



# On the Backward Error Incurred by the Compact Rational Krylov Linearization

Hongjia Chen<sup>1</sup> · Kuan Xu<sup>2</sup>

Received: 12 April 2021 / Revised: 7 July 2021 / Accepted: 23 July 2021 /  
Published online: 25 August 2021

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

## Abstract

One of the most successful methods for solving a polynomial (PEP) or rational eigenvalue problem (REP) is to recast it, by linearization, as an equivalent but larger generalized eigenvalue problem which can be solved by standard eigensolvers. In this work, we investigate the backward errors of the computed eigenpairs incurred by the application of the well-received compact rational Krylov (CORK) linearization. Our treatment is unified for the PEPs or REPs expressed in various commonly used bases, including Taylor, Newton, Lagrange, orthogonal, and rational basis functions. We construct one-sided factorizations that relate the eigenpairs of the CORK linearization and those of the PEPs or REPs. With these factorizations, we establish upper bounds for the backward error of an approximate eigenpair of the PEPs or REPs relative to the backward error of the corresponding eigenpair of the CORK linearization. These bounds suggest a scaling strategy to improve the accuracy of the computed eigenpairs. We show, by numerical experiments, that the actual backward errors can be successfully reduced by scaling and the errors, before and after scaling, are both well predicted by the bounds.

**Keywords** Backward error · Polynomial eigenvalue problem · Linearization · Taylor basis · Newton basis · Lagrange interpolation · Orthogonal polynomials · Chebyshev polynomials

**Mathematics Subject Classification** 15A18 · 15A22 · 65F15

---

HC is supported by the National Natural Science Foundation of China under grant No.12001262, No.61963028, and No.11961048, Natural Science Foundation of Jiangxi Province with No.20181ACB20001, and Double-Thousand Plan of Jiangxi Province No. jxsq2019101008. KX is supported by Anhui Initiative in Quantum Information Technologies under grant AHY150200.

---

✉ Kuan Xu  
kuanxu@ustc.edu.cn

Hongjia Chen  
chenhongjia@ncu.edu.cn

<sup>1</sup> Department of Mathematics, Nanchang University, 999 Xuefu Road, Nanchang 330031, Jiangxi, China

<sup>2</sup> School of Mathematical Sciences, University of Science and Technology of China, 96 Jinzhai Road, Hefei 230026, Anhui, China

### 1 Introduction

Given a matrix-valued function  $F(\lambda) : \Omega \subseteq \mathbb{C} \mapsto \mathbb{C}^{n \times n}$ , the nonlinear eigenvalue problem (NEP) is to find eigenvalues  $\lambda \in \Omega$  and nonzero right eigenvectors  $x$  and left eigenvectors  $y$  so that

$$F(\lambda)x = 0 \quad \text{and} \quad y^*F(\lambda) = 0. \tag{1}$$

In this article, we confine our discussion to regular NEPs, namely, those where  $\det F(\lambda)$  is not identically zero on  $\Omega$ .

One of the most widely-adopted numerical methods for solving such nonlinear eigenvalue problems is linearization, where the nonlinear eigenvalue problem (1) is first approximated by a proxy polynomial eigenvalue problem (PEP) or rational eigenvalue problem (REP)

$$F(\lambda) \approx R_m(\lambda) = D_0b_0(\lambda) + D_1b_1(\lambda) + \dots + D_mb_m(\lambda), \tag{2}$$

where  $D_j \in \mathbb{C}^{n \times n}$  are constant coefficient matrices,<sup>1</sup> before (2) is further reformulated as a generalized eigenvalue problem to which standard eigensolvers can be applied. In the rest of this paper, we will slightly abuse the notations by also denoting an eigenvalue, the associated right, and the associated left eigenvectors of  $R_m(\lambda)$  by  $\lambda$ ,  $x$ , and  $y$ , respectively. Here,  $b_j(\lambda)$ 's are polynomial basis functions of various kinds or rational basis functions [12,22] and satisfy recurrence relations as follows.

- When  $R_m(\lambda)$  is a degree  $m$  truncated Taylor series expansion of  $F(\lambda)$  at  $\lambda = \sigma$ ,  $b_j(\lambda)$ 's satisfy a two-term recurrence relation

$$b_{j+1}(\lambda) = \frac{\lambda - \sigma}{\beta_{j+1}} b_j(\lambda),$$

with  $b_0(\lambda) = 1/\beta_0$ , where  $\beta_j$  are nonzero scaling constants.

- When  $R_m(\lambda)$  is a Newton interpolant with nodes  $\sigma_0, \sigma_1, \dots, \sigma_m$ ,  $b_j(\lambda)$ 's satisfy a two-term recurrence relation

$$b_{j+1}(\lambda) = \frac{\lambda - \sigma_j}{\beta_{j+1}} b_j(\lambda), \tag{3}$$

with  $b_0(\lambda) = 1/\beta_0$ .

- When  $b_j(\lambda)$ 's are the Lagrange cardinal polynomials, that is,

$$b_j(\lambda) = \frac{1}{\beta_j} \prod_{\substack{i=0 \\ i \neq j}}^m (\lambda - \sigma_i),$$

defined by the interpolation nodes  $\sigma_0, \sigma_1, \dots, \sigma_m$ ,  $R_m(\lambda)$  becomes a Lagrange interpolant and the cardinal polynomials also satisfy a two-term recurrence relation

$$\beta_j(\lambda - \sigma_j)b_j(\lambda) = \beta_{j+1}(\lambda - \sigma_{j+1})b_{j+1}(\lambda).$$

- $R_m(\lambda)$  becomes a series of orthogonal polynomials when  $b_j(\lambda)$ 's are orthogonal polynomials defined by a three-term recurrence relation

$$b_{j+1}(\lambda) = \frac{\lambda b_j(\lambda) + \alpha_j b_j(\lambda) + \gamma_j b_{j-1}(\lambda)}{\beta_{j+1}} \tag{4}$$

for  $j \geq 0$  with  $b_{-1}(\lambda) = 0$  and  $b_0(\lambda) = 1/\beta_0$ .

<sup>1</sup> Throughout this paper, we assume that the coefficient matrix  $D_m$  of the highest order term is nonzero.

- Given interpolation nodes  $\{\sigma_j\}_{j=0}^m \in \Omega \subset \mathbb{C}$  and nonzero poles  $\{\xi_j\}_{j=0}^m \subset \overline{\mathbb{C}} \setminus \Omega$ , the recurrence relation

$$b_{j+1}(\lambda) = \frac{\lambda - \sigma_j}{\beta_{j+1}(1 - \lambda/\xi_{j+1})} b_j(\lambda) \quad \text{with } b_0(\lambda) = \frac{1}{\beta_0} \tag{5}$$

leads  $R_m(\lambda)$  to become a rational approximation.

If the first  $m + 1$  basis functions are collected in a column vector  $b(\lambda) = [b_0(\lambda), b_1(\lambda), \dots, b_m(\lambda)]^T$ , the recurrence relationships above can be written in a unified form

$$\lambda b^T(\lambda) K_m = b^T(\lambda) H_m, \tag{6}$$

where  $K_m, H_m \in \mathbb{C}^{(m+1) \times m}$  can be readily derived for each basis.

Infinitely many different linearizations can be constructed such that they share the same eigenstructure with the PEP or REP, in the sense that such a linearization has the same eigenvalues as the PEP or REP and the eigenvectors of the PEP or REP can be easily obtained from those of the linearization. The most commonly used linearizations include the companion [10] and arrowhead [1] linearizations. In the last decade, as a generalization of the NLEIGS method [13] the compact rational Krylov (CORK) linearization [22] has been proposed and has attracted much attention since it has a unified and simple structure that covers many well-established linearizations, e.g., the companion linearization. It also enjoys the absence of spurious eigenvalues that one may see in other linearizations, e.g., the arrowhead linearization, and, therefore, has a smaller dimension. Moreover, the Kronecker structures in the CORK linearization (see (7) below) allow speedup in the solution of linear systems arising in the shift-and-invert step in Krylov subspace methods [22].

Specifically, the CORK linearization is an  $mn \times mn$  pencil given by

$$\mathbf{L}_m(\lambda) = \mathbf{A}_m - \lambda \mathbf{B}_m,$$

where

$$\mathbf{A}_m = \begin{bmatrix} A_0 & A_1 & \cdots & A_{m-1} \\ \hline H_{m-1}^T & \otimes I_n & & \end{bmatrix} \quad \text{and} \quad \mathbf{B}_m = \begin{bmatrix} B_0 & B_1 & \cdots & B_{m-1} \\ \hline K_{m-1}^T & \otimes I_n & & \end{bmatrix}. \tag{7}$$

Here, the horizontal lines are used to separate the top block rows from the lower parts which are Kronecker-structured. The  $n \times n$  matrices  $\{A_j\}_{j=0}^{m-1}$  and  $\{B_j\}_{j=0}^{m-1}$  are obtained by rewriting  $R_m(\lambda)$  in (2) as

$$g(\lambda)R_m(\lambda) = \sum_{j=0}^{m-1} b_j(\lambda)(A_j - \lambda B_j), \tag{8}$$

where  $g(\lambda) = \lambda - \sigma_m$  and  $g(\lambda) = 1 - \lambda/\xi_m$  for the Lagrange and rational cases, respectively, and  $g(\lambda) = 1$  for all other bases. For a summary of  $A_j$  and  $B_j$  in various bases, see [22, Table 1]. We spell out  $\mathbf{A}_m$  and  $\mathbf{B}_m$  in Table 1 to facilitate the discussion in the rest of this article.

The last few decades have witnessed the success of linearization in the solution of PEPs and REPs. The study has been focused on the linearization of PEPs represented in the monomial basis, see, for example, [2, 8, 16, 19]. A key question to answer is how accurate an approximate eigenpair of the PEPs or REPs is when it is found by recovering from that of the linearization, since a little error in the latter may lead to a much larger error in the former. In a pioneering work [14], some upper bounds are constructed for the backward error of an approximate eigenpair of the the PEPs expressed in the monomial basis relative to that of the corresponding approximate eigenpair of certain companion linearizations. These bounds are shown to be able to identify the circumstances under which the linearizations do or do not lead to larger

**Table 1**  $A_m$  and  $B_m$  for various bases. For the Lagrange basis,  $\theta_j = \beta_j/\beta_{j+1}$

$b_j(\lambda)$	$A_m$	$B_m$
Taylor	$\begin{bmatrix} D_0 & D_1 & \dots & D_{m-2} & D_{m-1} - \frac{\sigma}{\beta_m} D_m \\ \sigma I_n & \beta_1 I_n & & & \\ \sigma I_n & \beta_2 I_n & \dots & & \\ & \dots & \dots & \sigma I_n & \beta_{m-1} I_n \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & \dots & 0 & -\frac{D_m}{\beta_m} \\ I_n & 0 & & & \\ & \dots & \dots & & \\ & & \dots & 0 & \\ & & & I_n & 0 \end{bmatrix}$
Newton	$\begin{bmatrix} D_0 & D_1 & \dots & D_{m-2} & D_{m-1} - \frac{\sigma_{m-1}}{\beta_m} D_m \\ \sigma_0 I_n & \beta_1 I_n & & & \\ \sigma_1 I_n & \beta_2 I_n & \dots & & \\ & \dots & \dots & \sigma_{m-2} I_n & \beta_{m-1} I_n \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & \dots & 0 & -\frac{D_m}{\beta_m} \\ I_n & 0 & & & \\ & \dots & \dots & & \\ & & \dots & 0 & \\ & & & I_n & 0 \end{bmatrix}$
Orthogonal	$\begin{bmatrix} D_0 & D_1 & D_2 & \dots & D_{m-3} & D_{m-2} + \frac{\gamma_{m-1}}{\beta_m} D_m & D_{m-1} + \frac{\alpha_{m-1}}{\beta_m} D_m \\ -\alpha_0 I_n & \beta_1 I_n & & & & & \\ -\gamma_1 I_n & -\alpha_1 I_n & \beta_2 I_n & \dots & & & \\ & \dots & \dots & \dots & \beta_{m-3} I_n & & \\ & & & & \dots & -\alpha_{m-3} I_n & \beta_{m-2} I_n \\ & & & & & -\gamma_{m-2} I_n & -\alpha_{m-2} I_n \\ & & & & & & \beta_{m-1} I_n \end{bmatrix}$	$\begin{bmatrix} 0 & 0 & \dots & 0 & -\frac{D_m}{\beta_m} \\ I_n & 0 & & & \\ & \dots & \dots & & \\ & & \dots & 0 & \\ & & & I_n & 0 \end{bmatrix}$

Table 1 continued

$b_j(\lambda)$	$A_m$	$B_m$
Lagrange	$\begin{bmatrix} \sigma_m D_0 & \sigma_m D_1 & \dots & \sigma_m D_{m-2} & \sigma_m D_{m-1} & \sigma_m D_{m-1} + \theta_m \sigma_{m-1} D_m \\ \sigma_0 \theta_1 I_n & -\sigma_1 I_n & & & & \\ \sigma_1 \theta_2 I_n & -\sigma_2 I_n & & & & \\ \dots & \dots & & & & \\ \sigma_{m-2} \theta_{m-1} I_n & & & & & -\sigma_{m-1} I_n \end{bmatrix}$	$\begin{bmatrix} D_0 & D_1 & \dots & D_{m-2} & D_{m-1} & D_{m-1} + \theta_m D_m \\ \theta_1 I_n & -I_n & & & & \\ \dots & \dots & & & & \\ \dots & \dots & & & & \\ \theta_{m-1} I_n & & & & & -I_n \end{bmatrix}$
Rational	$\begin{bmatrix} D_0 & D_1 & \dots & D_{m-2} & D_{m-1} & \frac{\sigma_{m-1}}{\beta_m} D_m \\ \sigma_0 I_n & \beta_1 I_n & & & & \\ \sigma_1 I_n & \beta_2 I_n & & & & \\ \dots & \dots & & & & \\ \sigma_{m-2} I_n & & & & & \beta_{m-1} I_n \end{bmatrix}$	$\begin{bmatrix} \frac{D_0}{\xi_m} & \frac{D_1}{\xi_m} & \dots & \frac{D_{m-2}}{\xi_m} & \frac{D_{m-1}}{\xi_m} & \frac{D_m}{\beta_m} \\ I_n & \beta_1 I_n & & & & \\ \dots & \dots & & & & \\ \dots & \dots & & & & \\ I_n & \beta_2 I_n & & & & \frac{\beta_{m-1}}{\xi_{m-1}} I_n \end{bmatrix}$

backward errors. However, the PEPs expressed in non-monomial bases have been proposed due to the advantages offered by the non-monomial basis functions [1,7] and the study of the accuracy of the eigenpairs obtained via linearization with the non-monomial basis functions are largely missing in the literature. The only investigation, so far, on the backward error associated to the PEPs in non-monomial bases is [17] which addresses the backward errors of the approximate eigenpairs incurred in the arrowhead linearization of the PEPs in the Lagrange basis. This paper aims at filling this gap.

A few preprocessing techniques have been proposed for the PEPs to gain improved accuracy, including the weighted balancing method in [4], the tropical scaling method due to [9], and the simple scaling method in [14], all of which deal with the monomial basis. So far, the only attempt to precondition the PEPs in a non-monomial basis is the work in [17], who proposed a two-sided balancing strategy based on block diagonal similarity transformations to improve the accuracy of an approximate eigenpair of the PEPs given in the Lagrange basis. Following a careful backward error analysis, we propose a diagonal scaling approach in this paper for the CORK linearization corresponding to PEPs expressed in various common non-monomial bases and our approach is shown to be effective in improving the accuracy of the computed eigenpairs.

Our discussion of the backward errors is based on the local backward error analysis first suggested in [14], for which we derive the one-sided factorization which are not found before for the aforementioned bases. This is the key to our analysis. We also note that the backward error analysis based on the global analysis framework [18] are, in general, not suitable for our work, since it was derived only for the right eigenpairs and the upper bounds obtained there can hardly help us develop a scaling strategy.

After introducing a few key definitions at the beginning of Sect. 2, we give the one-sided factorizations for the CORK linearization associated with the aforementioned basis functions. Based on these one-sided factorizations, we develop the bounds for the backward error of a computed eigenpair of  $R_m(\lambda)$  relative to that of the corresponding eigenpair of the CORK linearization. To improve accuracy, we discuss the scaling strategy in Sect. 3, which is justified by the improved error bounds. We verify the bounds and the idea of scaling in Sect. 4 with a few numerical experiments and conclude in Sect. 5 with a couple of remarks.

## 2 Backward Errors

In this section, we relate the backward error of an approximate eigenpair of  $R_m(\lambda)$  to that of the linearization  $L_m(\lambda)$  and further bound the former in terms of the latter. These bounds will help identify the cases where the eigenpairs of  $R_m(\lambda)$  can be computed accurately via its CORK linearization in the backward error sense.

### 2.1 Definitions

We first include the definition of the normwise backward error of an approximate eigenpair of the rational function  $R_m(\lambda)$  expressed in the aforementioned bases  $b_j(\lambda)$ . This definition is extended from the definition for the monomial basis given in [14]. The case of the basis  $\{b_j(\lambda)\}$  being the Lagrange polynomials is discussed in, for example, [17] for the arrowhead linearization.

The normwise backward error of an approximate right eigenpair  $(\lambda, x)$  of  $R_m(\lambda)$ , where  $\lambda$  is finite, is defined by

$$\eta_{R_m}(\lambda, x) = \min \{ \epsilon : (R_m(\lambda) + \Delta R_m(\lambda))x = 0, \|\Delta D_j\|_2 \leq \epsilon \|D_j\|_2, 0 \leq j \leq m \},$$

where  $\Delta R_m(\lambda) = \sum_{j=0}^m \Delta D_j b_j(\lambda)$ . Similarly, for an approximate left eigenpair  $(\lambda, y^*)$  of  $R_m(\lambda)$ , the normwise backward error is defined by

$$\eta_{R_m}(\lambda, y^*) = \min \{ \epsilon : y^*(R_m(\lambda) + \Delta R_m(\lambda)) = 0, \|\Delta D_j\|_2 \leq \epsilon \|D_j\|_2, 0 \leq j \leq m \}.$$

Analogues can also be drawn from the monomial case [20] to obtain equivalent but explicit expressions for the backward errors of approximate eigenpairs  $(\lambda, x)$  and  $(\lambda, y^*)$  of  $R_m(\lambda)$  in terms of the basis  $\{b_j(\lambda)\}$

$$\begin{aligned} \eta_{R_m}(\lambda, x) &= \frac{\|R_m(\lambda)x\|_2}{(\sum_{j=0}^m \|D_j\|_2 |b_j(\lambda)|) \|x\|_2}, \\ \eta_{R_m}(\lambda, y^*) &= \frac{\|y^* R_m(\lambda)\|_2}{(\sum_{j=0}^m \|D_j\|_2 |b_j(\lambda)|) \|y\|_2}, \end{aligned} \tag{9}$$

and those of the approximate eigenpairs  $(\lambda, v)$  and  $(\lambda, w^*)$  of  $\mathbf{L}_m(\lambda)$

$$\begin{aligned} \eta_{\mathbf{L}_m}(\lambda, v) &= \frac{\|\mathbf{L}_m(\lambda)v\|_2}{(|\lambda| \|\mathbf{B}_m\|_2 + \|\mathbf{A}_m\|_2) \|v\|_2}, \\ \eta_{\mathbf{L}_m}(\lambda, w^*) &= \frac{\|w^* \mathbf{L}_m(\lambda)\|_2}{(|\lambda| \|\mathbf{B}_m\|_2 + \|\mathbf{A}_m\|_2) \|w\|_2}. \end{aligned} \tag{10}$$

### 2.2 One-sided Factorizations for CORK Linearization

The key to estimating the backward error of an approximate eigenpair of  $R_m(\lambda)$  in terms of that of  $\mathbf{L}_m(\lambda)$  is to find an appropriate one-sided factorization that relates  $R_m(\lambda)$  and  $\mathbf{L}_m(\lambda)$ . Specifically, the first  $m$  equations in (6), when combined with (8), lead to

$$\mathbf{L}_m(\lambda)H(\lambda) = g(\lambda)e_1 \otimes R_m(\lambda), \tag{11}$$

where  $e_1 \in \mathbb{C}^{m \times 1}$  is the first unit vector and

$$H(\lambda) = b(\lambda) \otimes I_n = \begin{bmatrix} b_0(\lambda) \\ b_1(\lambda) \\ \vdots \\ b_{m-1}(\lambda) \end{bmatrix} \otimes I_n \in \mathbb{C}^{mn \times n},$$

where the notation  $b(\lambda)$  is reused here and in the rest of this paper to denote the same vector in (6) but with the last entry  $b_m(\lambda)$  dropped. Note that this right-sided factorization is valid for any of the bases mentioned in Sect. 1.

We also need the left-sided factorization

$$G(\lambda)\mathbf{L}_m(\lambda) = e_1^T \otimes (g(\lambda)R_m(\lambda)), \tag{12}$$

where  $G(\lambda) \in \mathbb{C}^{n \times mn}$ . Unlike  $H(\lambda)$ ,  $G(\lambda)$  does not have a uniform expression valid for all the aforementioned bases. We now give  $G(\lambda)$  below for each basis and to the best of our knowledge this is the first time it is constructed for these bases.

**Taylor basis:**

$$G(\lambda) = \left[ b_0(\lambda)I_n, -\frac{1}{\lambda-\sigma} \sum_{j=1}^m b_j(\lambda)D_j, \right. \\ \left. -\frac{1}{\lambda-\sigma} \frac{b_0(\lambda)}{b_1(\lambda)} \sum_{j=2}^m b_j(\lambda)D_j, \dots, -\frac{1}{\lambda-\sigma} \frac{b_0(\lambda)}{b_{m-2}(\lambda)} \sum_{j=m-1}^m b_j(\lambda)D_j \right].$$

**Newton Basis:**

$$G(\lambda) = \left[ b_0(\lambda)I_n, -\frac{1}{\lambda-\sigma_0} \sum_{j=1}^m b_j(\lambda)D_j, \right. \\ \left. -\frac{1}{\lambda-\sigma_1} \frac{b_0(\lambda)}{b_1(\lambda)} \sum_{j=2}^m b_j(\lambda)D_j, \dots, -\frac{1}{\lambda-\sigma_{m-2}} \frac{b_0(\lambda)}{b_{m-2}(\lambda)} \sum_{j=m-1}^m b_j(\lambda)D_j \right].$$

**Orthogonal basis:** To construct  $G(\lambda)$  in the case of orthogonal basis, we define a new sequence of orthogonal polynomials  $b_j^{(k)}(\lambda)$  that are associated with  $b_j(\lambda)$  given in (4) by

$$b_{j+1}^{(k)}(\lambda) = \frac{\lambda b_j^{(k)}(\lambda) + \alpha_{j+k} b_j^{(k)}(\lambda) + \gamma_{j+k} b_{j-1}^{(k)}(\lambda)}{\beta_{j+k+1}}, \tag{13}$$

where  $b_j^{(k)}(\lambda) \equiv 0$  for  $j < 0$  and  $b_0^{(k)}(\lambda) = 1$ . When  $k = 0$ ,  $b_j^{(k)}(\lambda)$  reduces to the original orthogonal polynomials defined in (4).<sup>2</sup> We have not seen any discussion in the literature about the orthogonal polynomials defined by (13) and will simply call them the *shifted orthogonal polynomials* generated by the original orthogonal sequence or the shifted orthogonal polynomials for short, since they are defined by a subset of the same recurrence coefficients but with indices shifted by  $k$ . The Chebyshev polynomial of the first kind (known as Chebyshev  $T$ ) and the second kind (known as Chebyshev  $U$  or as ultraspherical or Gegenbauer polynomial  $C^{(1)}$ ) serve as the simplest example. If we let  $b_j(\lambda) = T_j(\lambda)$ , where  $T_j(\lambda)$  is the  $j$ th Chebyshev polynomial of the first kind, the new orthogonal sequence  $b_j^{(1)}(\lambda) = U_j(\lambda)$ , i.e., the Chebyshev polynomial of the second kind, and the recurrence coefficients in (4) are

$$\alpha_k = 0 \text{ for all } k, \\ \beta_k = \begin{cases} 1, & k = 1, \\ 1/2, & k \in \mathbb{Z}^{\geq 2}, \end{cases} \\ \gamma_k = \begin{cases} 0, & k = 0, \\ -1/2, & k \in \mathbb{Z}^+.$$

It can be verified by induction (see Appendix A) that for the orthogonal basis

$$G(\lambda) = \left[ I_n, -\frac{1}{\beta_1} \sum_{j=1}^m b_{j-1}^{(1)}(\lambda)D_j, -\frac{1}{\beta_2} \sum_{j=2}^m b_{j-2}^{(2)}(\lambda)D_j, \dots, \right. \\ \left. -\frac{1}{\beta_{k-1}} \sum_{j=k-1}^m b_{j-(k-1)}^{(k-1)}(\lambda)D_j, \dots, -\frac{1}{\beta_{m-2}} \sum_{j=m-2}^m b_{j-(m-2)}^{(m-2)}(\lambda)D_j, \right. \\ \left. -\frac{1}{\beta_{m-1}} \sum_{j=m-1}^m b_{j-(m-1)}^{(m-1)}(\lambda)D_j \right]. \tag{14}$$

<sup>2</sup> In this case,  $\beta_0$  is identified with 1.



In practice, the most commonly used orthogonal polynomials for the purpose of function approximation are the Chebyshev polynomials, for which

$$G(\lambda) = \left[ I_n, -\sum_{j=1}^m U_{j-1}(\lambda)D_j, -2\sum_{j=2}^m U_{j-2}(\lambda)D_j, \dots, -2\sum_{j=k-1}^m U_{j-(k-1)}(\lambda)D_j, \dots, -2\sum_{j=m-2}^m U_{j-(m-2)}(\lambda)D_j, -2\sum_{j=m-1}^m U_{j-(m-1)}(\lambda)D_j \right].$$

**Lagrange basis:**

$$G(\lambda) = \left[ b_0(\lambda)I_n, \frac{\lambda - \sigma_m}{\lambda - \sigma_1} \frac{b_0(\lambda)}{b_1(\lambda)} \sum_{j=1}^m b_j(\lambda)D_j, \frac{\lambda - \sigma_m}{\lambda - \sigma_2} \frac{b_0(\lambda)}{b_2(\lambda)} \sum_{j=2}^m b_j(\lambda)D_j, \dots, \frac{\lambda - \sigma_m}{\lambda - \sigma_{m-1}} \frac{b_0(\lambda)}{b_{m-1}(\lambda)} \sum_{j=m-1}^m b_j(\lambda)D_j \right].$$

**Rational basis:**

$$G(\lambda) = \left[ b_0(\lambda)I_n, \frac{g(\lambda)}{\sigma_0 - \lambda} \sum_{j=1}^m b_j(\lambda)D_j, \frac{g(\lambda)}{\sigma_1 - \lambda} \frac{b_0(\lambda)}{b_1(\lambda)} \sum_{j=2}^m b_j(\lambda)D_j, \dots, \frac{g(\lambda)}{\sigma_{m-2} - \lambda} \frac{b_0(\lambda)}{b_{m-2}(\lambda)} \sum_{j=m-1}^m b_j(\lambda)D_j \right].$$

What is remarkable is the striking similarity shared by these  $G(\lambda)$ 's despite of basis functions. This enables us to develop bounds for the backward error ratios in a rather unified fashion in Theorems 2 and 3.

The following theorem serves as a generalization of its monomial counterpart given in [11] by associating the eigensystem of  $R_m(\lambda)$  and that of  $L_m(\lambda)$  when the one-sided factorizations (11) and (12) hold.

**Theorem 1** *Let  $R_m(\lambda)$  and  $L_m(\lambda)$  be matrix functions of dimensions  $n \times n$  and  $mn \times mn$ , respectively. Assume that (11) and (12) hold at  $\lambda \in \mathbb{C}$  with  $g(\lambda) \neq 0$  and  $b(\lambda) \neq 0$ .*

- (i) *If  $v$  is a right eigenvector of  $L_m(\lambda)$  with eigenvalue  $\lambda$ , then  $(g(\lambda)e_1^T \otimes I_n)v$  is a right eigenvector of  $R_m(\lambda)$  with eigenvalue  $\lambda$ ;*
- (ii) *if  $w^*$  is a left eigenvector of  $L_m(\lambda)$  with eigenvalue  $\lambda$ , then  $w^*(g(\lambda)e_1 \otimes I_n)$  is a left eigenvector of  $R_m(\lambda)$  with eigenvalue  $\lambda$  provided that it is nonzero.*

**Proof** Based on the relations (11) and (12), we have

$$G(\lambda)L_m(\lambda)v = e_1^T \otimes (g(\lambda)R_m(\lambda))v = R_m(\lambda)(g(\lambda)e_1^T \otimes I_n)v, \tag{15}$$

$$w^*L_m(\lambda)H(\lambda) = w^*(g(\lambda)e_1 \otimes R_m(\lambda)) = w^*(g(\lambda)e_1 \otimes I_n)R_m(\lambda). \tag{16}$$

□

Now we are in the position to investigate the backward error incurred by the CORK linearization using (15) and (16).

### 2.3 The Bounds for Backward Error Ratios

Now we have all the ingredients for bounding the backward error of an eigenpair of  $R_m(\lambda)$ . The following theorem gives bounds for the backward error ratios  $\eta_{R_m}(\lambda, x)/\eta_{L_m}(\lambda, v)$  and  $\eta_{R_m}(\lambda, y^*)/\eta_{L_m}(\lambda, w^*)$ .

**Theorem 2** Let  $(\lambda, x)$  and  $(\lambda, y^*)$  be approximate right and left eigenpairs of  $R_m(\lambda)$  respectively and  $(\lambda, v)$  and  $(\lambda, w^*)$  be the corresponding approximate right and left eigenpairs of  $L_m(\lambda)$ . If  $|b_0(\lambda)| = 1$ , the bound for the ratio of  $\eta_{R_m}(\lambda, x)$  to  $\eta_{L_m}(\lambda, v)$  is given by

$$\frac{\eta_{R_m}(\lambda, x)}{\eta_{L_m}(\lambda, v)} \leq \frac{|\lambda| \|B_m\|_2 + \|A_m\|_2}{|g(\lambda)| \left( \sum_{j=0}^m \|D_j\|_2 |b_j(\lambda)| \right)} \frac{G_U(\lambda) \|v\|_2}{\|x\|_2}, \tag{17}$$

where

$$G_U(\lambda) = \sqrt{m} \max \left( 1, \tilde{G}_U(\lambda) \right) \tag{18}$$

and

$$\tilde{G}_U(\lambda) = \begin{cases} \max_{j=1:m-1} \frac{1}{|\beta_j|} \times \sum_{j=1:m} \max_{i=1:m-1} |b_{j-1}^{(i)}(\lambda)| \times \max_{j=1:m} \|D_j\|_2, & \text{orthogonal,} \tag{19a} \\ Q(\lambda) \times \sum_{j=1}^m |b_j(\lambda)| \times \max_{j=1:m} \|D_j\|_2, & \text{all other bases,} \tag{19b} \end{cases}$$

with

$$Q(\lambda) = \begin{cases} \max_{j=0:m-2} \frac{1}{|\lambda - \sigma_j| |b_j(\lambda)|}, & \text{Taylor,} \tag{20a} \\ \max_{j=0:m-2} \frac{1}{|\lambda - \sigma_j| |b_j(\lambda)|}, & \text{Newton, rational,} \tag{20b} \\ \max_{j=1:m-1} \frac{|\lambda - \sigma_m|}{|\lambda - \sigma_j| |b_j(\lambda)|}, & \text{Lagrange.} \tag{20c} \end{cases}$$

The bound for the backward error of a left eigenpair  $(\lambda, y^*)$  relative to that of  $(\lambda, w^*)$  is given by

$$\frac{\eta_{R_m}(\lambda, y^*)}{\eta_{L_m}(\lambda, w^*)} \leq \frac{|\lambda| \|B_m\|_2 + \|A_m\|_2}{|g(\lambda)| \left( \sum_{j=0}^m \|D_j\|_2 |b_j(\lambda)| \right)} \frac{H_U(\lambda) \|w\|_2}{\|y\|_2}, \tag{21}$$

where

$$H_U(\lambda) = \sqrt{m} \max_{j=0:m-1} (|b_j(\lambda)|). \tag{22}$$

**Proof** From (16) and (15), we have

$$\begin{aligned} \|R_m(\lambda)x\|_2 &\leq \frac{1}{|g(\lambda)|} \|G(\lambda)\|_2 \|L_m(\lambda)v\|_2, \\ \|y^* R_m(\lambda)\|_2 &\leq \frac{1}{|g(\lambda)|} \|H(\lambda)\|_2 \|w^* L_m(\lambda)\|_2, \end{aligned}$$

which, along with (9) and (2.1), give

$$\begin{aligned} \frac{\eta_{R_m}(\lambda, x)}{\eta_{L_m}(\lambda, v)} &\leq \frac{|\lambda| \|B_m\|_2 + \|A_m\|_2}{|g(\lambda)| \left( \sum_{j=0}^m \|D_j\|_2 |b_j(\lambda)| \right)} \frac{\|G(\lambda)\|_2 \|v\|_2}{\|x\|_2}, \\ \frac{\eta_{R_m}(\lambda, y^*)}{\eta_{L_m}(\lambda, w^*)} &\leq \frac{|\lambda| \|B_m\|_2 + \|A_m\|_2}{|g(\lambda)| \left( \sum_{j=0}^m \|D_j\|_2 |b_j(\lambda)| \right)} \frac{\|H(\lambda)\|_2 \|w\|_2}{\|y\|_2}. \end{aligned} \tag{23}$$

By Lemma 3.5 in [14] we have

$$\|H(\lambda)\|_2 = \|b(\lambda) \otimes I_n\|_2 \leq H_U(\lambda),$$

which substituted into (23) leads to (21).

For the rational basis,

$$\|G(\lambda)\|_2 \leq \sqrt{m} \max \left( 1, \max_{j=0:m-2} \frac{1}{|\lambda - \sigma_j| |b_j(\lambda)|} \times \sum_{j=1}^m |b_j(\lambda)| \|D_j\|_2 \right),$$

which can be further relaxed to obtain the upper bound  $G_U(\lambda)$  given by (18), (19b), and (20b).  $G_U(\lambda)$  for all other bases except the orthogonal one can be derived analogously.

In the case of the orthogonal basis,

$$\|G(\lambda)\|_2 \leq \sqrt{m} \max \left( 1, \max_{j=1:m-1} \frac{1}{|\beta_j|} \times \sum_{j=1}^m \left( \max_{i=1:m-1} |b_{j-1}^{(i)}(\lambda)| \right) \|D_j\|_2 \right) \leq G_U(\lambda),$$

which leads to (18) and (19a). □

In particular, we furnish the following corollary which gives the bound for  $\eta_{R_m}(\lambda, x) / \eta_{L_m}(\lambda, v)$  when  $b_j(\lambda)$  are Chebyshev polynomials  $T_j(\lambda)$  for that is the most commonly used orthogonal basis in practice.

**Corollary 1** *Let  $(\lambda, x)$  and  $(\lambda, v)$  be approximate right eigenpairs of  $R_m(\lambda)$  and  $L_m(\lambda)$ , respectively. If  $\{b_j(\lambda)\}_{j=0}^{m-1}$  is the Chebyshev basis, the bound for the backward error ratio  $\eta_{R_m}(\lambda, x) / \eta_{L_m}(\lambda, v)$  is given by (17) with*

$$G_U(\lambda) = \sqrt{m} \max \left( 1, m(m+1) \max_{j=1:m} \|D_j\|_2 \right).$$

**Proof** Since  $|U_j(\lambda)| \leq j + 1$ ,

$$\|G(\lambda)\|_2 \leq \sqrt{m} \max \left( 1, 2 \sum_{j=1}^m |U_{j-1}(\lambda)| \times \max_{j=1:m} \|D_j\|_2 \right) \leq G_U(\lambda).$$

□

### 3 Scaling the Linearization

It is ideal if the ratios  $\eta_{R_m}(\lambda, x) / \eta_{L_m}(\lambda, v)$  and  $\eta_{R_m}(\lambda, y^*) / \eta_{L_m}(\lambda, w^*)$  are roughly 1. When this is the case, an eigenpair of  $R_m(\lambda)$  would have accuracy comparable to that of the linearization. In practice, however, the ratios  $\eta_{R_m}(\lambda, x) / \eta_{L_m}(\lambda, v)$  and  $\eta_{R_m}(\lambda, y^*) / \eta_{L_m}(\lambda, w^*)$  are often much larger than 1 and certain scaling techniques are employed to precondition the linearization so that the backward errors can be reduced. Based on the bounds derived in the last section, we now develop a simple diagonal scaling approach for the CORK linearization formed from PEPs and REPs expressed in any of the aforementioned bases. To see how it works, we let

$$S = \begin{bmatrix} 1 & & & \\ & s & & \\ & & \ddots & \\ & & & s \end{bmatrix} \otimes I_n \in \mathbb{C}^{mn \times mn},$$

where  $s = \max_{j=1:m} (\|D_j\|_2)$ . By (16) and (15), we have

$$\begin{aligned} \widehat{w}^* \widehat{\mathbf{L}}_m(\lambda) H(\lambda) &= y^* R_m(\lambda), \\ \widehat{G}(\lambda) \widehat{\mathbf{L}}_m(\lambda) v &= R_m(\lambda) x, \end{aligned}$$

where  $\widehat{G}(\lambda) = G(\lambda)S^{-1}$ ,  $\widehat{\mathbf{L}}_m(\lambda) = S\mathbf{L}_m(\lambda)$ ,  $x = (g(\lambda)e_1^T \otimes I_n)v$ ,  $y^* = w^*(g(\lambda)e_1 \otimes I_n)$  and  $\widehat{w}^* = w^*S^{-1}$ . Now we restate Theorem 2 with  $G(\lambda)$ ,  $\mathbf{L}_m(\lambda)$ , and  $w^*$  replaced by their hatted counterparts. The proof is omitted as it is similar to that of Theorem 2.

**Theorem 3** *The backward error  $\eta_{R_m}(\lambda, x)$  of an approximate right eigenpair  $(\lambda, x)$  of  $R_m(\lambda)$  relative to the backward error  $\eta_{R_m}(\lambda, v)$  of the corresponding approximate eigenpair  $(\lambda, v)$  of  $\widehat{\mathbf{L}}_m(\lambda)$  can be bounded by*

$$\frac{\eta_{R_m}(\lambda, x)}{\eta_{\widehat{\mathbf{L}}_m}(\lambda, v)} \leq \frac{(|\lambda| \|\widehat{\mathbf{B}}_m\|_2 + \|\widehat{\mathbf{A}}_m\|_2)}{|g(\lambda)| \left(\sum_{j=0}^m \|D_j\|_2 |b_j(\lambda)|\right)} \frac{G_S(\lambda) \|v\|_2}{\|x\|_2}, \tag{24}$$

where  $\widehat{\mathbf{A}}_m = S\mathbf{A}_m$ ,  $\widehat{\mathbf{B}}_m = S\mathbf{B}_m$ ,

$$G_S(\lambda) = \sqrt{m} \max\left(1, \tilde{G}_S(\lambda)\right)$$

and

$$\tilde{G}_S(\lambda) = \begin{cases} \max_{j=1:m-1} \frac{1}{|\beta_j|} \times \sum_{j=1}^m \max_{i=1:m-1} |b_{j-1}^{(i)}(\lambda)|, & \text{orthogonal,} \\ Q(\lambda) \times \sum_{j=1}^m |b_j(\lambda)|, & \text{all other bases.} \end{cases}$$

Here,  $Q(\lambda)$  is given in (20).

The backward error ratio  $\eta_{R_m}(\lambda, y^*)/\eta_{\widehat{\mathbf{L}}_m}(\lambda, \widehat{w}^*)$  of an approximate left eigenpair  $(\lambda, y^*)$  of  $R_m(\lambda)$  can be bounded by

$$\frac{\eta_{R_m}(\lambda, y^*)}{\eta_{\widehat{\mathbf{L}}_m}(\lambda, \widehat{w}^*)} \leq \frac{|\lambda| \|\widehat{\mathbf{B}}_m\|_2 + \|\widehat{\mathbf{A}}_m\|_2}{|g(\lambda)| \left(\sum_{j=0}^m \|D_j\|_2 |b_j(\lambda)|\right)} \frac{H_U(\lambda) \|\widehat{w}\|_2}{\|y\|_2}, \tag{25}$$

where  $H_U(\lambda)$  is given by (22).

**Remark 1** For the Chebyshev basis, the bound  $\eta_{R_m}(\lambda, x)/\eta_{\widehat{\mathbf{L}}_m}(\lambda, v)$  for the backward error ratio of a right eigenpair is similar to that given in Corollary 1 but with  $G_U(\lambda)$  replaced by  $G_S(\lambda) = \sqrt{m} \max(1, m(m+1))$ .

Let us now compare the ratios given by Theorems 2 and 3. For the backward error ratio of a right eigenpair, we consider the case with  $\tilde{G}_U(\lambda) > 1$ , which is very common in practice. The quantities  $\|\mathbf{A}_m\|_2$ ,  $\|\mathbf{B}_m\|_2$ , and  $G_U(\lambda)$  in (17) now become  $\|\widehat{\mathbf{A}}_m\|_2$ ,  $\|\widehat{\mathbf{B}}_m\|_2$ , and  $G_S(\lambda)$  in (24) respectively. If we bound  $\|\mathbf{A}_m\|_2$ ,  $\|\mathbf{B}_m\|_2$ ,  $\|\widehat{\mathbf{A}}_m\|_2$ , and  $\|\widehat{\mathbf{B}}_m\|_2$  from above, we have

$$\begin{aligned} \|\mathbf{A}_m\|_2 &\leq \sqrt{2} \max\left(\sqrt{m} \max_{j=0:m-1} \|A_j\|_2, \|H_{m-1}\|_2\right), \\ \|\mathbf{B}_m\|_2 &\leq \sqrt{2} \max\left(\sqrt{m} \max_{j=0:m-1} \|B_j\|_2, \|K_{m-1}\|_2\right), \end{aligned}$$

whereas

$$\begin{aligned} \|\widehat{\mathbf{A}}_m\|_2 &\leq \sqrt{2} \max \left( \sqrt{m} \max_{j=0:m-1} \|A_j\|_2, s \|H_{m-1}\|_2 \right), \\ \|\widehat{\mathbf{B}}_m\|_2 &\leq \sqrt{2} \max \left( \sqrt{m} \max_{j=0:m-1} \|B_j\|_2, s \|K_{m-1}\|_2 \right). \end{aligned} \tag{26}$$

When  $\sqrt{m} \max_{j=0:m-1} \|A_j\|_2$  and  $\sqrt{m} \max_{j=0:m-1} \|B_j\|_2$  prevail in the max operators in (26),  $\|\widehat{\mathbf{A}}_m\|_2$  and  $\|\widehat{\mathbf{B}}_m\|_2$  stay the same after scaling, as we often see in practice (see Sect. 4). In case of  $s \|H_{m-1}\|_2$  and  $s \|K_{m-1}\|_2$  being dominant,  $\|\widehat{\mathbf{A}}_m\|_2$  and  $\|\widehat{\mathbf{B}}_m\|_2$  would be larger than  $\|\mathbf{A}_m\|_2$  and  $\|\mathbf{B}_m\|_2$ , respectively, by a factor  $O(s)$  at most. On the other hand, the scaling reduces the norm of  $G(\lambda)$  by a factor  $O(s)$  for sure, as can be seen from the upper bounds of  $G_U(\lambda)$  and  $G_S(\lambda)$

$$\begin{aligned} G_U(\lambda) &\leq \sqrt{m} \sum_{j=0}^m |b_j(\lambda)| \max \left( 1, Q(\lambda) \times \max_{j=1:m} \|D_j\|_2 \right), \\ G_S(\lambda) &\leq \sqrt{m} \sum_{j=0}^m |b_j(\lambda)| \max \left( 1, Q(\lambda) \right). \end{aligned}$$

With  $\|\mathbf{A}_m\|_2$  and  $\|\mathbf{B}_m\|_2$  unchanged or amplified by a factor which can be canceled out by the reduction in  $G_U(\lambda)$ , the bound for the backward error ratio  $\eta_{R_m}(\lambda, x)/\eta_{\widehat{\mathbf{T}}_m}(\lambda, v)$  is usually reduced, as we hope. In the meantime, the term  $\|w\|_2/\|y\|_2$  in (21) is replaced by  $\|\widehat{w}\|_2/\|y\|_2$ , which is smaller than  $\|w\|_2/\|y\|_2$  when  $s > 1$ . Since it is also very common in practice that  $s > 1$ , scaling also reduces the bound of the backward error ratio for left eigenpairs. As the bounds for the ratios reduced for both left and right eigenpairs, we expect to see improved accuracy using the CORK linearization when it is equipped with scaling.

### 4 Experimental Results

In this section, we test the bounds for the backward errors of the approximate eigenpairs of  $R_m(\lambda)$  computed via the CORK linearization  $\mathbf{L}_m(\lambda)$  and the scaled one  $\widehat{\mathbf{L}}_m(\lambda)$  and the effectiveness of the scaling approach. In our experiments, the eigenpairs of the linearizations are computed using MATLAB’s `eig` function with the eigenvectors of  $R_m(\lambda)$  recovered by Theorem 1. The backward errors of the computed eigenpairs of  $R_m(\lambda)$ ,  $\mathbf{L}_m(\lambda)$ , and  $\widehat{\mathbf{L}}_m(\lambda)$  are obtained using (9) and (2.1), while the upper bounds for the ratios of the backward errors are calculated using (17), (21), (24), and (25).

#### 4.1 Newton Basis

Our first test problem is the quadratic eigenvalue problem from a linearly damped mass-spring system [20], drawn from the collection of nonlinear eigenvalue problems NLEVP [5]

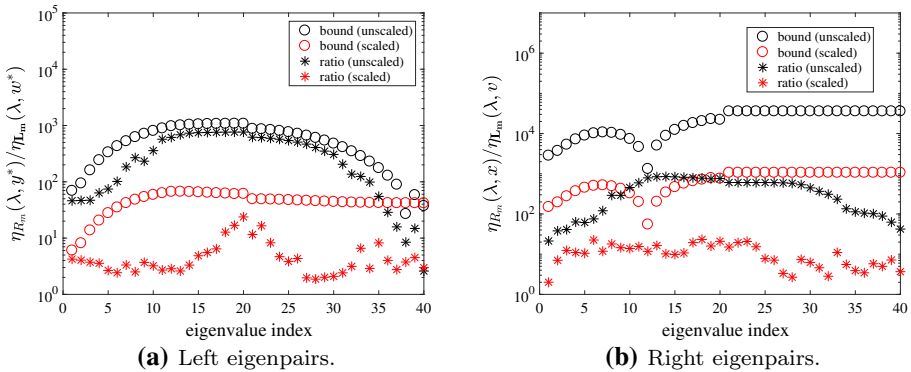
$$F(\lambda)x = (\lambda^2 A_2 + \lambda A_1 + A_0)x = 0,$$

where the  $20 \times 20$  matrices are  $A_0 = 5T$ ,  $A_1 = 10T$ ,  $A_2 = I$ , and

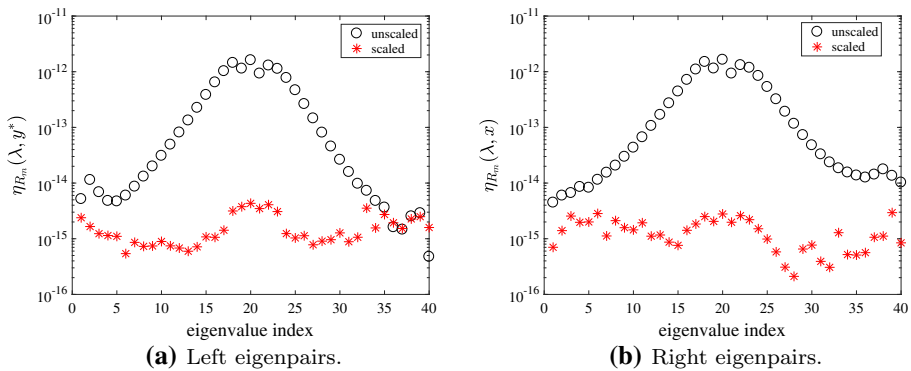
$$T = \begin{bmatrix} 3 & -1 & & & \\ -1 & \ddots & \ddots & & \\ & \ddots & \ddots & -1 & \\ & & -1 & 3 & \end{bmatrix}.$$

**Table 2** The spring problem: norms of the relevant quantities. The entries corresponding to the scaled  $\|\mathbf{A}_m\|_2$  and  $\|\mathbf{B}_m\|_2$  are  $\|\hat{\mathbf{A}}_m\|_2$  and  $\|\hat{\mathbf{B}}_m\|_2$ , respectively. Note the reduction in  $\max G_U(\lambda)$  and  $\max \|w\|_2/\|y\|_2$

	$\ \mathbf{A}_m\ _2$	$\ \mathbf{B}_m\ _2$	$\max G_U(\lambda)$	$\max H_U(\lambda)$	$\max \ v\ _2/\ x\ _2$	$\max \ w\ _2/\ y\ _2$
Unscaled	$2.0 \times 10^3$	$1.0 \times 10^0$	$2.3 \times 10^3$	$7.0 \times 10^0$	$5.1 \times 10^1$	$5.0 \times 10^1$
Scaled	$3.8 \times 10^3$	$6.5 \times 10^1$	$3.6 \times 10^1$	$7.0 \times 10^0$	$5.1 \times 10^1$	$1.2 \times 10^0$



**Fig. 1** The spring problem: actual backward error ratios  $\eta_{R_m}(\lambda, y^*)/\eta_{L_m}(\lambda, w^*)$  and  $\eta_{R_m}(\lambda, x)/\eta_{L_m}(\lambda, v)$  and their bounds



**Fig. 2** The spring problem: backward errors of the computed eigenpairs, before and after scaling

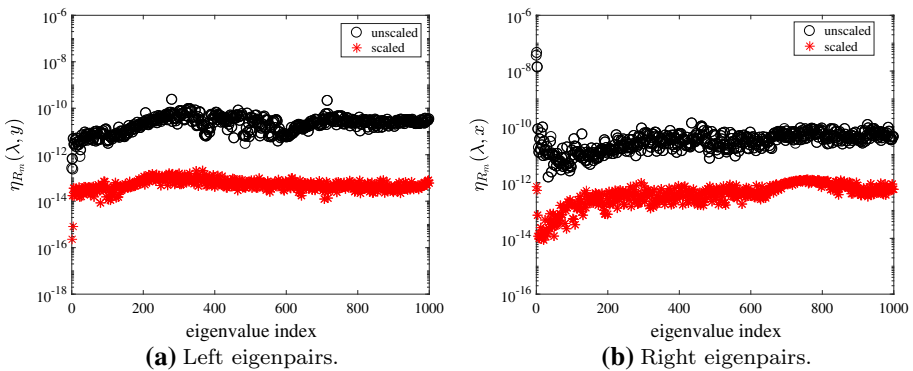
We interpolate  $F(\lambda)$  at the nodes  $\{-50, -25, 0\}$  using the Newton basis (3) where  $\beta_j = 1$  for  $j = 0, 1, 2$ . By the discussion in Sect. 3, we expect that the scaling reduces  $G_U(\lambda)$  and  $\|w\|_2/\|y\|_2$  while it keeps other quantities in the bounds relatively unchanged and this is confirmed by Table 2. The largest  $G_U(\lambda)$  and  $\|w\|_2/\|y\|_2$  for all the eigenvalues are reduced by a factor of  $10^2$  and  $10^1$ , respectively. These reductions lead us to expect smaller bounds for the ratios  $\eta_{R_m}(\lambda, x)/\eta_{L_m}(\lambda, v)$  and  $\eta_{R_m}(\lambda, y^*)/\eta_{L_m}(\lambda, w^*)$ . In Fig. 1, we display the bounds on these ratios for both the unscaled and the scaled linearizations, along with the actual ratios. The actual ratios are well predicted by the bounds. The reduced ratios in Fig. 1 further suggest improved accuracy and this is indeed what we see in Fig. 2, where we show the backward errors,

**Table 3** The 10-degree problem: norms of the coefficient matrices

$\ D_0\ _2$	$\ D_1\ _2$	$\ D_2\ _2$	$\ D_3\ _2$	$\ D_4\ _2$	$\ D_5\ _2$
$o(10^5)$	$o(10^2)$	$o(10^3)$	$o(10^2)$	$o(10^2)$	$o(10^1)$
$\ D_6\ _2$	$\ D_7\ _2$	$\ D_8\ _2$	$\ D_9\ _2$	$\ D_{10}\ _2$	
$o(10^2)$	$o(10^{-1})$	$o(10^2)$	$o(10^0)$	$o(10^1)$	

**Table 4** The 10-degree problem: norms of the relevant quantities

	$\ A_m\ _2$	$\ B_m\ _2$	$\max G_U(\lambda)$	$\max H_U(\lambda)$	$\max \ v\ _2/\ x\ _2$	$\max \ w\ _2/\ y\ _2$
Unscaled	$2.1 \times 10^7$	$1.8 \times 10^7$	$3.2 \times 10^{13}$	$8.4 \times 10^4$	$3.5 \times 10^4$	$1.3 \times 10^6$
Scaled	$1.7 \times 10^9$	$1.8 \times 10^8$	$1.7 \times 10^5$	$8.4 \times 10^4$	$3.5 \times 10^4$	$1.0 \times 10^0$



**Fig. 3** The 10-degree problem: backward errors of the computed eigenpairs, before and after scaling

with and without scaling. Scaling reduces the actual backward error ratios by 1 to 3 orders of magnitude.

### 4.2 Taylor Basis

Our second test problem is a PEP of degree 10 with randomly generated coefficient matrices

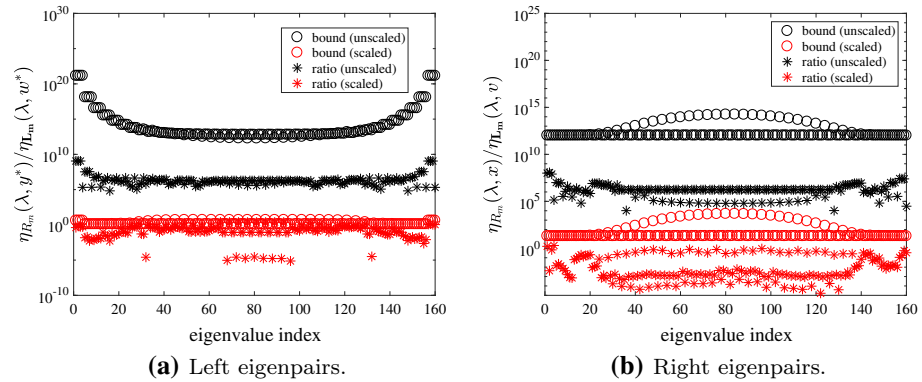
$$F(\lambda) = \sum_{k=0}^{10} D_k \lambda^k, \tag{27}$$

where the norms of the coefficient matrices  $D_k \in \mathbb{R}^{100 \times 100}$  are listed in Table 3. This example is featured by the fact that the norms vary by orders with  $\|D_0\|_2 \gg \|D_k\|_2$  for  $k > 0$ .

To perform the numerical experiment, we re-express (27) as a standard Taylor series about the origin. The effect of scaling is again reflected by the significant reduction of  $G_U(\lambda)$  and  $\|w\|_2/\|y\|_2$  in Table 4. Figure 3 clearly shows the improved accuracy brought about by scaling, where the backward errors are reduced approximately down to  $O(10^{-14})$ .

**Table 5** The orr\_sommerfeld problem: norms of the relevant quantities

	$\ A_m\ _2$	$\ B_m\ _2$	$\max G_U(\lambda)$	$\max H_U(\lambda)$	$\max \ v\ _2/\ x\ _2$	$\max \ w\ _2/\ y\ _2$
Unscaled	$2.6 \times 10^{10}$	$1.2 \times 10^{10}$	$3.9 \times 10^{11}$	$1.1 \times 10^2$	$1.2 \times 10^2$	$3.7 \times 10^{10}$
Scaled	$3.1 \times 10^{10}$	$2.5 \times 10^{10}$	$1.6 \times 10^1$	$1.1 \times 10^2$	$1.2 \times 10^2$	$1.8 \times 10^0$



**Fig. 4** The orr\_sommerfeld problem: actual backward error ratios  $\eta_{R_m}(\lambda, y^*)/\eta_{L_m}(\lambda, w^*)$  and  $\eta_{R_m}(\lambda, x)/\eta_{L_m}(\lambda, v)$  and their bounds

### 4.3 Orthogonal Basis

In the third example, we move to the CORK linearization when a PEP is represented in the orthogonal basis. Specifically, we choose  $b_j(\lambda)$  to be the Chebyshev polynomial of the first kind  $T_j(\lambda)$  and test on the quadratic eigenvalue problem arising in the spatial stability analysis of the Orr-Sommerfeld equation [21] which we, again, borrow from the collection of nonlinear eigenvalue problems [5]:

$$\left[ \left( \frac{d^2}{dz^2} - \lambda^2 \right)^2 - iR \left\{ (\lambda U - \omega) \left( \frac{d^2}{dz^2} - \lambda^2 \right) - \lambda U'' \right\} \right] \phi = 0.$$

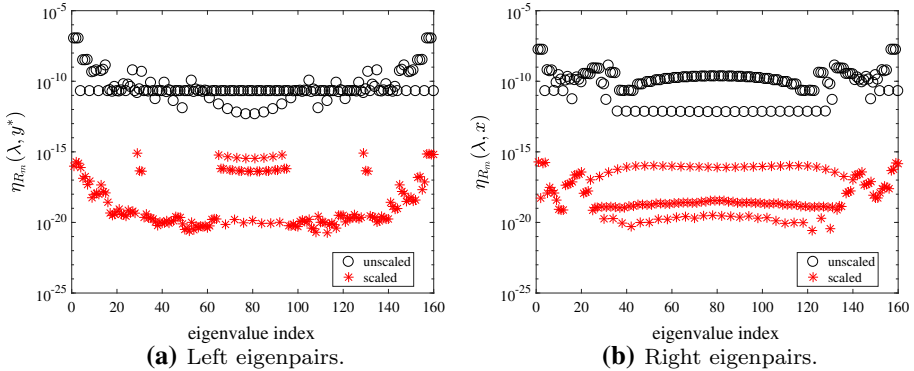
The choice of Chebyshev polynomials amounts to having  $\beta_1 = 1, \beta_2 = 1/2, \beta_3 = 1/2, \gamma_1 = \gamma_2 = -1/2$ , and  $\alpha_0 = \alpha_1 = \alpha_2 = 0$ . We sample the polynomial at Chebyshev points of the second kind, that is  $\sigma_j = \cos(j\pi/4), j = 0, 1, \dots, 4$ , in the interval  $[-1, 1]$  and the coefficient matrices  $D_j$  are computed using CHEBFUN [6].

As shown in Table 5, scaling brings  $\max G_U(\lambda)$  and  $\max \|w\|_2/\|y\|_2$  down by  $O(10^{10})$  while other key quantities are kept on the original order and this matches the reduction of the actual backward error ratios in Fig. 4. Figure 5 shows the backward errors are brought down to machine precision or below.

### 4.4 Lagrange Basis

This example is similar to the second one but now with the Lagrange basis, particularly with complex interpolation nodes. The matrix pencil  $F(\lambda) = \sum_{k=0}^{16} D_k \lambda^k$  is of degree 16 with the coefficient matrices  $D_k \in \mathbb{R}^{200 \times 200}$  randomly determined. Table 6 gives the norms of the coefficient matrices, which vary widely between  $O(10^{-1})$  to  $O(10^8)$  with  $\|D_3\|_2$  and  $\|D_5\|_2$



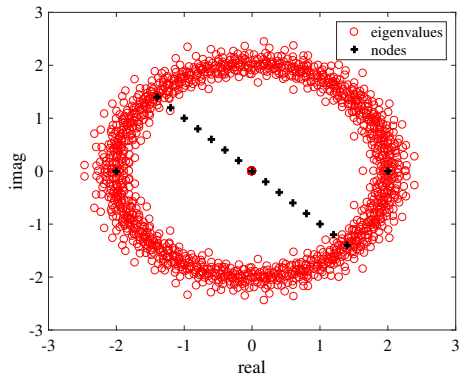


**Fig. 5** The `orr_sommerfeld` problem: backward errors of  $R_m(\lambda)$  of the computed eigenpairs, before and after scaling

**Table 6** The 16-degree problem: norms of the coefficient matrices

$\ D_0\ _2$	$\ D_1\ _2$	$\ D_2\ _2$	$\ D_3\ _2$	$\ D_4\ _2$	$\ D_5\ _2$	$\ D_6\ _2$	$\ D_7\ _2$	$\ D_8\ _2$
$O(10^0)$	$O(10^1)$	$O(10^{-1})$	$O(10^8)$	$O(10^{-2})$	$O(10^6)$	$O(10^1)$	$O(10^0)$	$O(10^3)$
$\ D_9\ _2$	$\ D_{10}\ _2$	$\ D_{11}\ _2$	$\ D_{12}\ _2$	$\ D_{13}\ _2$	$\ D_{14}\ _2$	$\ D_{15}\ _2$	$\ D_{16}\ _2$	
$O(10^{-3})$	$O(10^4)$	$O(10^5)$	$O(10^{-1})$	$O(10^6)$	$O(10^1)$	$O(10^0)$	$O(10^3)$	

**Fig. 6** The 16-degree problem: interpolation nodes (+) and locations of the eigenvalues (o)

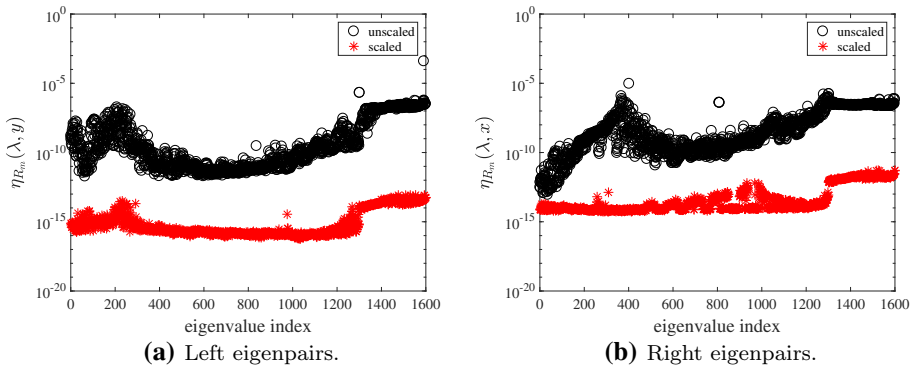


being the largest ones. We interpolate  $F(\lambda)$  using Lagrange interpolant with the interpolation nodes shown in Fig. 6 by plus signs.

Table 7 shows that the largest  $G_U(\lambda)$  and  $\|w\|_2/\|y\|_2$  are successfully reduced by  $10^8$  and  $10^9$ , respectively, due to the scaling, while the changes in other quantities are relatively negligible, which implies the significant reductions of the backward errors. Figure 7 confirms the effectiveness of scaling as the backward errors of the approximate eigenpairs of  $R_m(\lambda)$  are brought down to about machine precision, despite of the high degree of this problem and the extreme discrepancies in the magnitude of the coefficient matrices. The location of the eigenvalues computed with scaling are plotted in Fig. 6 as circles.

**Table 7** The 16-degree problem: norms of the relevant quantities with the Lagrange basis

	$\ A_m\ _2$	$\ B_m\ _2$	$\max G_U(\lambda)$	$\max H_U(\lambda)$	$\max \ v\ _2/\ x\ _2$	$\max \ w\ _2/\ y\ _2$
Unscaled	$1.8 \times 10^{10}$	$9.1 \times 10^9$	$6.2 \times 10^{17}$	$2.5 \times 10^7$	$1.2 \times 10^8$	$2.0 \times 10^9$
Scaled	$4.6 \times 10^{11}$	$2.3 \times 10^{11}$	$4.7 \times 10^9$	$2.5 \times 10^7$	$3.0 \times 10^8$	$1.0 \times 10^0$



**Fig. 7** The 16-degree problem: backward errors of computed eigenpairs, before and after scaling

**Table 8** The railtrack\_rep problem: norms of the relevant quantities

	$\ A_m\ _2$	$\ B_m\ _2$	$\max G_U(\lambda)$	$\max H_U(\lambda)$	$\max \ v\ _2/\ x\ _2$	$\max \ w\ _2/\ y\ _2$
Unscaled	$1.4 \times 10^{11}$	$2.1 \times 10^{10}$	$2.5 \times 10^{11}$	$1.8 \times 10^0$	$1.6 \times 10^0$	$8.1 \times 10^7$
Scaled	$6.1 \times 10^{11}$	$2.4 \times 10^{11}$	$1.4 \times 10^0$	$1.8 \times 10^0$	$1.6 \times 10^0$	$1.0 \times 10^0$

### 4.5 Rational Basis

We test the rational basis using the railtrack\_rep problem which is a rational eigenvalue problem arising from the study of the vibration of rail tracks [15]:

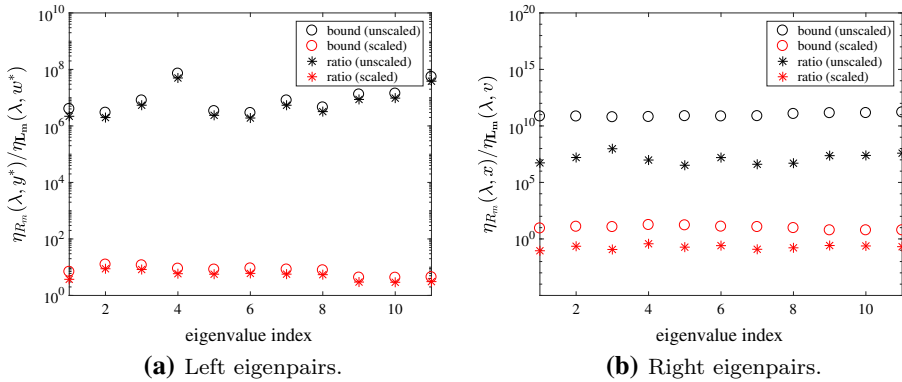
$$F(\lambda)x = (\lambda A^T + B + \lambda^{-1}A)x = 0,$$

where  $A$  and  $B$  are  $1005 \times 1005$  complex matrices and  $B = B^T$  is complex symmetric. In this example, we are particularly interested in eigenvalues with modulus between 3 and 10. In fact,  $F(\lambda)$  can be recast in the rational basis as

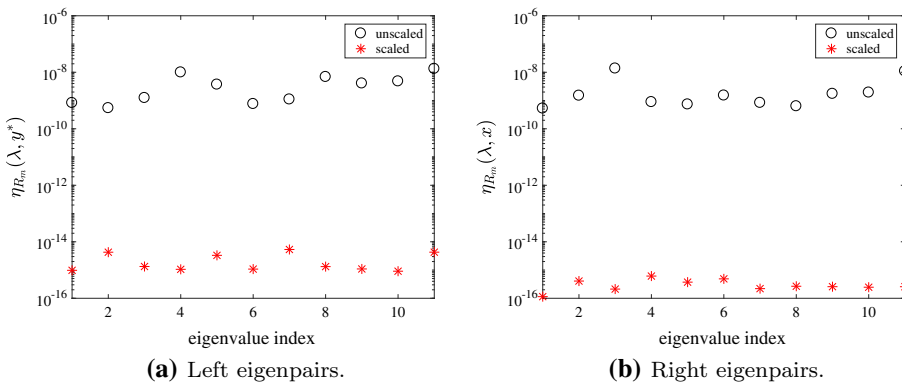
$$F(\lambda) \approx R_2(\lambda) = b_0(\lambda)D_0 + b_1(\lambda)D_1 + b_2(\lambda)D_2,$$

where  $b_j(\lambda)$ 's are given by (5). The pole locations  $\xi_1 = 1$  and  $\xi_2 = \infty$  and the parameters  $\beta_0 = 1$ ,  $\beta_1 = 0.78$ , and  $\beta_2 = 9.2$  are all found by using the RK Toolbox [3].

As seen from Table 8, the magnitudes of  $\max G_U(\lambda)$  and  $\max \|w\|_2/\|y\|_2$  fall down by orders of  $10^{11}$  and  $10^7$ , respectively. The reduction of the error ratios is displayed in Fig. 8, where we see the ratios are made roughly  $10^6$  times smaller by scaling for both the left and the right eigenpairs. The actual backward errors are brought down from somewhere between  $10^{-10}$  and  $10^{-8}$  to the level of machine precision, as shown in Fig. 9.



**Fig. 8** The railtrack\_rep problem: actual backward error ratios  $\eta_{R_m}(\lambda, y^*)/\eta_{L_m}(\lambda, w^*)$  and  $\eta_{R_m}(\lambda, x)/\eta_{L_m}(\lambda, v)$  and their bounds



**Fig. 9** The railtrack\_rep problem: backward errors of  $R_m(\lambda)$  of computed eigenpairs, before and after scaling

### 5 Closing Remarks

In this work, we have performed a backward error analysis for the PEPs and REPs solved by the CORK linearization, which has not been seen in the literature. This investigation has made two novel contributions. The first is that we established upper bounds for the ratios of the backward errors of approximate eigenpairs of  $R_m(\lambda)$  to those of the CORK linearization based on the one-sided factorization which are newly constructed and our treatment is given in a unified framework for all commonly-used bases. The second is a simple scaling approach, suggested by these bounds. Our analysis and this scaling approach are backed up by the numerical experiments in the last section which show an overall effectiveness and efficacy for reducing the backward errors.

**Acknowledgements** We are grateful to Françoise Tisseur for bringing this project to our attention and to Marco Fasoldini for his constructive commentaries which led us to improve our work. We thank Ren-Cang Li, Behnam Hashemi, and Yangfeng Su for very helpful discussions in various phases of this project. We also thank the anonymous reviewers for very helpful comments and suggestions.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Data Availability** Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

## A Proof of (14)

We now show that  $G(\lambda)$  given by (14) satisfies (12) with  $\mathbf{L}_m(\lambda)$  corresponding to the orthogonal basis and  $g(\lambda) = 1$ . To simplify the notation, we drop the argument  $\lambda$  of the basis function  $b_j^{(k)}(\lambda)$  in the rest of this proof.

We shall concentrate on showing that the product of  $G(\lambda)$  and the  $k$ th column of  $\mathbf{L}_m(\lambda)$  is  $k$ th column of  $e_1^T \otimes R_m(\lambda)$ , that is

$$D_{k-1} - \sum_{j=k-1}^m b_{j-(k-1)}^{(k-1)} D_j + \frac{\lambda + \alpha_{k-1}}{\beta_k} \sum_{j=k}^m b_{j-k}^{(k)} D_j + \frac{\gamma_k}{\beta_{k+1}} \sum_{j=k+1}^m b_{j-(k+1)}^{(k+1)} D_j = 0, \tag{28}$$

which is true if

$$b_0^{(k-1)} = 1, \tag{29a}$$

$$\frac{\lambda + \alpha_{k-1}}{\beta_k} b_{j-k}^{(k)} + \frac{\gamma_k}{\beta_{k+1}} b_{j-(k+1)}^{(k+1)} = b_{j-(k-1)}^{(k-1)}, \quad k \leq j \leq m, \tag{29b}$$

both hold.

Equation (29a) is given by the definition of the orthogonal basis. We now show (29b) by induction.

For  $j = k$ , (29b) reads

$$\frac{\lambda + \alpha_{k-1}}{\beta_k} b_0^{(k)} + \frac{\gamma_k}{\beta_{k+1}} b_{-1}^{(k+1)} = b_1^{(k-1)},$$

which reduces to

$$\frac{\lambda + \alpha_{k-1}}{\beta_k} b_0^{(k-1)} = b_1^{(k-1)}, \tag{30}$$

since  $b_{-1}^{(k+1)} = 0$  and  $b_0^{(k)} = b_0^{(k-1)} = 1$ . Equation (30) is the recurrence relation (13) with 0 and  $k - 1$  in place of  $j$  and  $k$ , respectively.

For  $j = k + 1$ , (29b) becomes

$$\frac{\lambda + \alpha_{k-1}}{\beta_k} b_1^{(k)} + \frac{\gamma_k}{\beta_{k+1}} b_0^{(k+1)} = b_2^{(k-1)}. \tag{31}$$

To show (31), we substitute  $b_1^{(k)} = \frac{\lambda + \alpha_k}{\beta_{k+1}} b_0^{(k)}$  into to have

$$\frac{\lambda + \alpha_k}{\beta_{k+1}} \left( \frac{\lambda + \alpha_{k-1}}{\beta_k} b_0^{(k-1)} \right) + \frac{\gamma_k}{\beta_{k+1}} b_0^{(k-1)} = b_2^{(k-1)},$$

where  $b_0^{(k-1)} = b_0^{(k)} = b_0^{(k+1)} = 1$  is used. Since the terms in the parentheses, by the recurrence relation, are just  $b_1^{(k-1)}$ , the task is boiled down to verify

$$\frac{\lambda + \alpha_k}{\beta_{k+1}} b_1^{(k-1)} + \frac{\gamma_k}{\beta_{k+1}} b_0^{(k-1)} = b_2^{(k-1)}.$$

This is, again, a variant of the recurrence relation of (13).

Now suppose that (29b) and

$$\frac{\lambda + \alpha_{k-1}}{\beta_k} b_{j+1-k}^{(k)} + \frac{\gamma_k}{\beta_{k+1}} b_{j+1-(k+1)}^{(k+1)} = b_{j+1-(k-1)}^{(k-1)} \quad (32)$$

both hold for any  $k - 1 \leq j \leq m - 2$ . By the recurrence relation (13), we have

$$(\lambda + \alpha_{j+1}) b_{j-k+2}^{(k-1)} + \gamma_{j+1} b_{j-k+1}^{(k-1)} = \beta_{j+2} b_{j-k+3}^{(k-1)},$$

where we now substitute (29b) and (32) for  $b_{j-(k-1)}^{(k-1)}$  and  $b_{j+1-(k-1)}^{(k-1)}$  respectively to have

$$\left( \frac{\lambda + \alpha_{k-1}}{\beta_k} b_{j-k+2}^{(k)} + \frac{\gamma_k}{\beta_{k+1}} b_{j-k+1}^{(k+1)} \right) \beta_{j+2} = b_{j-k+3}^{(k-1)} \beta_{j+2}, \quad (33)$$

where we have used two variants of the recurrence relation

$$\begin{aligned} (\lambda + \alpha_{j+1}) b_{j-k+1}^{(k)} + \gamma_{j+1} b_{j-k}^{(k)} &= \beta_{j+2} b_{j-k+2}^{(k)}, \\ (\lambda + \alpha_{j+1}) b_{j-k}^{(k+1)} + \gamma_{j+1} b_{j-k-1}^{(k+1)} &= \beta_{j+2} b_{j-k+1}^{(k+1)}. \end{aligned}$$

Dropping  $\beta_{j+2}$ 's in (33) yields

$$\frac{\lambda + \alpha_{k-1}}{\beta_k} b_{j+2-k}^{(k)} + \frac{\gamma_k}{\beta_{k+1}} b_{j+2-(k+1)}^{(k+1)} = b_{j+2-(k-1)}^{(k-1)}.$$

This, by induction, shows (29b), which, along with (29a), verifies (28).

The proofs for the first and the last two columns of  $G(\lambda) \mathbf{L}_m(\lambda)$  are essentially the same, though the structures of  $\mathbf{L}_m(\lambda)$  for these columns are slightly different from that of a general  $k$ th column.

## References

- Amiraslani, A., Corless, R.M., Lancaster, P.: Linearization of matrix polynomials expressed in polynomial bases. *IMA J. Numer. Anal.* **29**(1), 141–157 (2008)
- Antoniou, E.N., Vologianidis, S.: A new family of companion forms of polynomial matrices. *Electr. J. Linear Algebra* **11**(411), 78–87 (2004)
- Berljafa, M., Güttel, S.: A rational Krylov toolbox for Matlab, MIMS EPrint 2014.56. Manchester Institute for Mathematical Sciences, The University of Manchester, UK pp. 1–21 (2014)
- Betcke, T.: Optimal scaling of generalized and polynomial eigenvalue problems. *SIAM J. Matrix Anal. Appl.* **30**(4), 1320–1338 (2008)
- Betcke, T., Higham, N.J., Mehrmann, V., Schröder, C., Tisseur, F.: NLEVP: a collection of nonlinear eigenvalue problems. *ACM Trans. Math. Softw.* **39**(7), 1–28 (2013)
- Driscoll, T.A., Hale, N., Trefethen, L.N.: *Chebfun Guide*. Pafnuty Publications (2014). <http://www.chebfun.org/docs/guide/>
- Effenberger, C., Kressner, D.: Chebyshev interpolation for nonlinear eigenvalue problems. *BIT Numer. Math.* **52**(4), 933–951 (2012)
- Fiedler, M.: A note on companion matrices. *Linear Algebra App.* **372**, 325–331 (2003)
- Gaubert, S., Sharify, M.: Tropical scaling of polynomial matrices. In: *Positive systems*, pp. 291–303. Springer (2009)
- Gohberg, I., Lancaster, P., Rodman, L.: *Matrix Polynomials*. SIAM, New York (2009)
- Grammont, L., Higham, N.J., Tisseur, F.: A framework for analyzing nonlinear eigenproblems and parametrized linear systems. *Linear Algebra Appl.* **435**, 623–640 (2011)
- Güttel, S., Tisseur, F.: The nonlinear eigenvalue problem. *Acta Numerica* **26**, 1–94 (2017)
- Güttel, S., Van Beeumen, R., Meerbergen, K., Michiels, W.: NLEIGS: a class of fully rational Krylov methods for nonlinear eigenvalue problems. *SIAM J. Sci. Comput.* **36**(6), A2842–A2864 (2014)
- Higham, N.J., Li, R.C., Tisseur, F.: Backward error of polynomial eigen problems solved by linearization. *SIAM J. Matrix Anal. Appl.* **29**(4), 1218–1241 (2007)

15. Hilliges, A., Mehl, C., Mehrmann, V.: On the solution of palindromic eigenvalue problems. In: Proceedings of the 4th European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS). Jyväskylä, Finland (2004)
16. Karlsson, L., Tisseur, F.: Algorithms for Hessenberg-triangular reduction of Fiedler linearization of matrix polynomials. *SIAM J. Sci. Comput.* **37**(3), C384–C414 (2015)
17. Lawrence, P.W., Corless, R.M.: Backward error of polynomial eigenvalue problems solved by linearization of Lagrange interpolants. *SIAM J. Matrix Anal. Appl.* **36**(4), 1425–1442 (2015)
18. Lawrence, P.W., Van Barel, M., Van Dooren, P.: Backward error analysis of polynomial eigenvalue problems solved by linearization. *SIAM J. Matrix Anal. Appl.* **37**(1), 123–144 (2016)
19. Mackey, D.S., Mackey, N., Mehl, C., Mehrmann, V.: Vector spaces of linearizations for matrix polynomials. *SIAM J. Matrix Anal. Appl.* **28**(4), 971–1004 (2006)
20. Tisseur, F.: Backward error and condition of polynomial eigenvalue problems. *Linear Algebra Appl.* **309**(1–3), 339–361 (2000)
21. Tisseur, F., Higham, N.J.: Structured pseudospectra for polynomial eigenvalue problems, with applications. *SIAM J. Matrix Anal. Appl.* **23**(1), 187–208 (2001)
22. Van Beeumen, R., Meerbergen, K., Michiels, W.: Compact rational Krylov methods for nonlinear eigenvalue problems. *SIAM J. Matrix Anal. Appl.* **36**(2), 820–838 (2015)

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