



# Structural backward stability in rational eigenvalue problems solved via block Kronecker linearizations

Froilán M. Dopico<sup>1</sup> · María C. Quintana<sup>2</sup> · Paul Van Dooren<sup>3</sup>

Received: 22 January 2022 / Revised: 10 December 2022 / Accepted: 15 December 2022  
© The Author(s) under exclusive licence to Istituto di Informatica e Telematica (IIT) 2023

## Abstract

In this paper we study the backward stability of running a backward stable eigenstructure solver on a pencil  $S(\lambda)$  that is a strong linearization of a rational matrix  $R(\lambda)$  expressed in the form  $R(\lambda) = D(\lambda) + C(\lambda I_\ell - A)^{-1}B$ , where  $D(\lambda)$  is a polynomial matrix and  $C(\lambda I_\ell - A)^{-1}B$  is a minimal state-space realization. We consider the family of block Kronecker linearizations of  $R(\lambda)$ , which have the following structure

$$S(\lambda) := \begin{bmatrix} M(\lambda) & \widehat{K}_2^T C & K_2^T(\lambda) \\ B \widehat{K}_1 & A - \lambda I_\ell & 0 \\ K_1(\lambda) & 0 & 0 \end{bmatrix},$$

where the blocks have some specific structures. Backward stable eigenstructure solvers, such as the  $QZ$  or the staircase algorithms, applied to  $S(\lambda)$  will compute the exact eigenstructure of a perturbed pencil  $\widehat{S}(\lambda) := S(\lambda) + \Delta_S(\lambda)$  and the special structure of  $S(\lambda)$  will be lost, including the zero blocks below the anti-diagonal. In order to link this perturbed pencil with a nearby rational matrix, we construct in this paper a strictly equivalent pencil  $\widetilde{S}(\lambda) = (I - X)\widehat{S}(\lambda)(I - Y)$  that restores the original structure, and hence is a block Kronecker linearization of a perturbed rational matrix

---

The first and second authors were partially supported by “Ministerio de Economía, Industria y Competitividad (MINECO)” of Spain and “Fondo Europeo de Desarrollo Regional (FEDER)” of EU through grant MTM2015-65798-P, by the “Proyecto financiado por la Agencia Estatal de Investigación de España” (PID2019-106362GB-I00 / AEI / 10.13039/501100011033) and by the Madrid Government (Comunidad de Madrid-Spain) under the “Multiannual Agreement with Universidad Carlos III de Madrid in the line of Excellence of University Professors (EPUC3M23), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation)”. The second author was funded by the “contrato predoctoral” BES-2016-076744 of MINECO and by an Academy of Finland grant (Suomen Akatemian päätös 331240). This work was developed while the third author held a “Chair of Excellence UC3M - Banco de Santander” at Universidad Carlos III de Madrid in the academic year 2019–2020.

---

✉ Paul Van Dooren  
paul.vandooren@uclouvain.be

Extended author information available on the last page of the article

$\tilde{R}(\lambda) = \tilde{D}(\lambda) + \tilde{C}(\lambda I_\ell - \tilde{A})^{-1}\tilde{B}$ , where  $\tilde{D}(\lambda)$  is a polynomial matrix with the same degree as  $D(\lambda)$ . Moreover, we bound appropriate norms of  $\tilde{D}(\lambda) - D(\lambda)$ ,  $\tilde{C} - C$ ,  $\tilde{A} - A$  and  $\tilde{B} - B$  in terms of an appropriate norm of  $\Delta_S(\lambda)$ . These bounds may be, in general, inadmissibly large, but we also introduce a scaling that allows us to make them satisfactorily tiny, by making the matrices appearing in both  $S(\lambda)$  and  $R(\lambda)$  have norms bounded by 1. Thus, for this scaled representation, we prove that the staircase and the  $QZ$  algorithms compute the exact eigenstructure of a rational matrix  $\tilde{R}(\lambda)$  that can be expressed in exactly the same form as  $R(\lambda)$  with the parameters defining the representation very near to those of  $R(\lambda)$ . This shows that this approach is backward stable in a structured sense. Several numerical experiments confirm the obtained backward stability results.

**Keywords** Rational matrix · Rational eigenvalue problem · Linearization · Matrix pencils · Perturbations · Backward error analysis

**Mathematics Subject Classification** 65F15 · 15A18 · 15A22 · 15A54 · 93B18 · 93B20 · 93B60

## 1 Introduction

It has been known since the 1970s that the zeros of a rational matrix are also the eigenvalues of an appropriately defined pencil of matrices, i.e., a polynomial matrix of degree at most 1, and that its poles are the eigenvalues of a principal submatrix of such a pencil. This connection was established in the influential book of Rosenbrock [18]. About 10 years later numerical algorithms were proposed in [21, 22] to construct such a pencil in a numerically stable way. Not only the zeros and poles can be determined via these pencils, but also their structural indices, or partial multiplicities, as well as the minimal indices of the left and right null-spaces of the rational matrix, see e.g. [26]. Together, these are called the eigenstructure of the rational matrix, and the pencils considered in [26] are called *system matrices* of a *strongly irreducible generalized state-space realization*.

Polynomial matrices can be viewed as special cases of rational matrices, which happen to have all their poles at infinity. The notions of generalized state-space realizations and corresponding (strongly) irreducible system matrices therefore apply to polynomial matrices as well. But in the classic reference [10] a new notion of *strong linearization* is introduced for polynomial matrices which is consistent with that of *strongly irreducible system matrix* of [26] for the finite eigenvalues and their structural indices. But for the structural indices at infinity, these two definitions differ by a constant shift, which means that the structural indices at infinity can easily be recovered from one definition to the other. Moreover, the definition of strong linearization introduced in [10] does not guarantee any relationship between the minimal indices of the linearization and those of the polynomial matrix [5], in contrast with the pencils in [26] for which the minimal indices are equal.

Even though the definition of strong linearization in [10] was originally given for polynomial matrices, there have been several attempts to extend it to rational matrices

[2, 4], including in these extensions also the concept of (non-strong) linearization [1, 2]. Thus, inspired by previous results for polynomial matrices [7], a wide family of strong linearizations called *strong block minimal bases linearizations* is proposed in [2, Theorem 5.11] for any  $m \times n$  rational matrix  $R(\lambda)$  with coefficient matrices in an arbitrary field  $\mathbb{F}$ . These linearizations are based on the splitting of  $R(\lambda)$  into its strictly proper part  $R_p(\lambda)$  and its polynomial part  $D(\lambda)$  and in the representation :

$$R(\lambda) := R_p(\lambda) + D(\lambda) = C(\lambda I_\ell - A)^{-1}B + \sum_{i=0}^d D_i \lambda^i, \tag{1.1}$$

where  $C(\lambda I_\ell - A)^{-1}B$  is a minimal state-space realization of the strictly proper part  $R_p(\lambda)$ , represented in what follows by the triple  $\{A, B, C\}$ , and  $d > 1$  is the degree of the polynomial part. Then  $R(\lambda)$  is represented by the quadruple  $\{\lambda I_\ell - A, B, C, D(\lambda)\}$ . We emphasize that this representation of a rational matrix is one of the most commonly used in applications and that it is valid for any rational matrix [18]. Since in this paper we are analyzing perturbations related to backward errors of eigenvalue solvers of pencils with real or complex matrix coefficients, we restrict  $\mathbb{F}$  to be the real field  $\mathbb{R}$  or the complex field  $\mathbb{C}$ .

A particular case of the strong block minimal bases linearizations in [2, Theorem 5.11] of any  $m \times n$  rational matrix  $R(\lambda)$  represented as in (1.1) are (modulo block permutations) the pencils of the form

$$S(\lambda) := \begin{bmatrix} M(\lambda) & \widehat{K}_2^T C & K_2^T(\lambda) \\ B \widehat{K}_1 & A - \lambda I_\ell & 0 \\ K_1(\lambda) & 0 & 0 \end{bmatrix}, \tag{1.2}$$

with  $M(\lambda)$  a pencil whose properties are described below,

$$K_1(\lambda) := L_\epsilon(\lambda) \otimes I_n, \quad \widehat{K}_1 := \mathbf{e}_{\epsilon+1}^T \otimes I_n, \quad K_2(\lambda) := L_\eta(\lambda) \otimes I_m, \quad \widehat{K}_2 := \mathbf{e}_{\eta+1}^T \otimes I_m,$$

and where  $\otimes$  denotes the Kronecker product,  $\mathbf{e}_k = [0 \cdots 0 1]^T$  is the standard  $k$ th unit vector of dimension  $k$  and  $L_k(\lambda)$  is the classical Kronecker block of dimension  $k \times (k + 1)$

$$L_k(\lambda) := \begin{bmatrix} 1 - \lambda & & & & \\ & 1 - \lambda & & & \\ & & \ddots & \ddots & \\ & & & & 1 - \lambda \end{bmatrix}.$$

Moreover, the pencil  $M(\lambda)$  in (1.2) is related to the polynomial part  $D(\lambda)$  in (1.1) by the “dual basis” vector  $A_k(\lambda)$  of powers of  $\lambda$ ,

$$A_k^T(\lambda) := [\lambda^k \cdots \lambda^2 \lambda 1],$$

which satisfies  $L_k(\lambda)A_k(\lambda) = 0$  and also

$$D(\lambda) = (\Lambda_\eta(\lambda) \otimes I_m)^T M(\lambda) (\Lambda_\epsilon(\lambda) \otimes I_n).$$

Thus,  $d = \epsilon + \eta + 1$  (see [7, eq. (4.5)]). We emphasize that the values  $\epsilon = 0$  or  $\eta = 0$  are allowed in the definition of  $S(\lambda)$  in (1.2). If  $\epsilon = 0$ , then the last block row of  $S(\lambda)$  is omitted and  $\widehat{K}_1 = I_n$ , while if  $\eta = 0$ , then the last block column of  $S(\lambda)$  is omitted and  $\widehat{K}_2 = I_m$ . As we explain below, the parameters  $\epsilon$  and  $\eta$  determine how the minimal indices of  $S(\lambda)$  are related to those of  $R(\lambda)$ . The results in this paper will show that the backward error bounds are better when choosing  $\epsilon$  and  $\eta$  (almost) equal.

The strong linearizations (1.2) are inspired by the so-called “block Kronecker linearizations” that were introduced in [7, Section 4] for an arbitrary  $m \times n$  polynomial matrix  $D(\lambda)$ . Therefore, we use the same name in the rational setting. For polynomial matrices, such linearizations are obtained by omitting the second block row and the second block column in  $S(\lambda)$  and, for  $\epsilon = 0$  or  $\eta = 0$ , they include the classical second and first Frobenius companion forms [10], among many other linearizations. However, taking  $\epsilon = 0$  or  $\eta = 0$  does not allow to construct structure preserving linearizations of structured polynomial matrices. For this purpose, one should take  $\epsilon = \eta \neq 0$  [8]. The representation of  $R(\lambda)$  in (1.1) and the block Kronecker linearizations  $S(\lambda)$  of  $R(\lambda)$  (1.2) are the two fundamental ingredients of this paper.

As explained in [2, Section 3.1], the finite eigenvalues, together with their partial multiplicities, of  $S(\lambda)$  (resp.  $A - \lambda I_\ell$ ) coincide with the finite zeros (resp. poles) of  $R(\lambda)$ , together with their partial multiplicities. Moreover, the eigenvalue structure at infinity of  $S(\lambda)$  allows us to obtain via a simple shift rule the pole-zero structure at infinity of  $R(\lambda)$ .<sup>1</sup> In addition, as proved in [3, Section 6], the right (resp. left) minimal indices of  $S(\lambda)$  are those of  $R(\lambda)$  plus  $\epsilon$  (resp.  $\eta$ ). Thus,  $S(\lambda)$  comprises the complete eigenstructure of  $R(\lambda)$ . Observe that the application to  $S(\lambda)$  of the QZ algorithm [16], in the regular case, or of the staircase algorithm [21], in the singular case, gives the zeros and the minimal indices, in the singular case, of  $R(\lambda)$ , but not the poles, which are in  $A - \lambda I_\ell$ .

It is worth mentioning that although the families of block Kronecker linearizations of polynomial [7] and rational [2] matrices are very recent, some particular examples of strong linearizations in these families appeared much earlier in the literature. For instance, it was shown in [24] that a valid “realization” for the polynomial part  $D(\lambda)$  in (1.1) is given by the following minimal Rosenbrock polynomial system matrix [18]

$$S_D(\lambda) := \left[ \begin{array}{ccc|cc} I_n & -\lambda I_n & & & \\ & I_n & \ddots & & \\ & & \ddots & -\lambda I_n & \\ & & & I_n & -\lambda I_n \\ \hline \lambda D_d & \dots & \dots & \lambda D_2 & \lambda D_1 + D_0 \end{array} \right] := \left[ \begin{array}{c|c} T(\lambda) & -U(\lambda) \\ \hline V(\lambda) & W(\lambda) \end{array} \right],$$

<sup>1</sup> More precisely, according to [2, p. 1683] if  $r$  is the normal rank of  $R(\lambda)$  and  $e_1 \leq \dots \leq e_r$  are the  $r$  largest partial multiplicities at infinity of  $S(\lambda)$ , then  $e_1 - d \leq \dots \leq e_r - d$  are the structural indices at infinity of  $R(\lambda)$ .

which means that  $D(\lambda) = W(\lambda) + V(\lambda)T(\lambda)^{-1}U(\lambda)$ . It is easy to see that after moving the bottom block row of  $S_D(\lambda)$  to the top position, a block Kronecker linearization of  $D(\lambda)$  is obtained with  $K_2(\lambda)$  empty [7, Section 4]. Combining the minimal state-space realization  $C(\lambda I_\ell - A)^{-1}B$  and the polynomial system matrix  $S_D(\lambda)$  yields the following minimal polynomial system matrix for the rational matrix  $R(\lambda)$  in (1.1):

$$S_R(\lambda) := \left[ \begin{array}{c|c} \begin{array}{c} A - \lambda I_\ell \\ \hline I_n \quad -\lambda I_n \\ \quad \quad I_n \quad \ddots \\ \quad \quad \quad \ddots \quad -\lambda I_n \\ \quad \quad \quad \quad \quad I_n \end{array} & B \\ \hline C & \lambda D_d \quad \cdots \quad \lambda D_2 \quad \lambda D_1 + D_0 \end{array} \right] := \left[ \begin{array}{c|c} T_R(\lambda) & -U_R(\lambda) \\ \hline V_R(\lambda) & W_R(\lambda) \end{array} \right],$$

i.e.,  $R(\lambda) = W_R(\lambda) + V_R(\lambda)T_R(\lambda)^{-1}U_R(\lambda)$ . A pencil with a structure similar to  $S_R(\lambda)$  can also be found in [19]. Observe that if the bottom block row of  $S_R(\lambda)$  is moved to the top position and the leftmost block column is moved to the rightmost position, then the following reordered pencil is obtained

$$\left[ \begin{array}{c|c} \lambda D_d \quad \cdots \quad \lambda D_2 \quad \lambda D_1 + D_0 & C \\ \hline B & A - \lambda I_\ell \\ \hline I_n \quad -\lambda I_n & \\ \quad \quad I_n \quad \ddots & \\ \quad \quad \quad \ddots \quad -\lambda I_n & \\ \quad \quad \quad \quad \quad I_n \quad -\lambda I_n & \end{array} \right],$$

which is a particular case of the block Kronecker linearizations appearing in (1.2) for  $R(\lambda)$  with  $\eta = 0$ , i.e., with the last block column omitted,  $\widehat{K}_2 = I_m$ , and  $\epsilon = d - 1$ . In fact, it can be proved that what has been shown above for  $S_R(\lambda)$  holds for all the block Kronecker linearizations of  $R(\lambda)$  in (1.2), since all of them can be seen as minimal Rosenbrock polynomial system matrices of  $R(\lambda)$  when we permute them to

$$S_K(\lambda) := \left[ \begin{array}{c|c|c} A - \lambda I_\ell & 0 & B \widehat{K}_1 \\ \hline 0 & 0 & K_1(\lambda) \\ \hline \widehat{K}_2^T C & K_2^T(\lambda) & M(\lambda) \end{array} \right]$$

and, then, we partition them appropriately, since the bottom right submatrix is a linearization for  $D(\lambda)$ . This approach based on the block Kronecker linearizations for polynomial matrices, also contains the companion forms as a special case.

It was shown in [7] that perturbations of the block Kronecker linearizations of a polynomial matrix  $D(\lambda)$  can be mapped to perturbations of the coefficients of  $D(\lambda)$  without significant growth of the relative norms of the perturbations under mild assumptions that require to scale  $D(\lambda)$  to have norm equal to 1 and to use linearizations with the norm of  $M(\lambda)$  of the same order as the norm of  $D(\lambda)$  (see [7, Corollary 5.24]). As a

corollary of this perturbation result, we obtain that under such assumptions the computation of the eigenvalues and minimal indices of a polynomial matrix by applying the  $QZ$  or the staircase algorithm to one of its block Kronecker linearizations is a backward stable method from the point of view of the polynomial matrix. In this paper we show that this can be extended to rational matrices as well, considering as coefficients of the rational matrix those in the quadruple  $\{\lambda I_\ell - A, B, C, D(\lambda)\}$ . However, we emphasize that the perturbation analysis for block Kronecker linearizations of rational matrices is considerably more complicated than the one in [7] and, therefore, we limit ourselves to perform a first order analysis. We also remark that the scaling needed to get satisfactory perturbation bounds is more delicate than the one in [7]. As far as we know, this is the first structural backward error analysis of this type performed in the literature for linearizations of rational matrices.

We assume throughout the paper that  $\ell > 0$  since, otherwise,  $R(\lambda)$  in (1.1) is a polynomial matrix and this case was studied in [7]. Except in Sect. 4.6, we also assume that at least one of the parameters  $\epsilon$  and  $\eta$  in (1.2) is larger than zero since, otherwise, none of the blocks  $K_1(\lambda)$  and  $K_2(\lambda)$  appears and block Kronecker linearizations collapse to much simpler pencils. Note that  $\max(\eta, \epsilon) > 0$  implies that the degree  $d$  of the polynomial part  $D(\lambda)$  of  $R(\lambda)$  is larger than 1. The simple case  $d \leq 1$  is studied in Sect. 4.6.

In order to measure perturbations, we need to introduce appropriate norms for pencils, polynomial matrices and rational matrices expressed as in (1.1). For any pair of matrices  $X$  and  $Y$  of arbitrary dimensions (that might be different), we will use the following norms

$$\begin{aligned}\|(X, Y)\|_F &:= \left( \|X\|_F^2 + \|Y\|_F^2 \right)^{\frac{1}{2}} = \|\text{vec}(X)^T, \text{vec}(Y)^T\|_2, \\ \|(X, Y)\|_2 &:= \left( \|X\|_2^2 + \|Y\|_2^2 \right)^{\frac{1}{2}},\end{aligned}$$

where  $\|X\|_F$  and  $\|X\|_2$  are, respectively, the Frobenius and spectral matrix norms and  $\text{vec}(X)$  is the operator that stacks the columns of a matrix into one column vector [11]. For a pencil  $S(\lambda) := A - \lambda B$  we define the corresponding norms via the two matrix coefficients:

$$\|S(\lambda)\|_F := \|(A, B)\|_F, \quad \|S(\lambda)\|_2 := \|(A, B)\|_2.$$

More generally, for a polynomial matrix  $D(\lambda) := \sum_{i=0}^d D_i \lambda^i$ , we will use the norm

$$\|D(\lambda)\|_F := \sqrt{\sum_{i=0}^d \|D_i\|_F^2},$$

and for a list of polynomial matrices  $(D_1(\lambda), \dots, D_p(\lambda))$ , the norm

$$\|(D_1(\lambda), \dots, D_p(\lambda))\|_F := \sqrt{\sum_{i=1}^p \|D_i(\lambda)\|_F^2}.$$

Finally, for a rational matrix  $R(\lambda)$ , represented by a quadruple  $\{\lambda I_\ell - A, B, C, D(\lambda)\}$ , as in (1.1), we use the “norm”

$$\begin{aligned} \|R(\lambda)\|_F &:= \|(\lambda I_\ell - A, B, C, D(\lambda))\|_F \\ &= \sqrt{\ell + \|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2 + \sum_{i=0}^d \|D_i\|_F^2}. \end{aligned}$$

That is, the “norm” of a rational matrix  $R(\lambda)$  is defined as the norm of an associated polynomial system matrix  $P(\lambda)$ , in this case,

$$\|R(\lambda)\|_F := \|P(\lambda)\|_F \quad \text{where} \quad P(\lambda) := \begin{bmatrix} \lambda I_\ell - A & -B \\ C & D(\lambda) \end{bmatrix}. \quad (1.3)$$

We remark that  $\|R(\lambda)\|_F$  is not rigorously a “norm” for  $R(\lambda)$  because, for instance,  $R(\lambda)$  is zero if  $B = 0$  and  $D(\lambda) = 0$ , but  $\|R(\lambda)\|_F$  is not. Despite this fact, and with a clear abuse of nomenclature, we will use the terminology “norm of a rational matrix” in the sense explained above.

The paper is organized as follows. After this introductory section, we describe in Sect. 2 the basic systems of matrix equations we will use in this paper, and, in Sect. 3, some bounds for the singular values of certain matrices related to these systems of matrix equations. In Sect. 4 we explain how to restore the structure of block Kronecker linearizations of rational matrices after they suffer sufficiently small perturbations, and, in Sect. 5, we derive a scaling technique that allows us to guarantee structured backward stability for (regular or singular) rational eigenvalue problems solved via block Kronecker linearizations. Finally, in Sect. 6 we give a number of numerical results illustrating our theoretical bounds and, in Sect. 7, we establish some conclusions.

## 2 Generalized Sylvester equations

In order to restore the structure of perturbed block Kronecker linearizations of rational matrices, we will need to guarantee that some matrix equations have solutions and to bound the norm of their minimal norm solution. The matrix equations that we will encounter are particular cases of the generalized Sylvester equation for  $m_i \times n_i$  pencils of matrices  $A_i - \lambda B_i$ ,  $i = 1, 2$ , which is the following equation in the unknowns  $X$  and  $Y$ :

$$X(A_1 - \lambda B_1) + (A_2 - \lambda B_2)Y = H_a - \lambda H_b, \quad (2.1)$$

that is assumed to hold for any  $\lambda$ . It is easily seen to be equivalent to a linear system of equations, when rewriting it as

$$\begin{aligned} X A_1 + A_2 Y &= H_a, \\ X B_1 + B_2 Y &= H_b, \end{aligned}$$

or, when using Kronecker products and the  $\text{vec}(\cdot)$  notation, as

$$\left[ \begin{array}{c|c} A_1^T \otimes I_{m_2} & I_{n_1} \otimes A_2 \\ \hline B_1^T \otimes I_{m_2} & I_{n_1} \otimes B_2 \end{array} \right] \begin{bmatrix} \text{vec}(X) \\ \text{vec}(Y) \end{bmatrix} = \begin{bmatrix} \text{vec}(H_a) \\ \text{vec}(H_b) \end{bmatrix}. \tag{2.2}$$

The dimension of the unknowns  $X$  and  $Y$  are  $m_2 \times m_1$  and  $n_2 \times n_1$ , respectively, and those of the right hand sides  $H_a$  and  $H_b$  are each  $m_2 \times n_1$ . These equations will be used in this paper in two contexts, which we briefly recall here.

**Block elimination.** Let  $A_i - \lambda B_i$  be two  $m_i \times n_i$  pencils,  $i = 1, 2$ , that have respectively full column normal rank  $n_1$  and full row normal rank  $m_2$ . Then the problem of block anti-diagonalizing the pencil  $\begin{bmatrix} 0 & A_1 - \lambda B_1 \\ A_2 - \lambda B_2 & H_a - \lambda H_b \end{bmatrix}$ , that is, finding  $X$  and  $Y$  such that

$$\begin{bmatrix} I_{m_1} & 0 \\ -X & I_{m_2} \end{bmatrix} \begin{bmatrix} 0 & A_1 - \lambda B_1 \\ A_2 - \lambda B_2 & H_a - \lambda H_b \end{bmatrix} \begin{bmatrix} I_{n_2} & -Y \\ 0 & I_{n_1} \end{bmatrix} = \begin{bmatrix} 0 & A_1 - \lambda B_1 \\ A_2 - \lambda B_2 & 0 \end{bmatrix}, \tag{2.3}$$

amounts to finding a solution for the generalized Sylvester equation (2.1). It is known that there exists a solution  $(X, Y) \in \mathbb{F}^{m_2 \times m_1} \times \mathbb{F}^{n_2 \times n_1}$  for a particular right hand side  $(H_a, H_b) \in \mathbb{F}^{m_2 \times n_1} \times \mathbb{F}^{m_2 \times n_1}$  if and only if the pencils

$$\begin{bmatrix} 0 & A_1 - \lambda B_1 \\ A_2 - \lambda B_2 & H_a - \lambda H_b \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & A_1 - \lambda B_1 \\ A_2 - \lambda B_2 & 0 \end{bmatrix}$$

are strictly equivalent (i.e. have the same Kronecker structure) [6]. But in order to have a solution for any right hand side  $H_a - \lambda H_b$  one requires the stronger condition that the pencils  $A_1 - \lambda B_1$  and  $A_2 - \lambda B_2$  have no common generalized eigenvalues (see [23]). We recall here the result proven in [23] that is relevant for our work.

**Theorem 2.1** ([23]) *Let the pencils  $A_i - \lambda B_i$  of dimensions  $m_i \times n_i, i = 1, 2$ , be respectively of full column normal rank  $n_1 \leq m_1$  and of full row normal rank  $m_2 \leq n_2$ , and let these two pencils have no common generalized eigenvalues. Then there always exists a solution  $(X, Y)$  to the system of equations (2.3), for any perturbation  $H_a - \lambda H_b$ . Moreover, the generalized eigenvalues of the pencil (2.3) are the union of the generalized eigenvalues of the pencils  $A_i - \lambda B_i, i = 1, 2$ .*

The system is underdetermined if either of the two inequalities  $m_1 \geq n_1$  and  $n_2 \geq m_2$ , is strict. Under the hypotheses of Theorem 2.1, the system (2.2) must be compatible for any right hand side, and hence the Kronecker product matrix in the left hand side of (2.2) must have full row rank  $2m_2n_1$ . A bound for the minimum



Frobenius-norm solution  $(X, Y)$  is then obtained in terms of the smallest singular value  $\sigma_{2m_2n_1}$  of the matrix in (2.2):

$$\|(X, Y)\|_F \leq \frac{\|(H_a, H_b)\|_F}{\sigma_{2m_2n_1} \left( \left[ \begin{array}{c|c} A_1^T \otimes I_{m_2} & I_{n_1} \otimes A_2 \\ \hline B_1^T \otimes I_{m_2} & I_{n_1} \otimes B_2 \end{array} \right] \right)}. \tag{2.4}$$

**Equivalent pencils.** The second problem in this paper where a generalized Sylvester equation as in (2.1) arises is that of strictly equivalent pencils (see e.g. [9]). Let the pencils  $A_i - \lambda B_i$ ,  $i = 1, 2$ , be both of dimension  $m \times n$ , then they are strictly equivalent if and only if there exist invertible matrices  $S$  and  $T$  such that  $S(A_1 - \lambda B_1) = (A_2 - \lambda B_2)T$ . Such pencils must then have the same Kronecker canonical form [9]. We are interested in finding the solution where  $S$  and  $T$  are as close as possible to the identity matrix. This can be achieved by writing the transformation matrices as

$$S = I + X, \quad T = I - Y$$

and then minimizing the Frobenius norm of the pair  $(X, Y)$ . The corresponding equations are then

$$(I + X)(A_1 - \lambda B_1) = (A_2 - \lambda B_2)(I - Y)$$

or, when putting  $H_a - \lambda H_b := (A_2 - \lambda B_2) - (A_1 - \lambda B_1)$ , we finally obtain

$$X(A_1 - \lambda B_1) + (A_2 - \lambda B_2)Y = H_a - \lambda H_b, \tag{2.5}$$

which is again solved by using (2.2). We will use this to “restore” a slightly perturbed pencil  $(A_2 - \lambda B_2) := (A_1 - \lambda B_1) + (H_a - \lambda H_b)$  to its original form  $(A_1 - \lambda B_1)$  using a strict equivalence transformation

$$(I + X)^{-1}(A_2 - \lambda B_2)(I - Y) = A_1 - \lambda B_1 \tag{2.6}$$

that is very close to the identity, when we are sure that both pencils have the same Kronecker canonical form. The bounds for the norm of  $X$  and  $Y$  are in fact given by (2.4) for which we derive exact expressions in the next section. Notice that we can not apply Theorem 2.1 to prove existence of a solution for equation (2.5), since in this case both pencils must have the same generalized eigenvalues and the same normal rank. A sufficient condition for the consistency of (2.5) is that  $A_1 - \lambda B_1$  and  $A_2 - \lambda B_2$  have the same Kronecker canonical form.

The condition that the Kronecker canonical form of a pencil does not change under arbitrary sufficiently small perturbations only holds for very special pencils. In particular, it holds for the Kronecker product of Kronecker blocks times identity matrices, i.e., for  $L_k(\lambda) \otimes I_r$ . This is a consequence of the results in [25], because  $L_k(\lambda) \otimes I_r$  has full-Sylvester-rank by [25, Theorem 4.3(a)] and, then, [25, Theorem 6.6] guarantees that  $L_k(\lambda) \otimes I_r + (H_a - \lambda H_b)$  has the same Kronecker canonical form as  $L_k(\lambda) \otimes I_r$  for

all the perturbations  $(H_a, H_b)$  whose norms are smaller than the bounds in [25, Theorem 6.6]. Since we will solve (2.5)–(2.6) only in the case  $A_1 - \lambda B_1 = L_k(\lambda) \otimes I_r$ , these results prove that (2.5) has a solution for all sufficiently small perturbations  $(H_a, H_b)$  in the cases of interest in this paper.

### 3 Singular value bounds

In the analysis of Sect. 4, we will need upper bounds for the minimum norm solutions of the generalized Sylvester equation (2.1) for pairs of pencils  $(A_i - \lambda B_i)$ ,  $i = 1, 2$ , which all involve Kronecker blocks  $L_k(\lambda) := E_k - \lambda F_k$ , where the  $k \times (k + 1)$  matrices  $E_k$  and  $F_k$  are given by

$$E_k := \begin{bmatrix} 1 & 0 & & & \\ & 1 & 0 & & \\ & & \ddots & \ddots & \\ & & & 1 & 0 \end{bmatrix} \quad \text{and} \quad F_k := \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \end{bmatrix}.$$

To find such upper bounds is equivalent to find lower bounds for the singular values in the denominator of the right hand side of (2.4). We consider the generalized Sylvester equations for the following list of pencil pairs with their smallest singular value of the corresponding linear maps:

1.  $A_1 - \lambda B_1 = A - \lambda I_\ell$  and  $A_2 - \lambda B_2 = L_\epsilon(\lambda) \otimes I_n$ :

$$\omega_1 := \sigma_{2\ell en} \left[ \begin{array}{c|c} A^T \otimes I_{en} & I_\ell \otimes E_\epsilon \otimes I_n \\ \hline I_\ell \otimes I_{en} & I_\ell \otimes F_\epsilon \otimes I_n \end{array} \right]. \tag{3.1}$$

2.  $A_1 - \lambda B_1 = L_\eta^T(\lambda) \otimes I_m$  and  $A_2 - \lambda B_2 = A - \lambda I_\ell$ :

$$\omega_2 := \sigma_{2\eta ml} \left[ \begin{array}{c|c} E_\eta \otimes I_{m\ell} & I_{\eta m} \otimes A \\ \hline F_\eta \otimes I_{m\ell} & I_{\eta m} \otimes I_\ell \end{array} \right]. \tag{3.2}$$

3.  $A_1 - \lambda B_1 = L_\eta^T(\lambda) \otimes I_m$  and  $A_2 - \lambda B_2 = L_\epsilon(\lambda) \otimes I_n$ :

$$\omega_3 := \sigma_{2\eta m en} \left[ \begin{array}{c|c} E_\eta \otimes I_{m en} & I_{\eta m} \otimes E_\epsilon \otimes I_n \\ \hline F_\eta \otimes I_{m en} & I_{\eta m} \otimes F_\epsilon \otimes I_n \end{array} \right]. \tag{3.3}$$

4.  $A_1 - \lambda B_1 = L_k(\lambda) \otimes I_r$  and  $A_2 - \lambda B_2 = L_k(\lambda) \otimes I_r$ :

$$\omega_4 := \sigma_{2(k+1)rkr} \left[ \begin{array}{c|c} E_k^T \otimes I_{rkr} & I_{(k+1)r} \otimes E_k \otimes I_r \\ \hline F_k^T \otimes I_{rkr} & I_{(k+1)r} \otimes F_k \otimes I_r \end{array} \right]. \tag{3.4}$$

In Lemma 3.1 we analyze the first problem and give a lower bound for  $\omega_1$ .

**Lemma 3.1** *Let  $\omega_1$  be the singular value in (3.1). Then*

$$\omega_1 \geq \frac{1}{1 + 2\epsilon \max(1, \|A\|_2^\epsilon)}. \tag{3.5}$$

**Proof** It follows from the properties of singular values of Kronecker products that  $\omega_1$  is also equal to

$$\omega_1 = \sigma_{2\ell\epsilon} \left[ \begin{array}{c|c} A^T \otimes I_\epsilon & I_\ell \otimes E_\epsilon \\ \hline I_\ell \otimes I_\epsilon & I_\ell \otimes F_\epsilon \end{array} \right]$$

and using perfect shuffle permutations we also get

$$\omega_1 = \sigma_{2\epsilon\ell} \left[ \begin{array}{c|c} I_\epsilon \otimes A^T & E_\epsilon \otimes I_\ell \\ \hline I_\epsilon \otimes I_\ell & F_\epsilon \otimes I_\ell \end{array} \right].$$

The smallest singular value  $\sigma_{2\epsilon\ell}$  is larger than the smallest singular value of any  $2\epsilon\ell \times 2\epsilon\ell$  submatrix. Let us take for this the submatrix obtained by dropping the last block column:

$$M = \left[ \begin{array}{c|c} I_\epsilon \otimes A^T & I_\epsilon \otimes I_\ell \\ \hline I_\epsilon \otimes I_\ell & J_\epsilon \otimes I_\ell \end{array} \right], \quad \text{where } J_\epsilon := \begin{bmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \in \mathbb{F}^{\epsilon \times \epsilon}.$$

We can factorize this matrix as

$$M = \left[ \begin{array}{c|c} I_\epsilon \otimes A^T & I_\epsilon \otimes I_\ell \\ \hline I_\epsilon \otimes I_\ell & 0 \end{array} \right] \left[ \begin{array}{c|c} I_{\epsilon\ell} & 0 \\ \hline 0 & I_{\epsilon\ell} - J_\epsilon \otimes A^T \end{array} \right] \left[ \begin{array}{c|c} I_\epsilon \otimes I_\ell & J_\epsilon \otimes I_\ell \\ \hline 0 & I_\epsilon \otimes I_\ell \end{array} \right].$$

Therefore its inverse equals

$$\begin{aligned} M^{-1} &= \left[ \begin{array}{c|c} I_\epsilon \otimes I_\ell & -J_\epsilon \otimes I_\ell \\ \hline 0 & I_\epsilon \otimes I_\ell \end{array} \right] \left[ \begin{array}{c|c} I_{\epsilon\ell} & 0 \\ \hline 0 & (I_{\epsilon\ell} - J_\epsilon \otimes A^T)^{-1} \end{array} \right] \left[ \begin{array}{c|c} 0 & I_\epsilon \otimes I_\ell \\ \hline I_\epsilon \otimes I_\ell & -I_\epsilon \otimes A^T \end{array} \right] \\ &= \left[ \begin{array}{c|c} I_{\epsilon\ell} & \\ \hline 0 & \end{array} \right] [0 | I_{\epsilon\ell}] + \left[ \begin{array}{c|c} -J_\epsilon \otimes I_\ell & \\ \hline I_{\epsilon\ell} & \end{array} \right] (I_{\epsilon\ell} - J_\epsilon \otimes A^T)^{-1} [I_{\epsilon\ell} | -I_\epsilon \otimes A^T]. \end{aligned}$$

It then follows that

$$\|M^{-1}\|_2 \leq 1 + \sqrt{2} \sqrt{1 + \|A\|_2^2} \left[ 1 + \|A\|_2 + \|A\|_2^2 + \dots + \|A\|_2^{\epsilon-1} \right],$$

since

$$(I_{\epsilon\ell} - J_\epsilon \otimes A^T)^{-1} = \sum_{i=0}^{\epsilon-1} J_\epsilon^i \otimes A^{iT}.$$

In particular, for  $\|A\|_2 \leq 1$  we obtain the bound  $\|M^{-1}\|_2 \leq 1 + 2\epsilon$ , while for  $\|A\|_2 > 1$  we obtain the bound  $\|M^{-1}\|_2 \leq 1 + 2\epsilon\|A\|_2^\epsilon$ . This finally yields the inequality

$$\omega_1 \geq \frac{1}{1 + 2\epsilon \max(1, \|A\|_2^\epsilon)}.$$

□

The second generalized Sylvester equation is essentially a shuffled version of the first equation and the analysis is therefore completely analogous. This immediately yields Lemma 3.2.

**Lemma 3.2** *Let  $\omega_2$  be the singular value in (3.2). Then*

$$\omega_2 \geq \frac{1}{1 + 2\eta \max(1, \|A\|_2^\eta)}. \tag{3.6}$$

The third generalized Sylvester equation was analyzed in [7] and its associated smallest singular value is exactly equal to  $\omega_3 = 2 \sin(\pi/(4 \min(\epsilon, \eta) + 2))$  if  $\epsilon \neq \eta$ , and to  $2 \sin(\pi/4\eta)$  if  $\epsilon = \eta$ . Notice that we can assume  $\min(\epsilon, \eta) \geq 1$  since otherwise the equation is void. For  $\epsilon \neq \eta$  we then obtain  $\omega_3 \geq \frac{3}{2 \min(\epsilon, \eta) + 1}$  since  $\sin x \geq 3x/\pi$  for  $0 \leq x \leq \pi/6$ , and for  $\epsilon = \eta$  we then obtain  $\omega_3 \geq \frac{\sqrt{2}}{\eta}$  since  $\sin x \geq 2\sqrt{2}x/\pi$  for  $0 \leq x \leq \pi/4$ . We have also that  $2\eta = \epsilon + \eta$  if  $\epsilon = \eta$  and  $2 \min(\epsilon, \eta) + 1 \leq \epsilon + \eta$  if  $\epsilon \neq \eta$ , which finally yields the lower bound in Lemma 3.3 for  $\omega_3$ .

**Lemma 3.3** *Let  $\omega_3$  be the singular value in (3.3). Then*

$$\omega_3 \geq \frac{2\sqrt{2}}{\epsilon + \eta}. \tag{3.7}$$

In Lemma 3.4, we give a lower bound for the smallest singular value  $\omega_4$  corresponding to the fourth generalized Sylvester equation.

**Lemma 3.4** *Let  $\omega_4$  be the singular value in (3.4). Then*

$$\omega_4 \geq \frac{3}{4k - 1}. \tag{3.8}$$

**Proof** We prove first that  $\omega_4 = 2 \sin(\pi/(8k - 2))$ . This is obtained as follows. We can again use the properties of Kronecker products to prove that

$$\omega_4 = \sigma_{2k(k+1)} \left[ \begin{array}{c|c} E_k^T \otimes I_k & I_{(k+1)} \otimes E_k \\ \hline F_k^T \otimes I_k & I_{(k+1)} \otimes F_k \end{array} \right].$$

This matrix can be transformed by row and column permutations to the direct sum of smaller matrices:

$$M_1 \oplus M_1 \oplus M_3 \oplus M_3 \oplus \dots \oplus M_{2k-1} \oplus M_{2k-1} \oplus N_{2k},$$

see  $A$ , where the blocks

$$M_k := \begin{bmatrix} 1 & 1 & & & \\ & & 1 & \ddots & \\ & & & \ddots & 1 \\ & & & & & 1 \\ & & & & & & 1 \end{bmatrix} \in \mathbb{F}^{k \times k}, \quad N_k := \begin{bmatrix} 1 & 1 & & & \\ & & 1 & \ddots & \\ & & & \ddots & 1 \\ & & & & & 1 \\ & & & & & & 1 & 1 \end{bmatrix} \in \mathbb{F}^{k \times (k+1)} \quad (3.9)$$

have as smallest singular values  $2 \sin \frac{\pi}{4k+2}$  and  $2 \sin \frac{\pi}{2k+2}$ , respectively (see [7, Proof of Proposition B.4]). The smallest singular value therefore corresponds to  $M_{2k-1}$  and equals  $\omega_4 = 2 \sin(\pi/(8k - 2))$ . For  $k \geq 1$ , we use again that  $\sin x \geq 3x/\pi$  for  $0 \leq x \leq \pi/6$ , to obtain the bound  $\omega_4 \geq \frac{3}{4k-1}$ .  $\square$

### 4 Restoring the rational structure of the linearization after perturbations

We now consider perturbations of the following block Kronecker linearization introduced in (1.2)

$$S(\lambda) := \begin{bmatrix} S_{11}(\lambda) & S_{12}(\lambda) & S_{13}(\lambda) \\ S_{21}(\lambda) & S_{22}(\lambda) & 0 \\ S_{31}(\lambda) & 0 & 0 \end{bmatrix} := \begin{bmatrix} M(\lambda) & \widehat{K}_2^T C & K_2^T(\lambda) \\ B \widehat{K}_1 & A - \lambda I_\ell & 0 \\ K_1(\lambda) & 0 & 0 \end{bmatrix}, \quad (4.1)$$

where  $S_{13}(\lambda)$  is  $(\eta + 1)m \times \eta m$  and has full column rank  $\eta m$ ,  $S_{22}(\lambda)$  is  $\ell \times \ell$  and is a regular pencil,  $S_{31}(\lambda)$  is  $\epsilon n \times (\epsilon + 1)n$  and has full row rank  $\epsilon n$ , and where no two of these three pencils have common generalized eigenvalues. As explained in the introduction, if the state-space triple  $\{A, B, C\}$  is minimal, then  $S(\lambda)$  is a strong linearization of the  $m \times n$  rational matrix

$$R(\lambda) = C(\lambda I_\ell - A)^{-1} B + (\Lambda_\eta(\lambda) \otimes I_m)^T M(\lambda) (\Lambda_\epsilon(\lambda) \otimes I_n). \quad (4.2)$$

Except in Sect. 4.6, we assume in this section that  $\max(\eta, \epsilon) > 0$ . This means that the degree  $d = \epsilon + \eta + 1$  of the polynomial part  $D(\lambda) = (\Lambda_\eta(\lambda) \otimes I_m)^T M(\lambda) (\Lambda_\epsilon(\lambda) \otimes I_n)$  of  $R(\lambda)$  is greater than 1 and that at least one of the blocks  $K_1(\lambda)$  or  $K_2(\lambda)$  is not an empty matrix. The degenerate case in which  $\epsilon = 0$  and  $\eta = 0$  will be studied in Sect. 4.6.

Since  $S(\lambda)$  is a strong linearization of  $R(\lambda)$ ,  $S(\lambda)$  has the exact eigenstructure of the finite zeros of  $R(\lambda)$ , and its infinite zero structure as well as its left and right null-space structure can be correctly retrieved from the pencil via simple constant shifts, as explained in the introduction. In order to compute this eigenstructure, we make use of the staircase algorithm [21], followed by the  $QZ$  algorithm [16], on  $S(\lambda)$ . The backward stability of these two algorithms guarantees in fact that we computed the

exact eigenstructure of a slightly perturbed pencil

$$\widehat{S}(\lambda) := S(\lambda) + \Delta_S(\lambda), \quad \Delta_S(\lambda) := \begin{bmatrix} \Delta_{11}(\lambda) & \Delta_{12}(\lambda) & \Delta_{13}(\lambda) \\ \Delta_{21}(\lambda) & \Delta_{22}(\lambda) & \Delta_{23}(\lambda) \\ \Delta_{31}(\lambda) & \Delta_{32}(\lambda) & \Delta_{33}(\lambda) \end{bmatrix}, \quad (4.3)$$

where the pencil  $\Delta_S(\lambda)$  has a norm which is much smaller than the norm of  $S(\lambda)$ . More precisely,  $\|\Delta_S(\lambda)\|_F = O(\epsilon_M) \|S(\lambda)\|_F$ , where  $\epsilon_M$  is the machine precision of the computer. But even for very small perturbations, the structure of the pencil  $\widehat{S}(\lambda)$  is lost, and therefore also the connection between  $\widehat{S}(\lambda)$  and some rational matrix  $\widehat{R}(\lambda)$  is lost. In this section, we will show that this structure can be restored, without affecting the computed eigenstructure. For this, one needs only to find a strict equivalence transformation that is close to the identity and restores the structure of  $\widehat{S}(\lambda)$  to a new pencil  $\widetilde{S}(\lambda)$  that is a block Kronecker linearization, with the same parameters  $\epsilon$  and  $\eta$  as  $S(\lambda)$ , of a rational matrix  $\widetilde{R}(\lambda)$ :

$$\widetilde{S}(\lambda) := (I - X)(S(\lambda) + \Delta_S(\lambda))(I - Y) = \begin{bmatrix} \widetilde{M}(\lambda) & \widehat{K}_2^T \widetilde{C} & K_2^T(\lambda) \\ \widetilde{B} \widehat{K}_1 & \widetilde{A} - \lambda I_\ell & 0 \\ K_1(\lambda) & 0 & 0 \end{bmatrix}. \quad (4.4)$$

We will see that if  $\|\Delta_S(\lambda)\|_F$  is sufficiently small, then the perturbed system triple  $\{\widetilde{A}, \widetilde{B}, \widetilde{C}\}$  is very close to the unperturbed minimal one  $\{A, B, C\}$  and, so,  $\{\widetilde{A}, \widetilde{B}, \widetilde{C}\}$  is still minimal, since minimality is a generic property equivalent to the controllability matrix having full row rank and the observability matrix having full column rank [14, Chapter 6]. Observe that according to [2], or the discussion in the introduction,  $\widetilde{S}(\lambda)$  is a strong linearization of the  $m \times n$  rational matrix

$$\begin{aligned} \widetilde{R}(\lambda) &:= \widetilde{C}(\lambda I_\ell - \widetilde{A})^{-1} \widetilde{B} + (\Lambda_\eta(\lambda) \otimes I_m)^T \widetilde{M}(\lambda) (\Lambda_\epsilon(\lambda) \otimes I_n) \\ &=: \widetilde{C}(\lambda I_\ell - \widetilde{A})^{-1} \widetilde{B} + \widetilde{D}(\lambda). \end{aligned} \quad (4.5)$$

Since the eigenstructures of the pencils  $\widehat{S}(\lambda)$  and  $\widetilde{S}(\lambda)$  are identical, the results in this section prove that the computed finite eigenvalues of  $S(\lambda)$  and their partial multiplicities are the exact finite zeros and their partial multiplicities of  $\widetilde{R}(\lambda)$ , the computed right (resp. left) minimal indices of  $S(\lambda)$  minus  $\epsilon$  (resp.  $\eta$ ) are the exact right (resp. left) minimal indices of  $\widetilde{R}(\lambda)$ , and, if a number  $\nu_r$  of right minimal indices of  $S(\lambda)$  have been computed, then the computed  $n - \nu_r$  largest partial multiplicities at infinity of  $S(\lambda)$  minus  $d$  are the exact structural indices at infinity of  $\widetilde{R}(\lambda)$ . This is a very strong backward error result for the computation of the eigenstructure of  $R(\lambda)$  in the case we are able to prove that  $\|\widetilde{A} - A\|_F$ ,  $\|\widetilde{B} - B\|_F$ ,  $\|\widetilde{C} - C\|_F$  and  $\|\widetilde{D}(\lambda) - D(\lambda)\|_F$  are very small.

The restoration of the structure in  $\widehat{S}(\lambda)$  will be done in three steps, each of them involving a strict equivalence transformation close to the identity:

- **Step 1:** We restore the block anti-triangular structure of the perturbed pencil  $\widehat{S}(\lambda)$ , i.e., the blocks (2,3), (3,2) and (3,3) are transformed to become 0.

- **Step 2:** We take care of the anti-diagonal blocks (1,3), (2,2) and (3,1), by restoring their 0 and  $I$  block matrices.
- **Step 3:** We restore the special structure of the blocks (1,2) and (2,1).

At each step  $k$ , for  $k = 1, 2, 3$ , we obtain a pencil

$$\widehat{S}_k(\lambda) := (I - X_k)\widehat{S}_{k-1}(\lambda)(I - Y_k) := \widehat{S}_{k-1}(\lambda) + \Delta_k(\lambda), \tag{4.6}$$

where  $\widehat{S}_0(\lambda) := \widehat{S}(\lambda)$  and  $\Delta_0(\lambda) := \Delta_S(\lambda)$ :

$$S(\lambda) \xrightarrow{+\Delta_0(\lambda)} \widehat{S}(\lambda) = \widehat{S}_0(\lambda) \xrightarrow{+\Delta_1(\lambda)} \widehat{S}_1(\lambda) \xrightarrow{+\Delta_2(\lambda)} \widehat{S}_2(\lambda) \xrightarrow{+\Delta_3(\lambda)} \widehat{S}_3(\lambda) = \widetilde{S}(\lambda).$$

We will compute bounds for  $\|(X_k, Y_k)\|_F$  as a function of  $\|\widehat{S}_{k-1}(\lambda)\|_F$ , where the Frobenius norms are computed as defined in the introduction. Moreover, we define the cumulative errors:

$$\begin{aligned} \Delta_k^{old}(\lambda) &:= \sum_{i=0}^{k-1} \Delta_i(\lambda), \text{ and} \\ \Delta_k^{new}(\lambda) &:= \Delta_k^{old}(\lambda) + \Delta_k(\lambda) = \sum_{i=0}^k \Delta_i(\lambda), \end{aligned} \tag{4.7}$$

and we will also compute bounds for the Frobenius norm of these error pencils. In our analysis, we will assume that  $\delta := \frac{\|\Delta_S(\lambda)\|_F}{\|S(\lambda)\|_F}$  is very small, since in practice it is of the order of the machine precision  $\epsilon_M$ , and we will neglect, when appropriate, terms of order larger than 1 in  $\delta$  to simplify our bounds. Moreover, we will assume that  $\delta$  is sufficiently small for guaranteeing that all the steps in the analysis can be performed, for instance, for guaranteeing that some perturbed matrices are invertible. In particular, we have Lemma 4.1 for computing bounds of the growth of the cumulative errors  $\Delta_k^{new}(\lambda)$ .

**Lemma 4.1** *At each step  $k$  of our method, the perturbation  $\Delta_k^{new}(\lambda)$  can be bounded by*

$$\|\Delta_k^{new}(\lambda)\|_F \leq \sqrt{2}\|\widehat{S}_{k-1}(\lambda)\|_2\|(X_k, Y_k)\|_F + \|\Delta_k^{old}(\lambda)\|_F + \mathcal{O}(\delta^2),$$

assuming that  $\|(X_k, Y_k)\|_F$  is of the order of  $\|\Delta_S(\lambda)\|_F$ .

**Proof** At step  $k$ , we have  $\widehat{S}_k(\lambda) = (I - X_k)\widehat{S}_{k-1}(\lambda)(I - Y_k)$ . Therefore

$$\Delta_k^{new}(\lambda) = \Delta_k^{old}(\lambda) - X_k\widehat{S}_{k-1}(\lambda) - \widehat{S}_{k-1}(\lambda)Y_k + X_k\widehat{S}_{k-1}(\lambda)Y_k.$$

It then follows that the increment (up to  $\mathcal{O}(\delta^2)$  terms) is given by

$$-X_kS_a - S_aY_k + \lambda(X_kS_b + S_bY_k) + \mathcal{O}(\delta^2),$$

where  $S_a - \lambda S_b := \widehat{S}_{k-1}(\lambda)$ . We then use the inequalities

$$\|X_k S_a + S_a Y_k\|_F^2 \leq 2\|S_a\|_2^2 \|(X_k, Y_k)\|_F^2, \quad \|X_k S_b + S_b Y_k\|_F^2 \leq 2\|S_b\|_2^2 \|(X_k, Y_k)\|_F^2$$

and the definition for  $\|\widehat{S}_{k-1}(\lambda)\|_2$ , to finally get the required bound. □

### 4.1 Step 1: Restoring the block anti-triangular structure

For step 1, that is, restoring the block anti-triangular structure of  $S(\lambda)$  in the perturbed matrix pencil (4.3), we apply a strict equivalence transformation of the type:

$$\begin{bmatrix} I_{(\eta+1)m} & 0 & 0 \\ -X_{21} & I_\ell & 0 \\ -X_{31} & -X_{32} & I_{\epsilon n} \end{bmatrix} \widehat{S}(\lambda) \begin{bmatrix} I_{(\epsilon+1)n} & -Y_{12} & -Y_{13} \\ 0 & I_\ell & -Y_{23} \\ 0 & 0 & I_{\eta m} \end{bmatrix} \tag{4.8}$$

in order to eliminate the perturbations  $\Delta_{23}(\lambda)$ ,  $\Delta_{32}(\lambda)$  and  $\Delta_{33}(\lambda)$  of the error matrix pencil  $\Delta_0(\lambda)$ . The notation  $\widehat{S}_{ij}^a - \lambda \widehat{S}_{ij}^b := \widehat{S}_{ij} := \widehat{S}_{ij}(\lambda)$  will be used in this section to refer to sub-blocks of  $\widehat{S}_0(\lambda)$ . Let us write down the equations that we get by setting the blocks (2,3), (3,2) and (3,3) of the matrix in (4.8) equal to zero:

$$\begin{aligned} \Delta_{23}(\lambda) &:= \Delta_{23}^a - \lambda \Delta_{23}^b = X_{21} \widehat{S}_{13} + \widehat{S}_{21} Y_{13} + \widehat{S}_{22} Y_{23} - X_{21} \widehat{S}_{11} Y_{13} - X_{21} \widehat{S}_{12} Y_{23}, \\ \Delta_{32}(\lambda) &:= \Delta_{32}^a - \lambda \Delta_{32}^b = \widehat{S}_{31} Y_{12} + X_{31} \widehat{S}_{12} + X_{32} \widehat{S}_{22} - X_{31} \widehat{S}_{11} Y_{12} - X_{32} \widehat{S}_{21} Y_{12}, \\ \Delta_{33}(\lambda) &:= \Delta_{33}^a - \lambda \Delta_{33}^b = X_{31} \widehat{S}_{13} + \widehat{S}_{31} Y_{13} + X_{32} \Delta_{23} + \Delta_{32} Y_{23} \\ &\quad - X_{31} \widehat{S}_{11} Y_{13} - X_{32} \widehat{S}_{21} Y_{13} - X_{31} \widehat{S}_{12} Y_{23} - X_{32} \widehat{S}_{22} Y_{23}. \end{aligned} \tag{4.9}$$

This is a system of nonlinear matrix equations for the six matrix unknowns  $X_{21}, X_{31}, X_{32}, Y_{12}, Y_{13}$  and  $Y_{23}$ . We will show that it is consistent and that it has a solution for which the norms of the unknowns are of the order of  $\|\Delta_0(\lambda)\|_F$ , which implies that there are many terms in the above three equations that are of second order.

Using Kronecker product and the  $\text{vec}(\cdot)$  notation, the system of matrix equations (4.9) can be rewritten as:

$$\underbrace{\begin{bmatrix} \text{vec}(\Delta_{23}^a) \\ \text{vec}(\Delta_{23}^b) \\ \text{vec}(\Delta_{32}^a) \\ \text{vec}(\Delta_{32}^b) \\ \text{vec}(\Delta_{33}^a) \\ \text{vec}(\Delta_{33}^b) \end{bmatrix}}_{:=c} = (T + \Delta T) \underbrace{\begin{bmatrix} \text{vec}(X_{21}) \\ \text{vec}(Y_{23}) \\ \text{vec}(X_{32}) \\ \text{vec}(Y_{12}) \\ \text{vec}(X_{31}) \\ \text{vec}(Y_{13}) \end{bmatrix}}_{:=x} - \underbrace{\begin{bmatrix} \text{vec}(Z_1) \\ \text{vec}(Z_2) \\ \text{vec}(Z_3) \\ \text{vec}(Z_4) \\ \text{vec}(Z_5) \\ \text{vec}(Z_6) \end{bmatrix}}_{:=z}, \tag{4.10}$$

where

$$Z_1 := X_{21} \widehat{S}_{11}^a Y_{13} + X_{21} \widehat{S}_{12}^a Y_{23}, \quad Z_2 := X_{21} \widehat{S}_{11}^b Y_{13} + X_{21} \widehat{S}_{12}^b Y_{23},$$



$$\begin{aligned}
 Z_3 &:= X_{31} \widehat{S}_{11}^a Y_{12} + X_{32} \widehat{S}_{21}^a Y_{12}, & Z_4 &:= X_{31} \widehat{S}_{11}^b Y_{12} + X_{32} \widehat{S}_{21}^b Y_{12}, \\
 Z_5 &:= X_{31} \widehat{S}_{11}^a Y_{13} + X_{32} \widehat{S}_{21}^a Y_{13} + X_{31} \widehat{S}_{12}^a Y_{23} + X_{32} \widehat{S}_{22}^a Y_{23}, \\
 Z_6 &:= X_{31} \widehat{S}_{11}^b Y_{13} + X_{32} \widehat{S}_{21}^b Y_{13} + X_{31} \widehat{S}_{12}^b Y_{23} + X_{32} \widehat{S}_{22}^b Y_{23}, \\
 \Delta T &= \begin{bmatrix} \Delta_{13}^{aT} \otimes I_\ell I_{\eta m} \otimes \Delta_{22}^a & 0 & 0 & 0 & I_{\eta m} \otimes \Delta_{21}^a \\ \Delta_{13}^{bT} \otimes I_\ell I_{\eta m} \otimes \Delta_{22}^b & 0 & 0 & 0 & I_{\eta m} \otimes \Delta_{21}^b \\ 0 & 0 & \Delta_{22}^{aT} \otimes I_{\epsilon n} I_\ell \otimes \Delta_{31}^a & \Delta_{12}^{aT} \otimes I_{\epsilon n} & 0 \\ 0 & 0 & \Delta_{22}^{bT} \otimes I_{\epsilon n} I_\ell \otimes \Delta_{31}^b & \Delta_{12}^{bT} \otimes I_{\epsilon n} & 0 \\ 0 & I_{\eta m} \otimes \Delta_{32}^a & \Delta_{23}^{aT} \otimes I_{\epsilon n} & 0 & \Delta_{13}^{aT} \otimes I_{\epsilon n} I_{\eta m} \otimes \Delta_{31}^a \\ 0 & I_{\eta m} \otimes \Delta_{32}^b & \Delta_{23}^{bT} \otimes I_{\epsilon n} & 0 & \Delta_{13}^{bT} \otimes I_{\epsilon n} I_{\eta m} \otimes \Delta_{31}^b \end{bmatrix},
 \end{aligned}$$

and

$$T = \begin{bmatrix} T_{11} & T_{12} & 0 & 0 & 0 & T_{16} \\ T_{21} & T_{22} & 0 & 0 & 0 & 0 \\ & & T_{33} & T_{34} & T_{35} & 0 \\ & & T_{43} & T_{44} & 0 & 0 \\ & & & & T_{55} & T_{56} \\ & & & & T_{65} & T_{66} \end{bmatrix},$$

with

$$\begin{aligned}
 \begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} &:= \left[ \begin{array}{c|c} E_\eta \otimes I_{m\ell} & I_{\eta m} \otimes A \\ \hline F_\eta \otimes I_{m\ell} & I_{\eta m} \otimes I_\ell \end{array} \right], & \begin{bmatrix} T_{33} & T_{34} \\ T_{43} & T_{44} \end{bmatrix} &:= \left[ \begin{array}{c|c} A^T \otimes I_{\epsilon n} & I_\ell \otimes E_\epsilon \otimes I_n \\ \hline I_\ell \otimes I_{\epsilon n} & I_\ell \otimes F_\epsilon \otimes I_n \end{array} \right], \\
 \begin{bmatrix} T_{55} & T_{56} \\ T_{65} & T_{66} \end{bmatrix} &:= \left[ \begin{array}{c|c} E_\eta \otimes I_{m\epsilon n} & I_{\eta m} \otimes E_\epsilon \otimes I_n \\ \hline F_\eta \otimes I_{m\epsilon n} & I_{\eta m} \otimes F_\epsilon \otimes I_n \end{array} \right], & \begin{cases} T_{16} := I_{\eta m} \otimes e_{\epsilon+1}^T \otimes B \\ T_{35} := e_{\eta+1}^T \otimes C^T \otimes I_{\epsilon n} \end{cases}.
 \end{aligned}$$

We emphasize that the matrices in the two lines above are precisely those appearing in equations (3.2), (3.1) and (3.3), respectively. Observe also that the matrix  $\Delta T$  encodes those linear terms of the nonlinear system (4.9) related to blocks of the perturbation pencil  $\Delta_S(\lambda)$  in (4.3), while the matrix  $T$  encodes those linear terms of (4.9) related to blocks of the unperturbed block Kronecker linearization  $S(\lambda)$  in (4.3).

The smallest singular value of  $T$  and the 2–norm of  $\Delta T$  will be needed in the analysis of the bound for the structured backward errors. More precisely for proving that (4.9) is consistent and bounding the norm of one of its solutions. A lower bound for  $\sigma_{\min}(T)$  and an upper bound for  $\|\Delta T\|_2$  are given in Lemma 4.2 and Lemma 4.3, respectively.

**Lemma 4.2** *Let  $T$  be the matrix in (4.10). Let  $\alpha := 1 + 2\epsilon \max(1, \|A\|_2^\epsilon)$ ,  $\beta := 1 + 2\eta \max(1, \|A\|_2^\eta)$ ,  $\gamma := \frac{\epsilon+\eta}{2\sqrt{2}}$  and  $s := \max(\alpha, \beta, \gamma) + \gamma(\beta\|B\|_2 + \alpha\|C\|_2)$  then*

$$\sigma_{\min}(T) \geq \frac{1}{s}.$$

**Proof** If we partition the matrix  $T$  as a block triangular matrix

$$T = \begin{bmatrix} T_1 & 0 & T_B \\ & T_2 & T_C \\ & & T_3 \end{bmatrix},$$

then the diagonal blocks have full row ranks because their smallest singular values are strictly larger than zero according to Lemmas 3.2, 3.1 and 3.3, respectively. Therefore, they are right invertible, with Moore–Penrose pseudoinverses  $T_i^r$  satisfying  $T_i T_i^r = I$ , for  $i = 1, 2, 3$ . Moreover,  $\|T_1^r\|_2 = \omega_1^{-1}$ ,  $\|T_2^r\|_2 = \omega_2^{-1}$  and  $\|T_3^r\|_2 = \omega_3^{-1}$ , with  $\omega_1, \omega_2$  and  $\omega_3$  as in (3.1), (3.2) and (3.3). A right inverse  $T^r$  for  $T$  is given by

$$T^r = \begin{bmatrix} T_1^r & 0 & -T_1^r T_B T_3^r \\ & T_2^r & -T_2^r T_C T_3^r \\ & & T_3^r \end{bmatrix}$$

since  $TT^r = I$ . It then follows that the smallest singular value of  $T$  is lower bounded by  $\|T^r\|_2^{-1}$ . This right inverse can be written as the sum of three matrices (one of them being  $\text{diag}(T_1^r, T_2^r, T_3^r)$ ), and the 2-norm of each of them can be upper bounded using the results of Sect. 3 and the fact that  $\|T_B\|_2 = \|B\|_2$  and  $\|T_C\|_2 = \|C\|_2$ . We then obtain the bound:

$$\begin{aligned} \sigma_{\min}(T) &\geq 1 / \left[ \max(\omega_1^{-1}, \omega_2^{-1}, \omega_3^{-1}) + \omega_3^{-1}(\omega_2^{-1}\|B\|_2 + \omega_1^{-1}\|C\|_2) \right] \\ &\geq 1 / \left[ \max(\alpha, \beta, \gamma) + \gamma(\beta\|B\|_2 + \alpha\|C\|_2) \right], \end{aligned}$$

by taking into account inequalities (3.5), (3.6), (3.7). □

**Lemma 4.3** *Let  $\Delta T$  be the matrix in (4.10) and let  $\Delta_S(\lambda)$  be the pencil in (4.3). Then*

$$\|\Delta T\|_2 \leq \sqrt{3}\|\Delta_S(\lambda)\|_2.$$

**Proof** We consider a permutation matrix  $P$  such that

$$\begin{aligned} \Delta T &= \begin{bmatrix} \Delta_{13}^a T \otimes I_\ell & 0 & 0 & 0 & I_{\eta m} \otimes \Delta_{22}^a & I_{\eta m} \otimes \Delta_{21}^a \\ \Delta_{13}^b T \otimes I_\ell & 0 & 0 & 0 & I_{\eta m} \otimes \Delta_{22}^b & I_{\eta m} \otimes \Delta_{21}^b \\ 0 & I_\ell \otimes \Delta_{31}^a & \Delta_{22}^a T \otimes I_{\epsilon n} & \Delta_{12}^a T \otimes I_{\epsilon n} & 0 & 0 \\ 0 & I_\ell \otimes \Delta_{31}^b & \Delta_{22}^b T \otimes I_{\epsilon n} & \Delta_{12}^b T \otimes I_{\epsilon n} & 0 & 0 \\ 0 & 0 & \Delta_{23}^a T \otimes I_{\epsilon n} & \Delta_{13}^a T \otimes I_{\epsilon n} & I_{\eta m} \otimes \Delta_{32}^a & I_{\eta m} \otimes \Delta_{31}^a \\ 0 & 0 & \Delta_{23}^b T \otimes I_{\epsilon n} & \Delta_{13}^b T \otimes I_{\epsilon n} & I_{\eta m} \otimes \Delta_{32}^b & I_{\eta m} \otimes \Delta_{31}^b \end{bmatrix} P \\ &:= [T_1|T_2|T_3] P. \end{aligned}$$

Using properties of norms and Kronecker products (see [13, Chapter 4]) we have that  $\|T_i\|_2 \leq \|\Delta_S(\lambda)\|_2$  for  $i = 1, 2, 3$ . Finally, by [12, Lemma 3.5],

$$\|\Delta T\|_2 \leq \sqrt{3} \max\{\|T_1\|_2, \|T_2\|_2, \|T_3\|_2\} \leq \sqrt{3}\|\Delta_S(\lambda)\|_2.$$

□

In order to prove that the system of nonlinear matrix equations (4.9) is consistent, first, we remove quadratic terms in  $X_{ij}$  and  $Y_{ij}$  of these equations and we get the following system of linear equations:

$$\begin{aligned} \Delta_{23}(\lambda) &= X_{21}\widehat{S}_{13} + \widehat{S}_{21}Y_{13} + \widehat{S}_{22}Y_{23}, \\ \Delta_{32}(\lambda) &= \widehat{S}_{31}Y_{12} + X_{31}\widehat{S}_{12} + X_{32}\widehat{S}_{22}, \\ \Delta_{33}(\lambda) &= X_{31}\widehat{S}_{13} + \widehat{S}_{31}Y_{13} + X_{32}\Delta_{23} + \Delta_{32}Y_{23}. \end{aligned}$$

This linear system of matrix equations can be rewritten as the underdetermined linear system:

$$(T + \Delta T)x = c, \tag{4.11}$$

with the same notation as in (4.10). Next we prove that (4.11) is consistent for any right hand side if  $\Delta T$  is sufficiently small. From the minimum norm solution of (4.11), we obtain in Theorem 4.1 that there exists a solution for the quadratic system (4.10) under certain conditions and bound its norm.

**Lemma 4.4** *Let  $(T + \Delta T)x = c$  be the underdetermined linear system in (4.11), and let us assume that  $\sigma_{\min}(T) > \|\Delta T\|_2$ . Then  $(T + \Delta T)x = c$  is consistent and its minimum norm solution  $(X^0, Y^0) := (X_{21}^0, X_{31}^0, X_{32}^0, Y_{12}^0, Y_{13}^0, Y_{23}^0)$  satisfies*

$$\|(X^0, Y^0)\|_F \leq \frac{1}{\sigma} \|(\Delta_{23}(\lambda), \Delta_{32}(\lambda), \Delta_{33}(\lambda))\|_F,$$

where  $\sigma := \sigma_{\min}(T) - \|\Delta T\|_2$ .

**Proof** Analogous proof as for [7, Lemma 5.6]. □

The notation  $\sigma := \sigma_{\min}(T) - \|\Delta T\|_2$  has been chosen to remind that  $\sigma$  is a lower bound for the smallest singular value of  $T + \Delta T$ , since  $\sigma_{\min}(T + \Delta T) \geq \sigma_{\min}(T) - \|\Delta T\|_2$  by Weyl’s perturbation theorem for singular values [13, Theorem 3.3.16]. Lemma 4.5 gives a sufficient condition on  $\|\Delta_S(\lambda)\|_2$  that guarantees  $\sigma > 0$  and, hence, that allows us to apply Lemma 4.4.

**Lemma 4.5** *Consider the real number  $s$  defined as in Lemma 4.2. Let  $T$  and  $\Delta T$  be the matrices in (4.11), and let  $\Delta_S(\lambda)$  be the pencil in (4.3). If  $\|\Delta_S(\lambda)\|_2 < \frac{1}{2s}$  then*

$$\sigma = \sigma_{\min}(T) - \|\Delta T\|_2 > \frac{2 - \sqrt{3}}{2s} > 0.$$

**Proof** If  $\|\Delta_S(\lambda)\|_2 < \frac{1}{2s}$  we have, by Lemmas 4.2 and 4.3, that  $\sigma_{\min}(T) - \|\Delta T\|_2 \geq \frac{1}{s} - \sqrt{3}\|\Delta_S(\lambda)\|_2 > \frac{2 - \sqrt{3}}{2s} > 0$ . □

Theorem 4.1 establishes conditions in order the system of matrix equations (4.9) to have a solution as we announced. Moreover, it gives an upper bound for the Frobenius norm of this solution. We remark that Theorem 4.1 is similar to [7, Theorem 5.8], though the involved systems of matrix equations are very different from each other. Therefore, some details in the proof of Theorem 4.1 are omitted since they can be found in [7].

**Theorem 4.1** *There exists a solution  $(X, Y) := (X_{21}, X_{31}, X_{32}, Y_{12}, Y_{13}, Y_{23})$  of the quadratic system of equations (4.10) satisfying*

$$\|(X, Y)\|_F \leq 2 \frac{\theta}{\sigma},$$

whenever

$$\sigma > 0 \quad \text{and} \quad \frac{\theta\omega}{\sigma^2} < \frac{1}{4}, \tag{4.12}$$

where  $\omega := \|(M(\lambda), A - \lambda I_\ell, B, C)\|_F + \|\Delta_S(\lambda)\|_F$ ,  $\theta := \|(\Delta_{23}(\lambda), \Delta_{32}(\lambda), \Delta_{33}(\lambda))\|_F$ , and  $\sigma = \sigma_{\min}(T) - \|\Delta T\|_2$ .

**Proof** Since  $\sigma > 0$ , we can apply Lemma 4.4 and consider  $(X^0, Y^0)$  the minimum norm solution of (4.11). Let

$$x_0 := [\text{vec}(X_{21}^0)^T \text{vec}(Y_{23}^0)^T \text{vec}(X_{32}^0)^T \text{vec}(Y_{12}^0)^T \text{vec}(X_{31}^0)^T \text{vec}(Y_{13}^0)^T]^T.$$

Let us define the sequence  $\{(X^i, Y^i) := (X_{21}^i, X_{31}^i, X_{32}^i, Y_{12}^i, Y_{13}^i, Y_{23}^i)\}_{i=0}^\infty$  such that, for each  $i > 0$ ,  $(X^i, Y^i)$  is the minimum norm solution of the linear system

$$(T + \Delta T) \begin{bmatrix} \text{vec}(X_{21}^i) \\ \text{vec}(Y_{23}^i) \\ \text{vec}(X_{32}^i) \\ \text{vec}(Y_{12}^i) \\ \text{vec}(X_{31}^i) \\ \text{vec}(Y_{13}^i) \end{bmatrix} = c + \begin{bmatrix} \text{vec}(Z_1^{i-1}) \\ \text{vec}(Z_2^{i-1}) \\ \text{vec}(Z_3^{i-1}) \\ \text{vec}(Z_4^{i-1}) \\ \text{vec}(Z_5^{i-1}) \\ \text{vec}(Z_6^{i-1}) \end{bmatrix}, \tag{4.13}$$

where

$$\begin{aligned} Z_1^{i-1} &:= X_{21}^{i-1} \widehat{S}_{11}^a Y_{13}^{i-1} + X_{21}^{i-1} \widehat{S}_{12}^a Y_{23}^{i-1}, & Z_2^{i-1} &:= X_{21}^{i-1} \widehat{S}_{11}^b Y_{13}^{i-1} + X_{21}^{i-1} \widehat{S}_{12}^b Y_{23}^{i-1}, \\ Z_3^{i-1} &:= X_{31}^{i-1} \widehat{S}_{11}^a Y_{12}^{i-1} + X_{32}^{i-1} \widehat{S}_{21}^a Y_{12}^{i-1}, & Z_4^{i-1} &:= X_{31}^{i-1} \widehat{S}_{11}^b Y_{12}^{i-1} + X_{32}^{i-1} \widehat{S}_{21}^b Y_{12}^{i-1}, \\ Z_5^{i-1} &:= X_{31}^{i-1} \widehat{S}_{11}^a Y_{13}^{i-1} + X_{32}^{i-1} \widehat{S}_{21}^a Y_{13}^{i-1} + X_{31}^{i-1} \widehat{S}_{12}^a Y_{23}^{i-1} + X_{32}^{i-1} \widehat{S}_{22}^a Y_{23}^{i-1}, & \text{and} \\ Z_6^{i-1} &:= X_{31}^{i-1} \widehat{S}_{11}^b Y_{13}^{i-1} + X_{32}^{i-1} \widehat{S}_{21}^b Y_{13}^{i-1} + X_{31}^{i-1} \widehat{S}_{12}^b Y_{23}^{i-1} + X_{32}^{i-1} \widehat{S}_{22}^b Y_{23}^{i-1}. \end{aligned}$$

Note that the minimum norm solution of (4.13) is obtained by multiplying the right hand side of (4.13) by the Moore-Penrose pseudoinverse of  $T + \Delta T$ , denoted by  $(T + \Delta T)^\dagger$ , and that  $x_0 = (T + \Delta T)^\dagger c$ .

Now we assume that  $\frac{\theta\omega}{\sigma^2} < \frac{1}{4}$  holds. Then we can prove that the sequence  $\{(X^i, Y^i)\}_{i=0}^\infty$  converges to a solution  $(X, Y)$  of the quadratic system of equations (4.10) analogously as it is done in [7, Theorem 5.8]. For that, we have to take into account that, if  $\|(X^{i-1}, Y^{i-1})\|_F \leq \rho_{i-1}$ , then

$$\begin{aligned} & \|(X^i, Y^i)\|_F \\ & \leq \|(X^0, Y^0)\|_F + \|(T + \Delta T)^\dagger\|_2 \left\| \begin{bmatrix} X_{21}^{i-1} & 0 \\ X_{31}^{i-1} & X_{32}^{i-1} \end{bmatrix} \begin{bmatrix} \widehat{S}_{11} & \widehat{S}_{12} \\ \widehat{S}_{21} & \widehat{S}_{22} \end{bmatrix} \begin{bmatrix} Y_{12}^{i-1} & Y_{13}^{i-1} \\ 0 & Y_{23}^{i-1} \end{bmatrix} \right\|_F \\ & \leq \rho_0 + \sigma^{-1} \rho_{i-1}^2 \omega := \rho_i, \end{aligned}$$

where  $\|(X^0, Y^0)\|_F \leq \theta\sigma^{-1} := \rho_0$ . Therefore, we can define the same fixed point iteration as in the proof of [7, Theorem 5.8] and we obtain that the sequence is bounded, i.e.,  $\|(X^i, Y^i)\|_F \leq \rho$ , with  $\rho < 2\sigma^{-1}\theta$ , for all  $i \geq 0$ . In addition, if we define the sequence  $\{C_i := (X^{i+1}, Y^{i+1}) - (X^i, Y^i)\}_{i=0}^\infty$  then

$$\begin{aligned} \|C_i\|_F & \leq \|(T + \Delta T)^\dagger\|_2 \left( \left\| \begin{bmatrix} X_{21}^i & 0 \\ X_{31}^i & X_{32}^i \end{bmatrix} \begin{bmatrix} \widehat{S}_{11} & \widehat{S}_{12} \\ \widehat{S}_{21} & \widehat{S}_{22} \end{bmatrix} \begin{bmatrix} Y_{12}^i & Y_{13}^i \\ 0 & Y_{23}^i \end{bmatrix} \right. \right. \\ & \quad \left. \left. - \begin{bmatrix} X_{21}^{i-1} & 0 \\ X_{31}^{i-1} & X_{32}^{i-1} \end{bmatrix} \begin{bmatrix} \widehat{S}_{11} & \widehat{S}_{12} \\ \widehat{S}_{21} & \widehat{S}_{22} \end{bmatrix} \begin{bmatrix} Y_{12}^{i-1} & Y_{13}^{i-1} \\ 0 & Y_{23}^{i-1} \end{bmatrix} \right\|_F \right) \\ & \leq \|(T + \Delta T)^\dagger\|_2 \left( \left\| \begin{bmatrix} X_{21}^i - X_{21}^{i-1} & 0 \\ X_{31}^i - X_{31}^{i-1} & X_{32}^i - X_{32}^{i-1} \end{bmatrix} \begin{bmatrix} \widehat{S}_{11} & \widehat{S}_{12} \\ \widehat{S}_{21} & \widehat{S}_{22} \end{bmatrix} \begin{bmatrix} Y_{12}^i & Y_{13}^i \\ 0 & Y_{23}^i \end{bmatrix} \right\|_F \right. \\ & \quad \left. + \left\| \begin{bmatrix} X_{21}^{i-1} & 0 \\ X_{31}^{i-1} & X_{32}^{i-1} \end{bmatrix} \begin{bmatrix} \widehat{S}_{11} & \widehat{S}_{12} \\ \widehat{S}_{21} & \widehat{S}_{22} \end{bmatrix} \begin{bmatrix} Y_{12}^i - Y_{12}^{i-1} & Y_{13}^i - Y_{13}^{i-1} \\ 0 & Y_{23}^i - Y_{23}^{i-1} \end{bmatrix} \right\|_F \right) \\ & \leq 2\sigma^{-1} \rho \omega \|C_{i-1}\|_F. \end{aligned}$$

The above inequality implies that  $\{(X^i, Y^i)\}_{i=0}^\infty$  is a Cauchy sequence, since  $2\sigma^{-1} \rho \omega < 1$ . Thus, taking limits in both sides of (4.13), we see that  $\{(X^i, Y^i)\}_{i=0}^\infty$  converges to a solution  $(X, Y)$  of the system of equations in (4.10) with  $\|(X, Y)\|_F \leq \rho$ .  $\square$

Theorem 4.1, together with Lemma 4.5, allow us to prove in Theorem 4.2 that there exists a solution  $(X, Y)$  of (4.9) which is of the order of the perturbation  $\Delta_S(\lambda)$  whenever  $\|\Delta_S(\lambda)\|_F$  is properly upper bounded.

**Theorem 4.2** Consider the real number  $s$  defined as in Lemma 4.2. Let  $S(\lambda)$  be a block Kronecker linearization as in (4.1), and let  $\Delta_S(\lambda)$  be a perturbation of  $S(\lambda)$  as in (4.3) such that

$$\|\Delta_S(\lambda)\|_F < \left( \frac{2 - \sqrt{3}}{4s} \right)^2 \frac{1}{1 + \|(M(\lambda), A - \lambda I_\ell, B, C)\|_F}. \tag{4.14}$$

Then there exists a solution  $(X, Y) := (X_{21}, X_{31}, X_{32}, Y_{12}, Y_{13}, Y_{23})$  of the quadratic system of matrix equations in (4.9) that satisfies

$$\|(X, Y)\|_F \leq \frac{4s \|\Delta_S(\lambda)\|_F}{2 - \sqrt{3}}. \tag{4.15}$$

**Proof** We have

$$\|\Delta_S(\lambda)\|_F < \left(\frac{2 - \sqrt{3}}{4s}\right)^2 \frac{1}{1 + \|(M(\lambda), A - \lambda I_\ell, B, C)\|_F} \leq \frac{1}{2s}$$

since  $s \geq 1$ . Then, by Lemma 4.5,  $\sigma = \sigma_{\min}(T) - \|\Delta T\|_2 > \frac{2 - \sqrt{3}}{2s} > 0$ . In addition, using the same notation as in Theorem 4.1,

$$\frac{\theta\omega}{\sigma^2} \leq \frac{\|\Delta_S(\lambda)\|_F (\|(M(\lambda), A - \lambda I_\ell, B, C)\|_F + \|\Delta_S(\lambda)\|_F)}{\left(\frac{2 - \sqrt{3}}{2s}\right)^2} < \frac{1}{4},$$

by (4.14). Therefore, the conditions (4.12) hold and, by Theorem 4.1, there exists a solution  $(X, Y)$  of the system in (4.9) satisfying

$$\|(X, Y)\|_F \leq 2 \frac{\theta}{\sigma} \leq \frac{4s \|\Delta_S(\lambda)\|_F}{2 - \sqrt{3}}.$$

□

After restoring the block anti-triangular structure of  $S(\lambda)$ , we get the perturbation error  $\Delta_1^{new}(\lambda)$  defined in (4.7). The following first order bound for the norm of  $\Delta_1^{new}(\lambda)$  in Corollary 4.1 follows from Lemma 4.1 and Theorem 4.2.

**Corollary 4.1** *Let us define the scalar  $f_1 := \frac{4\sqrt{2}s}{2 - \sqrt{3}}$ . Then*

$$\begin{aligned} \|\Delta_1^{new}(\lambda)\|_F &\leq [1 + f_1 \|\widehat{S}_0(\lambda)\|_2] \|\Delta_S(\lambda)\|_F + \mathcal{O}(\delta^2) \\ &\leq [1 + f_1 \|S(\lambda)\|_2] \|\Delta_S(\lambda)\|_F + \mathcal{O}(\delta^2). \end{aligned}$$

**4.2 Step 2: Restoring the Kronecker blocks  $K_1(\lambda), K_2(\lambda)$  and the identity  $I_\ell$**

At this stage we have obtained a pencil  $\widehat{S}_1(\lambda) = S(\lambda) + \Delta_1^{new}(\lambda)$  of the type

$$\widehat{S}_1(\lambda) := \begin{bmatrix} \widehat{M}(\lambda) & \widehat{C}(\lambda) & \widehat{K}_2^T(\lambda) \\ \widehat{B}(\lambda) & \widehat{A} - \lambda \widehat{I}_\ell & 0 \\ \widehat{K}_1(\lambda) & 0 & 0 \end{bmatrix}, \tag{4.16}$$

where the zero blocks below the anti-diagonal are exact and  $\widehat{S}_1(\lambda)$  is strictly equivalent to  $\widehat{S}(\lambda)$ . In this subsection, we will use  $\Delta_{ij}^a - \lambda \Delta_{ij}^b$  to denote the corresponding blocks of the updated perturbation matrix  $\Delta_1^{new}(\lambda)$ . We assume that the norm of the perturbation  $\Delta_1^{new}(\lambda)$  is small enough for  $\widehat{K}_1(\lambda)$  and  $\widehat{K}_2(\lambda)$  to be also minimal bases with row degrees all equal to 1 and the row degrees of their dual minimal bases all equal to  $\epsilon$  and  $\eta$ , respectively [7, Corollary 5.15]. Thus,  $\widehat{K}_1(\lambda)$  and  $\widehat{K}_2(\lambda)$  have the same Kronecker canonical forms as  $K_1(\lambda)$  and  $K_2(\lambda)$ , respectively, and are strictly

equivalent to them. We will then perform step 2, that is, an updating block-diagonal strict equivalent transformation of the type

$$\begin{bmatrix} I_{(\eta+1)m} - X_{11} & 0 & 0 \\ 0 & I_\ell - X_{22} & 0 \\ 0 & 0 & I_{\epsilon n} - X_{33} \end{bmatrix} \widehat{S}_1(\lambda) \begin{bmatrix} I_{(\epsilon+1)n} - Y_{11} & 0 & 0 \\ 0 & I_\ell - Y_{22} & 0 \\ 0 & 0 & I_{\eta m} - Y_{33} \end{bmatrix} \tag{4.17}$$

such that

$$(I - X_{33})\widehat{K}_1(\lambda)(I - Y_{11}) = K_1(\lambda), \quad (I - X_{11})\widehat{K}_2^T(\lambda)(I - Y_{33}) = K_2^T(\lambda),$$

and

$$(I - X_{22})\widehat{I}_\ell(I - Y_{22}) = I_\ell.$$

In the last three equations the sizes of some identity matrices are not specified for simplicity. Clearly, these three problems are independent from each other and can be treated separately.

Let us first look at the equation restoring  $K_1(\lambda)$ . As pointed out in Sect. 2, this can be reduced to the solution of a Sylvester equation. Let

$$\begin{aligned} \widehat{K}_1(\lambda) &= K_1(\lambda) + \Delta_{K_1}(\lambda) := L_\epsilon(\lambda) \otimes I_n + \Delta_{K_1}(\lambda) \\ &:= (E_\epsilon - \lambda F_\epsilon) \otimes I_n + (\Delta_{31}^a - \lambda \Delta_{31}^b). \end{aligned}$$

Then, making the change of variables  $Y_{11} := Y$  and  $X_{33} := X(I + X)^{-1}$ , it suffices to solve

$$(K_1(\lambda) + \Delta_{K_1}(\lambda))Y + XK_1(\lambda) = \Delta_{K_1}(\lambda),$$

or, equivalently,

$$\begin{bmatrix} \frac{E_\epsilon^T \otimes I_{n\epsilon n}}{F_\epsilon^T \otimes I_{n\epsilon n}} \Big| \frac{I_{(\epsilon+1)n} \otimes (E_\epsilon \otimes I_n + \Delta_{31}^a)}{I_{(\epsilon+1)n} \otimes (F_\epsilon \otimes I_n + \Delta_{31}^b)} \end{bmatrix} \begin{bmatrix} \text{vec}(X) \\ \text{vec}(Y) \end{bmatrix} = \begin{bmatrix} \text{vec}(\Delta_{31}^a) \\ \text{vec}(\Delta_{31}^b) \end{bmatrix}. \tag{4.18}$$

By Lemma 3.4, the smallest singular value of the unperturbed problem satisfies

$$\sigma_{2\epsilon n(\epsilon+1)n} \left[ \frac{E_\epsilon^T \otimes I_{n\epsilon n}}{F_\epsilon^T \otimes I_{n\epsilon n}} \Big| \frac{I_{(\epsilon+1)n} \otimes E_\epsilon \otimes I_n}{I_{(\epsilon+1)n} \otimes F_\epsilon \otimes I_n} \right] \geq \frac{3}{4\epsilon - 1}.$$

Then, by using Weyl’s perturbation theorem for singular values [13, Theorem 3.3.16], one obtains the following bound for the minimum norm solution of (4.18)

$$\|(X, Y)\|_F \leq \left[ \frac{3}{4\epsilon - 1} - \|\Delta_{31}^a\|_2 - \|\Delta_{31}^b\|_2 \right]^{-1} \|(\Delta_{31}^a, \Delta_{31}^b)\|_F,$$

assuming that the perturbation is small enough for satisfying  $\frac{3}{4\epsilon-1} - \|\Delta_{31}^a\|_2 - \|\Delta_{31}^b\|_2 > 0$ . In addition,

$$\|(X_{33}, Y_{11})\|_F \leq \|(X, Y)\|_F / (1 - \|(X, Y)\|_F).$$

Since  $\|\Delta_{31}^a\|_2$  and  $\|\Delta_{31}^b\|_2$  are of the order of  $\delta$ , this finally yields

$$\|(X_{33}, Y_{11})\|_F \leq \frac{4\epsilon - 1}{3} \|(\Delta_{31}^a, \Delta_{31}^b)\|_F + \mathcal{O}(\delta^2), \tag{4.19}$$

by regrouping the quantities of the order of  $\mathcal{O}(\delta^2)$ .

The problem for restoring  $K_2(\lambda)$  is clearly dual to the problem of  $K_1(\lambda)$  and will therefore yield the bound

$$\|(X_{11}, Y_{33})\|_F \leq \frac{4\eta - 1}{3} \|(\Delta_{13}^a, \Delta_{13}^b)\|_F + \mathcal{O}(\delta^2). \tag{4.20}$$

The problem of restoring  $I_\ell$  amounts to solving  $(I_\ell - X_{22})(I_\ell + \Delta_{22}^b)(I_\ell - Y_{22}) = I_\ell$ , with  $\widehat{I}_\ell = I_\ell + \Delta_{22}^b$ . There are many possible solutions. A very simple one is to take  $Y_{22} = 0$  and  $I_\ell - X_{22} = (I_\ell + \Delta_{22}^b)^{-1}$ , assuming  $\Delta_{22}^b$  is small enough for the inverse to exist. This means that  $X_{22} = \Delta_{22}^b + \mathcal{O}(\|\Delta_{22}^b\|_F^2)$  and

$$\|(X_{22}, Y_{22})\|_F = \|\Delta_{22}^b\|_F + \mathcal{O}(\delta^2). \tag{4.21}$$

We summarize this discussion in the following Theorem.

**Theorem 4.3** *Let the pencil  $\widehat{S}_1(\lambda)$  have the block anti-triangular form given in (4.16). If  $\max(\epsilon, \eta) > 0$ , then the updating strict equivalence transformation  $(I - X)\widehat{S}_1(\lambda)(I - Y)$  detailed in (4.17) exists and can be bounded by*

$$\|(X, Y)\|_F \leq \frac{4 \max(\epsilon, \eta) - 1}{3} \|\Delta_1^{new}(\lambda)\|_F + \mathcal{O}(\delta^2).$$

**Proof** The bound for  $\|(X, Y)\|_F$  follows directly from the identity

$$\|(X, Y)\|_F^2 = \|(X_{11}, Y_{33})\|_F^2 + \|(X_{22}, Y_{22})\|_F^2 + \|(X_{33}, Y_{11})\|_F^2,$$

from the inequality

$$\|(\Delta_{13}^a, \Delta_{13}^b)\|_F^2 + \|\Delta_{22}^b\|_F^2 + \|(\Delta_{31}^a, \Delta_{31}^b)\|_F^2 \leq \|\Delta_1^{new}(\lambda)\|_F^2$$

and from the individual inequalities (4.19), (4.20) and (4.21). □

The following first order bound in Corollary 4.2 for the norm of the perturbation error  $\Delta_2^{new}(\lambda)$  follows from Lemma 4.1, Theorem 4.3 and Corollary 4.1.



**Corollary 4.2** *Let us define the scalar  $f_2 := \frac{\sqrt{2}(4 \max\{\epsilon, \eta\} - 1)}{3}$ . Then*

$$\begin{aligned} \|\Delta_2^{new}(\lambda)\|_F &\leq [1 + f_2 \|\widehat{S}_1(\lambda)\|_2] \|\Delta_1^{new}(\lambda)\|_F + \mathcal{O}(\delta^2) \\ &\leq [1 + f_2 \|S(\lambda)\|_2] \|\Delta_1^{new}(\lambda)\|_F + \mathcal{O}(\delta^2). \end{aligned}$$

**4.3 Step 3: Restoring the constant B and C matrices**

From steps 1 and 2, described in the previous subsections, we have obtained a pencil  $\widehat{S}_2(\lambda) = S(\lambda) + \Delta_2^{new}(\lambda)$  of the type

$$\widehat{S}_2(\lambda) := \begin{bmatrix} \widehat{M}(\lambda) & \widehat{C}(\lambda) & K_2^T(\lambda) \\ \widehat{B}(\lambda) & \widehat{A} - \lambda I_\ell & 0 \\ K_1(\lambda) & 0 & 0 \end{bmatrix} \tag{4.22}$$

strictly equivalent to  $\widehat{S}(\lambda)$ . We emphasize that the blocks  $\widehat{M}(\lambda)$ ,  $\widehat{B}(\lambda)$ ,  $\widehat{C}(\lambda)$  and the matrix  $\widehat{A}$  are obviously different in (4.22) and in (4.16). We use the same symbols for avoiding a cumbersome notation. In this subsection, we will use  $\Delta_{ij}(\lambda) = \Delta_{ij}^a - \lambda \Delta_{ij}^b$  to denote the corresponding blocks of the updated perturbation matrix  $\Delta_2^{new}(\lambda)$ . In this third step, we will restore the pencil  $\widehat{S}_2(\lambda)$  to one where the blocks

$$\widehat{B}(\lambda) = B\widehat{K}_1 + \Delta_{21}(\lambda), \quad \text{and} \quad \widehat{C}(\lambda) = \widehat{K}_2^T C + \Delta_{12}(\lambda)$$

are transformed to  $\widetilde{B}\widetilde{K}_1$  and  $\widetilde{K}_2^T \widetilde{C}$ , respectively. We recall that

$$K_1(\lambda) = L_\epsilon(\lambda) \otimes I_n, \quad \widehat{K}_1 = \mathbf{e}_{\epsilon+1}^T \otimes I_n, \quad K_2(\lambda) = L_\eta(\lambda) \otimes I_m, \quad \widehat{K}_2 = \mathbf{e}_{\eta+1}^T \otimes I_m,$$

where  $\mathbf{e}_k$  is the standard  $k$ th unit vector of dimension  $k$  and  $L_k(\lambda)$  is the classical Kronecker block of dimension  $k \times (k + 1)$ , as introduced below (1.2). We will construct for this a strict equivalence transformation of the type

$$\begin{aligned} &\begin{bmatrix} I_{m(\eta+1)} & -X_{12} & 0 \\ & I_\ell & -X_{23} \\ & & I_{n\epsilon} \end{bmatrix} \widehat{S}_2(\lambda) \begin{bmatrix} I_{n(\epsilon+1)} & & \\ -Y_{21} & I_\ell & \\ 0 & -Y_{32} & I_{m\eta} \end{bmatrix} \\ &= \begin{bmatrix} \widetilde{M}(\lambda) & \widetilde{K}_2^T \widetilde{C} & K_2^T(\lambda) \\ \widetilde{B}\widetilde{K}_1 & \widehat{A} - \lambda I_\ell & 0 \\ K_1(\lambda) & 0 & 0 \end{bmatrix} \end{aligned} \tag{4.23}$$

The problems for  $\widehat{B}(\lambda)$  and  $\widehat{C}(\lambda)$  can again be treated separately. Let us first focus on the subsystem

$$\begin{aligned} &\begin{bmatrix} I_{m(\eta+1)} & -X_{12} \\ & I_\ell \end{bmatrix} \begin{bmatrix} \widehat{C}(\lambda) & L_\eta^T(\lambda) \otimes I_m \\ \widehat{A} - \lambda I_\ell & 0 \end{bmatrix} \begin{bmatrix} I_\ell \\ -Y_{32} & I_{m\eta} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{e}_{\eta+1}^T \otimes \widetilde{C} & L_\eta^T(\lambda) \otimes I_m \\ \widehat{A} - \lambda I_\ell & 0 \end{bmatrix}. \end{aligned}$$

If we partition the matrices  $X_{12}$ ,  $Y_{32}$  and  $\widehat{C}(\lambda)$  as follows:

$$X_{12} := \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_\eta \\ E_{\eta+1} \end{bmatrix}, \quad Y_{32} := \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_\eta \end{bmatrix}, \quad \widehat{C}(\lambda) := \begin{bmatrix} C_{01} \\ C_{02} \\ \vdots \\ C_{0\eta} \\ C_{0(\eta+1)} \end{bmatrix} - \begin{bmatrix} C_{11} \\ C_{12} \\ \vdots \\ C_{1\eta} \\ C_{1(\eta+1)} \end{bmatrix} \lambda,$$

where all blocks have dimension  $m \times \ell$ , then we need to solve the following system of equations

$$\begin{aligned} & [ E_1 \ F_1 \ E_2 \ \dots \ F_\eta \ E_{\eta+1} ] (I_{(2\eta+1)\ell} + N) \\ & = [ C_{11} \ C_{01} \ C_{12} \ \dots \ C_{0\eta} \ C_{1(\eta+1)} ], \end{aligned}$$

where

$$I_{(2\eta+1)\ell} + N := \begin{bmatrix} I_\ell & \widehat{A} & & & \\ & I_\ell & I_\ell & & \\ & & I_\ell & \widehat{A} & \\ & & & \ddots & \ddots \\ & & & & I_\ell & I_\ell \\ & & & & & I_\ell \end{bmatrix},$$

and  $\widetilde{C} := C_{0(\eta+1)} - E_{\eta+1}\widehat{A}$ . Clearly

$$\begin{aligned} & \| [ E_1 \ F_1 \ E_2 \ \dots \ F_\eta \ E_{\eta+1} ] \|_F = \| (X_{12}, Y_{32}) \|_F, \\ & \| [ C_{11} \ C_{01} \ C_{12} \ \dots \ C_{0\eta} \ C_{1(\eta+1)} ] \|_F \leq \| \Delta_{12}(\lambda) \|_F, \end{aligned}$$

and, since the matrix  $N$  is nilpotent with  $N^{2\eta+1} = 0$ ,

$$(I_{(2\eta+1)\ell} + N)^{-1} = \sum_{i=0}^{2\eta} (-N)^i.$$

In addition,  $N$  has even powers  $N^{2i}$  of 2-norm  $\|\widehat{A}^i\|_2 \leq \|\widehat{A}\|_2^i$ , whereas the odd powers  $N^{2i-1}$  have 2-norm  $\max(\|\widehat{A}^{i-1}\|_2, \|\widehat{A}^i\|_2) \leq \max(\|\widehat{A}\|_2^{i-1}, \|\widehat{A}\|_2^i)$ . Since both of them can be bounded by  $\max(1, \|\widehat{A}\|_2^i)$ , it then follows that

$$\begin{aligned} \| (X_{12}, Y_{32}) \|_F & \leq \| \Delta_{12}(\lambda) \|_F (1 + 2 \max(1, \|\widehat{A}\|_2) + \dots + 2 \max(1, \|\widehat{A}\|_2^\eta)) \\ & \leq [1 + 2\eta \max(1, \|\widehat{A}\|_2^\eta)] \| \Delta_{12}(\lambda) \|_F. \end{aligned} \tag{4.24}$$

The discussion for the  $\widehat{B}(\lambda)$  block is clearly analogous and will yield the bound

$$\| (X_{23}, Y_{21}) \|_F \leq [1 + 2\epsilon \max(1, \|\widehat{A}\|_2^\epsilon)] \| \Delta_{21}(\lambda) \|_F. \tag{4.25}$$

We can thus summarize this discussion in the following Theorem.

**Theorem 4.4** *Let the pencil  $\widehat{S}_2(\lambda)$  have the anti-triangular form given in (4.22). Then the updating strict equivalence transformation  $(I - X)\widehat{S}_2(\lambda)(I - Y)$  detailed in (4.23) exists and can be bounded by*

$$\|(X, Y)\|_F \leq [1 + 2 \max(\eta, \epsilon) \max(1, \|\widehat{A}\|_2^{\max(\eta, \epsilon)})] \|\Delta_2^{new}(\lambda)\|_F.$$

**Proof** The bound for  $\|(X, Y)\|_F$  follows directly from the identity

$$\|(X, Y)\|_F^2 = \|(X_{12}, Y_{32})\|_F^2 + \|(X_{23}, Y_{21})\|_F^2,$$

from the inequality  $\|\Delta_{12}(\lambda)\|_F^2 + \|\Delta_{21}(\lambda)\|_F^2 \leq \|\Delta_2^{new}(\lambda)\|_F^2$  and from the individual inequalities (4.24) and (4.25). □

The following first order bound in Corollary 4.3 for the norm of the perturbation error  $\Delta_3^{new}(\lambda)$  follows from Lemma 4.1, Theorem 4.4 and Corollaries 4.1 and 4.2.

**Corollary 4.3** *Let us define  $f_3 := \sqrt{2}(1 + 2 \max(\eta, \epsilon) \max(1, \|\widehat{A}\|_2^{\max(\eta, \epsilon)}))$ . Then*

$$\begin{aligned} \|\Delta_3^{new}(\lambda)\|_F &\leq [1 + f_3 \|\widehat{S}_2(\lambda)\|_2] \|\Delta_2^{new}(\lambda)\|_F + \mathcal{O}(\delta^2) \\ &\leq [1 + f_3 \|S(\lambda)\|_2] \|\Delta_2^{new}(\lambda)\|_F + \mathcal{O}(\delta^2). \end{aligned}$$

### 4.4 Putting it all together

In this subsection, we combine the obtained results regarding the strict equivalence transformation that restores in  $\widehat{S}(\lambda)$  of (4.3) the special structure of the unperturbed block Kronecker linearization  $S(\lambda)$  defined in (1.2), in such a way that the eigenstructure of  $\widehat{S}(\lambda)$  can be linked to that of a particular rational matrix  $\widetilde{R}(\lambda)$  as in (4.5). The final goal is to bound the norms of the differences between the quadruples  $\{\lambda I_\ell - A, B, C, D(\lambda)\}$  and  $\{\lambda I_\ell - \widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}(\lambda)\}$  that are used for representing the unperturbed rational matrix  $R(\lambda)$  and the perturbed one  $\widetilde{R}(\lambda)$ , respectively.

Recall that we were given the pencil  $S(\lambda)$  of which we want to compute the eigenstructure, since it gives the one of the rational matrix  $R(\lambda)$  in (4.2). Instead, our backward stable algorithm applied to  $S(\lambda)$  computes the exact eigenstructure of a slightly perturbed pencil  $\widehat{S}(\lambda)$  with additive error  $\Delta_S(\lambda)$  which is induced by the eigenstructure algorithm and is bounded as:

$$\|\Delta_S(\lambda)\|_F \leq c(\ell, m\eta, n\epsilon) \cdot \epsilon_M \cdot \|S(\lambda)\|_F,$$

where  $\epsilon_M$  is the machine precision of the used computer, and  $c(\ell, m\eta, n\epsilon)$  is a moderate function depending only on the size of the matrix pencil. We then constructed in three steps a new modified block Kronecker linearization

$$\widetilde{S}(\lambda) := (I - X)\widehat{S}(\lambda)(I - Y) := (I - X_3)(I - X_2)(I - X_1)\widehat{S}(\lambda)(I - Y_1)(I - Y_2)(I - Y_3) \tag{4.26}$$

as in (4.4), strictly equivalent to  $\widehat{S}(\lambda)$ , where both  $\|X\|_F$  and  $\|Y\|_F$  are also of the order of the machine precision times some factors and such that the corresponding rational matrix  $\widetilde{R}(\lambda)$  (4.5) has a similar representation as  $R(\lambda)$ . Since  $\widehat{S}(\lambda)$  and  $\widetilde{S}(\lambda)$  are strictly equivalent pencils, they have *exactly* the same eigenstructure, which implies that we have computed the exact eigenstructure of the nearby rational matrix  $\widetilde{R}(\lambda)$ .

For convenience, the blocks of  $\widetilde{S}(\lambda)$  will be expressed in the sequel as  $\widetilde{M}(\lambda) := M(\lambda) + \Delta M(\lambda)$ ,  $\widetilde{A} := A + \Delta A$ ,  $\widetilde{B} := B + \Delta B$  and  $\widetilde{C} := C + \Delta C$ . In the previous subsections, we rewrote  $\widetilde{S}(\lambda)$  as an additive perturbation

$$\widetilde{S}(\lambda) = S(\lambda) + \Delta_3^{new}(\lambda)$$

and derived a first order bound for the norm of the error pencil  $\Delta_3^{new}(\lambda)$  in Corollaries 4.1, 4.2 and 4.3:

$$\|\Delta_3^{new}(\lambda)\|_F \leq (1 + f_1 \|S(\lambda)\|_2)(1 + f_2 \|S(\lambda)\|_2)(1 + f_3 \|S(\lambda)\|_2) \|\Delta_S(\lambda)\|_F + \mathcal{O}(\delta^2). \tag{4.27}$$

This implies, in particular, that if  $\|\Delta_S(\lambda)\|_F$  is sufficiently small, then the norms of the perturbations  $\Delta A$ ,  $\Delta B$  and  $\Delta C$  are sufficiently small to guarantee that  $\widetilde{C}(\lambda I_\ell - \widetilde{A})^{-1} \widetilde{B}$  is a minimal state-space realization, as announced. Then, according to [2],  $\widetilde{S}(\lambda)$  is indeed a strong linearization of the rational matrix  $\widetilde{R}(\lambda)$  in (4.5). Moreover, (4.27) also implies that if  $\|\Delta_S(\lambda)\|_F$  is sufficiently small, then  $\widetilde{D}(\lambda) := \sum_{i=0}^d (D_i + \Delta D_i) \lambda^i$  in (4.5) is a polynomial matrix with the same degree  $d = \eta + \epsilon + 1$  as the polynomial part  $D(\lambda)$  of  $R(\lambda)$  (recall that we are assuming that  $d$  is the degree of  $D(\lambda)$  or, equivalently, that  $D_d \neq 0$ ).

Notice that  $\widetilde{R}(\lambda)$  in (4.5) is the transfer function of the following perturbed polynomial system matrix

$$P(\lambda) + \Delta P(\lambda) := \begin{bmatrix} \lambda I_\ell - A & -B \\ C & D(\lambda) \end{bmatrix} + \begin{bmatrix} -\Delta A & -\Delta B \\ \Delta C & \sum_{i=0}^d \Delta D_i \lambda^i \end{bmatrix}, \tag{4.28}$$

where  $P(\lambda)$  is a polynomial system matrix of the original rational matrix  $R(\lambda)$ . Recall that  $\|R(\lambda)\|_F$  is defined in (1.3) as  $\|P(\lambda)\|_F$ . This motivates us to define the norm of the perturbation of  $R(\lambda)$  as

$$\|\Delta R(\lambda)\|_F := \|\Delta P(\lambda)\|_F = \sqrt{\|\Delta A\|_F^2 + \|\Delta B\|_F^2 + \|\Delta C\|_F^2 + \sum_{i=0}^d \|\Delta D_i\|_F^2}.$$

After this discussion, we present our main perturbation results in Theorems 4.5 and 4.6. The first one focuses on block Kronecker linearizations and the second one on the corresponding rational matrices.

**Theorem 4.5** *Let  $R(\lambda)$  be the  $m \times n$  rational matrix in (1.1) and let  $S(\lambda)$  be a block Kronecker linearization of  $R(\lambda)$  as in (1.2). Let us define  $\alpha := 1 + 2\epsilon \max(1, \|A\|_F^2)$ ,*

$\beta := 1 + 2\eta \max(1, \|A\|_2^\eta)$ ,  $\gamma := \frac{\epsilon + \eta}{2\sqrt{2}}$  and  $s := \max(\alpha, \beta, \gamma) + \gamma(\beta\|B\|_2 + \alpha\|C\|_2)$ . Assume that  $\max(\epsilon, \eta) > 0$  and consider the functions dependent on the initial data

$$\begin{aligned} f_1 &:= f_1(\epsilon, \eta, \|A\|_2, \|B\|_2, \|C\|_2) := \frac{4\sqrt{2}s}{2 - \sqrt{3}}, \\ f_2 &:= f_2(\epsilon, \eta) := \frac{\sqrt{2}(4 \max(\epsilon, \eta) - 1)}{3}, \\ f_3 &:= f_3(\epsilon, \eta, \|A\|_2) := \sqrt{2}[1 + 2 \max(\eta, \epsilon) \max(1, \|A\|_2^{\max(\eta, \epsilon)})]. \end{aligned}$$

Let  $\widehat{S}(\lambda) := S(\lambda) + \Delta_S(\lambda)$  be a perturbed pencil as in (4.3). If  $\|\Delta_S(\lambda)\|_F$  is sufficiently small, then  $\widehat{S}(\lambda)$  is strictly equivalent to a block Kronecker linearization  $\widetilde{S}(\lambda)$  as in (4.4) with the same parameters  $\epsilon$  and  $\eta$  as  $S(\lambda)$ , i.e., the transformation (4.26) exists. Moreover,  $\widetilde{S}(\lambda) = S(\lambda) + \Delta_3^{new}(\lambda)$  with

$$\|\Delta_3^{new}(\lambda)\|_F \leq (1 + f_1 \|S(\lambda)\|_2)(1 + f_2 \|S(\lambda)\|_2)(1 + f_3 \|S(\lambda)\|_2) \|\Delta_S(\lambda)\|_F + \mathcal{O}(\delta^2), \tag{4.29}$$

where  $\delta := \frac{\|\Delta_S(\lambda)\|_F}{\|S(\lambda)\|_F}$ .

**Proof** This follows directly from (4.27), except that we have replaced the 2-norm of  $\widehat{A}$  in  $f_3$  in Corollary 4.3 by that of  $A$ , because the difference can be absorbed in the  $\mathcal{O}(\delta^2)$  term. □

Theorem 4.5 does not provide directly bounds on the norms of the differences between the quadruples representing the rational matrices  $R(\lambda)$  and  $\widetilde{R}(\lambda)$  corresponding to the block Kronecker linearizations  $S(\lambda)$  and  $\widetilde{S}(\lambda)$ . The reason is that the polynomial parts  $D(\lambda) = (\Lambda_\eta(\lambda) \otimes I_m)^T M(\lambda) (\Lambda_\epsilon(\lambda) \otimes I_n)$  and  $\widetilde{D}(\lambda) = (\Lambda_\eta(\lambda) \otimes I_m)^T \widetilde{M}(\lambda) (\Lambda_\epsilon(\lambda) \otimes I_n)$  of  $R(\lambda)$  and  $\widetilde{R}(\lambda)$  are not directly visible in  $S(\lambda)$  and  $\widetilde{S}(\lambda)$ . For this reason, we will need Lemma 4.6, that follows from [7, Lemma 2.15, Theorem 4.4 and Lemma 5.23(b)].

**Lemma 4.6** *Let  $M(\lambda)$  be a  $m(\eta + 1) \times n(\epsilon + 1)$  pencil and let  $\Lambda_k(\lambda) := [\lambda^k \dots \lambda 1]^T$ . If we define the polynomial matrix  $Q(\lambda)$  as*

$$Q(\lambda) := (\Lambda_\eta(\lambda) \otimes I_m)^T M(\lambda) (\Lambda_\epsilon(\lambda) \otimes I_n), \tag{4.30}$$

then we can bound its norm as follows

$$\|Q(\lambda)\|_F \leq \sqrt{2 \min(\epsilon + 1, \eta + 1)} \|M(\lambda)\|_F.$$

Moreover, for every polynomial matrix  $Q(\lambda)$  of degree at most  $d = \epsilon + \eta + 1$ , there exist infinitely many pencils  $M(\lambda)$  satisfying (4.30). For each of these pencils  $\|M(\lambda)\|_F \geq \|Q(\lambda)\|_F / \sqrt{2d}$  and there exist pencils such that  $\|Q(\lambda)\|_F = \|M(\lambda)\|_F$ .

As commented in [7], Fiedler and proper generalized Fiedler pencils (modulo permutations) of a polynomial matrix  $Q(\lambda)$  satisfy  $\|Q(\lambda)\|_F = \|M(\lambda)\|_F$  in Lemma 4.6.

On the other hand, it might be worth to remind that there exist pencils  $M(\lambda)$  satisfying (4.30) with norm arbitrarily larger than the norm of  $Q(\lambda)$ .

We are finally in the position of proving the main perturbation result of this paper.

**Theorem 4.6** *Let  $R(\lambda) = C(\lambda I_\ell - A)^{-1}B + \sum_{i=0}^d D_i \lambda^i$  be an  $m \times n$  rational matrix, where  $C(\lambda I_\ell - A)^{-1}B$  is a minimal state-space realization of the strictly proper part of  $R(\lambda)$ , let  $S(\lambda)$  be a block Kronecker linearization of  $R(\lambda)$  as in (1.2) with  $\max(\epsilon, \eta) > 0$ , and let  $f_1, f_2, f_3$  be the functions defined in Theorem 4.5. Let  $\tilde{S}(\lambda) := S(\lambda) + \Delta_S(\lambda)$  be a perturbed pencil as in (4.3). If  $\|\Delta_S(\lambda)\|_F$  is sufficiently small, then  $\tilde{S}(\lambda)$  is strictly equivalent to a block Kronecker linearization  $\tilde{S}(\lambda)$  as in (4.4), with the same parameters  $\epsilon$  and  $\eta$  as  $S(\lambda)$ , of a rational matrix*

$$\tilde{R}(\lambda) = \tilde{C}(\lambda I_\ell - \tilde{A})^{-1} \tilde{B} + \sum_{i=0}^d \tilde{D}_i \lambda^i,$$

where  $\tilde{C}(\lambda I_\ell - \tilde{A})^{-1} \tilde{B}$  is a minimal state-space realization of the strictly proper part of  $\tilde{R}(\lambda)$ . Moreover, if  $\tilde{A} := A + \Delta A$ ,  $\tilde{B} := B + \Delta B$ ,  $\tilde{C} := C + \Delta C$  and  $\tilde{D}_i := D_i + \Delta D_i$ ,  $i = 0, 1, \dots, d$ , then

$$\frac{\sqrt{\|\Delta A\|_F^2 + \|\Delta B\|_F^2 + \|\Delta C\|_F^2 + \sum_{i=0}^d \|\Delta D_i\|_F^2}}{\|R(\lambda)\|_F} \leq K_{S,R} \frac{\|\Delta_S(\lambda)\|_F}{\|S(\lambda)\|_F} + \mathcal{O}(\delta^2), \tag{4.31}$$

where

$$K_{S,R} := \sqrt{2 \min(\epsilon + 1, \eta + 1)} (1 + f_1 \|S(\lambda)\|_2)(1 + f_2 \|S(\lambda)\|_2) (1 + f_3 \|S(\lambda)\|_2) \frac{\|S(\lambda)\|_F}{\|R(\lambda)\|_F}$$

and  $\delta = \frac{\|\Delta_S(\lambda)\|_F}{\|S(\lambda)\|_F}$ .

**Proof** Since  $\tilde{S}(\lambda)$  and  $S(\lambda)$  have the same structure according to Theorem 4.5,

$$\Delta_3^{new}(\lambda) = \tilde{S}(\lambda) - S(\lambda) = \begin{bmatrix} \tilde{M}(\lambda) - M(\lambda) & \widehat{K}_2^T (\tilde{C} - C) & 0 \\ (\tilde{B} - B) \widehat{K}_1 & \tilde{A} - A & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

and  $\|\Delta_3^{new}(\lambda)\|_F = \sqrt{\|\Delta A\|_F^2 + \|\Delta B\|_F^2 + \|\Delta C\|_F^2 + \|\tilde{M}(\lambda) - M(\lambda)\|_F^2}$ . Next, we combine this expression of  $\|\Delta_3^{new}(\lambda)\|_F$  with  $\sum_{i=0}^d D_i \lambda^i = (\Lambda_\eta(\lambda) \otimes I_m)^T M(\lambda) (\Lambda_\epsilon(\lambda) \otimes I_n)$ ,  $\sum_{i=0}^d \tilde{D}_i \lambda^i = (\Lambda_\eta(\lambda) \otimes I_m)^T \tilde{M}(\lambda) (\Lambda_\epsilon(\lambda) \otimes I_n)$  and Lemma 4.6, and we get

$$\sqrt{\|\Delta A\|_F^2 + \|\Delta B\|_F^2 + \|\Delta C\|_F^2 + \sum_{i=0}^d \|\Delta D_i\|_F^2} \leq \sqrt{2 \min(\epsilon + 1, \eta + 1)} \|\Delta_3^{new}(\lambda)\|_F.$$

The rest of the proof follows from (4.29). □

The strength of the new structured backward error analysis that we present in this paper for the computation of the eigenstructure of a rational matrix  $R(\lambda)$  by applying a backward stable generalized eigenvalue algorithm to a block Kronecker linearization  $S(\lambda)$  of  $R(\lambda)$  is that we can interpret the computed eigenstructure as the exact eigenstructure for a slightly perturbed rational matrix  $\tilde{R}(\lambda)$  corresponding to the nearby quadruple  $\{\lambda I_\ell - \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}(\lambda)\}$ , and that we have a bound on the error because we have a specific coordinate system in which we can describe both the original rational matrix  $R(\lambda)$  and its perturbed version  $\tilde{R}(\lambda)$ , namely by the quadruples  $\{\lambda I_\ell - A, B, C, D(\lambda)\}$  and  $\{\lambda I_\ell - \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}(\lambda)\}$ . It still remains to analyze under which conditions this bound is satisfactory. This is the purpose of the next subsection.

### 4.5 Sufficient conditions for structural backward stability

The goal of this section is to establish sufficient conditions on  $R(\lambda)$  and  $S(\lambda)$  that guarantee that  $K_{S,R}$  in (4.31) is moderate and, thus, that guarantee structural backward stability. We advance that these conditions are the following

$$\max(\|A\|_F, \|B\|_F, \|C\|_F, \|D(\lambda)\|_F) \leq 1 \quad \text{and} \quad \|M(\lambda)\|_F \approx \|D(\lambda)\|_F, \quad (4.32)$$

where the notation introduced in the previous section is used. Observe that the first condition is a condition on  $R(\lambda)$  while the second one is on  $S(\lambda)$ . According to Lemma 4.6, the second condition can be satisfied simply by choosing an adequate block Kronecker linearization  $S(\lambda)$ . In addition, we will see that the conditions (4.32) are essentially necessary for  $K_{S,R}$  to be moderate, though this does not mean that they are necessary for structural backward stability since (4.31) is an upper bound. For the sake of clarity, the discussion in this section focuses on identifying the key ingredients for structural backward stability instead of on providing precise bounds. There exist, obviously, rational matrices which do not satisfy the first condition in (4.32). We will discuss in Sect. 5 how to proceed in such cases.

In the first place observe that each of the essential four factors of  $K_{S,R}$ , that is,  $(1 + f_1\|S(\lambda)\|_2)$ ,  $(1 + f_2\|S(\lambda)\|_2)$ ,  $(1 + f_3\|S(\lambda)\|_2)$  and  $\frac{\|S(\lambda)\|_F}{\|R(\lambda)\|_F}$ , is larger than 1. This is obvious for the first three factors. For the fourth factor, it follows from the equalities

$$\begin{aligned} \|S(\lambda)\|_F^2 &= \|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2 + \|M(\lambda)\|_F^2 + \ell + 2(m\eta + n\epsilon) \quad \text{and} \\ \|R(\lambda)\|_F^2 &= \|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2 + \|D(\lambda)\|_F^2 + \ell. \end{aligned} \quad (4.33)$$

To find upper bounds for the three factors  $(1 + f_1\|S(\lambda)\|_2)$ ,  $(1 + f_2\|S(\lambda)\|_2)$ ,  $(1 + f_3\|S(\lambda)\|_2)$  of  $K_{S,R}$  requires to upper bound each  $f_i$  and  $\|S(\lambda)\|_2$ . For this purpose, we consider Lemmas 4.7 and 4.8. Lemma 4.7 provides a bound on the function  $f_1$  that allows us to identify its most relevant dependencies. Moreover, Lemma 4.7 emphasizes the key role of  $t := \max(\eta, \epsilon)$  in our perturbation analysis. Lemma 4.8 bounds  $\|S(\lambda)\|_2$ .

**Lemma 4.7** *Let us define  $\widehat{M}_a := \max(1, \|A\|_2)$ ,  $M_b := \max(\|B\|_2, \|C\|_2)$  and  $t := \max(\eta, \epsilon) > 0$  and consider the functions  $f_1, f_2$  and  $f_3$  in Theorem 4.5. Then*

$$1 \leq f_1 \leq 22(1 + 2tM'_a)(1 + \sqrt{2}tM_b), \quad 1 \leq f_2 \\ = \frac{\sqrt{2}}{3}(4t - 1), \quad 1 \leq f_3 = \sqrt{2}(1 + 2tM'_a).$$

**Proof** It follows by taking into account the inequalities  $\gamma \leq \frac{t}{\sqrt{2}}$  and  $s \leq (1 + 2tM'_a)(1 + \sqrt{2}tM_b)$ . □

**Lemma 4.8** *Let  $S(\lambda)$  be the block Kronecker linearization (1.2). Then*

$$\max(1, \|A\|_2, \|B\|_2, \|C\|_2, \|M(\lambda)\|_2) \leq \|S(\lambda)\|_2$$

and

$$\|S(\lambda)\|_2 \leq \sqrt{2} + \left\| \begin{bmatrix} M(\lambda) & \widehat{K}_2^T C \\ B\widehat{K}_1 & A \end{bmatrix} \right\|_2 \leq \sqrt{2} + \sqrt{\|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2 + \|M(\lambda)\|_F^2}.$$

**Proof** The first inequality follows from the definition of the 2-norm of a pencil given in the introduction and the fact that the 2-norm of a matrix is larger than or equal to the 2-norm of any of its submatrices. The second inequality follows from applying the triangular inequality to

$$S(\lambda) = \begin{bmatrix} M(\lambda) & \widehat{K}_2^T C & 0 \\ B\widehat{K}_1 & A & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & K_2^T(\lambda) \\ 0 & -\lambda I_\ell & 0 \\ K_1(\lambda) & 0 & 0 \end{bmatrix}.$$

Note that the 2-norm of a pencil as defined in the introduction is indeed a norm and, so, the triangular inequality can be applied. □

We remark that Lemmas 4.7 and 4.8 imply that the conditions (4.32) are essentially necessary for  $K_{S,R}$  to be moderate. This can be seen as follows. First, from Lemma 4.6, we have  $\|M(\lambda)\|_F \geq \|D(\lambda)\|_F / \sqrt{2(\epsilon + \eta + 1)}$ . Thus,  $\max(\|A\|_F, \|B\|_F, \|C\|_F, \|D(\lambda)\|_F) \gg 1$  implies  $\|S(\lambda)\|_2 \gg 1$ , which in turns implies  $K_{S,R} \gg 1$ , since  $f_i \geq 1$  for  $i = 1, 2, 3$ . Moreover, if  $\|M(\lambda)\|_F \gg \|D(\lambda)\|_F$ , then  $\|S(\lambda)\|_F / \|R(\lambda)\|_F \gg 1$  may happen, according to (4.33), and  $K_{S,R} \gg 1$  in that situation. We emphasize that the condition  $\|M(\lambda)\|_F \approx \|D(\lambda)\|_F$  was also used in the analysis in [7, Corollary 5.24].

Next, we prove the announced result that conditions (4.32) are sufficient for  $K_{S,R}$  to be moderate and, thus, for structural backward stability.

**Corollary 4.4** *Under the hypotheses and with the notation of Theorem 4.6, assume, in addition, that (4.32) holds and let  $t := \max(\eta, \epsilon) > 0$ . Then,*

$$K_{S,R} \leq g t^q \sqrt{m + n},$$



where  $q = 5$ , if  $\eta > 0$  and  $\epsilon > 0$ ,  $q = 9/2$ , if  $\eta = 0$  or  $\epsilon = 0$ , and  $g$  is a moderate number (a constant that does not depend on  $\eta, \epsilon, m, n, \ell$ ). Moreover

$$\frac{\sqrt{\|\Delta A\|_F^2 + \|\Delta B\|_F^2 + \|\Delta C\|_F^2 + \sum_{i=0}^d \|\Delta D_i\|_F^2}}{\|R(\lambda)\|_F} \leq g t^q \sqrt{m+n} \frac{\|\Delta_S(\lambda)\|_F}{\|S(\lambda)\|_F} + \mathcal{O}(\delta^2).$$

**Proof** Note that (4.32) and Lemmas 4.7 and 4.8 imply  $\|S(\lambda)\|_2 \lesssim 2 + \sqrt{2}$ ,  $f_1 \leq g_1 t^2$ ,  $f_2 \leq g_2 t$ , and  $f_3 \leq g_3 t$ , with  $g_1, g_2, g_3$  moderate numbers. Moreover, from (4.33), (4.32) and  $\|R(\lambda)\|_F \geq 1$ , we get that  $\|S(\lambda)\|_F^2 \approx \|R(\lambda)\|_F^2 + 2(m\eta + n\epsilon)$  and

$$\|S(\lambda)\|_F^2 \leq (1 + 2(m\eta + n\epsilon)) \|R(\lambda)\|_F^2 \leq 3(m+n)t \|R(\lambda)\|_F^2.$$

It only remains to analyze the factor  $\sqrt{2 \min(\epsilon + 1, \eta + 1)}$  of  $K_{S,R}$ , which is less than or equal to  $\sqrt{2(t+1)}$ , if  $\eta > 0$  and  $\epsilon > 0$ , or equal to  $\sqrt{2}$ , if  $\eta = 0$  or  $\epsilon = 0$ . Combining all these bounds with the fact that  $t \geq 1$ , the result follows as a corollary of Theorem 4.6. □

**Remark 4.1** Observe that (4.32) allow  $\max(\|A\|_F, \|B\|_F, \|C\|_F, \|D(\lambda)\|_F) \ll 1$ . However, since the rational matrix  $R(\lambda)$  in (1.1) can be multiplied by a nonzero number without affecting at all its eigenstructure, it is natural and convenient to use as sufficient conditions

$$\max(\|A\|_F, \|B\|_F, \|C\|_F, \|D(\lambda)\|_F) = 1 \quad \text{and} \quad \|M(\lambda)\|_F \approx \|D(\lambda)\|_F. \tag{4.34}$$

Such conditions would have appeared as sufficient in the analysis if we had defined the norm of  $R(\lambda)$  as

$$\|R(\lambda)\|_F := \sqrt{\|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2 + \sum_{i=0}^d \|D_i\|_F^2}, \tag{4.35}$$

instead as in (1.3) (observe that we have removed the  $\ell$  summand), depending only on the free parameters of the representation of  $R(\lambda)$  in (1.1). We have chosen to use (1.3) because, first, it identifies the informal “norm” of  $R(\lambda)$  with the formal norm of the polynomial system matrix  $P(\lambda)$  and, second, it corresponds to the particular case  $E = I_\ell$  of the more general representation  $R(\lambda) = C(\lambda E - A)^{-1}B + D(\lambda)$ , with  $E$  nonsingular, when taking as norm the one of the corresponding polynomial system matrix. Under the conditions (4.34), it is essentially equivalent to use (1.3) or (4.35) as “norm” of  $R(\lambda)$ . The use of representations  $R(\lambda) = C(\lambda E - A)^{-1}B + D(\lambda)$  for rational matrices is of interest in certain applications and the block Kronecker linearizations in this case are obtained just by replacing  $A - \lambda I_\ell$  by  $A - \lambda E$  in (1.2). We will consider the analysis of this general case in the future.

#### 4.6 Restoring the structure when the polynomial part of the rational matrix is linear

In this subsection, we consider the particular case of having a rational matrix with linear polynomial part. That is, the case of having a rational matrix that can be written in the form

$$R(\lambda) = C(\lambda I_\ell - A)^{-1}B + M(\lambda),$$

where  $C(\lambda I_\ell - A)^{-1}B$  is a minimal state-space realization and  $M(\lambda)$  is a matrix pencil. Then  $R(\lambda)$  can be strongly linearized using the following linear polynomial system matrix

$$S(\lambda) := \begin{bmatrix} M(\lambda) & C \\ B & A - \lambda I_\ell \end{bmatrix}. \quad (4.36)$$

Notice that, in this case, the linearization does not have the block anti-triangular structure as the block Kronecker linearization in (1.2) since  $K_1(\lambda)$  and  $K_2(\lambda)$  are empty matrices. The strong linearization (4.36) can be seen as the limit case of (1.2) when  $\epsilon = \eta = 0$ .

If we compute the eigenstructure of  $S(\lambda)$ , the backward stability of the staircase algorithm [21] and the  $QZ$  algorithm [16] guarantees that we computed the exact eigenstructure of a slightly perturbed pencil

$$\widehat{S}(\lambda) := S(\lambda) + \Delta_S(\lambda), \quad \Delta_S(\lambda) := \begin{bmatrix} \Delta_{11}(\lambda) & \Delta_{12}(\lambda) \\ \Delta_{21}(\lambda) & \Delta_{22}(\lambda) \end{bmatrix}. \quad (4.37)$$

The structure of (4.36) is lost in (4.37) since the off-diagonal blocks of  $\widehat{S}(\lambda)$  are not constant matrices and the identity block  $I_\ell$  is not preserved by the perturbation.

Notice that restoring in  $\widehat{S}(\lambda)$  the original structure of  $S(\lambda)$  is much simpler than in previous sections, as we do not have to restore any anti-triangular zero block nor the minimal bases  $K_1(\lambda)$  and  $K_2(\lambda)$  in (4.1). We only have to take care of restoring the identity matrix  $I_\ell$  and the constant matrices  $B$  and  $C$  to obtain in two steps a new strictly equivalent linear polynomial system matrix

$$\widetilde{S}(\lambda) := (I - X)\widehat{S}(\lambda)(I - Y) := (I - X_2)(I - X_1)\widehat{S}(\lambda)(I - Y_1)(I - Y_2) \quad (4.38)$$

of the form

$$\widetilde{S}(\lambda) := \begin{bmatrix} \widetilde{M}(\lambda) & \widetilde{C} \\ \widetilde{B} & \widetilde{A} - \lambda I_\ell \end{bmatrix}, \quad (4.39)$$

where  $\widetilde{M}(\lambda) := M(\lambda) + \Delta M(\lambda)$ ,  $\widetilde{A} := A + \Delta A$ ,  $\widetilde{B} := B + \Delta B$  and  $\widetilde{C} := C + \Delta C$ . For that, we consider the discussion in Sect. 4.2, for restoring  $I_\ell$ ; and a simplified version of the discussion in Sect. 4.3, for restoring the constant matrices  $B$  and  $C$ .

In particular, from the bound in (4.21) and a counterpart of Theorem 4.4 we get the following result.

**Theorem 4.7** *Let  $S(\lambda)$  be a minimal linear system matrix as in (4.36). The transformation  $(X, Y)$  in (4.38) exists and we can bound the corresponding perturbation  $\tilde{S}(\lambda) - S(\lambda)$  as follows:*

$$\|\tilde{S}(\lambda) - S(\lambda)\|_F \leq (1 + \sqrt{2}\|S(\lambda)\|_2)^2 \|\Delta_S(\lambda)\|_F + \mathcal{O}(\delta^2). \tag{4.40}$$

*In addition, if  $\|\tilde{S}(\lambda) - S(\lambda)\|_F$  is sufficiently small, then the perturbed pencil  $\tilde{S}(\lambda)$  is a minimal linear system matrix of the rational matrix  $\tilde{R}(\lambda) = \tilde{C}(\lambda I_\ell - A)^{-1} \tilde{B} + \tilde{M}(\lambda)$  and*

$$\frac{\sqrt{\|\Delta A\|_F^2 + \|\Delta B\|_F^2 + \|\Delta C\|_F^2 + \|\Delta M(\lambda)\|_F^2}}{\|R(\lambda)\|_F} \leq (1 + \sqrt{2}\|S(\lambda)\|_2)^2 \frac{\|\Delta_S(\lambda)\|_F}{\|S(\lambda)\|_F} + \mathcal{O}(\delta^2),$$

where  $\delta = \|\Delta_S(\lambda)\|_F / \|S(\lambda)\|_F$ .

The simplicity of the bound in Theorem 4.7 is also a consequence of  $\|S(\lambda)\|_F = \|R(\lambda)\|_F$ .

### 5 Scaling for obtaining structural backward stability

Once a block Kronecker linearization  $S(\lambda)$  in (1.2) of  $R(\lambda)$  in (1.1) satisfying  $\|M(\lambda)\|_F \approx \|D(\lambda)\|_F$  is chosen and the staircase or the *QZ* algorithm is applied to  $S(\lambda)$ , structural backward stability is guaranteed for the computed eigenstructure if the first condition in (4.32) holds. However, there exist rational matrices which do not satisfy  $\max(\|A\|_F, \|B\|_F, \|C\|_F, \|D(\lambda)\|_F) \leq 1$  and, therefore, the computation of their eigenstructure via a block Kronecker linearization might not be structurally backward stable. In this section, we study how to proceed in these cases.

First observe that the eigenstructure of the rational matrix  $R(\lambda)$  does not change at all if it is multiplied by a positive real constant  $d_R$ . Choosing appropriately  $d_R$ , we get easily a rational matrix such that  $\max(\|B\|_F, \|C\|_F, \|D(\lambda)\|_F) \leq 1$ . Even more, if  $d_R$  is an integer power of 2, this multiplication can be performed without introducing any rounding error. This indicates that the crucial point is how to deal with rational matrices with  $\|A\|_F > 1$ . For this, note that when representing a rational matrix  $R(\lambda)$  by a realization quadruple  $\{\lambda I_\ell - A, B, C, D(\lambda)\}$ , where  $D(\lambda)$  is polynomial,

$$R(\lambda) := C(\lambda I_\ell - A)^{-1} B + \sum_{i=0}^d D_i \lambda^i,$$

one can change the coordinate system of the state-space realization  $\{A, B, C\}$  of the strictly proper part of  $R(\lambda)$  by a diagonal similarity scaling  $T := \text{diag}(d_1, \dots, d_\ell)$ ,

$d_i > 0$ , without changing  $R(\lambda)$  since

$$C(\lambda I_\ell - A)^{-1}B = CT(\lambda I_\ell - T^{-1}AT)^{-1}T^{-1}B.$$

Thus, before multiplying  $R(\lambda)$  by  $d_R$ , we can choose  $T$  to balance  $A$ , i.e., to minimize its Frobenius norm under all diagonal similarities by making the 2-norms of the rows and columns of  $T^{-1}AT$  become equal [17]. Moreover, at the same time, the Frobenius norms of  $T^{-1}B$  and  $CT$  can be made equal by considering a positive scalar factor multiplying  $T$ . Observe, in addition, that if the entries of  $T$  are integer powers of 2, this process does not introduce rounding errors, though, in this case, the norm of  $T^{-1}AT$  is only approximately minimized. However, the effects of  $T$  are limited since  $\|T^{-1}AT\|_F \geq \sqrt{|\lambda_1|^2 + \dots + |\lambda_\ell|^2}$ , where  $\lambda_1, \dots, \lambda_\ell$  are the eigenvalues of  $A$ , for any invertible  $T$ , i.e., diagonal or not. Therefore, other approaches are needed for dealing with all instances of matrices  $A$  with large norms. It is important to emphasize at this point that the influence of a large norm matrix  $A$  on the bound (4.31) is huge, because it contributes to  $\|S(\lambda)\|_2$ , but also the factor  $\|A\|_2^{\max(\eta, \epsilon)}$  is present in both  $f_1$  and  $f_3$ .

The final solution comes from changing the variable  $\lambda$  to  $\widehat{\lambda} := d_\lambda \lambda$  and from combining this with the multiplication by the constant  $d_R$  and the diagonal scaling  $T$  discussed above. Note that the change of variable transforms the zeros and the poles of  $R(\lambda)$  in a very simple way, preserving their partial multiplicities, and that does not change at all its minimal indices [15, 20]. The combination of all these scalings yields a new transfer function

$$\widehat{R}(\widehat{\lambda}) := \widehat{D}(\widehat{\lambda}) + \widehat{C}(\widehat{\lambda}I_\ell - \widehat{A})^{-1}\widehat{B} := d_R R(\widehat{\lambda}/d_\lambda) \tag{5.1}$$

where

$$\widehat{A} := d_\lambda T^{-1}AT, \quad \widehat{B} := \sqrt{d_\lambda d_R} T^{-1}B, \quad \widehat{C} := \sqrt{d_\lambda d_R} CT \tag{5.2}$$

and

$$\widehat{D}_i := d_R d_\lambda^{-i} D_i, \quad \text{for all } i = 0, 1, \dots, d. \tag{5.3}$$

Then, we can choose  $d_\lambda := \min(1, \|T^{-1}AT\|_F^{-1})$ , such that  $\widehat{A}$  has norm smaller than or equal to 1. Note that the preliminary balancing will make this step milder, in the sense that  $d_\lambda$  will be closer to 1. Finally, based on (5.1), we summarize the following scaling procedure for obtaining a rational matrix  $\widehat{R}(\widehat{\lambda})$  with  $\max(\|\widehat{A}\|_F, \|\widehat{B}\|_F, \|\widehat{C}\|_F, \|\widehat{D}(\widehat{\lambda})\|_F) = 1$  from the data  $\{A, B, C, D_0, D_1, \dots, D_d\}$ :

**Step 1.** Compute  $T = \text{diag}(d_1, \dots, d_\ell)$  to balance  $A$  and to make equal the norms of  $T^{-1}B$  and  $CT$ .

**Step 2.** Choose  $d_\lambda := \min(1, \|T^{-1}AT\|_F^{-1})$ .

**Step 3.** Choose

$$d_R = \frac{1}{\max(\|\sqrt{d_\lambda} T^{-1} B\|_F^2, \|\sqrt{d_\lambda} C T\|_F^2, \sqrt{\sum_{i=0}^d \|d_\lambda^{-i} D_i\|_F^2})}$$

**Step 4.** Compute  $\{\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}_0, \widehat{D}_1, \dots, \widehat{D}_d\}$  as in (5.2)–(5.3).

This process can be easily arranged to use scale factors that are all integer powers of two and, thus, can be implemented without any rounding error. Moreover, this scaling can be applied directly to the pencil  $S(\lambda)$ . More precisely, the pencil

$$\widehat{S}(\widehat{\lambda}) := D_\ell S(\widehat{\lambda}/d_\lambda) D_r,$$

where the left and right diagonal scalings  $D_\ell$  and  $D_r$  are given by

$$D_\ell := \text{diag}(d_R^{\frac{1}{2}} d_\lambda^{-\eta} I_m, \dots, d_R^{\frac{1}{2}} d_\lambda^0 I_m, d_\lambda^{\frac{1}{2}} d_1^{-1}, \dots, d_\lambda^{\frac{1}{2}} d_\ell^{-1}, d_R^{-\frac{1}{2}} d_\lambda^\epsilon I_n, \dots, d_R^{-\frac{1}{2}} d_\lambda^1 I_n),$$

$$D_r := \text{diag}(d_R^{\frac{1}{2}} d_\lambda^{-\epsilon} I_n, \dots, d_R^{\frac{1}{2}} d_\lambda^0 I_n, d_\lambda^{\frac{1}{2}} d_1, \dots, d_\lambda^{\frac{1}{2}} d_\ell, d_R^{-\frac{1}{2}} d_\lambda^\eta I_m, \dots, d_R^{-\frac{1}{2}} d_\lambda^1 I_m),$$

is a block Kronecker linearization of the rational matrix  $\widehat{R}(\widehat{\lambda})$  in (5.1).

### 6 Numerical experiments

In this section, we describe three experiments illustrating that the potential sources of structural backward instability revealed by the bound (4.31) are indeed observed in practice. More precisely, the experiments will illustrate that if a rational matrix  $R(\lambda)$  as in (1.1) does not satisfy the first condition in (4.32), then the computation of the eigenstructure of  $R(\lambda)$  by applying the  $QZ$  algorithm to a block Kronecker linearization  $S(\lambda)$  of  $R(\lambda)$  that satisfies  $\|M(\lambda)\|_F = \|D(\lambda)\|_F$  is not structurally backward stable. Moreover, the experiments also illustrate that the scaling described in Sect. 5 is effective and leads to structured backward stability for the scaled rational matrices and linearizations.

A difficulty for performing fully reliable numerical experiments in this setting is that to estimate the actual global backward error for the *whole* computed eigenstructure, i.e., the left-hand side of (4.31), is a challenging optimization problem for which we do not know yet a solution. Therefore, we will limit ourselves to computing a lower bound for the backward error based on the “local” backwards errors of each computed zero of the rational matrix, as we explain below. This lower bound might severely underestimate the actual global backward error. Thus, we cannot check from our experiments the sharpness of the bound (4.31), which, on the other hand, was deduced through many potentially overestimating inequalities with the main goal of getting a bound as clear as possible instead of optimizing its sharpness.

For simplicity, we will restrict our numerical experiments to square and regular rational matrices  $R(\lambda)$  with a corresponding quadruple  $\{A, B, C, D(\lambda)\}$  of moderate

dimensions and degree of its polynomial part:  $m = n = 2, \ell = 5, d = 3$ . The block Kronecker pencil we choose for our computations is

$$S(\lambda) := \begin{bmatrix} \lambda D_3 + D_2 & 0 & 0 & I_2 \\ 0 & \lambda D_1 + D_0 & C & -\lambda I_2 \\ 0 & B & A - \lambda I_\ell & 0 \\ I_2 & -\lambda I_2 & 0 & 0 \end{bmatrix},$$

which has  $\eta$  and  $\epsilon$  equal to 1, size  $11 \times 11$  and satisfies  $\|M(\lambda)\|_F = \|D(\lambda)\|_F$ . We also will look at the polynomial system matrix

$$P(\lambda) := \begin{bmatrix} A - \lambda I_\ell & B \\ C & D(\lambda) \end{bmatrix}, \quad D(\lambda) := D_0 + \lambda D_1 + \lambda^2 D_2 + \lambda^3 D_3$$

of  $R(\lambda)$  because it allows us to estimate the backward errors of our algorithm as follows. We look for a rational matrix  $\tilde{R}(\lambda)$  corresponding to a quadruple  $\{A + \Delta A, B + \Delta B, C + \Delta C, (D + \Delta D)(\lambda)\}$  such that all its finite zeros are exactly all the computed finite eigenvalues obtained by applying the  $QZ$  algorithm to  $S(\lambda)$  and such that  $\|(\Delta A, \Delta B, \Delta C, (\Delta D)(\lambda))\|_F$  is as small as possible. As a consequence of the classical results of Rosenbrock [18], this is equivalent to find a perturbed polynomial system matrix  $P(\lambda) + \Delta P(\lambda)$  of  $\tilde{R}(\lambda)$ , whose finite zeros are the computed eigenvalues  $\lambda_i$  and such that  $\|(\Delta A, \Delta B, \Delta C, (\Delta D)(\lambda))\|_F$  is as small as possible. Therefore,  $\{\Delta A, \Delta B, \Delta C, \Delta D_0, \Delta D_1, \Delta D_2, \Delta D_3\}$  must have the property that *simultaneously*, at each computed eigenvalue  $\lambda_i$ , the matrix

$$P(\lambda_i) + \Delta P(\lambda_i) = P(\lambda_i) + \left[ \begin{array}{cc|cc|cc|cc} \Delta A & \Delta B & 0 & 0 & 0 & 0 & 0 & 0 \\ \Delta C & \Delta D_0 & 0 & \Delta D_1 & 0 & \Delta D_2 & 0 & \Delta D_3 \end{array} \right] \begin{bmatrix} I_{\ell+m} \\ \lambda_i I_{\ell+m} \\ \lambda_i^2 I_{\ell+m} \\ \lambda_i^3 I_{\ell+m} \end{bmatrix}$$

must be singular. To find the smallest possible Frobenius norm of all possible  $\{\Delta A, \Delta B, \Delta C, \Delta D_0, \Delta D_1, \Delta D_2, \Delta D_3\}$  that satisfy this property *for all* computed  $\lambda_i$  is not obvious, however to solve this problem *for only one* computed  $\lambda_i$  is easy. For this purpose, let  $\Delta^{(i)}$  be the minimum Frobenius norm matrix that makes  $P(\lambda_i) + \Delta^{(i)}$  singular. Note that  $\Delta^{(i)}$  can be computed through the singular value decomposition of  $P(\lambda_i)$  and that, generically, it is a rank one matrix with Frobenius norm equal to  $\sigma_{\min} P(\lambda_i)$ . Then, the linear system

$$\Delta^{(i)} := \begin{bmatrix} \Delta_{11}^{(i)} & \Delta_{12}^{(i)} \\ \Delta_{21}^{(i)} & \Delta_{22}^{(i)} \end{bmatrix} = \left[ \begin{array}{cc|cc|cc|cc} \Delta A & \Delta B & 0 & 0 & 0 & 0 & 0 & 0 \\ \Delta C & \Delta D_0 & 0 & \Delta D_1 & 0 & \Delta D_2 & 0 & \Delta D_3 \end{array} \right] \begin{bmatrix} I_{\ell+m} \\ \lambda_i I_{\ell+m} \\ \lambda_i^2 I_{\ell+m} \\ \lambda_i^3 I_{\ell+m} \end{bmatrix}$$

for the unknowns  $\{\Delta A, \Delta B, \Delta C, \Delta D_0, \Delta D_1, \Delta D_2, \Delta D_3\}$  is consistent and its minimum Frobenius norm solution is given by

$$\Delta A := \Delta_{11}^{(i)}, \quad \Delta B := \Delta_{12}^{(i)}, \quad \Delta C := \Delta_{21}^{(i)}, \quad \Delta D_k := \Delta_{22}^{(i)} \bar{\lambda}_i^k / g(\lambda_i), \quad k = 0, 1, 2, 3,$$

where  $g(\lambda_i) := (1 + |\lambda_i|^2 + |\lambda_i|^4 + |\lambda_i|^6)$ , and the Frobenius norm of this 7-tuple of matrices is given by

$$r(P, \lambda_i) := \left\| \begin{bmatrix} \Delta_{11}^{(i)} & \Delta_{12}^{(i)} \\ \Delta_{21}^{(i)} & \Delta_{22}^{(i)} / \sqrt{g(\lambda_i)} \end{bmatrix} \right\|_F.$$

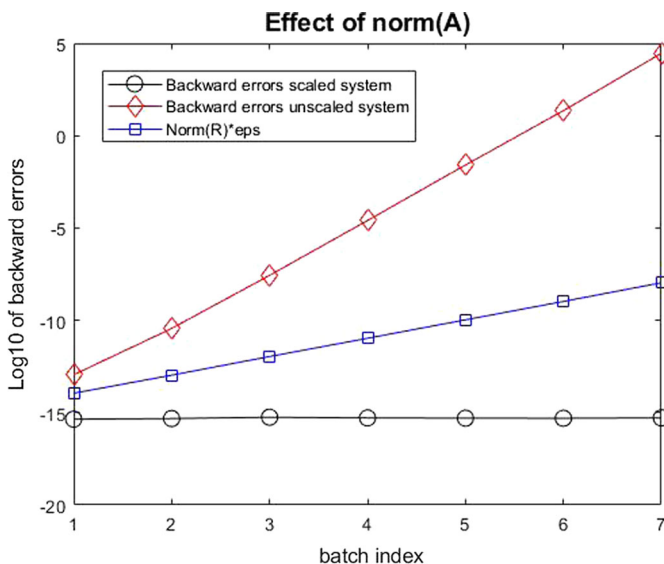
This leads us to use in our experiments

$$r(P) := \max_i r(P, \lambda_i) \tag{6.1}$$

as an estimate for the structured absolute backward error induced by our algorithm, i.e., as an estimate for the numerator of the left-hand side of (4.31). We emphasize that this is a lower bound for the actual global structured backward error, since it corresponds to a rational matrix that has only one of the computed eigenvalues as a finite zero.

In the first experiment, we investigate the behavior of the structured backward error for rational matrices with matrices  $A$  of increasing (large) norms, and with the rest of the matrices in the quadruple  $\{A, B, C, D(\lambda)\}$  having norms of order 1. The reason why we pay first particular attention to the norm of  $A$  is because according to the bound (4.31) the influence of  $A$  should be huge because it contributes to  $\|S(\lambda)\|_2$  and also to  $f_1$  and  $f_3$ . For this purpose, we generated with the Matlab function `randn`, 7 batches of samples of 50 random matrix-tuples  $\{A, B, C, D_0, D_1, D_2, D_3\}$ , and in each batch indexed with  $i$ , we multiplied the matrix  $A$  by  $10^i$ , with  $i$  going from 1 till 7, in each of the 50 runs of each batch. In each batch, we computed the average of the absolute backward error estimators (6.1) for both the original matrix-tuples and the scaled ones after applying the procedure in Sect. 5. In Fig. 1, we plot the results of these computations: the horizontal axis represents the index  $i$  defining each batch and the vertical axis the logarithm of the average absolute backward errors. Ideally, the absolute backward error should be of order  $\epsilon_M \|R(\lambda)\|_F$ , where  $\epsilon_M$  is the machine precision, and, so, we also plot this magnitude for the unscaled original data taking in each batch the average of all  $\|R(\lambda)\|_F$  (for the scaled data, this magnitude is always of order  $\epsilon_M$  and is not plotted). We observe that the absolute backward errors for the unscaled problem grow very strongly with the index  $i$ , i.e., with the norm of  $A$ , and that computing the zeros of a rational matrix by applying the  $QZ$  algorithm to the block Kronecker linearization  $S(\lambda)$  is highly structurally backward unstable for large norms of  $A$ , as predicted by the bound (4.31). In contrast, when applying the scaling procedure described in Sect. 5, this growth is absent and we get perfect structural backward stability for the scaled rational matrix, as predicted by (4.31).

In the second experiment, we investigate the behavior of the structured backward error for rational matrices with matrices  $A$  of norms of order 1, and with the rest



**Fig. 1** Experiment 1: behavior of absolute structured backward errors for increasing values of the norm of  $A$

of the matrices in the quadruple  $\{A, B, C, D(\lambda)\}$  having increasing (large) norms. The situation in this experiment is opposite to the one in the first experiment. The matrices are generated following the same pattern of the first experiment except by the fact that once the matrices  $\{A, B, C, D_0, D_1, D_2, D_3\}$  are generated with `randn`,  $B$  is multiplied by  $10^{i/2}$ ,  $C$  by  $10^{i/3}$ ,  $D_1$  by  $10^i$ ,  $D_2$  by  $10^{i/2}$  and  $D_3$  by  $10^{i/3}$ , for  $i = 1, \dots, 7$ . The results are plotted in Fig. 2 and the conclusions are the same as in the first experiment and are in agreement with our analysis. However, note that the growth of the absolute backward errors of the original unscaled data is much smaller than in the first experiment. This effect is qualitatively expected from the bound (4.31), since  $f_3$  does not depend on the norms of  $B, C$  and  $D(\lambda)$ , but the observed very large quantitative difference is not fully explained by (4.31). Possible reasons of this are that, as we have emphasized before, our backward error estimator is a lower bound that may underestimate severely the actual global backward error and/or that the bound in (4.31) overestimates the actual error.

The last experiment we present combines the scalings used in the first and second experiments. That is, once the matrices  $\{A, B, C, D_0, D_1, D_2, D_3\}$  are generated with `randn`,  $A$  is multiplied by the factor used in Experiment 1 and  $B, C, D_1, D_2$ , and  $D_3$  are multiplied by the factors used in Experiment 2. Taking into account that the function  $f_1$  appearing in the bound (4.31) includes a product of the norm of  $A$  times the norm of  $B$  and a product of the norm of  $A$  times the norm of  $C$ , we expect backward errors for the unscaled system larger than those of Experiment 1. The results are plotted in Fig. 3. The errors for the unscaled system (red line) are indeed larger than those in Fig. 1, but just a bit larger. The possible reasons of this small increment of the errors are the same as in the second experiment.



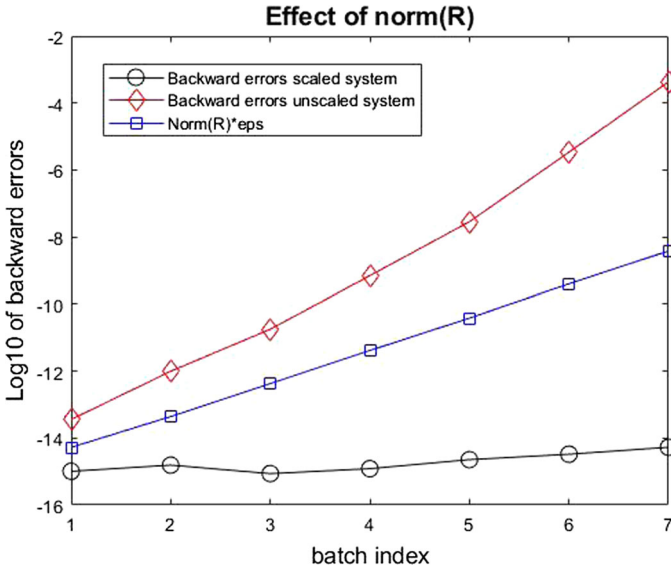


Fig. 2 Experiment 2: behavior of absolute structured backward errors for increasing values of the norms of  $B$ ,  $C$  and  $D(\lambda)$

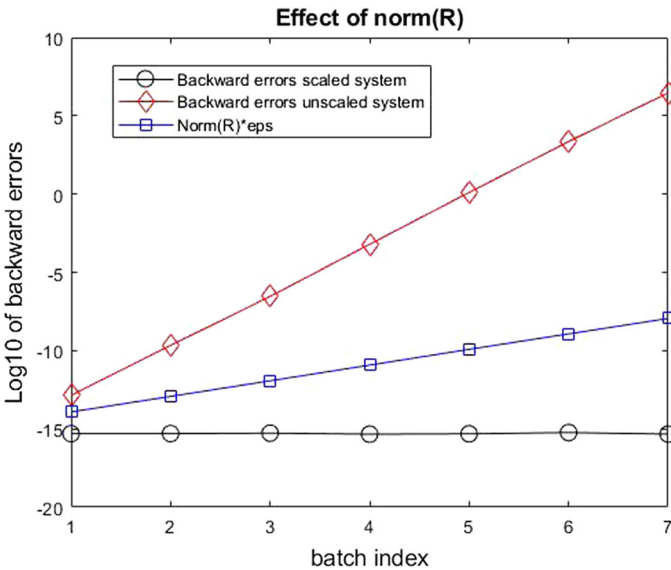


Fig. 3 Experiment 3: behavior of absolute structured backward errors for increasing values of the norms of  $A$ ,  $B$ ,  $C$  and  $D(\lambda)$

The main conclusion of this section is that our main *a priori* structured backward error bound (4.31) identifies correctly the sources of instability of computing the eigenstructure of a rational matrix by applying the  $QZ$  algorithm to its block Kronecker

linearizations and that the scaling proposed in Sect. 5 leads to structural backward stability.

### 7 Conclusions and future work

We have developed the first structured backward error analysis for an algorithm that computes the eigenstructure of a rational matrix. More precisely, the considered algorithm starts from a rational matrix expressed as in (1.1) and computes its eigenstructure by applying a backward stable generalized eigenproblem algorithm to its block Kronecker linearizations described in (1.2). As a consequence of this analysis, we have identified the simple sufficient conditions (4.32) for structural backward stability. In the case of rational matrices which do not satisfy these conditions, we have developed a scaling procedure that transforms the original matrix in another one for which structural backward stability is guaranteed. A number of numerical experiments confirming the predictions of the backward error analysis have been performed and discussed. The results in this paper open new research problems in the area of structured backward error analysis, since other representations used in applications of the given rational matrix should be considered in the future, as well as other families of linearizations.

### A Auxiliary result for Lemma 3.4

We prove in this appendix that the matrix

$$\left[ \begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] := \left[ \begin{array}{c|c} E_k^T \otimes I_k & I_{(k+1)} \otimes E_k \\ \hline F_k^T \otimes I_k & I_{(k+1)} \otimes F_k \end{array} \right]$$

appearing in the proof of Lemma 3.4 can be transformed by row and column permutations to the direct sum of the following matrices:

$$M_1 \oplus M_1 \oplus M_3 \oplus M_3 \oplus \dots \oplus M_{2k-1} \oplus M_{2k-1} \oplus N_{2k},$$

where the blocks  $M_k$  and  $N_k$  are as defined in (3.9). Let us take for example  $k = 3$ , then the matrix looks like

$$\left[ \begin{array}{c|ccc} I_3 & & & E_3 \\ & I_3 & & E_3 \\ & & I_3 & E_3 \\ \hline & & & F_3 \\ I_3 & & & F_3 \\ & I_3 & & F_3 \\ & & I_3 & F_3 \end{array} \right].$$

There are three submatrices  $M_1$ ,  $M_3$  and  $M_5$  that take elements  $a$ ,  $b$ ,  $c$  and  $d$  in the respective blocks  $A$ ,  $B$ ,  $C$  and  $D$ , as indicated below

$$M_1 = [b], \quad M_3 = \begin{bmatrix} b & a \\ & c & d \\ & & b \end{bmatrix}, \quad M_5 = \begin{bmatrix} b & a \\ & c & d \\ & & b \end{bmatrix}$$

and they each start with a leading element in one of the  $E_3$  blocks. For instance,  $M_1 = [b_{10,13}]$ ,  $M_3$  starts with the leading element  $b_{7,9}$  in the third  $E_3$  block, and  $M_5$  starts with the leading element in the second  $E_3$  block:

$$M_1 = [b_{10,13}], \quad M_3 = \begin{bmatrix} b_{7,9} & a_{7,7} & & & \\ & c_{10,7} & d_{10,14} & & \\ & & b_{11,14} & & \\ & & & & \end{bmatrix}, \quad M_5 = \begin{bmatrix} b_{4,5} & a_{4,4} & & & \\ & c_{7,4} & d_{7,10} & & \\ & & b_{8,10} & a_{8,8} & \\ & & & c_{11,8} & d_{11,15} \\ & & & & b_{12,15} \end{bmatrix}.$$

Notice that the  $[b \ a]$  and  $[c \ d]$  pairs have the same row index and that the  $\begin{bmatrix} a \\ c \end{bmatrix}$  and  $\begin{bmatrix} d \\ b \end{bmatrix}$  pairs have the same column index, which explains the permutation that has to be constructed to extract the matrix. Also the transitions

$$b_{7,9} \rightarrow b_{11,14}, \quad \text{and} \quad b_{4,5} \rightarrow b_{8,10} \rightarrow b_{12,15}$$

always go down to the next diagonal element in the next  $E_3$  block. In a similar fashion, one finds another set of submatrices  $M_1$ ,  $M_3$  and  $M_5$  that take elements  $a$ ,  $b$ ,  $c$  and  $d$  in the respective blocks  $A$ ,  $B$ ,  $C$  and  $D$  in a different order, as indicated below

$$M_1 = [d], \quad M_3 = \begin{bmatrix} d & c \\ & a & b \\ & & d \end{bmatrix}, \quad M_5 = \begin{bmatrix} d & c \\ & a & b \\ & & d \end{bmatrix}$$

and they each start with a trailing element in one of the first three  $F_3$  blocks. Finally, the remaining matrix  $N_6$  takes elements in the blocks  $A$ ,  $B$ ,  $C$  and  $D$  in the following order

$$N_6 = \begin{bmatrix} b & a & & & & \\ & c & d & & & \\ & & b & a & & \\ & & & c & d & \\ & & & & b & a \\ & & & & & c & d \end{bmatrix}$$

and starts with the leading element in the leading  $E_3$  block, and ends with the trailing element in the trailing  $F_3$  block.

## References




1. Alam, R., Behera, N.: Linearizations for rational matrix functions and Rosenbrock system polynomials. *SIAM J. Matrix Anal. Appl.* **37**(1), 354–380 (2016)
2. Amparan, A., Dopico, F.M., Marcaida, S., Zaballa, I.: Strong linearizations of rational matrices. *SIAM J. Matrix Anal. Appl.* **39**(4), 1670–1700 (2018)
3. Amparan, A., Dopico, F.M., Marcaida, S., Zaballa, I.: On minimal bases and indices of rational matrices and their linearizations. *Linear Algebra Appl.* **623**, 14–67 (2021)
4. Das, R., Alam, R.: Affine spaces of strong linearizations for rational matrices and the recovery of eigenvectors and minimal bases. *Linear Algebra Appl.* **569**, 335–368 (2019)
5. De Terán, F., Dopico, F.M., Mackey, D.S.: Spectral equivalence of matrix polynomials and the index sum theorem. *Linear Algebra Appl.* **459**, 264–333 (2014)
6. Dmytryshyn, A., Kågström, B.: Coupled Sylvester-type matrix equations and block diagonalization. *SIAM J. Matrix Anal. Appl.* **36**(2), 580–593 (2015)
7. Dopico, F.M., Lawrence, P.W., Pérez, J., Van Dooren, P.: Block Kronecker linearizations of matrix polynomials and their backward errors. *Numer. Math.* **140**, 373–426 (2018)
8. Dopico, F.M., Pérez, J., Van Dooren, P.: Structured backward error analysis of linearized structured polynomial eigenvalue problems. *Math. Comp.* **88**, 1189–1228 (2019)
9. Gantmacher, F.R.: *The Theory of Matrices*, Vols I and II. Chelsea, New York (1959)
10. Gohberg, I., Lancaster, P., Rodman, L.: *Matrix Polynomials*, SIAM Publications, Philadelphia, 2009. Academic Press, New York (1982)
11. Higham, N.J.: *Accuracy and Stability of Numerical Algorithms*, 2nd edn. SIAM Publications, Philadelphia (2002)
12. Higham, N.J., Li, R.-C., Tisseur, F.: Backward error of polynomial eigenproblems solved by linearization. *SIAM J. Matrix Anal. Appl.* **29**(4), 1218–1241 (2007)
13. Horn, R.A., Johnson, C.R.: *Topics in Matrix Analysis*. Cambridge University Press, Cambridge (1994). (**Corrected reprint of the 1991 original**)
14. Kailath, T.: *Linear Systems*. Prentice Hall, Englewood Cliffs, NJ (1980)
15. Mackey, D.S., Mackey, N., Mehl, C., Mehrmann, V.: Möbius transformations of matrix polynomials. *Linear Algebra Appl.* **470**, 120–184 (2015)
16. Moler, C., Stewart, G.W.: An algorithm for generalized matrix eigenvalue problems. *SIAM J. Numer. Anal.* **10**(2), 241–256 (1973)
17. Parlett, B.N., Reinsch, C.: Balancing a matrix for calculation of eigenvalues and eigenvectors. *Numer. Math.* **13**, 293–304 (1969)
18. Rosenbrock, H.: *State-Space and Multivariable Theory*. Thomas Nelson and Sons, London (1970)
19. Su, Y., Bai, Z.: Solving rational eigenvalue problems via linearization. *SIAM J. Matrix Anal. Appl.* **32**(1), 201–216 (2011)
20. Van Dooren, P.: *The Generalized Eigenstructure Problem: Applications in Linear System Theory*. PhD thesis, Katholieke Universiteit Leuven, Leuven, Belgium (1979)
21. Van Dooren, P.: The computation of Kronecker’s canonical form of a singular pencil. *Linear Algebra Appl.* **27**, 103–140 (1979)
22. Van Dooren, P.: The generalized eigenstructure problem in linear system theory. *IEEE Trans. Autom. Control.* **26**(1), 111–129 (1981)

23. Van Dooren, P.: Reducing subspaces: definitions, properties and algorithms. *Matrix Pencils*, Lecture Notes in Mathematics, Vol. **973**, Springer, pp. 58–73 (1983)
24. Van Dooren, P., Dewilde, P.: The eigenstructure of an arbitrary polynomial matrix: computational aspects. *Linear Algebra Appl.* **50**, 545–579 (1983)
25. Van Dooren, P., Dopico, F.M.: Robustness and perturbations of minimal bases. *Linear Algebra Appl.* **542**, 246–281 (2018)
26. Verghese, G., Van Dooren, P., Kailath, T.: Properties of the system matrix of a generalized state-space system. *Int. J. Control* **30**(2), 235–243 (1979)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

## Authors and Affiliations

Froilán M. Dopico<sup>1</sup>  · María C. Quintana<sup>2</sup>  · Paul Van Dooren<sup>3</sup> 

Froilán M. Dopico  
dopico@math.uc3m.es

María C. Quintana  
maría.quintanaponce@aalto.fi

- <sup>1</sup> Departamento de Matemáticas, Universidad Carlos III de Madrid, Avda. Universidad 30, Leganés, 28911 Madrid, Spain
- <sup>2</sup> Department of Mathematics and Systems Analysis, Aalto University, Otakaari 1, Aalto 00076, Finland
- <sup>3</sup> Department of Mathematical Engineering, Université catholique de Louvain, Avenue Georges Lemaître 4, Louvain-la-Neuve 1348, Belgium