n}$, i.e., $A \preceq B$ when the elements of the matrix $B - A$ are non-negative.

The matrix representation L of a linear matrix operator $\mathcal{L}$ is defined by $\text{vec}(\mathcal{L}(Z)) = L\text{vec}(Z)$ for all Z.

The notation $A(z) = O(a(z))$, $z \to 0$, for a matrix $A(z)$ and a scalar $a(z) \geq 0$ depending on a vector z means that there are positive constants c and C such that for every non-negative $\gamma \leq c$ it holds that $\|A(z)\| \leq Ca(z)$, provided that $\|z\| \leq \gamma$. Similarly, $A(z) = o(a(z))$, $z \to 0$, means that for every $\varepsilon > 0$ there exists $\gamma(\varepsilon) > 0$ such that $\|A(z)\| \leq \varepsilon a(z)$ for all z with $\|z\| \leq \gamma(\varepsilon)$. The abbreviation ':=' stands for 'equal by definition'.

2 Algebraic Riccati equations in descriptor form

In order to illustrate the general framework for the perturbation of matrix equations we present the full perturbation analysis for matrix Riccati equations in descriptor form. To demonstrate the use of this equation consider the stabilizable and detectable continuous-time control system

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t),\ t > 0,\ x(0) = x_0, \\ y(t) &= Cx(t), \end{aligned} \tag{1}$$

where $x(t) \in \mathbb{F}^n$, $u(t) \in \mathbb{F}^m$ and $y(t) \in \mathbb{F}^r$ are the state, control and output vectors, respectively, and $E, A \in \mathbb{F}^{n\times n}$, $B \in \mathbb{F}^{n\times m}$, $C \in \mathbb{F}^{r\times n}$ are constant matrices. It is assumed here that the matrix E is non-singular but it may be ill-conditioned with respect to inversion. It should be noted that the complete analysis for the case of a singular matrix E is an open problem.

We recall that a matrix pair $[F, G) \in \mathbb{F}^{n\times n}\times\mathbb{F}^{n\times m}$ is *stabilizable*, if there is a matrix $K \in \mathbb{F}^{m\times n}$ such the matrix $F + GK$ is *stable*, i.e., has its spectrum in the left open complex half-plane. The pair $(H, F] \in \mathbb{F}^{r\times n}\times\mathbb{F}^{n\times n}$ is *detectable* if the pair $[F^*, H^*)$ is stabilizable. System (1), identified with the triple $(C, E^{-1}A, E^{-1}B)$ (or with the triple (CE^{-1}, AE^{-1}, B)), such that $[AE^{-1}, B)$ is stabilizable and $(C, E^{-1}A]$ is detectable, is called *regular*.

Consider the minimization of the quadratic performance index

$$J(u, x_0) := \int_0^\infty (y^*(t)y(t) + u^*(t)u(t))\mathrm{d}t \to \min .$$

The control that minimizes $J(u, x_0)$ for every initial state $x_0 \in \mathbb{F}^n$ can be realized in the form of a state feedback $u(t) = -B^*X_0Ex(t)$, where $X_0 = X_0^*$ is the unique non-negative definite solution of the descriptor Riccati equation

$$Q + A^*XE + E^*XA - E^*XSXE = 0,\ \text{where}\ Q := C^*C,\ S := BB^*, \tag{2}$$

see [5, 46]. In this case $J(u, x_0) = x_0^*X_0x_0$. It follows from the regularity of (1) that equation (2) has a unique symmetric non-negative solution X_0 such that the matrix $AE^{-1} - SX_0$ is stable [44, 45]. At the same time equation (2) may have other solutions (which are not non-negative definite and not stabilizing), including non-symmetric ones.

The above results follow immediately from the theory of linear-quadratic optimization [30, 1] for standard systems (with $E = I_n$) setting $z(t) := Ex(t)$. In this case the closed loop system is $\dot{z}(t) = (AE^{-1} - SX_0)z(t)$.

Matrix Riccati equations of descriptor form arise also in many other areas of linear systems theory, e.g. in $\mathcal{H}_\infty$ control problems such equations appear without the assumption that $Q \geq 0$ and/or $S \geq 0$ [11, 60].

Closely related to standard Riccati equations are the *associated Riccati equations* [39]. For equation (2) the associated equation is

$$Q + A^* X E + E^* X^* A - E^* X^* S X E = 0. \tag{3}$$

We briefly discuss the results for this case as well.

3 Perturbation analysis for matrix equations

Equations (2) and (3) are particular examples of quadratic matrix equations (general quadratic matrix equations are considered in [42]). Consider the general matrix equation

$$F(P, X) = 0, \tag{4}$$

where $P := (P_1, \ldots, P_k)$ is a k–tuple of matrix parameters $P_i \in \mathbb{F}^{m_i \times n_i}$, $X \in \mathbb{F}^{p \times q}$ is the unknown matrix, $F : \mathcal{P} \times \mathcal{X} \to \mathbb{F}^{m \times n}$ is a continuous matrix valued function, $pq = mn$, and $\mathcal{P} \subset \mathbb{F}^{m_1 \times n_1} \times \cdots \mathbb{F}^{m_k \times n_k}$ and $\mathcal{X} \subset \mathbb{F}^{p \times q}$ are open sets. For simplicity we restrict ourselves to the case $p = q = m = n$ and $m_1 = n_1 = \cdots = m_k = n_k = n$. We also write P as $n \times kn$ matrix $[P_1, \ldots, P_k]$. Accordingly, a norm and a generalized norm of P are defined by

$$\|P\| := \|[P_1, \ldots, P_k]\| \in \mathbb{R}_+, \ \ |\!|\!|P|\!|\!| := [\|P_1\|, \ldots, \|P_k\|]^\top \in \mathbb{R}_+^k.$$

Suppose that for a given nominal value P of the data there exists an isolated solution X_0 of equation (4), i.e., for some $\varepsilon = \varepsilon(P, X_0) > 0$ there are no solutions X satisfying $0 < \|X - X_0\| < \varepsilon$. Suppose next that the matrix coefficients in (4) are subject to perturbations $P_i \mapsto P_i + \delta P_i$.

The *perturbed equation* is obtained from (4) replacing the nominal value P of the matrix data with $P + \delta P = (P_1 + \delta P_1, \ldots, P_k + \delta P_k)$:

$$F(P + \delta P, X_0 + \delta X) = 0. \tag{5}$$

The aim of *norm-wise perturbation analysis* is to find computable bounds for the norm

$$\delta_X := \|\delta X\|$$

of the perturbation in the solution X_0 as a function of the *perturbation vector*

$$\delta = [\delta_1, \ldots, \delta_k]^\top := |\!|\!|\delta P|\!|\!| \in \mathbb{R}_+^k, \ \ \delta_i = \|\delta P_i\|.$$

Using different norms for δX and δP_i different bounds will be obtained. However, we shall use the Frobenius norm for the perturbations due to the following reasons.

- The Frobenius–norm $\|Z\|_{\mathrm{F}}$ of a matrix Z is 'natural' when this matrix is identified with its vector representation $\mathrm{vec}(Z)$ since $\|\mathrm{vec}(Z)\|_2 = \|Z\|_{\mathrm{F}}$. On the other hand if $\|Z\|$ is some other matrix norm of Z, it is not immediately clear how to estimate a vector norm of $\mathrm{vec}(Z)$ in terms of $\|Z\|$.

- When the perturbations in the data arise from parameter and measurement errors and/or rounding errors during computatioins, they are usually estimated in terms of the Frobenius–norm. For example, if a perturbation δP_i in P_i is due to rounding errors in a finite arithmetic with rounding unit eps, then the only information about δP_i is that $\|\delta P_i\|_{\mathrm{F}} \le \mathrm{eps}\|P_i\|_{\mathrm{F}}$.

- If the perturbations in the data and the solution are considered in terms of spectral norms, then we can still interprete them in Frobenius norm using the inequalities $\|Z\|_2 \le \|Z\|_{\mathrm{F}} \le \sqrt{r}\|Z\|_2$, where r is the rank of Z.

- Last but not least, the use of Frobenius–norm allows to obtain explicit computable condition measures and local perturbation bounds in terms of Kronecker matrix products. This is computationally reliable when the involved matrices are of moderate size.

Because of these reasons we consider the problem of estimating $\|\delta X\|_{\mathrm{F}}$ in terms of the quantities $\delta_i := \|\delta P_i\|_{\mathrm{F}}$.

In the *component-wise perturbation analysis* it is necessary to estimate the absolute values of the elements of the perturbation δX as functions of the elements of the matrices δP_i. This is done by estimating $|\delta X|$ in terms of $|\delta P_i|$. There are also other formulations of the component-wise perturbation analysis [22, 23, 16, 25] which are not considered here.

A priori it is not clear whether, given a perturbation δP, the perturbed equation (5) has a solution at all. So formally we have to assume that a solution to (5) exists for the given δP. However, from the non-linear perturbation analysis presented below, conditions for solvability of equation (5) will emerge.

4 A general framework for perturbation analysis

In this section we present a very general approach aimed at obtaining tight perturbation bounds for the solution of general matrix equations, when the matrix parameters in the equation are subject to perturbations. We consider techniques for deriving condition numbers and local as well as non-local non-linear perturbation bounds.

Let the general matrix equation $F(P, X) = 0$ be given, where F is a continuous matrix valued function, $P = (P_1, \dots, P_k)$ is a collection of matrix parameters and X is the unknown matrix. Let X_0 be a known solution. Let the data be changed from P to $P + \delta P$. Then we obtain the perturbed equation $F(P + \delta P, X_0 + \delta X) = 0$, where δX is the perturbation in the solution. Here two main tasks arise.

- Find conditions which guarantee that the perturbed equation has a solution $\delta X = \Xi(\delta P)$, depending continuously on δP and such that $\Xi(0) = 0$.

- Derive computable bounds for the norm $\|\delta X\|$ or the matrix absolute value $|\delta X|$ of δX as a function of the perturbations $\delta_i = \|\delta P_i\|$ or $|\delta P_i|$ in P_i.

In turn, the perturbation bounds may be local (valid for δP asymptotically small) and non-local (valid in a finite domain for δP).

The general **framework for the perturbation analysis of general matrix equations** includes the following steps.

1. *Construction of an equivalent operator equation.* This is a matrix equation $\delta X = \Pi(\delta P, \delta X)$ for δX, where $\Pi(0, 0) = 0$. For this purpose the technique of Fréchet (pseudo) derivatives is used. Via an appropriate representation, the operator equation is then represented as an equivalent matrix equation. After this, the matrix equation is vectorized as $x = \pi(p, x)$, where $x = \mathrm{vec}(\delta X)$ and $p = (p_1, \dots, p_k)$, $p_i = \mathrm{vec}(\delta P_i)$. This is accomplished by using Kronecker products of matrices and the technique of outer matrix factors.

2. *Calculation of condition numbers.* The quantity $\pi(p,x)$ is represented as $\pi_{10}(p) + \pi_{20}(p) + \pi_2(p,x)$, where $\pi_{10}(p) = O(\|p\|)$, $p \to 0$, $\pi_{20}(p) = o(\|p\|)$, $p \to 0$, and $\pi_2(p,x) = o(\|p\| + \|x\|)$, $\|p\| + \|x\| \to 0$. When only one component p_i of p is non-zero the quantity $\|\pi_{10}(p)\|/\|p_i\|$ is asymptotically bounded by K_i, the absolute condition number for the solution X_0 relative to perturbations in the matrix P_i. Thus, K_i is the asymptotic Lipschitz constant of π_{10} in p_i (if π_{10} is not Lipschitz continuous in p_i, then the condition number relative to P_i does not exist). If F is Fréchet (pseudo) differentiable then the condition numbers (in Frobenius norm) are the spectral norms of certain matrices depending on the Fréchet (pseudo) derivatives of F. For symmetric algebraic equations and $X_0 = X_0^*$ the partial Fréchet derivative $\mathcal{L}_0 := F_X(P, X_0)$ is a Lyapunov operator, and when the perturbed equation is symmetric, then $\delta X = \delta X^*$. In this case the inverse of $\mathcal{L}_0$ is estimated in terms of the Lyapunov norm which is generally less than the norm induced by the spectral matrix norm.

3. *Derivation of local perturbation bounds and overall measures of conditioning.* Here the problem is to estimate the maximum of $\|\pi_{10}(p)\|$ under the constraints $\|p_i\| \leq \delta_i$, $i = 1, \ldots, k$. Thus, local perturbation bounds and overall measures of conditioning are solutions to complicated optimization problems. For the solution of these problems, simple and easily–computable upper bounds are derived using techniques for estimating the norm of a linear combination of vectors as well as Lyapunov norms of Lyapunov operators. Local component-wise perturbation bounds are also described. An important feature of the proposed computable local perturbation bounds is that they are asymptotically exact.

4. *Construction, analysis and solution of Lyapunov majorant equations.* Setting $\delta = (\delta_1, \ldots, \delta_k)$, then the *Lyapunov majorant* for the operator $\pi(p,\cdot)$ is a function $(\delta,\rho) \mapsto h(\delta,\rho)$ such that $\|\pi(p,x)\|_2 \leq h(\delta,\rho)$, provided that $\|p_i\|_2 \leq \delta_i$ and $\|x\|_2 \leq \rho$. Under certain conditions on h and δ the majorant equation $\rho = h(\delta,\rho)$ has a solution $\rho_0 = f(\delta)$, where f is continuous and $f(0) = 0$. An inclusion of the type $\delta \in \Omega$, where Ω is a certain set (possibly small but finite) guarantees that such a solution exists. The Lyapunov majorant is constructed by the technique of outer matrix products and using the local perturbation bounds derived in the previous step.

   In most cases the majorant equation $\rho = h(\delta,\rho)$ is, or can be reduced to, an algebraic equation of degree $d \geq 1$. If $d \leq 2$, then the root $\rho_0 = f(\delta)$ can easily be found in explicit form. When $d > 2$ there are two approaches to solve the problem. First, for a given δ the majorant equation may be solved numerically and ρ_0 may be defined as the smallest root for this case. If there are no positive solutions then δ is too 'large' and the method of Lyapunov majorants does not produce non-local perturbation bounds. This also may indicate that the perturbed equation has no solutions δX vanishing together with δP. If, however, there are positive roots it may not be clear whether the computed root ρ_0 is on a continuous path f with $f(0) = 0$. To avoid such situations a second approach is used. Here $h(\delta,\rho)$ is majorized by a new Lyapunov majorant $\widehat{h}(\delta,\rho)$ for which the equation $\rho = \widehat{h}(\delta,\rho)$ has a convenient closed form solution $\widehat{\rho}_0 = \widehat{f}(\delta)$ with $\widehat{f}$ continuous and $\widehat{f}(0) = 0$. This guarantees that the initial majorant equation has a solution $\rho_0 \leq \widehat{f}(\delta)$.

5. *Topological fixed point principles and non-local perturbation bounds.* If we have a small solution $f(\delta)$ of the majorant equation (or some of its upper bounds $\widehat{f}(\delta)$), then the fixed point principles of Schauder (or Brouwer) and Banach can be used to prove that

the equivalent vector equation has a 'small' solution x in the central, closed ball of radius $f(\delta)$. In view of the identity $\|\delta X\|_{\rm F} = \|x\|_2$ this gives the non-local estimate $\|\delta X\|_{\rm F} \le f(\delta) \le \widehat{f}(\delta)$, $\delta \in \Omega$.

5 Construction of an equivalent operator equation

5.1 Analysis of the perturbed equation

We may rewrite (5) as an equivalent matrix equation for the perturbation δX. This may not be a routine task even for quadratic equations. For example, the expression $F(P + \delta P, X_0 + \delta X)$ for equation (2) (or (3)) contains 50 terms, since a product of l perturbed matrices produces 2^l terms. Even more subtle is the case of fractional-affine equations such as the discrete-time Riccati equation. However, some terms in the perturbed equation may be augmented using the technique of outer matrix products presented later on.

The construction of the equivalent equation is based on the following scheme. For real equations and for some complex equations the function F is Fréchet differentiable in all its arguments at the point (P, X_0). We recall that a matrix valued function $\Phi : \mathcal{X} \to \mathbb{F}^{n\times n}$, where $\mathcal{X}$ is an open subset of $\mathbb{F}^{n\times n}$, is Fréchet differentiable at the point X_0 if there is a linear operator $\Phi_X(X_0) : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ such that $\Phi(X_0 + Z) = \Phi(X_0) + \Phi_X(X_0)(Z) + o(\|Z\|)$, $Z \to 0$. If Φ is a function of several matrix arguments, then we may define the partial Fréchet derivative in any of these arguments.

If F is Fréchet differentiable in P and X at (P, X_0), then we have

$$\begin{aligned} F(P + \delta P, X_0 + Z) &= F(P, X_0) + F_X(P, X_0)(Z) + F_P(P, X_0)(\delta P) + G(P, X_0)(\delta P, Z) \\ &= F_X(P, X_0)(Z) + F_P(P, X_0)(\delta P) + G(P, X_0)(\delta P, Z), \qquad (6) \end{aligned}$$

where $F_X(P, X_0) : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ is the partial Fréchet derivative of F in X calculated at the point (P, X_0). Similarly,

$$F_P(P, X_0)(\delta P) = \sum_{i=1}^{k} F_{P_i}(P, X_0)(\delta P_i),$$

where $F_{P_i}(P, X_0) : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ is the partial Fréchet derivative of F in P_i at (P, X_0).

Example 5.1 For the descriptor equation (2) we have $F(P, X) = Q + A^*XE + E^*XA - E^*XSXE$, $P = (Q, E, A, S)$, and, for $X_0 = X_0^*$,

$$F_X(P, X_0)(Z) = A_0^* ZE + E^* ZA_0, \ A_0 := A - SX_0E. \qquad (7)$$

We also have $F_Q(P, X_0)(Z) = Z$ and $F_S(P, X_0)(Z) = -E^*X_0ZX_0E$. In the real case it holds that

$$F_E(P, X_0)(Z) = A_0^\top X_0 Z + Z^\top X_0 A_0, \ F_A(P, X_0)(Z) = E^\top X_0 Z + Z^\top X_0 E.$$

For the associated equation (3) we have $F(P, X) = Q + A^*XE + E^*X^*A - E^*X^*SXE$. Then, for certain X_0 (which may not be symmetric), it holds that $F_Q(P, X_0)(Z) = Z$ and $F_S(P, X_0)(Z) = -E^*X_0^*ZX_0E$. If the associated equation is real then $F_X(P, X_0)(Z) = A_0^\top ZE + E^\top Z^\top A_0$ and

$$F_E(P, X_0)(Z) = A_0^\top X_0 Z + Z^\top X_0 A_0, \ F_A(P, X_0)(Z) = E^\top X_0^\top Z + Z^\top X_0 E,$$

where $A_0 := A - SX_0E$.

In the complex case similarily defined operators $F_{P_i}(P, X_0)$ (and even $F_X(P, X_0)$) may not be Fréchet derivatives, since F may not be Fréchet differentiable in some of its matrix arguments. In this case they are non-linear additive operators [38] constructed as follows. Suppose that $F(P, X)$ is written in the form $H(P_1, \overline{P}_1, \ldots, P_k, \overline{P}_k, X)$ and, for X_0 fixed, consider the function

$$(Y_1, Z_1, \ldots, Y_k, Z_k) \mapsto H(Y_1, Z_1, \ldots, Y_k, Z_k, X_0).$$

Suppose that the partial Fréchet derivatives $H_{Y_i}(P, X_0)$ and $H_{Z_i}(P, X_0)$ of this function in Y_i and Z_i, respectively, exist. Then we set $F_{P_i}(P, X_0)(\delta P) := H_{Y_i}(P, X_0)(\delta P_i) + H_{Z_i}(P, X_0)(\delta \overline{P}_i)$ and

$$F_P(P, X_0)(\delta P) = \sum_{i=1}^{k} \left(H_{Y_i}(P, X_0)(\delta P_i) + H_{Z_i}(P, X_0)(\delta \overline{P}_i) \right).$$

The operator $F_P(P, X_0) : \mathbb{C}^{n\times n} \times \cdots \times \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$, called a *Fréchet pseudo derivative* [39], is additive in the sense that $F_P(P + \widetilde{P}, X_0) = F_P(P, X_0) + F_P(\widetilde{P}, X_0)$ but it is not homogeneous, since $F_P(\lambda P, X_0) \neq \lambda F_P(P, X_0)$ for $\lambda \in \mathbb{C}\backslash\mathbb{R}$, see also [38].

Similarly, suppose that $F(P, X)$ is not Fréchet differentiable in X at (P, X_0) but $F(P, X) = D(P, X, \overline{X})$, where $D(P, X, Y)$ is Fréchet differentiable in X and Y at $(P, X_0, \overline{X}_0)$. Then we may set $F_X(P, X_0)(Z) = D_X(P, X_0, \overline{X}_0)(Z) + D_Y(P, X_0, \overline{X}_0)(\overline{Z})$. The operator $F_X(P, X_0) : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is again a Fréchet pseudo derivative. The general form of such an additive operator is $Z \mapsto \sum_i (M_i Z N_i + R_i \overline{Z} T_i)$ or $Z \mapsto \sum_i (M_i Z N_i + R_i Z^{\mathrm{H}} T_i)$.

Example 5.2 For the complex descriptor equation (2) with $X_0 = X_0^{\mathrm{H}}$ we have

$$F_E(P, X_0)(Z) = A_0^{\mathrm{H}} X_0 Z + Z^{\mathrm{H}} X_0 A_0,\ F_A(P, X_0)(Z) = E^{\mathrm{H}} X_0 Z + Z^{\mathrm{H}} X_0 E.$$

In turn, for the complex associated equation (3) it holds that

$$F_X(P, X_0)(Z) = A_0^{\mathrm{H}} Z E + E^{\mathrm{H}} Z^{\mathrm{H}} A_0,\ A_0 := A - S X_0 E \tag{8}$$

and

$$F_E(P, X_0)(Z) = A_0^{\mathrm{H}} X_0 Z + Z^{\mathrm{H}} X_0 A_0,\ F_A(P, X_0)(Z) = E^{\mathrm{H}} X_0^{\mathrm{H}} Z + Z^{\mathrm{H}} X_0 E.$$

The term

$$\begin{aligned} G(P, X_0)(\delta P, Z) &:= F(P + \delta P, X_0 + Z) - F(P, X_0) - F_X(P, X_0)(Z) - F_P(P, X_0)(\delta P) \\ &= F(P + \delta P, X_0 + Z) - F_X(P, X_0)(Z) - F_P(P, X_0)(\delta P) \end{aligned}$$

in (6) contains higher order terms in δP, Z,

$$\|(G(P, X_0)(\delta P, Z)\| = o(\|\delta P\| + \|Z\|),\ \|\delta P\| + \|Z\| \to 0.$$

This quantity can also be represented as

$$G(P, X_0)(\delta P, Z) = G_1(P, X_0)(\delta P) + G_2(P, X_0)(\delta P, Z),$$

where

$$\begin{aligned} G_1(P, X_0)(\delta P) &:= G(P, X_0)(\delta P, 0) = F(P + \delta P, X_0) - F_P(P, X_0)(\delta P) \\ &= o(\|\delta P\|),\ \delta P \to 0, \\ G_2(P, X_0)(\delta P, Z) &:= G(P, X_0)(\delta P, Z) - G_1(P, X_0)(\delta P) \\ &= F(P + \delta P, X_0 + Z) - F(P + \delta P, X_0) - F_X(P, X_0)(Z) \\ &= o(\|\delta P\| + \|Z\|),\ \|\delta P\| + \|Z\| \to 0. \end{aligned}$$

In the following we shall abbreviate $F_Z(P, X_0)$, $G(P, X_0)$ and $G_i(P, X_0)$ as F_Z^0, G^0 and G_i^0, respectively, omitting the dependence on the fixed quantities P, X_0 whenever appropriate.

Example 5.3 Setting $\widetilde{X} := X_0 + Z$, $\widetilde{S} := S + \delta S$ in the descriptor equation (2) we have

$$\begin{aligned} G^0(\delta P, Z) \quad = \quad & A_0^* Z \delta E + \delta E^* Z A_0 + E^* Z \delta A + \delta A^* Z E + \delta A^* \widetilde{X} \delta E + \delta E^* \widetilde{X} \delta A \\ & - E^* X_0 \delta S \widetilde{X} \delta E - \delta E^* \widetilde{X} \delta S X_0 E - E^* Z \widetilde{S} \widetilde{X} \delta E - \delta E^* \widetilde{X} \widetilde{S} Z E \\ & - E^* Z \widetilde{S} Z E - E^* X_0 \delta S Z E - E^* Z \delta S X_0 E - \delta E^* \widetilde{X} \widetilde{S} \widetilde{X} \delta E \end{aligned}$$

and

$$G_1^0(\delta P) = \delta A^* X_0 \delta E + \delta E^* X_0 \delta A - E^* X_0 \delta S X_0 \delta E - \delta E^* X_0 \delta S X_0 E - \delta E^* X_0 \widetilde{S} X_0 \delta E.$$

For the associated equation (3) the expressions for G^0 and G_1^0 are similar.

5.2 Equivalent matrix equation

Suppose that F is Fréchet differentiable in X at (P, X_0) and that the operator $\mathcal{L}_0 := F_X^0$ is invertible, i.e., its matrix representation $L_0 \in \mathbb{F}^{n^2 \times n^2}$ is invertible. Then we may rewrite the perturbed equation (5) as an equivalent matrix equation

$$\delta X = \Pi(\delta P, \delta X) := -\mathcal{L}_0^{-1} \circ F_P^0(\delta P) - \mathcal{L}_0^{-1} \circ G^0(\delta P, \delta X). \tag{9}$$

If, in the complex case, F is not Fréchet differentiable but $\mathcal{L}_0 := F_X^0 : \mathbb{C}^{n\times n} \to \mathbb{C}^{n\times n}$ is an additive operator, then an associated $n^2 \times n^2$ complex matrix of $\mathcal{L}_0$ is not defined. Instead, one can use the real $2n^2 \times 2n^2$ matrix Λ_0 of the real vectorized embedding of $\mathcal{L}_0$ (see [39, 38]), which is already a linear operator $\mathbb{R}^{2n^2} \to \mathbb{R}^{2n^2}$. For this purpose we recall some facts about non-linear additive operators.

Let $V = V_0 + \imath V_1$, $W = W_0 + \imath W_1$ be complex $m \times n$ matrices with V_0, V_1, W_0, W_1 real, and $z = z_0 + \imath z_1$ be a complex n–vector with z_0, z_1 real. Then we have

$$\max\{\|Vz + W\overline{z}\|_2 : \|z\|_2 \le 1\} = \|\Theta(V, W)\|_2,$$

where

$$\Theta(V, W) := \begin{bmatrix} V_0 + W_0 & W_1 - V_1 \\ V_1 + W_1 & V_0 - W_0 \end{bmatrix} \in \mathbb{R}^{2m \times 2n}. \tag{10}$$

We define the *real embedding* of the vector $z \in \mathbb{C}^n$ as $z^{\mathbb{R}} := \left[z_0^\top, z_1^\top\right]^\top \in \mathbb{R}^{2n}$. This gives $(Vz)^{\mathbb{R}} := V^{\mathbb{R}} z^{\mathbb{R}} \in \mathbb{R}^{2m}$, where $V^{\mathbb{R}} := \begin{bmatrix} V_0 & -V_1 \\ V_1 & V_0 \end{bmatrix} \in \mathbb{R}^{2m \times 2n}$ is the *real embedding* of the complex matrix V. Note that $(Vz + W\overline{z})^{\mathbb{R}} = \Theta(V, W) z^{\mathbb{R}}$ and $\Theta(V, 0) = V^{\mathbb{R}}$. Thus $\Theta(V, W)$ is the matrix of the real embedding of the complex additive operator $z \mapsto Vz + W\overline{z}$.

Returning to the case of complex pseudo differentiable functions F, one may require that the matrix of the real embedding of the vectorized complex additive operator $\mathcal{L}_0$ is invertible. Here the equivalent matrix equation is constructed by preliminarily taking the real embeddings of the corresponding complex quantities. For some equations such as (3) the operator $\mathcal{L}_0$ is not invertible but its restriction to a certain subset $\mathcal{C}$ of $\mathbb{C}^{n\times n}$ is invertible. In this case only perturbations δP which guarantee that $-F(P + \delta P, X_0 + Z) \in \mathcal{C}$ for all $Z \in \mathcal{C}$ are considered, see the next example as well as [39] for more details.

Example 5.4 Let $X_0 = X_0^*$ be a solution of the descriptor equation (2). Then the linear operator $\mathcal{L}_0 := F_X^0 : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$, defined by (7), has a matrix representation

$$L_0 := E^\top \otimes A_0^* + A_0^\top \otimes E^* \in \mathbb{F}^{n^2\times n^2}, \text{ where } A_0 := A - SX_0E.$$

Thus, see [14], the operator $\mathcal{L}_0$ is invertible if and only if the matrix $(A - SX_0E)E^{-1} = AE^{-1} - SX_0$ has no eigenvalues of opposite sign, i.e., $\lambda_i(AE^{-1} - SX_0) + \lambda_j(AE^{-1} - SX_0) \neq 0$, $i, j = 1, 2, \ldots, n$. In particular if the system (1) is regular then there is a (unique) solution $X_0 \geq 0$ such that the matrix $AE^{-1} - SX_0$ is stable and hence the operator $\mathcal{L}_0$ is invertible [46].

For the associated desciptor equation (3) with a possibly non-symmetric solution X_0, the operator F_X^0, defined by (8), is not invertible, since $F_X(Z) = (F_X(Z))^*$. However, its restriction to the set $\mathcal{C}$ of symmetric matrices is invertible if and only if the matrix A_0 is non-singular [39] (we recall that E is already assumed non-singular).

Note that $F_P^0(0) = 0$ and $G^0(0, 0) = 0$. This guarantees that for small δP equation (9) has a solution $\delta X = \Xi(\delta P)$ such that $\Xi(0) = 0$. Moreover, under certain additional conditions this solution is 'small' in the sense that $\delta X = -\mathcal{L}_0^{-1} \circ F_P^0(\delta P) + o(\|\delta P\|) = O(\|\delta P\|)$, $\delta P \to 0$.

We make use of the following version of the implicit function theorem [47, 7].

Theorem 5.1 *Let the function F be continuous in an open neighbourhood D of (P, X_0) and let the operator F_X^0 be invertible. Then there is an open neighbourhood D_1 of the point $\delta P = 0$ such that the following assertions hold.*

1. *For $\delta P \in D_1$ the perturbed equation has a solution $\delta X = \Xi(\delta P)$, where $\Xi : D_1 \to \mathbb{F}^{n\times n}$ is a continuous function such that $\Xi(0) = 0$.*

2. *If F is of class $C^k(D)$, $k \in \{\mathbb{N}, +\infty\}$, then the function Ξ is of class $C^k(D_1)$. If F is analytic on D then Ξ is analytic on D_1.*

3. *If $\mathbb{F} = \mathbb{C}$ and the function F is not Fréchet differentiable in either P and/or X but its real and imaginary parts are of class C^k (or analytic) as functions of the real and imaginary parts of P and X, then the real and imaginary parts of δX are also of class C^k (or analytic) as functions of the real and imaginary parts of δP.*

Example 5.5 Consider the scalar version $Q + (A\overline{E} + \overline{A}E)X - S|E|^2X^2 = 0$ of the complex equation (2) for $Q = E = S = 1$ and $A = 0$ and its solution $X_0 = 1$. Let $\delta A = a + \imath b \in \mathbb{C}$. Then $\delta X = \Xi(\delta A) := a + \sqrt{a^2 + 1} - 1$. We see that Ξ is not differentiable as a function $\mathbb{C} \to \mathbb{C}$ but its real part (Ξ itself) is analytic as a function of a, b.

In some cases the operator F_X^0 exists and is invertible but for some (or all) i the function F is neither Fréchet differentiable nor pseudo differentiable in P_i. Then assertion 1 of Theorem 5.1 still guarantees that the perturbed equation (5) has a solution δX which depends continuously on the data δP and vanishes at $\delta P = 0$. Here the equivalent operator equation is taken in the form

$$\delta X = \Gamma(\delta P, \delta X) := -\mathcal{L}_0^{-1} \circ F(P + \delta P, X_0) - \mathcal{L}_0^{-1} \circ E^0(\delta P, \delta X), \quad (11)$$

where $E^0(\delta P, Z) := F(P + \delta P, X_0 + Z) - F_X^0(Z) - F(P + \delta P, X_0)$. In this case we have $E^0(0, 0) = 0$ and $E^0(\delta P, Z) = o(\|Z\|)$, $Z \to 0$.

In the following we assume that the function F is continuous in an open neighbourhood of (P, X_0) and that the operator F_X^0 is invertible.

5.3 Equivalent vector equation

It is convenient to rewrite the equivalent matrix equation (9) in vector form. Taking the vec operation on both sides of (9) we obtain the vector equation $x = \pi(p,x)$, with

$$x := \mathrm{vec}(\delta X) \in \mathbb{F}^{n^2},\ p := \mathrm{vec}(\delta P) = \left[p_1^\top, \dots, p_k^\top\right]^\top \in \mathbb{F}^{kn^2},\ p_i := \mathrm{vec}(\delta P_i) \in \mathbb{F}^{n^2}, \quad (12)$$

and $\pi(p,x) := \mathrm{vec}(\Pi(\mathrm{vec}^{-1}(p), \mathrm{vec}^{-1}(x)))$ (of course, x in (12) is not the one from (1)). Note also that the inverse vec operations are well-defined, since the integers n and k are known. We have

$$\pi(p,x) = \pi_{10}(p) + \pi_{20}(p) + \pi_2(p,x), \quad (13)$$

where

$$\begin{aligned}
\pi_{10}(p) &:= -\mathrm{vec}\left(\mathcal{L}_0^{-1} \circ F_P^0(\mathrm{vec}^{-1}(p))\right) = -L_0^{-1}\mathrm{vec}\left(F_P^0(\mathrm{vec}^{-1}(p))\right), \quad (14)\\
\pi_{20}(p) &:= -\mathrm{vec}\left(\mathcal{L}_0^{-1} \circ G_1^0(\mathrm{vec}^{-1}(p))\right) = -L_0^{-1}\mathrm{vec}\left(G_1^0(\mathrm{vec}^{-1}(p))\right),\\
\pi_2(p,x) &:= -\mathrm{vec}\left(\mathcal{L}_0^{-1} \circ G_2^0(\mathrm{vec}^{-1}(p), \mathrm{vec}^{-1}(x))\right) = -L_0^{-1}\mathrm{vec}\left(G_2^0(\mathrm{vec}^{-1}(p), \mathrm{vec}^{-1}(x))\right).
\end{aligned}$$

In case of algebraic equations the equivalent vector equation (12) may be written in several different forms using Kronecker products of matrices [4, 17, 20, 57]. In doing so the formulae

$$\begin{aligned}
\mathrm{vec}(AYB) &= (B^\top \otimes A)\mathrm{vec}(Y),\\
(A \otimes B)(Y \otimes Z) &= (AY) \otimes (BZ), \quad (15)\\
(B \otimes A)\Pi_{n^2} &= \Pi_{n^2}(A \otimes B) \quad (16)
\end{aligned}$$

are frequently used, where all standard matrix products are well-defined and the third formula is valid for $A, B \in \mathbb{F}^{n\times n}$. We also use the relations

$$(A \otimes B)^{\mathrm{H}} = A^{\mathrm{H}} \otimes B^{\mathrm{H}},\ \|A \otimes B\|_2 = \|A\|_2\|B\|_2,\ (A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \quad (17)$$

(in the third relation A and B are assumed to be non-singular), as well as

$$\|AB\|_{\mathrm{F}} \le \min\{\|A\|_2\|B\|_{\mathrm{F}}, \|A\|_{\mathrm{F}}\|B\|_2\}. \quad (18)$$

In order to obtain tight perturbation bounds we use the basic relations (16), (17), (18) in combination with the technique of *outer matrix products*, presented below.

Suppose that we have a product $R = R_1 \cdots R_l$ of l matrices R_i which are either the nominal values P_j of the matrix parameters in (4) or their perturbations δP_s and that there is at least one term δP_s. Let in addition $\mathcal{L} : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ be an invertible linear matrix operator with a matrix representation $L \in \mathbb{F}^{n^2\times n^2}$. Then the problem that we are interested in is to vectorize $\mathcal{L}^{-1}(R)$ and to estimate $\|\mathcal{L}^{-1}(R)\|_{\mathrm{F}}$.

Suppose that the first l_1 and the last l_2 factors in R are not perturbations, while the factors R_{l_1+1} and R_{l-l_2} are perturbations. If the first (last) factor is already a perturbation we formally set $l_1 = 0$ ($l_2 = 0$). Then we can use the outer non-perturbed factors $R_1 \cdots R_{l_1}$ and $R_{l-l_2+1} \cdots R_l$ in order to obtain

$$\mathrm{vec}(\mathcal{L}^{-1}(R)) = L^{-1}\left((R_{l-l_2+1} \cdots R_l)^\top \otimes (R_1 \cdots R_{l_1})\right)\mathrm{vec}(R_{l_1+1} \cdots R_{l-l_2})$$

and

$$\|\mathcal{L}^{-1}(R)\|_{\mathrm{F}} = \|L^{-1}\mathrm{vec}(R)\|_2 \le \left\|L^{-1}\left((R_{l-l_2+1} \cdots R_l)^\top \otimes (R_1 \cdots R_{l_1})\right)\right\|_2 \|R_{l_1+1} \cdots R_{l-l_2}\|_{\mathrm{F}}. \quad (19)$$

Example 5.6 If $R = P_1P_3(\delta P_6)(\delta P_4)P_5P_7(\delta P_2)$ then $l = 7$, $l_1 = 2$, $l_2 = 0$ and

$$\mathrm{vec}(\mathcal{L}^{-1}(R)) = L^{-1}(I_n \otimes P_1P_3)\mathrm{vec}(R_3 \cdots R_7) = L^{-1}(I_n \otimes P_1P_3)\mathrm{vec}(\delta P_6 \delta P_4 P_5 P_7 \delta P_2).$$

Suppose further that the product $R_{l_1+1} \cdots R_{l-l_2}$ contains r unperturbed factors $S_1, \ldots, S_r$ and s perturbations $R_{j_1} = \delta P_{k_1}, \ldots, R_{j_s} = \delta P_{k_s}$, where any two unperturbed factors are separated by a product of perturbed factors. Then

$$\|R_{l_1+1} \cdots R_{l-l_2}\|_{\mathrm{F}} \leq \prod_{\alpha=1}^{r} \|S_\alpha\|_2 \prod_{\beta=1}^{s} \|\delta P_{k_\beta}\|_{\mathrm{F}}.$$

Substituting this bound in (19) we finally get

$$\|\mathcal{L}^{-1}(R)\|_{\mathrm{F}} \leq \left\| L^{-1}\left((R_{l-l_2+1} \cdots R_l)^\top \otimes (R_1 \cdots R_{l_1}) \right) \right\|_2 \prod_{\alpha=1}^{r} \|S_\alpha\|_2 \prod_{\beta=1}^{s} \|\delta P_{k_\beta}\|_{\mathrm{F}}.$$

Example 5.7 In Example 5.6 we have $S_1 = P_5P_7$, $j_1 = 3$, $k_1 = 6$, $j_2 = 4$, $k_2 = 4$, $j_3 = 7$, $k_3 = 2$ and $R_3 = \delta P_6$, $R_4 = \delta P_4$, $R_7 = \delta P_2$. Hence $\|R_3 \cdots R_7\|_{\mathrm{F}} \leq \|P_5\|_2 \|P_7\|_2 \|\delta P_2\|_{\mathrm{F}} \|\delta P_4\|_{\mathrm{F}} \|\delta P_6\|_{\mathrm{F}}$ and

$$\|\mathcal{L}^{-1}(R)\|_{\mathrm{F}} \leq \left\| L^{-1}(I_n \otimes P_1P_3) \right\|_2 \|P_5\|_2 \|P_7\|_2 \|\delta P_2\|_{\mathrm{F}} \|\delta P_4\|_{\mathrm{F}} \|\delta P_6\|_{\mathrm{F}}.$$

Example 5.8 Consider equation (2) with a symmetric solution X_0. In the real case the terms $\pi_{10}(p)$ and $\pi_{20}(p)$ in (13) are

$$\begin{aligned}
\pi_{10}(p) &= M_1p_1 + M_2p_2 + M_3p_3 + M_4p_4, \qquad (20)\\
\pi_{20}(p) &= M_{2,3}\mathrm{vec}\left(\delta A^\top X_0 \delta E\right) + M_{2,4}\mathrm{vec}(\delta S X_0 \delta E) + M_{2,2,4}\mathrm{vec}\left(\delta E^\top X_0 \delta S X_0 \delta E\right),
\end{aligned}$$

where $p_1 := \mathrm{vec}(\delta Q)$, $p_2 := \mathrm{vec}(\delta E)$, $p_3 := \mathrm{vec}(\delta A)$, $p_4 := \mathrm{vec}(\delta S)$ and

$$\begin{aligned}
M_1 &:= -L_0^{-1},\ M_2 := -L_0^{-1}(I_{n^2} + \Pi_{n^2})\left(I_n \otimes A_0^\top\right), \qquad (21)\\
M_3 &:= -L_0^{-1}(I_{n^2} + \Pi_{n^2})\left(I_n \otimes X_0E\right),\ M_4 := L_0^{-1}\left(E^\top X_0 \otimes E^\top X_0\right),\\
M_{2,3} &:= -L_0^{-1}(I_{n^2} + \Pi_{n^2}),\ M_{2,4} := M_3,\ M_{2,2,4} := M_1.
\end{aligned}$$

In the complex case we have $L_0 = E^\top \otimes A_0^{\mathrm{H}} + A_0^\top \otimes E^{\mathrm{H}}$, $A_0 := A - SX_0E$ and

$$\begin{aligned}
\pi_{10}(p) &= N_1p_1 + N_{2,0}p_2 + N_{2,1}\overline{p}_2 + N_{3,0}p_3 + N_{3,1}\overline{p}_3 + N_4p_4, \qquad (22)\\
\pi_{20}(p) &= N_{2,3}\left(\mathrm{vec}\left(\delta A^* X_0 \delta E\right) + \mathrm{vec}\left(\delta E^* X_0 \delta A\right)\right)\\
&\quad + N_{2,4,0}\mathrm{vec}(\delta S X_0 \delta E) + N_{2,4,1}\mathrm{vec}(\overline{\delta S X_0 \delta E}) + N_{2,2,4}\mathrm{vec}\left(\delta E^{\mathrm{H}} X_0 \delta S X_0 \delta E\right),
\end{aligned}$$

where

$$\begin{aligned}
N_1 &:= -L_0^{-1},\ N_{2,0} := -L_0^{-1}\left(I_n \otimes A_0^{\mathrm{H}}\right),\ N_{2,1} := -L_0^{-1}\left(A_0^\top \otimes I_n\right)\Pi_{n^2}, \qquad (23)\\
N_{3,0} &:= -L_0^{-1}\left(I_n \otimes E^* X_0\right),\ N_{3,1} := -L_0^{-1}\Pi_{n^2}\left(I_n \otimes E^\top X_0^\top\right),\\
N_4 &:= L_0^{-1}\left(E^\top X_0^\top \otimes E^* X_0\right),\ N_{2,3} := N_1,\ N_{2,4,0} := -L_0^{-1}\left(I_n \otimes E^* X_0\right),\\
N_{2,4,1} &:= -L_0^{-1}\left(I_n \otimes X_0 E\right),\ N_{2,2,4} := N_1.
\end{aligned}$$

Before we continue with the general framework we include two sections on norms.

6 Norms of linear combinations of vectors

Consider the vector $\pi_{10}(p)$ (or $\pi_{20}(p)$), defined by (13), (14), which contains the first order terms $O(\|p\|)$, $p \to 0$, in p in the expression $\pi(p,x)$. If all Fréchet derivatives $F^0_{P_i} := F_{P_i}(P, X_0)$ exist as in (20), then we have

$$\pi_{1,0}(p) = M_1 p_1 + \cdots + M_k p_k,\ p_i := \mathrm{vec}(\delta P_i), \tag{24}$$

where M_i is the matrix of the linear operator $-\mathcal{L}_0^{-1} \circ F^0_{P_i}$. Hence $M_i = L_0^{-1} F_i$, where $F_i \in \mathbb{F}^{n\times n}$ is the matrix of $F^0_{P_i}$. The problem that arises is to estimate $\|\pi_{10}(p)\|_2$ under the constraints $\|p_i\|_2 \le \delta_i$.

When in the complex case some (or all) of the Fréchet derivatives $F^0_{P_i}$ do not exist but the function F is pseudo Fréchet differentiable in P as in (22), then we have

$$\pi_{1,0}(p) = M_{1,0} p_1 + M_{1,1} \overline{p}_1 + \cdots + M_{k,0} p_k + M_{k,1} \overline{p}_k. \tag{25}$$

Here $M_{i,0}$ and $M_{i,1}$ are the matrices of the linear operators $-\mathcal{L}_0^{-1} \circ H^0_{Y_i}$ and $-\mathcal{L}_0 \circ H^0_{Z_i}$, where $H^0_{Y_i}$ and $H^0_{Z_i}$ are defined in Section 5.1. Here again the task is to estimate $\|\pi_{1,0}(p)\|_2$ for $\|p_i\|_2 \le \delta_i$. In this case we can pass to the real embedding of (25),

$$\pi^{\mathbb{R}}_{1,0}(p) = \sum_{i=1}^{k} \Theta(M_{i,0}, M_{i,1}) p_i^{\mathbb{R}}, \tag{26}$$

see Section 5.2 and (10). Thus, it suffices to consider linear combinations of the form

$$z = \sum_{i=1}^{k} C_i y_i \in \mathbb{F}^n, \tag{27}$$

where $C_i \in \mathbb{F}^{n\times n_i}$ and $y_i \in \mathbb{F}^{n_i}$. The problem then is to find (or to estimate) the supremum of $\|z\|_2$ for y_i satisfying $\|y_i\|_2 \le \delta_i$, $i = 1, \ldots, k$ (only the non-trivial case $k > 1$ will be considered). Since the vectors y_i vary over closed bounded sets in $\mathbb{F}^{n_i}$, the supremum is in fact a maximum and is achievable for a certain collection of vectors y_i. Denote this maximum as

$$\eta(C_1, \ldots, C_k; \delta_1, \ldots, \delta_k) := \max\left\{ \left\| \sum_{i=1}^{k} C_i y_i \right\|_2 : \|y_i\|_2 \le \delta_i, i = 1, \ldots, k \right\}. \tag{28}$$

Setting $C := [C_1, \ldots, C_k] \in \mathbb{F}^{n\times\nu}$, $\nu = n_1 + \cdots + n_k$, and $\delta = [\delta_1, \ldots, \delta_k]^\top$, we also use the shorter expression $\eta(C;\delta)$ for $\eta(C_1, \ldots, C_k; \delta_1, \ldots, \delta_k)$. Furthermore, we assume that all matrices C_i are non-zero and all elements of δ are positive, since this can always be achieved by removing, if necessary, some of the vectors $C_i y_i$ from the linear combination z. We also note that the problem of finding or estimating $\eta(C;\delta)$ can be reduced to the corresponding problem with $\delta_1 = \cdots = \delta_k = 1$. Indeed, for a k–vector α with positive elements α_i set $D_\alpha := \mathrm{diag}(\alpha_1, \ldots, \alpha_k) \in \mathbb{R}_+^{k\times k}$ and $\Delta_\alpha := \mathrm{diag}(\alpha_1 I_{n_1}, \ldots, \alpha_k I_{n_k}) \in \mathbb{R}_+^{\nu\times\nu}$. Then

$$\eta(C\Delta_\alpha; D_\alpha^{-1}\delta) = \eta(C;\delta) \tag{29}$$

and, setting $\alpha = \delta$ we get $\eta(C;\delta) = \eta(C\Delta_\delta; 1, \ldots, 1)$. However, we prefer to denote the dependence of η on C and δ as $\eta(C;\delta)$. Alternatively, one may assume that $\|C_1\|_2 =$

$\cdots = \|C_k\|_2 = 1$ setting $\alpha_i = 1/\|C_i\|_2$, $i = 1, \ldots, k$. Then $\eta(C;\delta) = \eta(\widetilde{C};\widetilde{\delta})$, where $\widetilde{C}_i := C_i/\|C_i\|_2$, $\widetilde{\delta}_i = \|C_i\|_2\delta_i$ and $\|\widetilde{C}_i\|_2 = 1$.

The determination of $\eta(C;\delta)$ is a difficult task, the more so in explicit form. It is easy to show that the function η is continuous and homogeneous in the sense that $\eta(\gamma C,\delta) = |\gamma|\eta(C,\delta)$, $\gamma \in \mathbb{F}$, and $\eta(C, c\delta) = c\eta(C,\delta)$, $c \geq 0$. However, in general it is not differentiable neither in C nor in δ. The next example gives an idea how $\eta(C;\delta)$ may look.

Example 6.1 Consider the simplest non-trivial case $k = 2$, $n_1 = n_2 = 1$, i.e., $z = C_1y_1 + C_2y_2$, where C_i are unit vectors and y_i are scalars satisfying $|y_i| \leq \delta_i$. Denoting by $s \in [-1,1]$ the real part of $C_1^*C_2$, we have $\|z\|_2^2 = y_1^2 + 2sy_1y_2 + y_2^2$. It is immediately clear that, e.g.,

$$\eta(C;\delta) = \eta_0(s;\delta) := \sqrt{\delta_1^2 + 2s\delta_1\delta_2 + \delta_2^2} \tag{30}$$

if $s \geq 0$, and $\eta(C;\delta) = \max\{\delta_1,\delta_2\}$ if $s = -1$. A more detailed analysis shows that

$$\eta(C;\delta) = \eta_*(s;\delta) := \begin{cases} \max\{\delta_1,\delta_2\} & \text{if } 2s + \varepsilon(\delta) \leq 0, \\ \eta_0(s;\delta) & \text{if } 2s + \varepsilon(\delta) > 0, \end{cases} \tag{31}$$

where $\varepsilon(\delta) := \min\{\delta_1,\delta_2\}/\max\{\delta_1,\delta_2\} \in (0,1]$. The function $\eta_* : [-1,1] \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$, defined by (30) and (31), is continuous but not differentiable. However, this function is analytic for $2s + \varepsilon(\delta) > 0$ as well as for $2s + \varepsilon(\delta) < 0$, $\delta_1 \neq \delta_2$. Similar explicit results are valid for $k = 3$ and $n_1 = n_2 = n_3 = 1$.

Since the determination of $\eta(C;\delta)$ in closed form is a hopeless task in general, the problem is to find good upper boundsfor this quantity.

Definition 6.1 *A bound* $\mathrm{est}(C;\delta)$ *for* $\eta(C;\delta)$ *in (28) is said to be* asymptotically exact *if there is a vector* δ^+ *with positive elements such that* $\mathrm{est}(C;\delta^+) = \eta(C;\delta^+)$.

We see that if a bound $\mathrm{est}(C;\delta)$ is asymptotically exact in the sense of Definition 6.1, then it is equal to the estimated quantity $\eta(C;\delta)$ for any $\delta = a\delta^+$, $a > 0$.

The simplest upper bound for $\eta(C;\delta)$ is

$$\mathrm{est}_1(C;\delta) := \sum_{i=1}^{k} c_i\delta_i,\ c_i := \|C_i\|_2. \tag{32}$$

The application of this approach to the perturbation analysis of equation (4) gives the widely used condition number based bounds with c_i being the *absolute individual condition number* with respect to perturbations in the data matrix P_i. We can also write

$$\mathrm{est}_1(C;\delta) = c^\top\delta,\ c := [c_1, \ldots, c_k]^\top, \tag{33}$$

where $c \in \mathbb{R}_+^k$ can be regarded as an *absolute vector condition measure*. These bounds, however, are not always tight enough. In particular it is not clear (except for simple linear equations) when such a bound is asymptotically exact.

Another upper bound for $\eta(C;\delta)$ is based on the representation $z = Cy$, where $y := \left[y_1^\top, \ldots, y_k^\top\right]^\top \in \mathbb{F}^\nu$, namely

$$\mathrm{est}_2(C;\delta) := \sigma\|\delta\|_2,\ \sigma := \|C\|_2. \tag{34}$$

Note that this bound is generically achievable and hence asymptotically exact. Let $v = \left[v_1^\top, \ldots, v_k^\top\right]^\top \in \mathbb{F}^\nu$, $v_i \in \mathbb{F}^{n_i}$, be the right singular vector of the matrix $C \in \mathbb{F}^{n\times\nu}$ corresponding to its largest singular value σ. Then for every $\lambda \in \mathbb{F}$ and $w = \lambda v \in \mathbb{F}^\nu$ we have $\|Cw\|_2 = |\lambda|\sigma$. Suppose now that all v_i are non-zero and $\delta_i = a\|v_i\|_2$ for some constant $a > 0$. Then $\eta(C;\delta) = \sigma\|\delta\|_2$, which in turn is the bound $\mathrm{est}_2(C;\delta)$.

The bounds $\mathrm{est}_1(C;\delta)$ and $\mathrm{est}_2(C;\delta)$ are *alternative* in the sense that both inequalities $\mathrm{est}_1(C;\delta) < \mathrm{est}_2(C;\delta)$ and $\mathrm{est}_1(C;\delta) \geq \mathrm{est}_2(C;\delta)$ are possible, see Example 6.2 below.

Example 6.2 Consider the expression $z = C_1y_1 + C_2y_2$ from Example 6.1. We have

$$\mathrm{est}_1(C;\delta) = \delta_1 + \delta_2,\ \mathrm{est}_2(C;\delta) = \sqrt{1+|s|}\sqrt{\delta_1^2+\delta_2^2}.$$

The two bounds are alternative with $\mathrm{est}_2(C;\delta) < \mathrm{est}_1(C;\delta)$ when $0 < |s| < 1$ and

$$\frac{1}{|s|} - \sqrt{\frac{1}{s^2}-1} < \frac{\delta_1}{\delta_2} < \frac{1}{|s|} + \sqrt{\frac{1}{s^2}-1}.$$

A simple calculation shows that

$$\frac{c_{\min}}{\sigma} < \frac{\mathrm{est}_1(C;\delta)}{\mathrm{est}_2(C;\delta)} \leq \frac{\|c\|_2}{\sigma}, \tag{35}$$

where $c_{\min} := c_j := \min\{c_1, \ldots, c_k\}$. The right inequality in (35) is achievable as equality (take $\delta = c$), while the left inequality may be achieved as approximate quality with an arbitrary accuracy. Indeed, take $\delta_j = 1$ and set the remaining $k-1$ elements of δ to be arbitrarily small.

Note finally that the bounds $\mathrm{est}_1(C;\delta)$ and $\mathrm{est}_2(C;\delta)$ are easily computable from the data C and δ.

An optimization procedure can be applied in order to get improved bounds of type est_2. Indeed, using (29) we have $\eta(C;\delta) \leq \psi(C;\delta;\gamma) := \left\|C\Delta_\gamma^{-1}\right\|_2 \|D_\gamma\delta\|_2$, where γ is a k–vector with positive elements γ_i. Now we can try to minimize $\psi(C;\delta;\gamma)$ over the set of all such vectors γ. This gives the estimate

$$\eta(C;\delta) \leq \mathrm{est}_2^0(C;\delta), \tag{36}$$

where

$$\mathrm{est}_2^0(C;\delta) := \inf\{\psi(C;\delta;\gamma) : \gamma_1, \ldots, \gamma_k > 0\}. \tag{37}$$

Note that $\mathrm{est}_2(C;\delta) = \psi(C;\delta;1,\ldots,1)$ and hence $\mathrm{est}_2^0(C;\delta) \leq \mathrm{est}_2(C;\delta)$.

The estimate (36) reduces the optimization problem for determining $\mathrm{est}(C;\delta)$, defined in $\mathbb{F}^\nu$, to an approximate optimization problem for determining $\mathrm{est}_2^0(C;\delta)$, defined in $\mathbb{R}^{k-1}$ (since one of the elements of γ may be fixed as 1, say $\gamma_1 = 1$). The use of the estimate (36) is justified by the following theorem.

Theorem 6.1 *There exists a vector $\gamma^0 = \left[1, \gamma_2^0, \ldots, \gamma_k^0\right]^\top$ with positive elements such that*

$$\mathrm{est}_2^0(C;\delta) = \psi(C;\delta;\gamma^0).$$

Proof. The idea of the proof is to show that in minimizing

$$\widetilde{\psi}(\gamma_2,\dots,\gamma_k) := \psi(C;\delta;1,\gamma_2,\dots,\gamma_k) = \|[C_1, C_2/\gamma_2,\dots,C_k/\gamma_k]\|_2 \sqrt{\delta_1^2+\gamma_2^2\delta_2^2+\cdots+\gamma_k^2\delta_k^2}$$

the quantities γ_i cannot be too large or too small. Setting $a := \|c\|_2\|\delta\|_2$ and noting that $\widetilde{\psi}(1,\dots,1) = \mathrm{est}_2(C;\delta) \le a$, we see that only values of γ_i such that $\widetilde{\psi}(\gamma_2,\dots,\gamma_k) \le a$ should be considered. On the other hand for every $i \in \{2,\dots,k\}$ it is fulfilled that $\widetilde{\psi}(\gamma_2,\dots,\gamma_k) \ge (c_i/\gamma_i)\delta_1$ and $\widetilde{\psi}(\gamma_2,\dots,\gamma_k) \ge c_1(\gamma_i\delta_i)$. Hence γ_i must satisfy $c_i\delta_1/\gamma_i \le a$ and $c_1\gamma_i\delta_i \le a$, or equivalently,

$$\frac{c_i\delta_1}{a} \le \gamma_i \le \frac{a}{c_1\delta_i}. \tag{38}$$

In view of the definition of a, we have $c_1c_i\delta_1\delta_i \le a^2$ and the inequalitites (38) are both achievable. Hence, γ_i must vary in closed bounded intervals, where, according to the Weierstrass theorem, the function $\widetilde{\psi}$ achieves its minimum. □

Theorem 6.1 shows that the infimum in (37) is in fact a minimum.

There is also a third easily computable bound for $\eta(C;\delta)$, which is obtained by noting that

$$\|z\|_2^2 = y^*C^*Cy = \sum_{i,j=1}^{k} y_i^*C_i^*C_jy_j.$$

Hence,

$$\|z\|_2^2 \le \sum_{i,j=1}^{k} \|C_i^*C_j\|_2\,\|y_i\|_2\|y_j\|_2 \le \sum_{i,j=1}^{k} \|C_i^*C_j\|_2\,\delta_i\delta_j,$$

which gives

$$\|z\|_2 \le \mathrm{est}_3(C,\delta) := \sqrt{\delta^\top C_0\delta}, \tag{39}$$

where $C_0 = [c_{ij}] \in \mathbb{R}_+^{k\times k}$ is the symmetric matrix with elements $c_{ij} := \|C_i^*C_j\|_2$, $i,j = 1,\dots,k$. Note that the matrix C_0 may be indefinite (i.e., it may also have some negative eigenvalues).

Example 6.3 For the problem of Example 6.1 we have $\mathrm{est}_3(C;\delta) = \sqrt{\delta_1^2+2|s|\delta_1\delta_2+\delta_2^2}$. This bound is obviously superior to $\mathrm{est}_1(C;\delta)$ and $\mathrm{est}_2(C;\delta)$. Moreover, here the bound $\mathrm{est}_3(C;\delta)$ is equal to the exact value of the estimated quantity $\eta(C;\delta)$ if $s \ge 0$. Furthermore, the bounds $\mathrm{est}_1(C;\delta)$ and $\mathrm{est}_3(C;\delta)$ are equal if $|s| = 1$. In turn, $\mathrm{est}_2(C;\delta) = \mathrm{est}_3(C;\delta)$ if $s = 0$, or if $s \ne 0$ but $\delta_1 = \delta_2$.

Theorem 6.2 *The relationship between the bounds* $\mathrm{est}_1, \mathrm{est}_2, \mathrm{est}_3$ *is given by*

$$\eta(C;\delta) \le \mathrm{est}(C;\delta), \tag{40}$$

where

$$\mathrm{est}(C;\delta) := \min\{\mathrm{est}_2(C;\delta), \mathrm{est}_3(C;\delta)\}. \tag{41}$$

Proof. The bound $\mathrm{est}_3(C,\delta)$ is always not worse than $\mathrm{est}_1(C,\delta)$. Indeed, we have $c_{ij} \le c_ic_j$ and hence

$$\mathrm{est}_3(C;\delta) \le \sqrt{\sum_{i,j=1}^{k} c_ic_j\delta_i\delta_j} = \sum_{i=1}^{k} c_i\delta_i = \mathrm{est}_1(C;\delta),$$

which completes the proof. $\square$

Theorem 6.2 gives an easily computable bound for the norm of the linear combination (27). Note that the function $\text{est}(C;\cdot) : \mathbb{R}_+^k \to \mathbb{R}$, defined in (41), is continuous but only piecewise differentiable.

The bounds $\text{est}_i(C,\delta)$ $i = 1,2,3$ can be compared quantitatively as follows. The matrix C_0 is non-negative with positive diagonal elements $c_{ii} = c_i^2$. Hence, according to the Perron–Frobenius theorem [14], its norm $\sigma_0 := \|C_0\|_2$ is equal to its largest (positive) eigenvalue. Moreover, there is a an eigenvector v_0 of C_0 with positive elements such that $C_0 v_0 = \sigma_0 v_0$, see [6]. Furthermore, we have $\sqrt{\gamma^\top C_0 \gamma} \geq c_{\min}\|\gamma\|_2$, $\gamma \in \mathbb{R}_+^k$, and equality is achievable (take $\gamma_j = 1$ and $\gamma_i = 0$ for $i \neq j$). Hence

$$c_{\min}\|\delta\|_2 < \text{est}_3(C;\delta) \leq \sqrt{\sigma_0}\|\delta\|_2.$$

It follows from these considerations that

$$\frac{c_{\min}}{\sigma} < \frac{\text{est}_3(C;\delta)}{\text{est}_2(C;\delta)} \leq \frac{\sqrt{\sigma_0}}{\sigma}. \tag{42}$$

The right inequality is achievable as equality (taking $\delta = v_0$), while the left inequality can be achieved as approximate equality with any prescribed accuracy as for the comparison (35). Since $c_{\min} < \sigma \leq \sqrt{\sigma_0}$ and the inequality $\sigma < \sqrt{\sigma_0}$ is possible, we see that the bounds $\text{est}_2(C;\delta)$ and $\text{est}_3(C;\delta)$ are alternatives, i.e., which one is better depends on the particular data C, δ.

Finally we have

$$1 \leq \frac{\text{est}_1(C;\delta)}{\text{est}_3(C;\delta)} < \frac{\|c\|_2}{c_{\min}}. \tag{43}$$

The left inequality is achievable as an equality (see Example 6.2). Unfortunately, we do not know whether the right inequality is sharp. Thus, the problem remains to estimate the maximum of $\text{est}_1(C;\delta)/\text{est}_3(C;\delta)$ over the vectors δ with positive elements. This maximum (depending only on C) is something between, $\|c\|_2^2/\sqrt{c^\top C_0 c}$ and $\|c\|_2/c_{\min}$.

7 Norms of Lyapunov operators

In this section we briefly describe Lyapunov matrix operators and their norms. More details can be found in [33].

For a linear operator $\mathcal{L} : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ define its Frobenius–norm by

$$\|\mathcal{L}\|_{\mathrm{F}} = \max\{\|\mathcal{L}(Z)\|_{\mathrm{F}} : \|Z\|_{\mathrm{F}} = 1\} = \|L\|_2,$$

where $L \in \mathbb{F}^{n^2\times n^2}$ is the matrix representation of $\mathcal{L}$. Then for every $Z \in \mathbb{F}^{n\times n}$ the inequality $\|\mathcal{L}(Z)\|_{\mathrm{F}} \leq \|\mathcal{L}\|_{\mathrm{F}}\|Z\|_{\mathrm{F}}$ holds. If, however, the operator $\mathcal{L}$ has certain symmetry properties and acts on a certain subset of $\mathbb{F}^{n\times n}$, then another norm may give better estimates for the quantity $\|\mathcal{L}(Z)\|_{\mathrm{F}}$.

Definition 7.1 *A linear matrix operator $\mathcal{L} : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ is called a* Lyapunov operator *if $\mathcal{L}^*(Z) = \mathcal{L}(Z^*)$ for all $Z \in \mathbb{F}^{n\times n}$.*

Taking the vec operation on both sides of the identity $\mathcal{L}^*(Z) = \mathcal{L}(Z^*)$ we get $\Pi_{n^2}\overline{L}\text{vec}(\overline{Z}) = L\Pi_{n^2}\text{vec}(\overline{Z})$. Hence we have the following simple characterization of the set of Lyapunov operators in $\mathbb{F}^{n\times n}$.

Proposition 7.1 *The linear matrix operator $\mathcal{L} : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ is a Lyapunov operator if and only if its matrix L satisfies $\Pi_{n^2}\overline{L} = L\Pi_{n^2}$.*

We also have the following useful result.

Theorem 7.1 *The following assertions hold.*

1. *The composition $\mathcal{L}\circ\mathcal{M}$ of two Lyapunov operators $\mathcal{L}, \mathcal{M} : \mathbb{F}^{n\times n} \to \mathbb{F}^{n\times n}$ is a Lyapunov operator.*
2. *If $\mathcal{L}$ is an invertible Lyapunov operator, then its inverse $\mathcal{L}^{-1}$ is again a Lyapunov operator.*

Proof. 1. First we note that $\mathrm{vec}(\mathcal{L}\circ\mathcal{M}(Z)) = L\mathrm{vec}(\mathcal{M}(Z)) = LM\mathrm{vec}(Z)$ and hence the matrix of the composition $\mathcal{L}\circ\mathcal{M}$ of the linear operators $\mathcal{L}$ and $\mathcal{M}$ is LM. Now, if $\mathcal{L}$ and $\mathcal{M}$ are Lyapunov operators, then we have $\Pi_{n^2}(\overline{LM}) = (\Pi_{n^2}\overline{L})\overline{M} = (L\Pi_{n^2})\overline{M} = LM\Pi_{n^2}$ which proves that $\mathcal{L}\circ\mathcal{M}$ is a Lyapunov operator.

2. For an invertible linear operator $\mathcal{L}$ the matrix of its inverse $\mathcal{L}^{-1}$ is L^{-1}. If in addition $\mathcal{L}$ is a Lyapunov operator we can multiply the equality $\Pi_{n^2}\overline{L} = L\Pi_{n^2}$ on the left by L^{-1} and on the right by $\overline{L}^{-1}$. The result $L^{-1}\Pi_{n^2} = \Pi_{n^2}\overline{L}^{-1}$ shows that $\mathcal{L}^{-1}$ is indeed a Lyapunov operator. □

In other words, the set of Lyapunov operators is a semi-group, and the set of invertible Lyapunov operators is a group relative to the composition law.

The set of Lyapunov operators in $\mathbb{F}^{2\times 2}$ is described in the next example.

Example 7.1 Based on Proposition 7.1 we see that the matrix L of a complex Lyapunov operator $\mathcal{L} : \mathbb{C}^{2\times 2} \to \mathbb{C}^{2\times 2}$ is

$$L = \begin{bmatrix} a_1 & a_4 + \imath b_2 & a_4 - \imath b_2 & a_8 \\ a_2 + \imath b_1 & a_5 + \imath b_3 & a_6 - \imath b_4 & a_9 + \imath b_6 \\ a_2 - \imath b_1 & a_6 + \imath b_4 & a_5 - \imath b_3 & a_9 - \imath b_6 \\ a_3 & a_7 + \imath b_5 & a_7 - \imath b_5 & a_{10} \end{bmatrix}, \tag{44}$$

where $a_1, \ldots, a_{10}, b_1, \ldots, b_6 \in \mathbb{R}$. Thus the set of such operators is isomorphic to $\mathbb{C}^6 \times \mathbb{R}^4 \simeq \mathbb{R}^{16}$. Similarly, the matrix L of a real Lyapunov operator $\mathcal{L} : \mathbb{R}^{2\times 2} \to \mathbb{R}^{2\times 2}$ is obtained from (44) setting $b_1 = \cdots = b_6 = 0$.

It is shown in [33] that the space of real Lyapunov operators is isomorhic to $\mathbb{R}^{n^2(n^2+1)/2}$, while the set of complex Lyapunov operators is isomorphic to $\mathbb{R}^{n^4}$.

Every Lyapunov operator $\mathcal{L}$ has a continuous-time representation $\mathcal{L}(Z) = \sum_{i=1}^{l_c} A_i^* Z B_i + B_i^* Z A_i$ as well as a discrete-time representation $\mathcal{L}(Z) = \sum_{j=1}^{l_d} \varepsilon_j C_j^* Z C_j$ with a minimum number of terms, where $A_i, B_i, C_j \in \mathbb{F}^{n\times n}$ and $\varepsilon_j = \pm 1$. We note that the sets of symmetric ($Z^* = Z$) and skew-symmetric ($Z^* = -Z$) matrices from $\mathbb{F}^{n\times n}$ are invariant under the action of a Lyapunov operator, see also [9]. This fact allows to define the *Lyapunov singular values* and *Lyapunov norm* of a Lyapunov operator [33].

Definition 7.2 *The quantity*

$$\|\mathcal{L}\|_{\mathrm{Lyap}} := \max\left\{\|\mathcal{L}(Z)\|_{\mathrm{F}} : \|Z\|_{\mathrm{F}} = 1, Z = Z^*\right\}$$

is called Lyapunov norm *of the Lyapunov operator $\mathcal{L}$.*

Explicit expressions for Lyapunov singular values and Lyapunov norm of a real or complex Lyapunov operator $\mathcal{L}$ in terms of its matrix L are given in [33]. Let $Q_n \in \mathbb{R}_+^{n^2\times n(n+1)/2}$ be any matrix such that $Q_n^\top Q_n = I_{n(n+1)/2}$ and $(P_{n_2} - I_{n^2})Q_n = 0$. Similarly, let $R_n \in \mathbb{R}_+^{n^2\times n(n-1)/2}$ be any matrix such that $R_n^\top R_n = I_{n(n-1)/2}$ and $J_{n^2}R_n = 0$, where $J_{n^2} = I_{n^2} + \Pi_{n^2}$. Note that in this case the matrix $[Q_n, R_n]$ is orthogonal.

Example 7.2 For $n = 2$ one can choose

$$Q_n = \begin{bmatrix} 1 & 0 & 0 \\ 0 & q & 0 \\ 0 & q & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad R_n = \begin{bmatrix} 0 \\ q \\ -q \\ 0 \end{bmatrix}, \quad q = 1/\sqrt{2}.$$

The Lyapunov norm of a Lyapunov operator can be computed according to the following theorem [33].

Theorem 7.2 *The Lyapunov norm of a Lyapunov operator $\mathcal{L}$ is*

$$\|\mathcal{L}\|_{\mathrm{Lyap}} = \|LQ_n\|_2 = \frac{1}{2}\|LJ_{n^2}\|_2$$

in the real case and

$$\|\mathcal{L}\|_{\mathrm{Lyap}} = \left\|L^{\mathbb{R}}\mathrm{diag}(Q_n, R_n)\right\|_2 = \frac{1}{2}\left\|L^{\mathbb{R}}\mathrm{diag}(J_{n^2}, I_{n^2} - \Pi_{n^2})\right\|_2$$

in the complex case.

We have $\|\mathcal{L}\|_{\mathrm{Lyap}} \le \|\mathcal{L}\|_{\mathrm{F}}$ and strict inequality is possible for some Lyapunov operators, see the next example.

Example 7.3 Let the matrix L of the Lyapunov operator $\mathcal{L} : \mathbb{R}^{2\times 2} \to \mathbb{R}^{2\times 2}$ be

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1+a & -a & 0 \\ 0 & -a & 1+a & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $a \ge 0$. The operator $\mathcal{L}$ is invertible ($\det(L) = 1+2a \ge 1$) and $\|\mathcal{L}\|_{\mathrm{F}} = 1+2a$. At the same time $\|\mathcal{L}\|_{\mathrm{Lyap}} = \|LQ_2\|_2 = \sqrt{2}$. Hence the ratio $\|\mathcal{L}\|_{\mathrm{Lyap}}/\|\mathcal{L}\|_{\mathrm{F}}$ can be made arbitrary small.

This observation shows that the use of Lyapunov norms may give better results when estimating the action of Lyapunov operators on sets of symmetric matrices.

8 Calculation of condition numbers

We continue with the general framework and derive in this section asymptotic condition numbers for general matrix equations in the sense proposed by Rice [50], see also [8]. With regard to the local perturbation analysis both real and complex equations are treated similarly.

Consider the perturbed equation (5). Under the assumptions of Theorem 5.1 there exists a vector $d \in \mathbb{R}^k_+$ with positive elements d_i such that for $\delta P \in B_d$, where

$$B_d := \{U = [U_1, \ldots, U_k] \in \mathbb{F}^{n\times kn} : \|U\| \preceq d\} \subset \mathbb{F}^{n\times kn},$$

the perturbed equation has a solution $\delta X = \Xi(\delta P)$. Here $\Xi : B_d \to \mathbb{F}^{n\times n}$ is a continuous function satisfying $\Xi(0) = 0$.

For fixed $i \in \{1, \ldots, k\}$ denote by $U_{(i)} = [U_1, \ldots, U_k] \in \mathbb{F}^{n\times kn}$ the matrix with blocks $U_j \in \mathbb{F}^{n\times n}$, such that $U_j = 0$ if $j \neq i$. Suppose that the function Ξ is Lipschitz continuous in δP_i in the sense that

$$l_i(P, X_0, d_i) := \sup \left\{ \frac{\|\Xi(U_{(i)})\|_{\mathrm{F}}}{\|U_i\|_{\mathrm{F}}} : 0 < \|U_i\|_{\mathrm{F}} \leq d_i \right\} < \infty.$$

Definition 8.1 *The quantity*

$$L_i = L_i(P, X_0) = \lim_{\alpha \to +0} l_i(P, X_0, \alpha) \tag{45}$$

is called the absolute condition number *for the solution X_0 with respect to perturbations in the matrix P_i.*

Note that the limit in (45) exists, since the function $l_i(P, X_0, \cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ is non-negative and non-increasing.

In view of the above considerations and having in mind (24) and (26), we obtain the following result for the condition numbers in the Frobenius–norm.

Theorem 8.1 *Consider the general matrix equation (4) and the perturbed matrix equation (5) and let i be a fixed integer from the set $\{1, \ldots, k\}$. Then the following assertions hold.*

1. *Suppose that the partial Fréchet derivative $F^0_{P_i}$ exist. Then the absolute condition number K_i of the solution X_0 with respect to perturbations in the matrix P_i is $\|M_i\|_2$, where M_i is the matrix of the linear operator $-\mathcal{L}_0^{-1} \circ F^0_{P_i}$.*

2. *Suppose that the partial Fréchet derivatives $H^0_{Y_i}$ and $H^0_{Z_i}$ exist, where*

$$F(P_1, \ldots, P_k, X_0) =: H(Y_1, Z_1, \ldots, Y_k, Z_k, X_0), \;\; Y_i := P_i, \;\; Z_i := \overline{P}_i.$$

Then the absolute condition number K_i of the solution X_0 with respect to perturbations in the matrix P_i is $\|\Theta(M_{i,0}, M_{i,1})\|_2$, where $M_{i,0}$ and $M_{i,1}$ are the matrices of the linear operators $-\mathcal{L}_0^{-1} \circ H^0_{Y_i}$ and $-\mathcal{L}_0^{-1} \circ H^0_{Y_i}$, respectively.

When $P_i \neq 0$ and $X_0 \neq 0$, then the *relative condition numbers* are $k_i := K_i \frac{\|P_i\|_{\mathrm{F}}}{\|X_0\|_{\mathrm{F}}}$ which are used in estimating the relative perturbation $\varepsilon_X := \delta_X / \|X_0\|_{\mathrm{F}}$ in the solution in terms of the relative perturbations $\varepsilon_i := \delta_i / \|P_i\|_{\mathrm{F}}$ in the coefficient matrices P_i.

If *only* one matrix P_i is perturbed, then we have $\delta_X \leq K_i \delta_i + o(\delta_i)$, $\delta_i \to 0$, and $\varepsilon_X \leq k_i \varepsilon_i + o(\varepsilon_i)$, $\varepsilon_i \to 0$. These condition number based estimates usually give good results.

Returning to the descriptor equation (2) we have the following corollaries of Theorem 8.1.

Corollary 8.1 *The absolute condition numbers for the solution X_0 of the real equation (2) with respect to the matrix coefficients $Z = Q, E, A, S$ are*

$$\begin{aligned} K_Q &= \|L_0^{-1}\|_2, \ K_E = \left\| L_0^{-1} J_{n^2} \left(I_n \otimes (A - SX_0E)^\top X_0 \right) \right\|_2, \\ K_A &= \left\| L_0^{-1} J_{n^2} \left(I_n \otimes E^\top X_0 \right) \right\|_2, \ K_S = \left\| L_0^{-1} \left(E^\top X_0 \otimes E^\top X_0 \right) \right\|_2, \end{aligned}$$

where $L_0 = E^\top \otimes A_0^\top + A_0^\top \otimes E^\top$ and $A_0 = A - SX_0S$. If in addition $\delta Q^\top = \delta Q$ and $\delta S^\top = \delta S$ then $K_Q = \|L_0^{-1}\|_{\mathrm{Lyap}}$ and $K_S = \left\| L_0^{-1} \left(E^\top X_0 \otimes E^\top X_0 \right) \right\|_{\mathrm{Lyap}}$.

Corollary 8.2 *The absolute condition numbers for the solution X_0 of the complex equation (2) with respect to the matrix coefficients $Z = Q, E, A, S$ are*

$$\begin{aligned} K_Q &= \|L_0^{-1}\|_2, \ K_E = \|\Theta(N_{20}, N_{21})\|_2, \\ K_A &= \|\Theta(N_{30}, N_{31})\|_2, \ K_S = \left\| L_0^{-1} \left(E^\top \overline{X}_0 \otimes E^{\mathrm{H}} X_0 \right) \right\|_2, \end{aligned}$$

where $L_0 = E^\top \otimes A_0^{\mathrm{H}} + A_0^\top \otimes E^{\mathrm{H}}$, $A_0 = A - SX_0S$ and the matrices N_{ij} are given in (23). If in addition $\delta Q^{\mathrm{H}} = \delta Q$ and $\delta S^{\mathrm{H}} = \delta S$, then $K_Q = \|L_0^{-1}\|_{\mathrm{Lyap}}$ and $K_S = \left\| L_0^{-1} \left(E^\top \overline{X}_0 \otimes E^{\mathrm{H}} X_0 \right) \right\|_{\mathrm{Lyap}}$.

Finally, we consider the case when the function $F(\cdot, X)$ is only Lipschitz continuous.

Suppose that the operator F_X^0 is invertible and the function $F(\cdot, X)$ is neither Fréchet nor pseudo Fréchet differentiable but is only Lipschitz continuous. Then we can again determine the condition numbers without solving the perturbed equation. Indeed, it follows from (11) and $F(P, X_0) = 0$ that

$$l_i(P, X_0, d_i) = \sup \left\{ \frac{\left\| \mathcal{L}_0^{-1} \circ F(P + U_{(i)}, X_0) \right\|_{\mathrm{F}}}{\|U_i\|_{\mathrm{F}}} : 0 < \|U_i\|_{\mathrm{F}} \leq d_i \right\} < \infty. \tag{46}$$

As a result for a non-differentiable (and non-pseudo differentiable) function $F(\cdot, X_0)$ we have the following result.

Theorem 8.2 *Consider the general matrix equation (4) and the perturbed matrix equation (5) and suppose that inequality (46) is valid. Then the quantity L_i, defined by (45), is the absolute condition number for the solution X_0 with respect to perturbations in the matrix P_i.*

Of course, Theorem 8.2 is also valid when the function $U_i \mapsto F(P + U_{(i)}, X_0)$ is Fréchet differentiable or Fréchet pseudo differentiable. In this case it reduces to Theorem 8.1.

9 Local perturbation bounds and overall measures of conditioning

Using the results from Section 6 we obtain the following result.

Theorem 9.1 *Consider the general matrix equation (4) and the perturbed matrix equation (5). Then the following local perturbation estimates are valid.*

1. *If the function F is Fréchet differentiable in P at (P, X_0) then*

$$\delta_X \leq \text{est}(M_1, \dots, M_k; \delta) + o(\|\delta\|), \ \delta \to 0. \tag{47}$$

2. *If in the complex case the function F is Fréchet pseudo differentiable in P at (P, X_0) then*

$$\delta_X \leq \text{est}(\Theta(M_{1,0}, M_{1,1}), \dots, \Theta(M_{k,0}, M_{k,1}); \delta) + o(\|\delta\|), \ \delta \to 0. \tag{48}$$

Corollary 9.1 *For the real descriptor equation (2) the local estimate*

$$\delta_X \leq \text{est}(M_1, \dots, M_4; \delta) + o(\|\delta\|), \ \delta \to 0,$$

is valid, where the matrices M_i are defined in (21).

For the complex descriptor equation (2) the local estimate

$$\delta_X \leq \text{est}(C_1, \dots, C_4; \delta) + o(\|\delta\|), \ \delta \to 0,$$

is valid, where $C_1 = N_1^{\mathbb{R}}$, $C_2 = \Theta(N_{2,0}, N_{2,1})$, $C_3 = \Theta(N_{3,0}, N_{3,1})$, $C_4 = N_4^{\mathbb{R}}$ and the matrices N_i, $N_{i,j}$ are defined in (23).

In many applications it is convenient to have a scalar measure of the relative conditioning of the problem. Such a measure may be derived in several ways. Suppose that the perturbations in the coefficient matrices P_i satisfy $\delta_i = \varepsilon\|P_i\|_{\text{F}}$ for some $\varepsilon > 0$ and $i = 1, \dots, k$. This will be, for example, the case when these perturbations are due to rounding of the data when storing it in computer memory and ε is (a small multiple of) the rounding unit. In this case $\|\delta P\| = \varepsilon\|P\|$.

Set

$$d_j(P, X_0) := \frac{\text{est}_j(N_1, \dots, N_k; \|P\|)}{\|X_0\|_{\text{F}}}, \ j = 1, 2, 3, \tag{49}$$

where N_i is M_i or $\Theta(M_{i,0}, M_{i,1})$. Then the quantities $\varepsilon d_j(P, X_0)$ are first order bounds for the relative perturbation ε_X in the solution X_0.

Definition 9.1 *The quantity $d(P, X_0) := \min\{d_2(P, X_0), d_3(P, X_0)\}$, with d_j defined by (49), is called the* overall relative condition measure *of the solution X_0.*

Usually the quantity $d_1(P, X_0)$ is used as an overall condition measure. However, since $d_3(P, X_0) \leq d_1(P, X_0)$, we see that the measure $d(P, X_0)$ generally gives better results.

Suppose that we have information about the size of the perturbations in the elements of the coefficient matrices δP_i. Then, if F is Fréchet differentiable it follows from (24) that the component-wise estimate

$$|x| \preceq \sum_{i=1}^{k} |M_i|\,|p_i| + o(\|\delta P\|), \ \delta P \to 0,$$

holds, where $p_i = \text{vec}(\delta P_i)$.

Similarly, if F in the complex case is Fréchet pseudo differentiable then (25) yields

$$|x| \preceq \sum_{i=1}^{k} (|M_{i,0}| + |M_{i,1}|)|p_i| + o(\|\delta P\|), \ \delta P \to 0.$$

Finally, if element-wise information for the real part $p_{i,0}$ and imaginary part $p_{i,1}$ of p_i is available, then we may use (26) to obtain

$$|x| \preceq \sum_{i=1}^{k} |\Theta(M_{i,0}, M_{i,1})| \left| p_i^{\mathbb{R}} \right| + o(\|\delta P\|), \ \delta P \to 0,$$

where $p_i^{\mathbb{R}} = \left[p_{i,0}^\top, p_{i,1}^\top \right]^\top \in \mathbb{R}_+^{2n^2}$ and the matrices $\Theta(M_{i,0}, M_{i,1})$ are given by (10).

The bounds given in Theorem 9.1 may be very tight. Suppose that F is Fréchet differentiable in P and that the vector p is proportional to the singular vector of the matrix $[M_1, \ldots, M_k]$, corresponding to its 2-norm. Then $\|\pi_{10}(p)\|_2$ is equal to $\mathrm{est}_2(M_1, \ldots, M_k; \delta)$ and hence to $\mathrm{est}(M_1, \ldots, M_k; \delta)$ within first order terms. The same argument is applicable when F is Fréchet pseudo differentiable in P.

10 Construction, analysis and solution of Lyapunov majorant equations

Lyapunov majorant functions (or, briefly, Lyapunov majorants) [18, 36] are a very useful tool in studying existence, uniqueness and perturbation problems for operator equations, including equations in abstract spaces. In this section we briefly describe the method of scalar Lyapunov majorants for operator equations in finite dimensional spaces. Lyapunov majorants are used to estimate the norm or the generalized norm of (linear or non-linear) operators, depending on parameters, which are small in a sense to be precisely defined later on. Using Lyapunov majorants, a so called majorant equation is constructed. Under certain conditions the solution of the majorant equation is a small quantity which is a bound for the solution of the equivalent operator equation. The method of Lyapunov majorants is applied in combination with topological fixed point principles such as the principles of Schauder (or Brouwer) and Banach, see [47, 7].

10.1 Definitions and properties

Consider the perturbed equation (5) together with its vectorized version (24). Then we can define the following bound $\varphi(\delta, \rho) := \max\{\|\pi(p, x)\|_2 : \|\!|p|\!\| \preceq \delta, \ \|x\| \le \rho\}$ for the norm of the vector $\pi(p, x)$. In certain cases $\varphi(\delta, \rho)$ is defined only for $\delta \preceq \delta^0$ and $\rho \le \rho^0$, or for all $\delta \in \mathbb{R}_+^k$ and $\rho < \psi(\delta)$, where ψ is a positive function. It follows from the continuity of π that φ is also continuous and, moreover, $\varphi(0, 0) = 0$.

For a given δ consider the equation $\rho = \varphi(\delta, \rho)$. Suppose that for δ sufficiently small there is a solution $\rho = \rho(\delta)$ of this equation, where $\rho(\cdot)$ is a continuous function, satisfying $\rho(0) = 0$. This also means that the operator $\pi(\delta, \cdot)$ transforms the ball $\mathcal{B}_{\rho(\delta)}$ of radius $\rho(\delta)$ into itself. Since $\pi(\delta, \cdot)$ is continuous and $\mathcal{B}_\rho$ is closed and convex, according to the Schauder fixed point principle, there is a solution of the operator equation $x = \pi(\delta, x)$ (in the finite dimensional case the application of the Brouwer principle suffices). Now the quantity $\rho(\delta)$ is a bound for $\|x\|_2$ and hence for $\|\delta X\|_{\mathrm{F}}$.

Unfortunately, the construction of φ in closed form is a hopeless task even for simple equations $F(P, X) = 0$. In fact, we cannot find an explicit expression even for the first order term $\pi_{10}(p)$ of $\pi(p, x)$, see also Section 9. The idea to apply fixed point principles for the solution of the perturbation problem can nevertheless be realized by using Lyapunov majorants.

Definition 10.1 *Consider the operator equation (12). A continuous function $h : \mathbb{R}_+^k \times \mathbb{R}_+ \to \mathbb{R}_+$ is called a* Lyapunov majorant *for the operator π if it satisfies the following conditions.*

1. *If $\|p\| \preceq \delta$ and $\|x\|_2 \leq \rho$ then $\|\pi(p,x)\|_2 \leq h(\delta,\rho)$.*

2. *The function h is non-decreasing in all of its arguments in the sense that $0 \preceq \delta \preceq \delta'$ and $\rho \leq \rho'$ imply $h(\delta,\rho) \leq h(\delta',\rho')$.*

3. *It is fulfilled that $h(0,0) = 0$ and*

$$h_1(\delta) := \lim_{\alpha \to +0} \sup \left\{ \frac{h(\delta,\rho) - h_0(\delta)}{\rho} : 0 < \rho \leq \alpha \right\} < 1, \tag{50}$$

   where $h_0(\delta) := h(\delta, 0)$.

Note that if h is differentiable in ρ at the point $(\delta, 0)$, then condition (50) reduces to $h_1(\delta) := h_\rho(\delta, 0) < 1$. In practice the quantity $h(\delta,\rho)$ is determined as a computable bound for $\varphi(\delta,\rho)$.

For a given Lyapunov majorant h we define the *majorant equation*

$$\rho = h(\delta,\rho). \tag{51}$$

The following theorem reveals the properties of the majorant equation.

Theorem 10.1 *Let h be a Lyapunov majorant. Then there exists a closed domain $\Omega \subset \mathbb{R}_+^k$, bounded by the non-negative coordinate $(1/k)$-planes and a certain surface $\mathcal{S} \subset \mathbb{R}_+^k$, with the following properties.*

1. *For $\delta \in \Omega \backslash \mathcal{S}$ the majorant equation (51) has a solution $\rho_0 = f(\delta)$, where $f : \Omega \backslash \mathcal{S} \to \mathbb{R}_+$ is a continuous non-negative and non-decreasing function such that $f(0) = 0$. For $\delta \notin \Omega$ this equation has no non-negative roots.*

2. *If for any non-zero $\delta \in \mathbb{R}_+^k$ the function $h(\delta, \cdot)$ is strictly convex, then the function f from 1. is defined for the whole domain Ω. Moreover, (i) for $\delta \in \Omega \backslash \mathcal{S}$ the majorant equation (51) has two non-negative roots, the smaller one being $f(\delta)$; (ii) for $\delta \in \mathcal{S}$ this equation has a double positive root $f(\delta)$; (iii) for $\delta \notin \Omega$ the equation has no non-negative roots.*

Proof. For a proof see [18, 36]. □

We are interested in the solution $\rho_0 = f(\delta)$ with f continuous and $f(0) = 0$. Since δ is usually small, in the sense that $\|\delta\|/\|P\| \ll 1$, then by a continuity argument $f(\delta)$ will also be small. We shall further on refer to $f(\delta)$ as the *small solution* of (51).

Having a Lyapunov majorant, the next step is to solve the majorant equation and in particular to find its small solution ρ_0 (whenever it exists). It would be desirable to do this by finding the dependence of ρ_0 in δ in an explicit form, but in general for a fixed δ the majorant equation has to be solved numerically. Then if it has two roots $0 \leq \widehat{\rho}_0 \leq \widehat{\rho}_1$ one chooses $\widehat{\rho}_0$ as a candidate for the small solution vanishing together with δ. This 'numerical' approach may or may not work. The problem is that it is not clear whether for the computed solution $\widehat{\rho}_0$ there is indeed a continuous function f with $\widehat{\rho}_0 \simeq f(\delta)$ and $f(0) = 0$ (i.e., whether a small solution exists). The next example shows that the numerical approach may be misleading.

Example 10.1 Let

$$h(\delta,\rho) = 6\delta + \frac{\delta\rho^2}{1-\rho}, \ \ \delta \in \mathbb{R}_+. \tag{52}$$

Then the majorant equation is formally equivalent to the quadratic equation

$$(1+\delta)\rho^2 - (1+6\delta)\rho + 6\delta = 0, \ \ \rho \neq 1.$$

For $\delta \leq (3-\sqrt{6})/6 \simeq 0.09175$ the small solution

$$\rho_0 = f(\delta) = \frac{12\delta}{1+6\delta+\sqrt{12\delta^2 - 12\delta + 1}}$$

is of order 6δ and indeed tends to zero with $\delta \to 0$. However, for $\delta \geq (3+\sqrt{6})/6 \simeq 0.90825$ the quadratic equation has roots which are not small because δ cannot be small. For example, if $\delta = 1$ then the roots are 1.5 and 1. Of course, the latter case should in fact be excluded from consideration, since $h(\delta,\rho)$ in (52) is defined only for $\rho < 1$. But in practice, cases like this may cause problems.

In many applications the expression $h(\delta,\rho)$ has the form

$$h(\delta,\rho) = g_1(\delta,\rho) + \frac{g_2(\delta,\rho)}{g(\delta) - g_3(\delta,\rho)},$$

where $g_i(\delta,\rho)$ are polynomials in ρ, $g_i(\delta,\rho) = \sum_{j=0}^{r_i} a_{i,j}(\delta)\rho^j$, $i = 1,2,3$. Here the coefficients $a_{i,j}(\delta)$ and $g(\delta)$ are polynomials in $\delta \in \mathbb{R}_+^k$ with non-negative coefficients, and $g(0) > a_{3,0}(0)$. Also, according to Definition 10.1 we must have $a_{1,0}(0) = a_{2,0}(0) = 0$ and $a_{1,1}(0) + a_{2,1}(0) < 1$. In this case $h(\delta,\rho)$ is well defined if $\delta \in \mathbb{R}_+^k$ and $\rho < \psi(\delta)$, where $\psi(\delta)$ is the smallest positive root of the algebraic equation $g(\delta) = g_3(\delta,\rho)$.

Furthermore, the majorant equation can be reduced to an algebraic equation of degree $r := \max\{r_2, r_1 + r_3, r_3 + 1\}$ in ρ, namely

$$d(\delta,\rho) := \sum_{j=0}^{r} d_j(\delta)\rho^j = 0, \ \ \rho < \psi(\delta). \tag{53}$$

Note that the coefficients $d_j(\delta)$ may not be non-negative and/or non-decreasing in δ.

Here the surface $\mathcal{S} \subset \mathbb{R}_+^k$ is defined by the equation $\Delta(\delta) = 0$, where $\Delta(\delta)$ is the discriminant of d. In this case, equation (53) (and hence the majorant equation) has a double non-negative root. The discriminant of d may be constructed by different schemes (whenever appropriate we omit the dependence of d and d_j on their arguments). Let $r \geq 2$ and consider the derivative $d_\rho(\delta,\rho) = \sum_{j=0}^{r-1}(j+1)d_{j+1}\rho^j$ of d in ρ which must be zero at the double root. Multiplying d by $\rho, \ldots, \rho^{r-2}$ and d_ρ by $\rho, \ldots, \rho^{r-1}$ in view of $d = 0$ and $d_\rho = 0$, we obtain $2r-1$ homogeneous linear equations in the quantities $1, \rho, \ldots, \rho^{2r-1}$, which can be written as a vector equation

$$T^{(r)}b^{(r)} = 0, \ T^{(r)} = \left[t_{ij}^{(r)}\right] \in \mathbb{R}_+^{(2r-1)\times(2r-1)}, \ b^{(r)} := [1, \rho, \ldots, \rho^{2r-1}]^\top \in \mathbb{R}_+^{2r-1}.$$

Here the elements $t_{ij}^{(r)}$ of T are given by

$$t_{ij}^{(r)} := \begin{cases} 0 & \text{if } 1 \leq i \leq r-1 & \text{and } j < i, \\ d_{j-i} & \text{if } 1 \leq i \leq r-1 & \text{and } i \leq j \leq r+i, \\ 0 & \text{if } 1 \leq i \leq r-1 & \text{and } j > r+i, \\ 0 & \text{if } r \leq i \leq 2r-1 & \text{and } j < i-r+1, \\ (j-i+r)d_{j-i+r} & \text{if } r \leq i \leq 2r-1 & \text{and } i-r+1 \leq j \leq i, \\ 0 & \text{if } r \leq i \leq 2r-1 & \text{and } j > i. \end{cases}$$

Example 10.2 For $r = 2$ and $r = 3$ the equations for $b^{(2)}$ and $b^{(3)}$ are

$$T^{(2)}b^{(2)} = \begin{bmatrix} d_0 & d_1 & d_2 \\ d_1 & 2d_2 & 0 \\ 0 & d_1 & 2d_2 \end{bmatrix} \begin{bmatrix} 1 \\ \rho \\ \rho^2 \end{bmatrix} = 0, \; T^{(3)}b^{(3)} = \begin{bmatrix} d_0 & d_1 & d_2 & d_3 & 0 \\ 0 & d_0 & d_1 & d_2 & d_3 \\ d_1 & 2d_2 & 3d_3 & 0 & 0 \\ 0 & d_1 & 2d_2 & 3d_3 & 0 \\ 0 & 0 & d_1 & 2d_2 & 3d_3 \end{bmatrix} \begin{bmatrix} 1 \\ \rho \\ \rho^2 \\ \rho^3 \\ \rho^4 \end{bmatrix} = 0. \tag{54}$$

The discriminant of d is $\Delta = \det(T^{(r)})$. Since $b^{(r)} \neq 0$, and having in mind that $T^{(r)}$ depends on δ, it follows that $\mathcal{S} = \left\{\delta \in \mathbb{R}_+^k : \det(T^{(r)}(\delta)) = 0\right\}$.

10.2 Polynomial Lyapunov majorants

In the important particular case of polynomial or pseudo polynomial [39] matrix equations $F(P, X) = 0$ the Lyapunov majorant is a polynomial in ρ,

$$h(\delta, \rho) = \sum_{j=0}^{r} a_j(\delta)\rho^j, \tag{55}$$

where a_i are continuous, non-negative and non-decreasing functions of $\delta \in \mathbb{R}_+^k$ and $a_r(\delta) > 0$ for some $\delta \in \mathbb{R}_+^k$. In fact, $a_i(\delta)$ are often polynomials in δ with non-negative coefficients. In this case, according to Definition 10.1, the conditions $a_0(0) = 0$ and $a_1(0) < 1$ are fulfilled (in most applications we even have $a_1(0) = 0$).

Considering the majorant equation

$$\rho = \sum_{j=0}^{r} a_j(\delta)\rho^j, \tag{56}$$

we can always solve this equation numerically for a given $\delta \in \mathbb{R}_+^k$. Let the computed solutions be $\widehat{\rho}_0 \leq \widehat{\rho}_1$. Then we can take $\widehat{\rho}_0$ as the small solution lying on a continuous path to zero. Despite of this possible approach, it is still convenient to have (approximate) closed form solutions. Next we consider techniques to construct such solutions.

We denote by $\Omega_r \subset \mathbb{R}_+^k$ the set of all δ such that equation (56) has a small solution ρ_0, denoted as $f_r(\delta)$, where the function f_r is continuous and $f_r(0) = 0$. Upper bounds for f_r, defined for $\delta \in \widehat{\Omega}_r$, are denoted as $\widehat{f}_r$. As we demonstrate below, Ω_1 is bounded but not closed, while for $r > 1$ the set Ω_r is compact. Obviously we have $f_{j+1}(\delta) \leq f_j(\delta)$ and $\Omega_{j+1} \subset \Omega_j$, $j = 1, 2, \ldots$.

The case $r = 1$. Here the function $h(\delta, \cdot)$ is not strictly convex. Equation (56) has a unique solution

$$f_1(\delta) := \frac{a_0(\delta)}{1 - a_1(\delta)}, \; \delta \in \Omega_1 \backslash \mathcal{S}_1,$$

where $\Omega_1 := \left\{\delta \in \mathbb{R}_+^k : a_1(\delta) \leq 1\right\}$ and $S_1 := \left\{\delta \in \mathbb{R}_+^k : a_1(\delta) = 1\right\}$. This case arises in studying linear algebraic equations.

The case $r = 2$. Here the function $h(\delta, \cdot)$ is strictly convex for some δ. The domain for δ is

$$\Omega_2 = \left\{\delta \in \mathbb{R}_+^k : a_1(\delta) + 2\sqrt{a_0(\delta)a_2(\delta)} \leq 1\right\}, \tag{57}$$

and the surface $\mathcal{S}_2 \subset \Omega_2$ is obtained by replacing the inequality in (57) by equality. For $\delta \in \Omega_2 \backslash \mathcal{S}_2$ the majorant equation has two roots, the smaller one being

$$f_2(\delta) := \frac{2a_0(\delta)}{1 - a_1(\delta) + \sqrt{(1 - a_1(\delta))^2 - 4a_0(\delta)a_2(\delta)}}.$$

For $\delta \in \mathcal{S}_2$ the majorant equation has a double root $f_2(\delta) = 2a_0(\delta)/(1 - a_1(\delta))$, $\delta \in \mathcal{S}_2$.

Similar results hold for the case when

$$h(\delta, \rho) = a_{10}(\delta) + a_{11}(\delta)\rho + \frac{a_{20} + a_{21}(\delta)\rho + a_{22}\rho^2}{g(\rho) - a_{31}(\delta)\rho}.$$

The case $r = 3$. Here the majorant equation is cubic. The surface $\mathcal{S}_3$ is obtained by $\det(T^{(3)}(\delta)) = 0$, where the matrix $T^{(3)}$ is defined by (54). For this case there are closed form solutions, given by the Cardano formula. But we are interested in the case when the equation has two non-negative solutions (and hence one negative solution as well). This is the so called irreducible case when the explicit form solution is not very practical. So we determine an approximate closed form solution.

Suppose that for a given δ such that $a_1(\delta) < 1$, equation (56) has two non-negative solutions. Suppose also that $a_3(\delta) > 0$, since otherwise the majorant equation is of order less than 3.

For the small solution $\rho_0 = f_3(\delta)$ it holds that $\rho_0 \le \tau_3$, where τ_3 is the unique solution of the equation $1 = h_\rho(\delta, \rho)$, i.e., $1 = a_1 + 2a_2\tau_3 + 3a_3\tau_3^2$. Hence

$$\tau_3 = \tau_3(\delta) = \frac{1 - a_1(\delta)}{a_2(\delta) + \sqrt{a_2^2(\delta) + 3a_3(\delta)(1 - a_1(\delta))}}.$$

Furthermore, for $\rho \le \tau_3$ we have

$$\rho \le a_0 + a_1\rho + (a_2 + a_3\tau_3)\rho^2. \tag{58}$$

The right hand side of (58) is again a Lyapunov majorant in the form of a second degree polynomial in ρ. So we can apply the estimates already obtained for $r = 2$ above. As a result we get the estimate

$$f_3(\delta) \le \widehat{f}_3(\delta) := \frac{2a_0(\delta)}{1 - a_1(\delta) + \sqrt{(1 - a_1(\delta))^2 - 4a_0(\delta)\widehat{a}_2(\delta)}}, \ \delta \in \widehat{\Omega}_3, \tag{59}$$

where

$$\widehat{\Omega}_3 = \left\{\delta \in \mathbb{R}_+^k : a_1(\delta) + 2\sqrt{a_0(\delta)\widehat{a}_2(\delta)} \le 1\right\} \tag{60}$$

and $\widehat{a}_2(\delta) := a_2(\delta) + a_3(\delta)\tau_3(\delta)$.

We recall that the estimate (59), (60) is valid under the assumption that the majorant equation $\rho = a_0 + a_1\rho + a_2\rho^2 + a_3\rho^3$ posesses non-negative solutions. And, of course, the inclusion $\delta \in \widehat{\Omega}_3$ by no means guarantees that such solutions exist (in general Ω_3 may be a proper subset of $\widehat{\Omega}_3$). Fortunately, here the existence of non-negative solutions is easily checked by the inequality

$$\widehat{f}_3(\delta) \le \tau_3(\delta), \tag{61}$$

involving already computed quantities. In particular the equality in (61) is equivalent to $\det(T^{(3)}(\delta)) = 0$ or $\delta \in \mathcal{S}_3$. More precisely, the following result holds.

Theorem 10.2 *For a cubic majorant equation the following assertions hold:*

1. *If (61) is fulfilled then the majorant equation has a small solution $f_3(\delta) \le \widehat{f}_3(\delta)$. If (61) is violated, then the majorant equation has no non-negative solutions.*

2. *The case of equality in (61) describes the surface $\mathcal{S}_3 \subset \mathbb{R}^k_+$ on which the discriminant of the majorant equation vanishes and this equation has a double non-negative root.*

Proof. The case of equality in (61) means that the quantity $\widehat{f}_3(\delta)$ satisfies both the majorant equation $\rho = h(\delta,\rho)$ and the equation $1 = h_\rho(\delta,\rho)$. Hence $\widehat{f}_3(\delta)$ is a double root. □

Note that inequality (61) is equivalent to $h(\delta,\tau_3(\delta)) \le \tau_3(\delta)$ as well as to $h_\rho(\delta,\tau_3(\delta)) \le 1$. If $h(\delta,\widehat{f}_3(\delta)) < \widehat{f}_3(\delta)$ then we can construct better approximations by the scheme

$$\rho^{(q+1)} = \frac{\rho^{(q)} a_0(\delta)}{\rho^{(q)} - h(\delta,\rho^{(q)}) + a_0(\delta)},\ q = 1,2,\ldots,$$

where $\rho^{(0)} = \widehat{f}_3(\delta)$.

Example 10.3 Consider the majorant equation $\rho = h(\delta,\rho) := \delta(1+\rho+\rho^2+\rho^3)$, where $\delta \ge 0$ is a scalar. Here the interval $[0,\mathcal{S}_3]$ for δ is easily obtained noting that $\mathcal{S}_3$ is the maximum of the expression $\rho/(1+\rho+\rho^2+\rho^3)$ in $\rho > 0$. This maximum is achieved for the positive root of the equation $2\rho^3+\rho^2-1=0$ and is $\mathcal{S}_3 \simeq 0.27695$. We have $\tau_3(\delta) = (1-\delta)/(\delta+\sqrt{3\delta-2\delta^2})$ and

$$\widehat{f}_3(\delta) = \frac{2\delta}{1-\delta+\sqrt{1-2\delta-\delta^2(3+\tau_3(\delta))}}.$$

The results for the exact small solution $f_3(\delta)$ and its bound $\widehat{f}_3(\delta)$ are given in Table 1. The cases when the solution does not exist are marked by a double asterisk. The bound does not exist in the case marked by asterisk. We see that the bound $\widehat{f}_3(\delta)$ is good whenever applicable, i.e., for $\delta \le \mathcal{S}_3$. But it is correct also for $\delta = 0.28$, although for this value of δ the majorant equation does not have a small solution.

Table 1: Solutions and bounds for a cubic majorant equation

δ	f_3	$\widehat{f}_3$
0.03000	0.03096	0.03105
0.09000	0.09999	0.10148
0.15000	0.18350	0.18968
0.21000	0.29601	0.31388
0.27000	0.52607	0.57302
$\mathcal{S}_3 \simeq 0.27695$	0.65730	0.65730
0.28000	**	0.74925
0.29000	**	*

The case $r > 3$. For $r = 4$ there is a closed form solution, which is not very suitable for practical implementation. For $r > 4$ in general there are no closed form solutions. That is why for $r > 3$ we shall construct closed form approximations for the small solution of the majorant equation as in the case $r = 3$.

Suppose that for a given δ such that $a_1(\delta) < 1$ and $a_r(\delta) > 0$ equation (56) has two non-negative solutions. For the small solution $\rho_0 = f_r(\delta)$ we have $\rho_0 \leq \tau_r$, where $\tau_r = \tau_r(\delta)$ is the unique solution of the equation $1 = h_\rho(\delta, \rho)$,

$$1 = \sum_{j=0}^{r-1} (j+1) a_{j+1} \tau_r^j. \tag{62}$$

This equation has a unique solution. Indeed, $1 > a_1 = h_\rho(\delta, 0)$. On the other hand for ρ sufficiently large (take $r a_r \rho^{r-1} > 1$) we have $1 < h_\rho(\delta, \rho)$. Hence, there is a solution τ_r of equation (62). That τ_r is unique follows from the fact that the function $h_\rho(\delta, \cdot)$ is increasing.

We have $\rho_0 \leq g(\delta, \tau_r(\delta), \rho_0) := a_0(\delta) + a_1(\delta)\rho_0 + a_2(\delta)\rho_0^2 + b(\delta, \tau_r(\delta))\rho_0^2$, where

$$b(\delta, \tau) := \sum_{j=2}^{r-1} a_{j+1}(\delta) \tau^{j-1}.$$

Here $\widehat{g}(\delta, \rho) := g(\delta, \tau_3(\delta), \rho)$ is a new Lyapunov majorant. Since for $r > 3$ there is no convenient closed form expression for τ_r we shall find an upper bound $\widehat{b}(\delta)$ for $b(\delta, \tau_r(\delta))$. It follows from (62) that $(j+1) a_{j+1} \tau_r^j \leq 1 - a_1$ and

$$\tau_r \leq \left(\frac{1 - a_1}{(j+1) a_{j+1}} \right)^{1/j}, \ j = 2, \ldots, r-1.$$

Hence

$$a_{j+1} \tau_r^{j-1} \leq \alpha_{j+1} := a_{j+1}^{1/j} \left(\frac{1 - a_1}{j+1} \right)^{1-1/j}, \ j = 2, \ldots, r-1$$

and

$$b(\delta, \tau_r(\delta)) \leq \widehat{b}(\delta) := \sum_{j=2}^{r-1} \alpha_{j+1}(\delta).$$

As a result we have $\rho \leq a_0(\delta) + a_1(\delta)\rho + (a_2(\delta) + \widehat{b}(\delta))\rho^2$ and

$$f_r(\delta) \leq \widehat{f}_r(\delta) := \frac{2 a_0(\delta)}{1 - a_1(\delta) + \sqrt{(1 - a_1(\delta))^2 - 4 a_0(\delta)(a_2(\delta) + \widehat{b}(\delta))}} \tag{63}$$

provided that

$$\delta \in \widehat{\Omega}_r := \left\{ \delta \in \mathbb{R}_+^k : a_1(\delta) + 2\sqrt{a_0(\delta)(a_2(\delta) + \widehat{b}(\delta))} \leq 1 \right\}. \tag{64}$$

Thus we have proved the following result.

Theorem 10.3 *Consider the majorant equation (56) for* $r > 3$. *If the inequality* $\widehat{f}_r(\delta) \leq \tau_r(\delta)$ *is fulfilled, then the majorant equation has a small solution* $f_r(\delta)$ *for which the estimate (63) holds.*

Table 2: Solutions and bounds for a cubic majorant equation

δ	f_3	φ_3
0.03000	0.03096	0.03106
0.09000	0.09999	0.10181
0.15000	0.18350	0.19190
0.21000	0.29601	0.32554
0.27000	0.52607	*

Example 10.4 The bound (63) is applicable for $r = 3$ as well (in this case we denote the bound as $\varphi_3(\delta)$) although it will give slightly worse results than the bound (59), (60). Consider again the majorant equation from Example 10.3. Here $\varphi_3(\delta)$ is the small solution of the equation $(\delta + \alpha_3(\delta))\rho^2 - (1-\delta)\rho + \delta = 0$, see Table 2. In the case marked by an asterisk the bound φ_3 does not exist.

Example 10.5 Consider the majorant equation $\rho = h(\delta, \rho) := \delta(1+\rho+\rho^2+\rho^3+\rho^4)$, where $\delta \geq 0$ is a scalar. The interval $[0, \mathcal{S}_4]$ for δ is obtained by noting that $\mathcal{S}_4$ is the maximum of $\rho/(1+\rho+\rho^2+\rho^3+\rho^4)$. This maximum is achieved for the positive root of the equation $2\rho^3+\rho^2-1=0$ and is $\mathcal{S}_3 \simeq 0.27695$. We have

$$\widehat{b}(\delta) = \alpha_3(\delta) + \alpha_4(\delta) = \delta^{1/2}\left(\frac{1-\delta}{3}\right)^{1/2} + \delta^{1/3}\left(\frac{1-\delta}{4}\right)^{2/3}$$

and

$$\widehat{f}_4(\delta) = \frac{2\delta}{1-\delta+\sqrt{(1-\delta)-4\delta(\delta+\widehat{b}(\delta))}}.$$

The results for the small solution $f_4(\delta)$ and its bound $\widehat{f}_4(\delta)$ are given in Table 3. The cases when the solution does not exist are marked by a double asterisk. The bound does not exist in the case marked by an asterisk. The bound $\widehat{f}_4(\delta)$ is satisfactory whenever applicable. We also see that the bound ceases to to exist before the critical value $\mathcal{S}_4$ for δ.

Table 3: Solutions and bounds for a quartic majorant equation

δ	f_4	$\widehat{f}_4$
0.02000	0.02042	0.02047
0.08000	0.08769	0.09004
0.14000	0.16831	0.18215
0.20000	0.27568	0.34254
0.26000	0.53064	*
$\mathcal{S}_4 \simeq 0.26079$	0.56774	*
0.26100	**	*

We conclude this subsection by justifying certain 'cheap' perturbation bounds. An interesting feature of these bounds is that while they are valid for any $r > 2$, only the first two or three terms $a_j\rho^j$ of h are taken into account explicitely. The influence of higher order terms is implicit by the requirement $\delta \in \Omega_r$.

Theorem 10.4 *Consider the majorant equation (56) for $r > 2$ and let $\delta \in \Omega_r \backslash \mathcal{S}_1$. Then*

$$f_r(\delta) \le b_2(\delta) \le b_1(\delta), \tag{65}$$

where

$$b_1(\delta) := \frac{2a_0}{1 - a_1(\delta)}, \; b_2(\delta) := \frac{3a_0(\delta)}{1 - a_1(\delta) + \sqrt{(1 - a_1(\delta))^2 - 3a_0(\delta)a_2(\delta)}}. \tag{66}$$

Proof. We note first that the relation $\delta \in \Omega_r \subset \Omega_2$ guarantees that $a_1 + 2\sqrt{a_0 a_2} \le 1$ and hence the quantities b_j are correctly defined by (66). Consider now the second estimate $f_r \le b_1$ in (65). Recall that τ_r satisfies (62). Setting $c_l(\delta, \rho) := a_l(\delta)\rho^{j-l} + \cdots + a_r(\delta)\rho^{r-l}$, where $l = 2, 3$, we see that $a_1(\delta) + 2\tau_r(\delta)c_2(\delta, \tau_r(\delta)) \le 1$ and hence $c_2(\delta, \tau_r(\delta)) \le (1 - a_1(\delta))/(2\tau_r(\delta))$. On the other hand for every $\rho \le f_r(\delta)$ we have

$$\rho \le a_0(\delta) + a_1(\delta)\rho + c_2(\delta, \tau_r(\delta))\rho^2 \le a_0(\delta) + a_1(\delta)\rho + (1 - a_1(\delta))\frac{\rho^2}{2\tau_r(\delta)}.$$

Since $\rho \le \tau_r(\delta)$, we get that $\rho \le a_0(\delta) + a_1(\delta)\rho + (1 - a_1(\delta))\rho/2$ and hence $\rho \le b_1(\delta)$. Now the first inequality in (65) follows, since ρ may be chosen as $f_r(\delta)$.

Consider next the first bound $f_r \le b_2$ in (65). We have

$$a_1(\delta) + 2a_2(\delta)\tau_r(\delta) + 3\tau_r^2(\delta)c_3(\delta, \tau_r(\delta)) \le 1$$

and hence $c_3(\delta, \tau_r(\delta)) \le (1 - a_1(\delta) - 2a_2(\delta)\tau_r(\delta))/(3\tau_r^2(\delta))$. For every $\rho \le f_r(\delta)$ it is fulfilled that

$$\begin{aligned} \rho &\le a_0(\delta) + a_1(\delta)\rho + a_2(\delta)\rho^2 + c_3(\delta, \tau_r(\delta))\rho^3 \\ &\le a_0(\delta) + a_1(\delta)\rho + a_2(\delta)\rho^2 + (1 - a_1(\delta) - 2a_2(\delta)\tau_r(\delta))\frac{\rho^3}{3\tau_r^2(\delta)}. \end{aligned}$$

Now $\rho \le \tau_r(\delta)$ yields

$$\rho \le a_0(\delta) + a_1(\delta)\rho + a_2(\delta)\rho^2 + (1 - a_1(\delta) - 2a_2(\delta)\rho)\rho/3$$

and $0 \le 3a_0(\delta) - 2(1 - a_1(\delta))\rho + a_2(\delta)\rho^2$. Thus $\rho \le b_2(\delta)$ for all $\rho \le f_r(\delta)$.

Finally the inequality $b_2(\delta) \le b_1(\delta)$ is verified by direct calculation. This completes the proof. ◻

Of course, in applying the cheap estimates (65) one has to check whether $\delta \in \Omega_r$. A sufficient condition for $f_r(\delta) \le b_i(\delta)$ to be valid is $h(\delta, b_i(\delta)) \le b_i(\delta)$.

We conclude the consideration of cheap bounds with the following remarks. For $\delta \to 0$ the small solution $f_r(\delta)$ is of asymptotic order $\alpha(\delta) + o(\|\delta\|)$, where $\alpha(\delta) := a_0(\delta)/(1 - a_1(0))$. At the same time the bound $b_2(\delta)$ is of order $\frac{3}{2}\alpha(\delta) + o(\|\delta\|)$, while $b_1(\delta)$ is of order $2\alpha(\delta) + o(\|\delta\|)$. We note finally that $b_1(\delta) = b_2(\delta)$ if and only if $\delta \in \mathcal{S}_2$, i.e., $a_1(\delta) + 2\sqrt{a_0(\delta)a_2(\delta)} = 1$.

11 Topological fixed point principles and non-local perturbation bounds

The local perturbation estimates developed in Section 9 (Theorem 9.1) are usually applied in the chopped form $\delta_X \le \mathrm{est}(N;\delta)$ (here $N_i = M_i$ or $N_i = \Theta(M_{i,0}, M_{i,1})$), simply neglecting the $o(\|\delta\|)$ terms in (47) or (48). At the same time the local estimates from Theorem 9.1 are valid for small perturbations but this theorem gives no quantitative indications about how small the quantity $\|\delta\|$ must be so that the chopped estimate $\delta_X \le \mathrm{est}(N;\delta)$ is valid, or at least not violated drastically. Note that in some problems we have $\delta_X > \mathrm{est}(N;\delta)$ for all $\delta \ne 0$ (see Example 12.1 from Section 12), so the chopped bound is violated.

It may even happen that the perturbed equation (5) has a finite escape $\|\delta X\| \to +\infty$ when δ approaches a certain set $\mathcal{R} \subset \mathbb{R}_+^k$, such that for $\delta \in \mathcal{R}$ the solution δX ceazes to exist.

Example 11.1 *Consider the scalar real equation (2) with $Q = E = S = 1$, $A = 0$, $X_0 = 1$ and $\delta Q = \delta A = \delta S = 0$, $\delta S = -\delta$, $\delta \in [0, 1)$. Then $\delta X = \delta/(1-\delta)$ while the local bound is* $\mathrm{est}(N;\delta) = \delta$. *We see that $\delta_X > \mathrm{est}(N;\delta)$ for $\delta > 0$. Moreover, δ_X tends to $+\infty$ when δ approaches $\mathcal{R} = 1$, while the chopped estimate is $\delta_X < 1$.*

It follows that chopped local estimates as well as condition numbers and overall measures of conditioning are useful as qualitative indicators for the sensitivity of the solution with respect to perturbations in the data. At the same time, they may not produce rigorous upper bounds for the actual perturbation in the solution.

To avoid the disadvantages of local bounds one can apply the methods of non-local perturbation analysis. As a result one gets non-local (and in general non-linear) perturbation bounds of the form

$$\delta_X \le f(\delta), \ \ \delta \in \Omega, \tag{67}$$

where Ω is a bounded set of positive measure in $\mathbb{R}_+^k$ and f is a continuous function with $f(0) = 0$. The inclusion $\delta \in \Omega$ guarantees that the perturbed equation (5) indeed has a solution δX for which the estimate (67) holds.

A desirable property of non-local bounds is that

$$f(\delta) = \mathrm{est}(\delta) + o(\|\delta\|), \ \ \delta \to 0. \tag{68}$$

This motivates the following concept of asymptotic exactness for non-local bounds. We recall that $\delta X = \Xi(\delta P)$, where Ξ is a continuous, matrix–valued function defined in an open neighbourhood of $\delta P = 0$ and satisfying $\Xi(0) = 0$.

Definition 11.1 *The bound $f : \Omega \to \mathbb{R}_+$ from (67) is said to be* asymptotically exact *if there exists a vector δ^+ with positive elements from the interior of Ω such that*

$$\|\Xi(\delta P)\|_{\mathrm{F}} = \varepsilon f(\delta^+) + o(\varepsilon), \ \ \varepsilon \to 0,$$

for all δP with $\| \! | \delta P | \! \| \preceq \varepsilon\delta^+$, where $\varepsilon > 0$ and $\varepsilon\delta^+ \in \Omega$.

We see that a bound $f(\delta)$ satisfying (68) is asymptotically exact, since so is the local bound $\mathrm{est}(\delta)$.

A possible disadvantage of a non-local bound is that its domain of applicability Ω may be small. This seems to be the price of having a rigorous perturbation bound.

The non-local perturbation analysis is based on the method of Lyapunov majorants [18, 36] and fixed point principles [31, 47, 7], see Section 10. The aim is to show that, for δ from a certain set Ω, the equivalent operator $\pi(p,\cdot)$ in (12) maps a closed convex set $\mathcal{B} \subset \mathbb{F}^{n^2}$ into itself. The set $\mathcal{B}$ is small, of diameter $f(\delta) = O(\|\delta\|)$, $\delta \to 0$. Then, according to the Schauder fixed point principle, there exists a solution $\xi \in \mathcal{B}$ of (12) and hence $\delta_X = \|\xi\|_2 \le f(\delta)$. It even turns out that for $\delta \in \Omega\backslash\mathcal{S}$, where $\mathcal{S}$ is a part of the boundary of Ω, the operator $\pi(p,\cdot)$ is a contraction and according to the Banach fixed point principle the solution to the perturbed equation is unique.

Having a Lyapunov majorant $h(\delta,\rho)$ for the operator $\pi(p,\cdot)$ the problem is to decide whether, for a given $\delta \in \mathbb{R}_+^k$, the majorant equation $\rho = h(\delta,\rho)$ has a small solution $\rho_0 = f(\delta)$ vanishing together with δ. The next problem is to find the small solution or to construct a good approximation $\widehat{f}(\delta) \ge f(\delta)$ for this solution. Then the non-local perturbation estimate becomes $\delta_X \le f(\delta)$ or $\delta_X \le \widehat{f}(\delta)$.

In the following we present the corresponding results for the descriptor equation (2). We have four possible cases depending on whether the equation is real or complex and the perturbations δQ, δS are symmetric or non-symmetric. Consider the case of a real equation and symmetric perturbations in Q, S. This is the case when $Q = C^\top C$ and $S = BB^\top$, or $S = B\mathrm{diag}(I,-I)B^\top$ in the sign-indefinite case.

Set $\delta = [\delta_1,\delta_2,\delta_3,\delta_4]^\top := [\delta_Q,\delta_E,\delta_A,\delta_S]^\top$. We have

$$\delta X = \Pi(\delta P,\delta X) := \Pi_{1,0}(\delta P) + \Pi_{2,0}(\delta P) + \Pi_{1,1}(\delta P,\delta X) + \Pi_{0,2}(\delta P,\delta X), \tag{69}$$

where $\Pi_{i,j}(\cdot) := \mathcal{L}_0^{-1}(U_{i,j}(\cdot))$ and $U_{i,0}(T) = O(\|T\|^i)$, $T \to 0$, $i = 1,2$,

$$U_{1,1}(T,Z) = O(\|T\|\,\|Z\|),\ T \to 0,\ Z \to 0;\ U_{0,2}(T,Z) = O(\|Z\|^2),\ Z \to 0.$$

Here

$$\begin{aligned} U_{1,0}(\delta P) \;&:=\; -\,\delta Q - \delta E^\top X_0 A_0 - A_0^\top X \delta E - \,\delta A^\top X_0 E - E^\top X_0 \delta A + E^\top X_0 \delta S X_0 E,\\ U_{2,0}(\delta P) \;&:=\; -\,\delta A^\top X_0 \delta E - \delta E^\top X_0 \delta A + \delta E^\top X_0 S X_0 \delta E\\ &\quad\;+\,\delta E^\top X_0 \delta S X_0 E + E^\top X_0 \delta S X_0 \delta E + \delta E^\top X_0 \delta S X_0 \delta E \end{aligned}$$

and

$$\begin{aligned} U_{1,1}(\delta P,Z) \;&:=\; -\,\delta E^\top Z A - A^\top Z \delta E - \delta A^\top Z E - E^\top Z \delta A\\ &\quad\;+\,E^\top (Z(S+\delta S)X_0 + X_0(S+\delta S)Z)\delta E\\ &\quad\;+\,\delta E^\top (Z(S+\delta S)X_0 + X_0(S+\delta S)Z)E\\ &\quad\;+\,E^\top (Z\delta S X_0 + X_0 \delta S Z)E + \delta E^\top (Z(S+\delta S)X_0 + X_0(S+\delta S)Z)\delta E,\\ U_{0,2}(\delta P,Z) \;&:=\; \delta E^\top Z(S+\delta S)ZE + E^\top Z(S+\delta S)Z\delta E\\ &\quad\;+\,E^\top Z(S+\delta S)ZE + \delta E^\top Z(S+\delta S)Z\delta E. \end{aligned}$$

Consider the vectorized form of (69)

$$x = \pi(p,x) := \pi_{1,0}(p) + \pi_{2,0}(p) + \pi_{1,1}(p,x) + \pi_{0,2}(p,x), \tag{70}$$

where $\pi_{i,j} = L_0^{-1}(\mathrm{vec}(U_{i,j}))$, see also Example 5.8. The quantity $\|\pi_{1,0}(p)\|_2$ has already been bounded from above by $\mathrm{est}(M;\delta)$. The vector $\pi_{2,0}(p)$ can be represented as $\pi_{2,0}(p) = \psi_{2,0}(p) + L_0^{-1}\mathrm{vec}(\delta E^\top X_0 \delta S X_0 \delta E)$, where

$$\begin{aligned} \psi_{2,0}(p) \;&=\; -L_0^{-1} J_{n^2}\mathrm{vec}(\delta A^\top X_0 \delta E) + L_0^{-1} J_{n^2}\left(I_n \otimes E^\top X_0\right)\mathrm{vec}(\delta S X_0 \delta E)\\ &\quad +\; L_0^{-1}\mathrm{vec}(\delta E^\top X_0 S X_0 \delta E). \end{aligned}$$

Hence, we have two possible estimates for $\|\psi_{2,0}(p)\|_2$. The first one is based on the observation that $\|L_0^{-1}J_{n^2}\|_2 = 2\|L_0^{-1}\|_{\mathrm{Lyap}}$ and that the matrix $\delta E^\top X_0 \delta S X_0 \delta E$ is symmetric,

$$\|\psi_{2,0}(p)\|_2 \leq e_1(\delta) := \delta_E \left\|L_0^{-1}\right\|_{\mathrm{Lyap}} (2\|X_0\|_2\delta_A + 2\|X_0E\|_2\delta_S + \|X_0SX_0\|_2\delta_E)\,. \tag{71}$$

The other estimate is based on the estimates from Section 6,

$$\|\psi_{2,0}(p)\|_2 \leq e_2(\delta) := \delta_E\, \mathrm{est}(T_2, T_3, T_4; \|X_0SX_0\|_2\delta_E, \|X_0\|_2\delta_A, \|X_0\|_2\delta_S), \tag{72}$$

where

$$T_2 := L_0^{-1},\ T_3 := -L_0^{-1}J_{n^2},\ T_3 := L_0^{-1}J_{n^2}\left(I_n \otimes E^\top X_0\right). \tag{73}$$

Setting $e(\delta) := \min\{e_1(\delta), e_2(\delta)\}$ we have $\|\psi_{2,0}(p)\|_2 \leq e(\delta) + \|L_0^{-1}\|_{\mathrm{Lyap}}\|X_0\|_2^2\delta_E^2\delta_S$ and $\|\pi_{1,0}(\delta) + \pi_{2,0}(p)\|_2 \leq a_0(\delta)$, where

$$a_0(\delta) := \mathrm{est}(M;\delta) + e(\delta) + \|L_0^{-1}\|_{\mathrm{Lyap}}\|X_0\|_2^2\delta_E^2\delta_S. \tag{74}$$

If $\|Z\|_{\mathrm{F}} \leq \rho$, then there are different ways to estimate the norms of the vectors $\pi_{1,1}(p,x)$ and $\pi_{0,2}(p,x)$ as $\|\pi_{1,1}(p,x)\|_2 \leq a_1(\delta)\rho$ and $\|\pi_{0,2}(p,x)\|_2 \leq a_2(\delta)\rho^2$, e.g.,

$$a_1(\delta) := \alpha_2(\delta)\delta_E + \alpha_3\delta_A + \alpha_4\delta_S + \alpha_{22}(\delta)\delta_E^2,\ a_2(\delta) := (\|S\|_2 + \delta_S)\left(\beta_0 + \beta_1\delta_E + \beta_2\delta_E^2\right). \tag{75}$$

Here

$$\begin{aligned}
\alpha_2(\delta) &:= \left\|L_0^{-1}J_{n^2}(I_n \otimes A^\top)\right\|_2 + 2\|X_0\|\,\left\|L_0^{-1}J_{n^2}(I_n\otimes E^\top)\right\|_2 (\|S\|_2 + \delta_S), \qquad (76)\\
\alpha_3 &:= \left\|L_0^{-1}J_{n^2}(I_n \otimes E^\top)\right\|_2,\ \alpha_4(\delta) := 2\|X_0\|\,\left\|L_0^{-1}(E^\top\otimes E^\top)\right\|_2,\\
\alpha_{2,2}(\delta) &:= 2\|X_0\|\,\left\|L_0^{-1}\right\|_{\mathrm{Lyap}}(\|S\|_2 + \delta_S)
\end{aligned}$$

and

$$\beta_0 := \left\|L_0^{-1}(E^\top \otimes E^\top)\right\|_2,\ \beta_1 := \left\|L_0^{-1}J_{n^2}(I_n\otimes E^\top)\right\|_2,\ \beta_2 := \left\|L_0^{-1}\right\|_{\mathrm{Lyap}}. \tag{77}$$

Consider the vector operator equation (12) and suppose that $\|\xi\|_2 = \delta_X \leq \rho$ for some $\rho > 0$. Estimating the right-hand side of (12) we get

$$\|\pi(p,x)\|_2 \leq h(\delta,\rho) := a_0(\delta) + a_1(\delta)\rho + a_2(\delta)\rho^2.$$

The function h is a quadratic Lyapunov majorant for the operator equation (12). Consider the domain

$$\Omega := \left\{\delta \in \mathbb{R}_+^4 : a_1(\delta) + 2\sqrt{a_0(\delta)a_2(\delta)} \leq 1\right\}. \tag{78}$$

Since $a_0(0) = a_1(0) = 0$, a continuity argument shows that for some δ with positive elements it holds that $a_1(\delta) + 2\sqrt{a_0(\delta)a_2(\delta)} < 1$. Hence the set $\Omega \subset \mathbb{R}_+^3$ is well defined and has a non-empty interior. This set is bounded by the non-negative coordinate $(1/k)$-planes and by part of the surface $\mathcal{S} \subset \mathbb{R}^3$ given by $(1 - a_1(\delta))^2 = 4a_0(\delta)a_2(\delta)$. Due to the non-linearity of a_0 the set Ω may have a complex geometry, in particular, it may not be convex. However, if $\delta \in \Omega$ and $0 \preceq \widehat{\delta} \preceq \delta$, then $\widehat{\delta} \in \Omega$.

Corollary 11.1 *Let $\delta \in \Omega$, where Ω is given in (78). Then the non-local perturbation bound $\delta_X \leq f(\delta)$ is valid for the real equation (2), where the quantities $f(\delta)$ and $a_i(\delta)$ are determined by (79) and (71), (72), (73), (74), (75), (76), (77).*

Proof. If $\delta \in \Omega$ then the majorant equation $\rho = h(\delta, \rho)$, or, equivalently,

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

has a root

$$\rho(\delta) = f(\delta) := \frac{2a_0(\delta)}{1 - a_1(\delta) + \sqrt{(1 - a_1(\delta))^2 - 4a_0(\delta)a_2(\delta)}}. \tag{79}$$

Hence, for $\delta \in \Omega$ the operator $\pi(p, \cdot)$ maps the set $\mathcal{B}_{g(\delta)}$ into itself. Applying the Schauder fixed point principle we have the desired result. □

12 Examples of Riccati equations in descriptor form

In order to illustrate the presented perturbation estimates, in this section we consider some simple cases of our model problem, a matrix Riccati equation in descriptor form. This demonstrates how the proposed perturbation bounds work in the simplest case of scalar equations or of matrix equations which are diagonalizable. In the first example we study the sensitivity of the scalar quadratic equation.

Example 12.1 Consider the scalar version of (2) $Q + 2EAX - E^2SX^2 = 0$. If $E, S > 0$, then the non-negative solution is $X_0 = (A + \sqrt{A^2 + QS})/(ES)$. Setting $\omega := |A|/\sqrt{A^2 + QS} \in [0, 1)$, we see that the individual relative condition numbers for the solution X_0 are $k_Q = (1 - \omega\mathrm{sign}(A))/2$, $k_E = 1$, $k_A = \omega$, $k_S = (1 + \omega\mathrm{sign}(A))/2$, while the overall relative condition measure is $k_Q + k_E + k_A + k_S = 2 + \omega < 3$. Thus, the non-negative solution of the scalar descriptor equation is very well conditioned. It may be observed that the relative conditioning of the solution X_0 does not depend on E.

Let the nominal values of the parameters be $Q = E = S = 1$ and $A = 0$, which gives the positive solution $X_0 = 1$. We have $L_0 = -2$, $M_1 = 0.5$, $M_2 = -1$, $M_3 = 1$, $M_4 = -0.5$ and hence $\mathrm{est}_1(\delta) = \mathrm{est}_3(\delta) = 0.5(\delta_1 + \delta_4) + \delta_2 + \delta_3$, $\mathrm{est}_2(\delta) = \sqrt{2.5}\|\delta\|_2$. Here the bound $\mathrm{est}_3(\delta)$ from (39) is always superior (or equal to) the bound $\mathrm{est}_2(\delta)$ from (34) as is the case for scalar equations. The two bounds are equal only when $\delta_1 = \delta_4 = \delta_2/2 = \delta_3/2$. Also, $d_2(P, X_0) = \sqrt(7.5) \simeq 2.7386$, $d_3(P, X_0) = 2$ and hence $d(P, X_0) = 2$.

Let the perturbations be taken as $\delta Q = s \geq 0$, $\delta E = -2s$, $\delta A = 2s \geq 0$ and $\delta S = -s$, i.e., $\delta P = (s, -2s, 2s, -s)$ and $\delta = [s, 2s, 2s, s]^\top$. For $s < 0.5$ the positive solution of the perturbed equation $1 + s + 4s(1 - 2s)(1 + \delta X) - (1 - s)(1 - 2s)^2(1 + \delta X)^2 = 0$ is

$$\delta X = \frac{2s + \sqrt{1 + 3s^2}}{(1 - s)(1 - 2s)} - 1.$$

At the same time $a_0(\delta) = 5s$, $a_1(\delta) = s(7 + 8s + 4s^2)$, $a_2(s) = (1 + s)(0.5 + 2s + 2s^2)$. Thus the local and non-local bounds are $\mathrm{est}(\delta) = 5s$ and

$$f(\delta) = \frac{10s}{1 - s(7 + 8s + 4s^2) + \sqrt{d(s)}},$$

respectively, where $d(s) := 1 - 24s - 17s^2 + 24s^3 + 80s^4 + 64s^5 + 16s^6$. The non-linear bound works for $s < s_0$, where $s_0 \simeq 0.0406$ is the smaller positive root of the equation $d(s) = 0$.

Since, for $\delta < 0.5$, $\delta X = 5s + 4s^2 + O(s^3)$, $s \to 0$, we see that the local bound always *underestimates* the true perturbation for this particular structure of the perturbations.

Table 4: Perturbation bounds for a scalar equation

s	δ_X	local	non-local
0.005	0.0254	0.0250	0.0263
0.010	0.0515	0.0500	0.0556
0.015	0.0784	0.0750	0.0887
0.020	0.1061	0.1000	0.1271
0.025	0.1346	0.1250	0.1731
0.030	0.1648	0.1500	0.2311
0.035	0.1943	0.1750	0.3126
0.040	0.2255	0.2000	0.4834
0.045	0.2577	0.2250	*
0.050	0.2909	0.2500	*

In Table 4 we give the exact perturbation $\delta_X = \delta X$, the local bound est and the non-local bound f from Theorem 11.1 as functions of $s \geq 0$. The cases when the non-local bound does not exist are marked by an asterisk.

We see the main drawback of the non-local bounds – their relatively small domain of applicability. On the other hand in this case the local bound is *not* an upper bound for the perturbation in the solution but only gives information for its order of magnitude.

The next two examples are for 2×2 descriptor Riccati equations.

Example 12.2 Consider the 2×2 descriptor Riccati equation with matrices

$$Q = \begin{bmatrix} 0 & 0 \\ 0 & q \end{bmatrix}, \; E = \begin{bmatrix} e_1 & 0 \\ 0 & e_2 \end{bmatrix}, \; A = \begin{bmatrix} 0 & 0 \\ a & 0 \end{bmatrix}, \; S = \begin{bmatrix} s & 0 \\ 0 & 0 \end{bmatrix},$$

where $q, e_1, e_2, a, s > 0$. Setting $X_0 = \begin{bmatrix} x_1 & x \\ x & x_2 \end{bmatrix}$ the element-wise version of the equation becomes

$$\begin{bmatrix} e_1(2ax - e_1 s x_1^2) & e_2(ax_2 - e_1 s x_1) \\ e_2(ax_2 - e_1 s x x_1) & q - e_2^2 s x^2 \end{bmatrix} = 0.$$

The positive definite solution is given by $x_1 = \sqrt{2}\, q^{1/4} e_1^{-1/2} e_2^{-1/2} a^{1/2} s^{-3/4}$, $x = q^{1/2} e_2^{-1} s^{-1/2}$, $x_2 = \sqrt{2}\, q^{3/4} e_1^{1/2} e_2^{-3/2} a^{-1/2} s^{-1/4}$. Note that $x_1 x_2 - x^2 = q e_2^{-1} s^{-1}$ and the matrix $(AE^{-1} - SX_0) = \begin{bmatrix} -sx_1 & -sx \\ a/e_1 & 0 \end{bmatrix}$ has eigenvalues $\sqrt{2}\, q^{1/4} e_1^{-1/2} e_2^{-1/2} a^{1/2} s^{1/4} (-1 \pm \imath)$.

We choose nominal values of the data as $q = e_1 = e_2 = s = 1$, $a = 2$, which gives $X_0 = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$, $A - SX_0E = \begin{bmatrix} -2 & -1 \\ 2 & 0 \end{bmatrix}$ and

$$M_1 = \frac{1}{8} \begin{bmatrix} 2 & 0 & 0 & 4 \\ 0 & 2 & -2 & 4 \\ 0 & -2 & 2 & 4 \\ 1 & -2 & -2 & 6 \end{bmatrix}, \; M_2 = \frac{1}{2} \begin{bmatrix} -2 & 0 & -4 & -2 \\ 0 & 0 & -4 & -2 \\ 0 & 0 & -4 & -2 \\ 1 & 1 & -4 & -3 \end{bmatrix},$$

$$M_3 = \frac{1}{4}\begin{bmatrix} 4 & 2 & 4 & 4 \\ 0 & 0 & 4 & 4 \\ 0 & 0 & 4 & 4 \\ 0 & -1 & 2 & 4 \end{bmatrix}, \quad M_4 = -\frac{1}{8}\begin{bmatrix} 12 & 8 & 8 & 6 \\ 4 & 6 & 2 & 4 \\ 4 & 2 & 6 & 4 \\ 2 & 2 & 2 & 3 \end{bmatrix}.$$

The absolute and relative individual condition numbers are $K_Q = 1.1860$, $K_E = 4.6075$, $K_A = 2.7375$, $K_S = 2.6492$ and $k_Q = 0.4530$, $k_E = 2.4889$, $k_A = 2.0913$, $k_S = 1.0119$. The matrix M_0 with elements $\left\|M_i^\top M_j\right\|_2$ is

$$M_0 = \begin{bmatrix} 1.4065 & 5.3707 & 3.0415 & 2.5759 \\ 5.3707 & 21.2291 & 12.4075 & 10.8332 \\ 3.0415 & 12.4075 & 7.4939 & 6.8922 \\ 2.5759 & 10.8332 & 6.8922 & 7.0183 \end{bmatrix}$$

and $\|[M_1, M_2, M_3, M_4]\|_2 = 5.9781$. For the condition measures (49) we obtain $d_2(P, X_0) = 6.4585$, $d_3(P, X_0) = 5.9391$ and hence, the overall conditioning is $d(P, X_0) = 5.9391$ which is close to $d_1(P, X_0) = 6.0450$.

Let the perturbations in the data be $\delta q = s \geq 0$, $\delta e_1 = \delta e_2 = \delta s = -s$, $\delta a = 2s$, which gives $\delta = s[1, \sqrt{2}, 2, 1]^\top$. Then the perturbation in the solution is $\delta X = \begin{bmatrix} \delta x_1 & \delta x \\ \delta x & \delta x_2 \end{bmatrix}$, where

$$\delta x = (1+s)^{1/2}(1-s)^{-3/2} - 1,\ \delta x_1 = 2(1+s)^{3/4}(1-s)^{-7/4} - 2,\ \delta x_2 = (1+s)^{1/4}(1-s)^{-5/4} - 1.$$

For the local estimates we have that $\mathrm{est}_1(\delta) = 15.8261s$, $\mathrm{est}_2(\delta) = 16.9087s$, $\mathrm{est}_3(\delta) = 15.5487s$. Thus $\mathrm{est}(\delta) = \mathrm{est}_3(\delta) = 15.5487s$. The results for these perturbations are given in Table 5.

Table 5: Perturbation bounds for a 2×2 descriptor Riccati equation

s	δ_X	local	non-local
0.001	0.0059	0.0155	0.0163
0.002	0.0119	0.0311	0.0343
0.003	0.0179	0.0466	0.0544
0.004	0.0239	0.0622	0.0774
0.005	0.0299	0.0777	0.1043
0.006	0.0359	0.0933	0.1372
0.007	0.0420	0.1088	0.1807
0.008	0.0480	0.1244	0.2552
0.009	0.0541	0.1399	*
0.010	0.0602	0.1555	*

Example 12.3 Consider the descriptor equation from Example 12.2 with the same nominal data and perurbations

$$\delta Q = s\begin{bmatrix} 3 & -2 \\ -2 & 19 \end{bmatrix},\ \delta E = -s\begin{bmatrix} 6 & 67 \\ -4 & 37 \end{bmatrix},\ \delta A = s\begin{bmatrix} 10 & 29 \\ 3 & 33 \end{bmatrix},\ \delta S = -s\begin{bmatrix} 42 & 0 \\ 0 & 0 \end{bmatrix},$$

where $s > 0$ is a small parameter. Hence $\delta = s[19.4422, 76.8765, 45.1553, 42.0]^\top$ and $\|\delta\| = 100.454s$. We have $\mathrm{est}_1(\delta) = 612.1449\ s$, $\mathrm{est}_2(\delta) = 600.5259\ s$, $\mathrm{est}_3(\delta) = 601.2477\ s$ and thus $\mathrm{est}(\delta) = \mathrm{est}_2(\delta) = 600.5259\ s$. The results are given in Table 6.

Table 6: Perturbation bounds for a 2×2 descriptor Riccati equation

s	δ_X	local	non-local
0.00002	0.0112	0.0120	0.0125
0.00004	0.0225	0.0240	0.0261
0.00006	0.0338	0.0360	0.0411
0.00008	0.0452	0.0480	0.0577
0.00010	0.0566	0.0601	0.0764
0.00012	0.0681	0.0721	0.0979
0.00014	0.0797	0.0841	0.1233
0.00016	0.0913	0.0961	0.1550
0.00018	0.1030	0.1081	0.1989
0.00020	0.1147	0.1201	*

13 Further comments and conlusion

We have presented a general framework to derive perturbation results for general matrix equations and have applied this framework to algebraic Riccati equations in descriptor form.

Tighter bounds may be obtained when considering structured perturbations, e.g. when restricting the perturbations to be symmetric for symmetric Riccati equations. see e.g. [33].

The local and non-local bounds in this paper have the traditional features of these objects. The local bounds give satisfactory results for small perturbations in the data, but it is not clear how small they must be. For larger perturbations they may underestimate the perturbation in the solution, thus being not upper bounds in the strict sense. As a whole, the local bounds can give a good idea for the order of magnitude in the perturbation in the solution. On the other hand the non-local bounds hold indeed as upper bounds for the perturbation in the solution but their domain of applicability may be small.

In all cases the local and non-local bounds are equal in first order approximations. Both bounds may be very accurate for certain sufficiently small perturbations in the data. This, for example, will be the case when the vectorized data perturbation is approximately proportional to the singular vector of a certain matrix corresponding to its maximum singular value.

There is a number of papers devoted to the perturbation analysis of continuous-time Riccati equations arising in linear systems theory [41, 42, 12, 32, 34, 56]. Until recently, however, the complex case had not been analyzed in a sufficient extent. Here the treatment in [56] for the standard Riccati equation should be complemented with the analysis from [40].

References

[1] B.D.O. Anderson and J.B. Moore. *Linear Optimal Control*. Prentice-Hall, Englewood Cliffs, NJ, 1971.

[2] R. Aripirala and V.L. Syrmos. Sensitivity analysis of stable generalized Lyapunov equations. In *Proceedings of the 32nd IEEE Conference on Decision and Control*, pages 3144–3149, San Antonio, TX, December 1993.

[3] R. Aripirala and V.L. Syrmos. Sensitivity analysis and computable bounds for the generalized algebraic Riccati equation. In *Proceedings of the 1994 American Control Conference*, pages 2680–2684, Baltimore, MD, June 1994.

[4] R. Bellman. Kronecker products and the second method of Lyapunov. *Mathematische Nach.*, 20:17–19, 1959.

[5] D.J. Bender and A.J. Laub. The linear-quadratic optimal regulator for descriptor systems. *IEEE Trans. Automat. Control*, AC-32:672–688, 1987.

[6] A. Berman and R.J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences*. SIAM, Philadelphia, 1994.

[7] G.E. Bredon. *Topology and Geometry*. Springer-Verlag, New York, 1993.

[8] R. Byers. Numerical condition of the algebraic Riccati equation. *Contemp. Math.*, 47:35–49, 1985.

[9] R. Byers and S. Nash. On the singular "vectors" of the Lyapunov operator. *SIAM J. Algebraic Discrete Methods*, 8:59–66, 1987.

[10] J.W. Demmel. On condition numbers and the distance to the nearest ill-posed problem. *Numer. Math.*, 51:251–289, 1987.

[11] P. Gahinet and A. J. Laub. Numerically reliable computation of optimal performance in singular H_∞ control. *SIAM J. Cont. Optim.*, 35:1690–1710, 1997.

[12] P. Gahinet and A.J. Laub. Computable bounds for the sensitivity of the algebraic Riccati equation. *SIAM J. Cont. Optim.*, 28:1461–1480, 1990.

[13] P.M. Gahinet, A.J. Laub, C.S. Kenney, and G.A. Hewer. Sensitivity of the stable discrete-time Lyapunov equation. *IEEE Trans. Automat. Control*, 35:1209–1217, 1990.

[14] F.R. Gantmacher. *Theory of Matrices*, volume 1. Chelsea, New York, 1959.

[15] A.R. Ghavimi and A.J. Laub. Backward error, sensitivity and refinement of computed solutions of algebraic Riccati equations. *Numer. Alg. Appl.*, 2:29–49, 1995.

[16] I. Gohberg and I. Koltracht. Mixed, componentwise, and structured condition numbers. *SIAM J. Matrix Anal. Appl.*, 14:688–704, 1993.

[17] A. Graham. *Kronecker Products and Matrix Calculus with Applications*. Wiley, New York, 1981.

[18] E.A. Grebenikov and Yu.A. Ryabov. *Constructive Methods for Analysis of Nonlinear Systems*. Nauka, Moscow, 1979. In Russian.

[19] C. h. Chen. Perturbation analysis for solutions of algebraic Riccati equations. *J. Comput. Math.*, 6:336–347, 1988.

[20] H.V. Henderson and S.R. Searle. The vec-permutation matrix, the vec operator and Kronecker products: A review. *Lin. Multilin. Alg.*, 9:271–288, 1981.

[21] G. Hewer and C. Kenney. The sensitivity of the stable Lyapunov equation. *SIAM J. Cont. Optim.*, 26:321–344, 1988.

[22] D.J. Higham and N.J. Higham. Backward error and condition of structured linear systems. *SIAM J. Matrix Anal. Appl.*, 13:162–175, 1992.

[23] D.J. Higham and N.J. Higham. Componentwise perturbation theory for linear systems with multiple right-hand sides. *Linear Algebra Appl.*, 174:111–129, 1992.

[24] N. J. Higham. Perturbation theory and backward error for $AX - XB = C$. *BIT*, 33:124–136, 1993.

[25] N.J. Higham. A survey of componentwise perturbation theory in numerical linear algebra. In W. Gautchi, editor, *Mathematics of Computation 1943-1993: A Half Century of Computational Mathematics*, volume 48 of *Proceedings of Symposia in Applied Mathematics*, pages 49–77. Amer. Math. Soc., Providence, RI, USA, 1994.

[26] N.J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, Philadelphia, PA, 1996.

[27] E.A. Jonckheere. New bound on the sensitivity of the solution of the Lyapunov equation. *Linear Algebra Appl.*, 60:57–64, 1984.

[28] B. Kågström. A perturbation analysis of the generalized Sylvester equation. *SIAM J. Matrix Anal. Appl.*, 15:1045–1060, 1994.

[29] B. Kågström and L. Westin. Generalized Schur methods with condition estimators for solving the generalized Sylvester equation. *IEEE Trans. Automat. Control*, 34:745–751, 1989.

[30] R.E. Kalman. Contributions to the theory of optimal control. *Boletin Sociedad Matematica Mexicana*, 5:102–119, 1960.

[31] L. V. Kantorovich and G. P. Akilov. *Functional Analysis in Normed Spaces*. Pergamon, New York, 1964.

[32] C. Kenney and G. Hewer. The sensitivity of the algebraic and differential Riccati equations. *SIAM J. Cont. Optim.*, 28:50–69, 1990.

[33] M. Konstantinov, V. Mehrmann, and P. Petkov. On properties of general Sylvester and Lyapunov operators. *Linear Algebra Appl.*, 312:35–71, 2000.

[34] M. Konstantinov, P. Petkov, and D.W. Gu. Improved perturbation bounds for general quadratic matrix equations. *Numer. Func. Anal. Optim.*, 20:717–736, 1999.

[35] M. Konstantinov, P. Petkov, D.W. Gu, and V. Mehrmann. Sensitivity of general Lyapunov equations. Technical Report 98–15, Department of Engineering, Leicester University, Leicester, UK, Aug 1998.

[36] M. Konstantinov, P. Petkov, D.W. Gu, and I. Postlethwaite. Perturbation analysis in finite dimensional spaces. Technical Report 96–18, Department of Engineering, Leicester University, Leicester, UK, June 1996.

[37] M. Konstantinov, P. Petkov, A. Linnemann, J. Kawelke, D.W. Gu, and I. Postlethwaite. Sensitivity of system norms. *Int. J. Control*, 72:84–95, 1998.

[38] M. Konstantinov, P. Petkov, V. Mehrmann, and D. Gu. Additive matrix operators. In *Proc. 30 Spring Conf. of Union of Bulgarian Mathematicians*, pages 169–175, Borovetz, Bulgaria, April 2001.

[39] M. Konstantinov, M. Stanislavova, and P. Petkov. Perturbation bounds and characterisation of the solution of the associated algebraic Riccati equation. *Linear Algebra Appl.*, 285:7–31, 1998.

[40] M.M. Konstantinov and P.Hr. Petkov. A note on perturbation theory for algebraic Riccati equations. *SIAM J. Matrix Anal. Appl.*, 21:327, 1999.

[41] M.M. Konstantinov, P.Hr. Petkov, and N.D. Christov. Perturbation analysis of the continuous and discrete matrix Riccati equations. In *Proc. 1986 American Control Conference*, pages 636–639, Seattle, WA, June 1986.

[42] M.M. Konstantinov, P.Hr. Petkov, and N.D. Christov. Perturbation analysis of matrix quadratic equations. *SIAM J. Sci. Statist. Comput.*, 11:1159–1163, 1990.

[43] M.M. Konstantinov, P.Hr. Petkov, and N.D. Christov. Perturbation analysis of the discrete Riccati equation. *Kybernetica* (Prague), 29:18–29, 1993.

[44] V. Kučera. A contribution to matrix quadratic equations. *IEEE Trans. Automat. Control*, AC-17:344–347, 1972.

[45] V. Kučera. On nonnegative definite solutions to matrix quadratic equations. *Automatica*, 8:413–423, 1972.

[46] V. Mehrmann. *The Autonomous Linear Quadratic Control Problem, Theory and Numerical Solution.* Number 163 in Lecture Notes in Control and Information Sciences. Springer-Verlag, Heidelberg, 1991.

[47] J. Ortega and W. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables.* Academic Press, New York, 1970.

[48] S. Peng and C.E. de Souza. Bounds on the solution of the algebraic matrix Riccati equation under perturbations in the coefficients. *Syst. Contr. Lett.*, 15:175–181, 1990.

[49] S. Peng and C.E. de Souza. On bounds for perturbed discrete-time algebraic Riccati equations. In H. Kimura and S. Kodama, editors, *Mathematical Theory of Systems, Control, Networks and Signal Processing. Proceedings of the International Symposium MTNS-91, Kobe, Japan, June 1991*, pages 9–14. Mita Press, Tokyo, 1992.

[50] J.R. Rice. A theory of condition. *SIAM J. Numer. Anal.*, 3:287–310, 1966.

[51] S.M. Rump. Estimation of the sensitivity of linear and nonlinear algebraic problems. *Linear Algebra Appl.*, 153:1–34, 1991.

[52] G.W. Stewart and J.-G. Sun. *Matrix Perturbation Theory.* Academic Press, New York, 1990.

[53] J.-G. Sun. Sensitivity analysis of the discrete-time algebraic Riccati equation. Technical Report UMINF 96.08, Department of Computing Science, University of Umeå, Umeå, Sweden, 1996.

[54] J.-G. Sun. Backward error for the discrete-time algebraic Riccati equation. *Linear Algebra Appl.*, 259:183–208, 1997.

[55] J.-G. Sun. Residual bounds of approximate solutions of the algebraic Riccati equation. *Numer. Math.*, 76:249–263, 1997.

[56] J.-G. Sun. Perturbation theory for algebraic Riccati equations. *SIAM J. Matrix Anal. Appl.*, 19:39–65, 1998.

[57] R.J. Weidner and R.J. Mulholland. Kronecker product representation for the solution of the general linear matrix equation. *IEEE Trans. Automat. Control*, AC-25:563–564, 1980.

[58] A. Weinmann. *Uncertain Models and Robust Control.* Springer-Verlag, Wien, 1991.

[59] S. Xu. Sensitivity analysis of the algebraic Riccati equations. *Numer. Math.*, 75:121–134, 1996.

[60] K. Zhou, J.C. Doyle, and K. Glover. *Robust and Optimal Control.* Prentice-Hall, Englewood Cliffs, NJ, 1996.