Chebyshev rootfinding via computing eigenvalues of colleague matrices: when is it stable?

Noferini, Vanni and Perez Alvaro, Javier

2015

MIMS EPrint: **2015.24**

ISSN 1749-9097

Chebyshev rootfinding via computing eigenvalues of colleague matrices: when is it stable?

Vanni Noferini[*] and Javier Pérez[†]

March 23, 2015

Abstract

Computing the roots of a scalar polynomial, or the eigenvalues of a matrix polynomial, expressed in the Chebyshev basis $\{T_k(x)\}$ is a fundamental problem that arises in many applications. In this work, we analyze the backward stability of the polynomial rootfinding problem solved with colleague matrices. In other words, given a scalar polynomial $p(x)$ or a matrix polynomial $P(x)$ expressed in the Chebyshev basis, the question is to determine whether the whole set of computed eigenvalues of the colleague matrix, obtained with a backward stable algorithm, like the QR algorithm, are the set of roots of a nearby polynomial or not. In order to do so, we derive a first order backward error analysis of the polynomial rootfinding algorithm using colleague matrices adapting the geometric arguments in [A. Edelman and H. Murakami, *Polynomial roots for companion matrix eigenvalues*, Math. Comp. 210, 763–776, 1995] to the Chebyshev basis. We show that, if the absolute value of the coefficients of $p(x)$ (respectively, the norm of the coefficients of $P(x)$) are bounded by a moderate number, computing the roots of $p(x)$ (respectively, the eigenvalues of $P(x)$) via the eigenvalues of its colleague matrix using a backward stable eigenvalue algorithm is backward stable. This backward error analysis also expands on the very recent work [Y. Nakatsukasa and V. Noferini, *On the stability of computing polynomial roots via confederate linearizations*, To appear in Math. Comp.] that already showed that this algorithm is not backward normwise stable if the coefficients of the polynomial $p(x)$ do not have moderate norms.

Keywords: polynomial, roots, Chebyshev basis, matrix polynomial, colleague matrix, backward stability, polynomial eigenvalue problem, Arnold trasnversality theorem

MSC classification: 65H04, 65H17, 65F15, 65G50

1 Introduction

A popular way to compute the roots of a monic polynomial expressed in the monomial basis is via the eigenvalues of its companion matrix. This is, for instance, the way followed by the MATLAB command `roots`, that, after balancing the companion matrix, uses the QR-algorithm to get its eigenvalues. The numerical properties of this method for computing roots of polynomials have been extensively studied [7, 8, 14, 24], in particular with respect to conditioning and backward errors. It has been shown that, in practice, if the companion matrix is balanced [20], the rootfinding method using companion matrices is numerically stable, in the sense that the computed roots are the exact roots of a nearby polynomial. However, as it was made famous by Wilkinson [21, 25, 26], roots of polynomials whose roots lie on a real interval can be highly sensitive to perturbations in the coefficients when the monomial basis is used. So, even perturbations in the coefficients of order of the machine precision may produce a catastrophically large forward error. In practice, rootfinding on a real interval is a very frequent and important situation, and one way to circumvent this problem is to use, instead, a polynomial basis such that the roots of a polynomial expressed in that basis are better conditioned functions of its coefficients, like the *Chebyshev basis*.

[*]School of Mathematics, The University of Manchester, Manchester, England, M13 9PL (`vanni.noferini@maths.manchester.ac.uk`)

[†]School of Mathematics, The University of Manchester, Manchester, England, M13 9PL (`javier.perezalvaro@manchester.ac.uk`)

Chebyshev polynomials are a family of polynomials, orthogonal with respect to the weight function $w(x) = (1 - x^2)^{-1/2}$ on the interval $[-1, 1]$, which may be computed using the following recurrence relation [1, Chapter 22]:

$$\begin{aligned} &T_0(x) = 1, \\ &T_1(x) = x, \quad \text{and}, \\ &T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x), \quad \text{for } k \geq 2. \end{aligned} \tag{1}$$

Moreover, the Chebyshev polynomials $T_0(x), T_1(x), \ldots, T_n(x)$ form a basis for the vector space of polynomials of degree at most n with real coefficients $\mathbb{R}_n[x]$, so, any real polynomial $p(x) \in \mathbb{R}_n[x]$ can be written uniquely as $p(x) = \sum_{k=0}^{n} a_k T_k(x)$.

Chebyshev polynomials are widely used in many areas of numerical analysis, and in particular approximation theory [22]. In fact, a common approach, as done in Chebfun [23], for computing the roots of a nonlinear smooth function $f(x)$ on an interval is to approximate first $f(x)$ by a polynomial $p(x)$ expressed in the Chebyshev basis via Chebyshev interpolation and then compute the roots of $p(x)$ as the eigenvalues of its colleague matrix [10].

In this paper, we are interested in the backward stability of the rootfinding problem solved via colleague matrices and a backward stable eigenvalue algorithm. Our work is motivated by [17], which addresses related issues for confederate matrices (the colleague matrix is a particular example of a confederate matrix [4, 16]). Also, similar backward error analysis may be found in [7, 12, 13]. In [7], the authors study the backward stability of rootfinding methods using Fiedler companion matrices of monic polynomials expressed in the monomial basis; in [12], the authors study the backward stability of rootfinding methods using a suitable companion matrix of polynomials expressed in barycentric form; in [13], several bases are analyzed at once, for nonstandard linearizations of larger size with respect to the colleague or the companion.

Given a *monic scalar polynomial in the Chebyshev basis* of degree n

$$p(x) = T_n(x) + \sum_{k=0}^{n-1} a_k T_k(x), \quad \text{with } a_k \in \mathbb{R}, \quad \text{for } k = 0, 1, \ldots, n-1,$$

where by monic in the Chebyshev basis we mean that the coefficient of $T_n(x)$ is equal to 1, the following matrix

$$C_T = \frac{1}{2} \begin{bmatrix} -a_{n-1} & -a_{n-2}+1 & -a_{n-3} & \cdots & -a_2 & -a_1 & -a_0 \\ 1 & 0 & 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & 1 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 & 0 & 1 \\ 0 & \cdots & \cdots & \cdots & 0 & 2 & 0 \end{bmatrix} \in \mathbb{R}^{n \times n},$$

is known as the *colleague matrix* of $p(x)$ [10]. If we write $T_n(x) = \nu_n x^n + \cdots$, the colleague matrix has the property that $\det(xI - C_T) = \frac{1}{\nu_n} p(x)$, so the roots of $p(x)$ may be computed as the eigenvalues of C_T using, for instance, the QR algorithm. The QR algorithm is a backward stable algorithm, this means than the computed eigenvalues are the exact eigenvalues of a matrix $C_T + E$, where E is a (possibly dense) matrix such that

$$\|E\| = O(u)\|C_T\|,$$

for some matrix norm, where u denotes the machine precision. However, the previous equation does not guarantee that the computed eigenvalues are the roots of a nearby polynomial of $p(x)$ or, in other words, that this rootfinding method is backward stable. In order for the method to be backward stable (in a normwise sense), the computed eigenvalues should be the exact roots of a polynomial $\widetilde{p}(x) = T_n(x) + \sum_{k=0}^{n-1} \widetilde{a}_k T_k(x)$, such that

$$\frac{\|\widetilde{p} - p\|}{\|p\|} = O(u),$$

for some polynomial norm.

The backward stability of the polynomial rootfinding in degree-graded basis using confederate matrices is studied in [17]. In particular (see [17, Theorem 4.2]), the authors prove that if C_T is the colleague matrix of a polynomial $p(x)$ and $E \in \mathbb{R}^{n\times n}$ is any matrix, then the eigenvalues of $C_T + E$ are the exact roots of a polynomial $\widetilde{p}(x)$ such that

$$\widetilde{p}(x) - p(x) = \sum_{i=0}^{n-1} \delta_i(p, E) T_i(x) + O(\|E\|_2^2), \tag{2}$$

where, for $i = 0, 1, \dots, n-1$, the quantity $\delta_i(p, E)$ is an affine function of the coefficients of $p(x)$, and, separately, of the entries of E.

Equation (2) implies that if the roots of $p(x)$ are computed as the eigenvalues of its colleague matrix C_T using a backward stable eigenvalue algorithm, then, the computed roots will be the exact roots of a polynomial $\widetilde{p}(x)$ such that

$$\frac{\|\widetilde{p} - p\|}{\|p\|} = \kappa(n) O(u) \|p\|,$$

for some constant $k(n)$. The previous equation shows, first, that this method is not backward stable if $\|p\| \gg 1$, and, second, that this method is backward stable if the following two conditions are satisfied: (i) the quantity $\kappa(n)$ is a low-degree polynomial in n with moderate coefficients; and, (ii) the norm $\|p\|$ is moderate. As it is observed in [17], writing $\delta_i(p, E) = \sum_{i,j,\ell} \beta_{ij\ell} a_\ell E_{ij}$, since it is not clear what exactly are the constants $\beta_{ij\ell}$ involved, in principle it could happen that $|\beta_{ij\ell}| >> 1$, implying that $\kappa(n)$ might not be a polynomial in n with moderate coefficients. However, in this work we show that, in fact, $|\beta_{ij\ell}| \leq 4$, and, so,

$$\frac{\|\widetilde{p} - p\|}{\|p\|} = O(u) \|p\|,$$

holds. The previous equation implies that computing the roots of $p(x)$ via the eigenvalues of its colleague matrix using a backward stable eigenvalue algorithm is a backward stable rootfinding algorithm, provided that $\|p\| \lesssim 1$.

Computing the eigenvalues of matrix polynomials in the Chebyshev basis is becoming an important problem [9]. Given a $p \times p$ *monic matrix polynomial in the Chebyshev basis* of degree n

$$P(x) = I_p T_n(x) + \sum_{k=0}^{n-1} A_k T_k(x), \quad \text{with } A_k \in \mathbb{R}^{p\times p}, \quad \text{for } k = 0, 1, \dots, n-1, \tag{3}$$

where by monic in the Chebyshev basis we mean that the coefficient of $T_n(x)$ is equal to I_p (the $p \times p$ identity matrix), the *polynomial eigenvalue problem* consists of finding the eigenvalues of $P(x)$, that is, finding the roots of the scalar polynomial $\det(P(x))$ (note that the monicity of $P(x)$ implies its regularity), that is, $\det(P(x))$ is not identically zero). A common approach to solve the polynomial eigenvalue problem for $P(x)$ is to use the *block colleague matrix*

$$C_T = \frac{1}{2} \begin{bmatrix} -A_{n-1} & -A_{n-2} + I_p & -A_{n-3} & \cdots & -A_2 & -A_1 & -A_0 \\ I_p & 0 & I_p & 0 & \cdots & \cdots & 0 \\ 0 & I_p & 0 & I_p & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & I_p & 0 & I_p \\ 0 & \cdots & \cdots & \cdots & 0 & 2I_p & 0 \end{bmatrix} \in \mathbb{R}^{np\times np}, \tag{4}$$

since it is known (see [2]) that the eigenvalues of (4) coincide with the eigenvalues of $P(x)$.

The backward stability question arises also in the matrix polynomial case, since the eigenvalues of C_T may be computed using a backward stable eigenvalue algorithm, like the QR algorithm. As in the scalar polynomial case, the backward stability of the QR algorithm implies that the computed eigenvalues are the exact eigenvalues of the perturbed matrix $C_T + E$, with $\|E\| = O(u)\|C_T\|$. However, again, the backward stability of the QR algorithm does not guarantee that the computed eigenvalues are the eigenvalues of a nearby matrix polynomial of $P(x)$ or, in other words, that the method is backward stable. In order for the method to be backward stable (in a normwise sense), the computed eigenvalues should be the exact eigenvalues of a matrix polynomial $\widetilde{P}(x) = I_pT_n(x) + \sum_{k=0}^{n-1} \widetilde{A}_kT_k(x)$, with $\widetilde{A}_k \in \mathbb{R}^{p\times p}$, such that

$$\frac{\|\widetilde{P} - P\|}{\|P\|} = O(u),$$

for some matrix polynomial norm.

Nevertheless, we will prove using some arguments inspired by [3, 8, 14, 15] that if the eigenvalues of a matrix polynomial $P(x)$ are computed as the eigenvalues of its colleague matrix using a backward stable eigenvalue algorithm, then, the computed eigenvalues are the exact eigenvalues of a monic matrix polynomial in the Chebyshev basis $\widetilde{P}(x)$ such that

$$\frac{\|\widetilde{P} - P\|}{\|P\|} = O(u)\|P\|.$$

The previous equation implies that this method is backward stable if $\|P\|$ is moderate.

The paper is organized as follows. In Section 2 we present Arnold transversality theorem for colleague matrices, which will be the main tool to study the polynomial backward stability of the rootfinding method using colleague matrices. In Section 3, to give a flavor of our algebraic approach based on interpreting the companion or colleague linearizations as multiplication-by-x operators in certain quotient modules (see also [18, Sec. 5]), we first review the backward error analysis of the rootfinding method using companion matrices in [8, 14]. Then, in Section 4 we prove Arnold transversality theorem for colleague matrices, and we use this theorem to study the backward stability of the rootfinding method using colleague matrices.

Throughout this paper, for a $p \times p$ matrix polynomial $P(x) = \sum_{k=0}^{n} A_kT_k(x)$, non necessarily monic, $\|P\|_F$ is the norm on the vector space of $p \times p$ matrix polynomials of degree less than or equal to n defined as

$$\|P\|_F = \sqrt{\sum_{k=0}^{n} \|A_k\|_F^2}.$$

Notice that, since we are going to deal with monic polynomials in the Chebyshev basis, $A_n = I_p$. Also notice that for a scalar polynomial $p(x) = \sum_{k=0}^{n} a_kT_k(x)$, that is, for $p = 1$, this norm reduces to the usual 2-norm:

$$\|p\|_F = \|p\|_2 = \sqrt{\sum_{k=0}^{n-1} |a_k|^2}.$$

2 Arnold transversality theorem for colleague matrices

Arnold transversality theorem will be the main tool in Section 4 to study what kind of polynomial backward stability is provided by matrix backward stability when the roots of scalar polynomials or the eigenvalues of matrix polynomials are computed as the eigenvalues of its colleague matrix with a backward stable eigenvalue algorithm. This theorem was first stated in [3] for companion matrices, and later generalized in [17] to confederate matrices of scalar polynomials.

Following [3, 7, 8, 14, 17], we consider the Euclidian matrix space $\mathbb{R}^{n\times n}$ with the usual Frobenius inner product

$$< A, B >:= \operatorname{tr}(AB^T),$$

where M^T denotes the transpose of $M \in \mathbb{R}^{n\times n}$. In this space, the set of matrices similar to a given matrix $A \in \mathbb{R}^{n\times n}$ is a differentiable manifold in $\mathbb{R}^{n\times n}$. This manifold is called the orbit of A under the action of similarity:

$$\mathcal{O}(A) := \{SAS^{-1} : S \in \mathbb{R}^{n\times n} \text{ and } \det(S) \neq 0\}.$$

A first-order expansion shows that the tangent space of $\mathcal{O}(A)$ at A is the set

$$T_A\mathcal{O}(A) := \{AX - XA \quad \text{for some } X \in \mathbb{R}^{n\times n}\}.$$

We also consider the vector subspace of *"first block row matrices"*, denoted by $\mathcal{BFR}_{n,p} \subset \mathbb{R}^{np\times np}$, which is defined as those $n \times n$ block matrices $[X_{ij}]$, with $X_{ij} \in \mathbb{R}^{p\times p}$, whose block rows are all zero except (possibly) the first:

$$\mathcal{BFR}_{n,p} := \left\{ X = \begin{bmatrix} I_P & 0 & \cdots & 0 \end{bmatrix}^T \begin{bmatrix} X_1 & X_2 & \cdots & X_n \end{bmatrix}, \text{ for some } X_1, X_2, \ldots, X_n \in \mathbb{R}^{p\times p} \right\} \subset \mathbb{R}^{np\times np}.$$

Note that taking $p = 1$ the space $\mathcal{BFR}_{n,p}$ reduces to the vector subspace $\mathcal{FR}_n$ of *"first row matrices"* introduced in [17].

Arnold transversality theorem for a block colleague matrix C_T of a monic matrix polynomial $P(x)$ in the Chebyshev basis states that any matrix $E \in \mathbb{R}^{np\times np}$ may be decomposed as

$$E = F_0 + T,$$

where $F_0 \in \mathcal{BFR}_{n,p}$ is a first block row matrix and $T \in T_{C_T}\mathcal{O}(C_T)$. Notice that taking $p = 1$ in the previous decomposition, this "block" version of Arnold transversality theorem reduces to a special case of [17, Theorem 4.1].

In Section 4.2 we present a proof of Arnold transversality theorem, different to the one in [17], extending (for the important case of the Chebyshev basis) [17, Theorem 4.1] to the more complicated case of matrix polynomials. The new approach allows us to compute explicitly the matrix F_0. Then, using this explicit expression, we study the polynomial backward stability of the rootfinding method using colleague matrices. However, before doing that, to give a flavor of our approach, we first treat the much easier case of the monomial basis (see [8, 14]).

3 Arnold transversality theorem for companion matrices and backward error analysis

The backward stability of the polynomial rootfinding using companion matrices is studied in [8]. In this section we review their results in the matrix polynomial case, since results for scalar polynomials may be seen as corollaries of them.

Given a $p \times p$ monic matrix polynomial expressed in the monomial basis

$$P(x) = I_p x^n + \sum_{k=0}^{n-1} B_k x^k, \quad \text{with } B_k \in \mathbb{R}^{p\times p}, \quad \text{for } k = 0, 1, \ldots, n-1,$$

its eigenvalues may be computed as the eigenvalues of the *block companion matrix* (see [11, Chapter 14])

$$C_1 := \begin{bmatrix} -B_{n-1} & -B_{n-2} & \cdots & -B_1 & -B_0 \\ I_p & 0 & \cdots & 0 & 0 \\ 0 & I_p & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & I_p & 0 \end{bmatrix},$$

using a backward stable eigenvalue algorithm.

In [8], it is shown how a dense perturbation E to the block companion matrix leads to first order perturbations in the coefficients B_k. To do so, they use Arnold transversality theorem for companion matrices, that is, they show how any matrix $E \in \mathbb{R}^{np\times np}$ may be decomposed as

$$E = E_0 + T = E_0 + C_1X - XC_1, \tag{5}$$

where $E_0 \in \mathcal{BFR}_{n,p}$ is a first block row matrix and $T = XC_1 - C_1X \in T_{C_1}\mathcal{O}(C_1)$ is a matrix in the tangent space of the orbit of C_1. As a consequence of the previous decomposition we have that if E is a small perturbation of C_1, then, to first order in E, the matrix $C_1 + E$ is similar to $C_1 + E_0$. In other words,

$$C_1 + E = (I + X)^{-1}(C_1 + E_0)(I + X) + O(\|E\|_2^2),$$

for some matrix $X \in \mathbb{R}^{np\times np}$. Noting that $C_1 + E_0$ is in turn a block companion matrix of another matrix polynomial, we see that a small perturbation of the block companion matrix of $P(x)$ is similar, to first order in the norm of the perturbation, to a block companion matrix of a perturbed polynomial $\widetilde{P}(x)$.

In order to compute E_0 in (5), define the matrices

$$N_k := \sum_{j=k}^{n} C_1^{n-j}(I_n \otimes B_j), \quad \text{for } k = 1, 2, \ldots, n, \tag{6}$$

where $B_n := I_p$ and $\otimes$ denotes the Kronecker product, which satisfy, for $k = 1, 2, \ldots, n-1$,

$$N_k = N_{k+1}C_1 + I_n \otimes B_k, \quad \text{with } N_n = I_n \otimes I_p. \tag{7}$$

Using the recurrence relation (7), the matrix N_k can be computed explicitly [8, Lemma 4.1]:

$$N_k = \left[\begin{array}{ccccccc}
0 & -B_{k-1} & \cdots & -B_1 & -B_0 & & \\
\vdots & & \ddots & \ddots & \ddots & \ddots & \\
0 & & & -B_{k-1} & \cdots & -B_1 & -B_0 \\
\hline
I_p & B_{n-1} & \cdots & B_k & & & \\
0 & I_p & B_{n-1} & \cdots & B_k & & \\
\vdots & & \ddots & \ddots & \ddots & \ddots & \\
0 & & & I_p & B_{n-1} & \cdots & B_k
\end{array}\right], \quad \text{for } k = 1, 2, \ldots, n, \tag{8}$$

where the first block-row contains $n - k$ block rows, and the second block row contains k block rows.

Writing the matrix E and the matrix N_k as $n\times n$ block matrices $E = [E_{ij}]$ and $N_k = [(N_k)_{ij}]$, respectively, with $E_{ij}, (N_k)_{ij} \in \mathbb{R}^{p\times p}$, the authors in [8] prove that the matrix E_0 in (5) is given by

$$E_0 = \begin{bmatrix}
\sum_{i,j=1}^n E_{ij}(N_n)_{ji} & \cdots & \sum_{i,j=1}^n E_{ij}(N_2)_{ji} & \sum_{i,j=1}^n E_{ij}(N_1)_{ji} \\
0 & \cdots & 0 & 0 \\
\vdots & & \vdots & \vdots \\
0 & \cdots & 0 & 0
\end{bmatrix}.$$

Therefore, the matrix $C_1 + E$ is similar, to first order in the norm of E, to a block companion matrix associated with the monic matrix polynomial

$$\widetilde{P}(x) = I_p x^n + \sum_{k=0}^{n-1} \widetilde{B}_k x^k = I_p x^n + \sum_{k=0}^{n-1}(B_k - \sum_{ij=1}^{n} E_{ij}(N_{k+1})_{ji}))x^k.$$

From the previous equation, together with (8), we get that the eigenvalues of $C_1 + E$ are the eigenvalues of a monic matrix polynomial $\widetilde{P}(x) = I_p x^n + \sum_{k=0}^{n-1} \widetilde{B}_k x^k$ such that, to first order in the norm of E,

$$\|\widetilde{B}_k - B_k\|_F = \|\sum_{s=k+1}^{n}\left(\sum_{j=n-k}^{n} B_s E_{j-s+k+1,j}\right) - \sum_{s=0}^{k}\left(\sum_{j=1}^{n-k-1} B_s E_{j-s+k+1,j}\right)\|_F, \tag{9}$$

for $k = 0, 1, \ldots, n-1$. Then, if the eigenvalues of the block companion matrix C_1 are computed with a backward stable eigenvalue algorithm, it may be proved from (9) that, to first order in E, the computed eigenvalues are the exact eigenvalues of a polynomial $\widetilde{P}(x)$ such that

$$\frac{\|\widetilde{P} - P\|_F}{\|P\|_F} = O(u)\|P\|_F. \tag{10}$$

Note that (10) implies that if the norm of the coefficients of $P(x)$ are moderate, then computing its eigenvalues via the eigenvalues of its block companion matrix C_1 is a backward stable method.

In the following, we present a different approach to the matrices N_k, for $k = 1, 2, \ldots, n$, in (8) in order to better understand them and the block Toeplitz structure of their two blocks. First, by direct multiplication, it may be easily checked (see also [18, Section 5] and [19, Section 9]) that the block companion matrix C_1 satisfies

$$x \begin{bmatrix} x^{n-1} \\ \vdots \\ x \\ 1 \end{bmatrix} \otimes I_p = C_1 \begin{bmatrix} x^{n-1} \\ \vdots \\ x \\ 1 \end{bmatrix} \otimes I_p + e_1 \otimes P(x), \tag{11}$$

where e_1 denotes the first column of the $n \times n$ identity matrix. Then, define the degree k Horner shift of $P(x)$ as the matrix polynomial (see [6, Definition 4.1])

$$P_k(x) := I_p x^k + \sum_{j=0}^{k-1} B_{n-k+j} x^j. \tag{12}$$

The Horner shifts of $P(x)$ satisfy, for $k = 1, 2, \ldots, n-1$,

$$P_k(x) = x P_{k-1}(x) + B_{n-k}, \quad \text{with } P_0(x) = I_p. \tag{13}$$

Using (13), together with (7) and (11), it may be proved that the matrix N_k in (6) is the unique matrix satisfying

$$\begin{bmatrix} x^{n-1} \\ \vdots \\ x \\ 1 \end{bmatrix} \otimes P_{n-k}(x) = N_k \begin{bmatrix} x^{n-1} \\ \vdots \\ x \\ 1 \end{bmatrix} \otimes I_p + \begin{bmatrix} r_{1k}(x) \\ \vdots \\ r_{n-1,k}(x) \\ r_{nk}(x) \end{bmatrix} \otimes P(x) \tag{14}$$

for some scalar polynomials $r_{1k}(x), \ldots, r_{nk}(x)$. In words, the ith block row of N_k contains the coefficients of the unique matrix polynomial $Q_{ik}(x)$ of degree less than or equal to $n-1$ such that

$$x^{n-i} P_{n-k}(x) = Q_{ik}(x) + r_{ik}(x) P(x), \tag{15}$$

for some scalar polynomial $r_{ik}(x)$.

Observing that the Horner shifts in (12) satisfy

$$x^k P_{n-k}(x) = P(x) - B_{k-1} x^{k-1} - \cdots - B_1 x - B_0, \quad \text{for } k = 1, 2, \ldots, n-1, \tag{16}$$

it is immediate to check that the polynomial $Q_{ik}(x)$ in (15) is given by

$$Q_{ik}(x) = \begin{cases} \sum_{j=k}^{n} B_j x^{n+j-i-k} & \text{if } i \geq n-k+1, \quad \text{and} \\ \sum_{j=0}^{k-1} -B_j x^{n+j-i-k} & \text{if } i \leq n-k, \end{cases} \tag{17}$$

for $i, k = 1, 2, \ldots, n$, where we set $B_n := I_p$. Then, from (15) and (17) we get

$$\begin{bmatrix} x^{n-1} \\ \vdots \\ x^k \\ x^{k-1} \\ \vdots \\ 1 \end{bmatrix} \otimes P_{n-k}(x) = \begin{bmatrix} -\sum_{j=0}^{k-1} B_j x^{n-1+j-k} \\ \vdots \\ -\sum_{j=0}^{k-1} B_j x^j \\ \sum_{j=k}^{n} B_j x^{j-1} \\ \vdots \\ \sum_{j=l}^{n} B_j x^{j-k} \end{bmatrix} + \begin{bmatrix} r_{1k}(x) \\ \vdots \\ r_{n-k,k}(x) \\ r_{n-k+1,k}(x) \\ \vdots \\ r_{nk}(x) \end{bmatrix} \otimes P(x) \tag{18}$$

and comparing (14) with (18), we recover the expression in (8) for the matrix N_k, and the Toeplitz structure of its two blocks is also immediately explained.

In Section 4.2 we will use an approach similar to (14) in the case of the Chebyshev basis. However, there are more technicalities to be addressed with respect to the monomial basis.

4 Arnold transversality theorem for colleague matrices and backward error analysis

4.1 Clenshaw shifts and Clenshaw matrices

In this section we introduce some matrix polynomials and some matrices, named here as *Clenshaw shifts* and *Clenshaw matrices*, respectively, associated with a monic matrix polynomial in the Chebyshev basis $P(x)$, that will be used through Section 4.2 and will be key in the following developments. Clenshaw shifts and Clenshaw matrices are the generalization of the Horner shifts in (12) and the matrices $N_1, N_2, \ldots, N_k$ in (6), respectively, when the polynomial $P(x)$ is expressed in the Chebyshev basis.

Associated with the $p \times p$ monic matrix polynomial in the Chebyshev basis $P(x)$ in (3), we define the following $p \times p$ matrix polynomials:

$$\begin{aligned}
&H_0(x) = 2I_p, \\
&H_1(x) = 2xH_0(x) + 2A_{n-1}, \\
&H_k(x) = 2xH_{k-1}(x) - H_{k-2}(x) + 2A_{n-k}, \quad \text{for } k = 2, 3, \ldots, n-2, \\
&H_{n-1}(x) = xH_{n-2}(x) - H_{n-3}(x)/2 + A_1.
\end{aligned} \tag{19}$$

We will refer, for $k = 1, 2, \ldots, n$, to the matrix polynomial $H_k(x)$ as the degree k *Clenshaw shifts* of $P(x)$, since for $p = 1$ they coincide with the well known Clenshaw shifts associated with a scalar polynomial expressed in the Chebyshev basis [5]. Clenshaw shifts are related with the polynomial $P(x)$ through the following equation [5]:

$$2P(x) = 2xH_{n-1}(x) - H_{n-2}(x) + 2A_0. \tag{20}$$

In Theorem 4.1, given the Chebyshev polynomial $T_{n-i}(x)$ and the Clenshaw shift $H_{n-k}(x)$, we show how to express $T_{n-i}(x)H_{n-k}(x)$ uniquely as $Q_{ij}(x) + r_{ik}(x)P(x)$, where $Q_{ij}(x)$ is a $p \times p$ matrix polynomial of degree less than or equal to $n-1$ and $r_{ik}(x)$ is a scalar polynomial. This result is the Chebyshev analogue of (15) and (17). The proof of Theorem 4.1 is elementary but rather technical, so we leave it to the appendix. In order to write down a reasonably simple formula for $T_{n-i}(x)H_{n-k}(x)$, we define the following quantities

$$\begin{aligned}
&\Gamma_{2k+1} = \Gamma_{2k-1} + 2A_{n-2k-1}, \quad \text{for } k = 1, 2, \ldots, \left\lfloor \frac{n}{2} \right\rfloor - 1, \quad \text{with } \Gamma_0 = 2I_p, \quad \text{and} \\
&\Gamma_{2k} = \Gamma_{2(k-1)} + 2A_{n-2k}, \quad \text{for } k = 1, 2, \ldots, \left\lceil \frac{n}{2} \right\rceil - 1, \quad \text{with } \Gamma_1 = 2A_{n-1}.
\end{aligned} \tag{21}$$

Notice that in Γ_k only appear coefficients of $P(x)$ with indices of the same parity.

Theorem 4.1. *Let $P(x) = I_pT_n(x) + \sum_{k=0}^{n-1} A_kT_k(x)$ be a $p \times p$ monic matrix polynomial in the Chebyshev basis of degree n, let $T_{n-i}(x)$ and $H_{n-k}(x)$ be, respectively, the degree $n-i$ Chebyshev polynomial and the degree $n-k$ Clenshaw shift of $P(x)$, with $i, k \in \{1, 2, \ldots, n\}$. Then, there exist a unique $p \times p$ matrix polynomial $Q_{ik}(x)$ of degree less than or equal to $n-1$ and a unique scalar polynomial $r_{ik}(x)$ such that*

$$T_{n-i}(x)H_{n-k}(x) = Q_{ik}(x) + r_{ik}(x)P(x),$$

where,

- *if $i \geq n-k+1$ and $k \geq 2$,*

$$Q_{ik}(x) = \sum_{\ell=0}^{n-k-1} \Gamma_\ell (T_{2n-i-k-\ell}(x) + T_{|k+\ell-i|}(x)) + \Gamma_{n-k}T_{n-i}(x); \tag{22}$$

- *if $i = n$ and $k = 1$,*

$$Q_{ik}(x) = \sum_{\ell=0}^{n-2} \Gamma_\ell T_{n-1-\ell}(x) + \frac{\Gamma_{n-1}}{2} T_0(x); \tag{23}$$

- *if $i \le n-k$ and $n-1 \ge k \ge 2$,*

$$Q_{ik}(x) = \sum_{\ell=0}^{i-2} \Gamma_\ell (T_{i+k-2-\ell}(x) + T_{|k+\ell-i|}(x)) + \Gamma_{i-1} T_{k-1}(x) - \sum_{\ell=1}^{n-k+1-i} \sum_{r=1}^{k-1+\ell} 2A_{k-1+\ell-r} T_{|n-i+1-\ell-r|}(x); \tag{24}$$

- *if $i \le n-k$ and $k = 1$*

$$Q_{ik}(x) = \sum_{\ell=0}^{i-2} \Gamma_\ell T_{i-1-\ell}(x) + \frac{\Gamma_{i-1}}{2} T_0(x) - \sum_{\ell=1}^{n-i} \sum_{r=1}^{\ell} A_{\ell-r} T_{|n-i+1-\ell-r|}(x); \tag{25}$$

where Γ_ℓ, for $\ell = 0, 1, 2, \ldots$, is defined in (21).

From Theorem 4.1, it is clear that there exists a unique $n \times n$ block matrix $M_k = [(M_k)_{ij}]$, with $(M_k)_{ij} \in \mathbb{R}^{p \times p}$, such that

$$\begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes H_{n-k}(x) = M_k \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes I_p + \begin{bmatrix} r_{1k}(x) \\ \vdots \\ r_{n-1,k}(x) \\ r_{nk}(x) \end{bmatrix} \otimes P(x), \quad \text{for } k = 1, 2, \ldots, n, \tag{26}$$

where $\otimes$ denotes the Kronecker product, for some scalar polynomials $r_{1k}(x), \ldots, r_{nk}(x)$. We will refer to the matrix M_k in (26) as the kth *Clenshaw matrix* of $P(x)$.

By direct multiplication, it may be easily checked that the block colleague matrix C_T satisfies

$$x \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes I_p = C_T \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes I_p + e_1 \otimes P(x), \tag{27}$$

Then, using (26) and (27), in Proposition 4.2 we show that the Clenshaw matrices $M_1, M_2, \ldots, M_n$ in (26) satisfy a simple recurrence relation. This recurrence relation is the analogue of (13) when the matrix polynomial $P(x)$ is expressed in the Chebyshev basis instead of the monomial basis.

Proposition 4.2. *Let $P(x) = I_p T_n(x) + \sum_{k=0}^{n-1} A_k T_k(x)$ be a $p \times p$ monic matrix polynomial in the Chebyshev basis of degree n, let C_T be the block colleague matrix of $P(x)$, and let $M_1, M_2, \ldots, M_n$ be the Clenshaw matrices in* (26)*. Then,*

$$\begin{aligned} &M_n = I_n \otimes 2I_p, \\ &M_{n-1} = 2M_n C_T + I_n \otimes 2A_{n-1}, \\ &M_k = 2M_{k+1} C_T - M_{k+2} + I_n \otimes 2A_k, \quad \textit{for } k = n-2, \ldots, 3, 2, \quad \textit{and} \\ &M_1 = M_2 C_T - M_3/2 + I_n \otimes A_1. \end{aligned} \tag{28}$$

Proof. The proof proceeds by induction on k (backwards from $k = n$). First, we prove that the result is true for $k = n$. From (19), we have

$$\begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes H_0(x) = \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes 2I_p = \begin{bmatrix} 2I_p & & & \\ & \ddots & & \\ & & 2I_p & \\ & & & 2I_p \end{bmatrix} \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes I_p.$$

Comparing the previous equation with (26), we deduce that $M_n = I_n \otimes 2I_p$.

Second, we prove that the result is true for $k = n-1$. From (19), we have

$$\begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes H_1(x) = \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes (2xH_0(x) + 2A_{n-1}) = 2x \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes H_0(x) + \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes 2A_{n-1}.$$

Using the inductive hypothesis, together with (27), we get

$$\begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes H_1(x) = (2M_nC_T + I_n \otimes 2A_{n-1}) \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} + \begin{bmatrix} r_{1,n-1}(x) \\ \vdots \\ r_{n-1,n-1}(x) \\ r_{n,n-1}(x) \end{bmatrix} \otimes P(x),$$

for some scalar polynomials $r_{1,n-1}(x), \ldots, r_{n,n-1}(x)$. Comparing the previous equation with (26), we deduce that $M_{n-1} = 2M_nC_T + I_n \otimes 2A_{n-1}$.

Third, assume that the result is true for $M_n, M_{n-1}, \ldots, M_{k+1}$, with $n-2 \geq k \geq 2$. From (19), we have

$$\begin{aligned} \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes H_{n-k}(x) &= \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes (2xH_{n-k-1}(x) - H_{n-k-2}(x) + 2A_k) \\ &= 2x \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes H_{n-k-1}(x) - \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes H_{n-k-2}(x) + \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes 2A_k. \end{aligned}$$

Using that the result is true for $k+1$ and $k+2$, together with (27), we get

$$\begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} \otimes H_{n-k}(x) = (2C_TM_{k+1} - M_{k+2} + I_n \otimes 2A_k) \begin{bmatrix} T_{n-1}(x) \\ \vdots \\ T_1(x) \\ T_0(x) \end{bmatrix} + \begin{bmatrix} r_{1k}(x) \\ \vdots \\ r_{n-1,k}(x) \\ r_{n,k}(x) \end{bmatrix} \otimes P(x),$$

for some scalar polynomials $r_{1k}(x), \ldots, r_{nk}(x)$. Comparing the previous equation with (26), we deduce that $M_k = 2C_TM_{k+1} - M_{k+2} + I_n \otimes 2A_k$.

Finally, the proof of the last case ($k = 1$) is similar to the proof for the previous cases ($n-2 \geq k \geq 2$), but using $H_{n-1}(x) = xH_{n-2}(x) - H_{n-3}(x)/2 + A_1$, so we omit it. $\square$

In contrast with (8), the Clenshaw matrices $M_1, M_2, \ldots, M_n$ have a complicated structure. For example, for $n = 6$ and $k = 3$, it is easy to check using (28) that the matrix M_k is equal to

$$\begin{bmatrix} 0 & -2A_2 & -2A_3 - 2A_1 & 2I_p - 2A_4 - 2A_2 - 4A_0 & -2A_3 - 4A_1 & -2A_2 - 2A_0 \\ 0 & 0 & 2I_p - 2A_2 & 2A_5 - 2A_3 - 2A_1 & 2I_p - 2A_2 - 4A_0 & -2A_1 \\ 0 & 2I_p & 2A_5 & 2I_p + 2A_4 - 2A_2 & 2A_5 - 2A_1 & 2I_p - 2A_0 \\ 2I_p & 2A_5 & 2I_p + 2A_4 & 2A_5 + 2A_3 & 4I_p + 2A_4 & 2A_5 \\ 0 & 2I_p & 2A_5 & 4I_p + 2A_4 & 4A_5 + 2A_3 & 2A_4 + 2I_p \\ 0 & 0 & 4I_p & 4A_5 & 4I_p + 4A_4 & 2A_5 + 2A_3 \end{bmatrix}. \tag{29}$$

Two observations about the matrix (29) are: (i) its first block column is equal to $e_{n-k+1} \otimes 2I_p$, where e_ℓ denotes the ℓth column of the $n \times n$ identity matrix; and, (ii) if we set $A_n := I_p$, each block entry has the form $\sum_{i=0}^{n} c_iA_i$, where $|c_i| \leq 4$. In Theorem 4.3, we show that the two previous observations are true for any n and k. Property (i) will be key to prove Arnold transversality theorem, and property (ii) will be key to study what kind of backward stability of a linearization-based algorithm for the polynomial eigenvalue problem is provided by the backward stability of an eigensolver for the linearized problem.

Theorem 4.3. *Let $P(x) = I_pT_n(x) + \sum_{k=0}^{n-1} A_kT_k(x)$ be a $p \times p$ monic matrix polynomial in the Chebyshev basis of degree n, and let M_k, for $k = 1, 2, \ldots, n$, be the kth Clenshaw matrix in (26). Then, the following statements hold:*

(a) *The first block column of M_k is equal to $e_{n-k+1} \otimes 2I_p$, where e_ℓ denotes the ℓth column of the $n \times n$ identity matrix.*

(b) *For $i, j = 1, 2, \ldots, n$, the (i, j)th block entry of M_k satisfies $(M_k)_{ij} = \sum_{t=0}^{n} \alpha_{t,ijk}A_t$ with $|\alpha_{t,ijk}| \leq 4$, where we set $A_n := I_p$.*

Proof. From Theorem 4.1 together with (26), we have $H_{n-k}(x)T_{n-i}(x) = \sum_{j=1}^{n}(M_k)_{ij}T_{n-j}(x)+r_{ik}(x)P(x)$. Therefore, to prove part (a) it is enough to show that

$$T_{n-i}(x)H_{n-k}(x) = 2I_pT_{n-1}(x) + \cdots + r_{ik}(x)P(x), \tag{30}$$

if $i = n - k + 1$, and that

$$T_{n-i}(x)H_{n-k}(x) = (M_k)_{i\nu}T_\nu(x) + \cdots + r_{ik}(x)P(x), \tag{31}$$

with $\nu < n - 1$, if $i \neq n - k + 1$, where the dots correspond to Chebyshev polynomials with lower indices.

First, suppose that $i \geq n-k+1$. We will prove that $T_{n-i}(x)H_{n-k}(x)$ is of the form (30) when $i = n-k+1$ and it is of the form (31) otherwise. We need to distinguish several cases. First, let $k = n$. From (22) we get that $T_{n-i}(x)H_0(x) = \Gamma_0T_{n-i}(x) = 2I_pT_{n-i}(x)$. Since the index $n - i$ is equal to $n - 1$ if and only if $i = 1$, the result is true in this case. Then, consider the case $n - 1 \geq k \geq 2$. There are three kinds of indices of Chebyshev polynomials in (22). The first is $2n - i - k - \ell$, which is equal to $n - 1$ if and only if $\ell = 0$ and $i = n - k + 1$. This gives a contribution $\Gamma_0T_{n-1}(x) = 2I_pT_{n-1}(x)$ only when $i = n - k + 1$. The second one is $|k + \ell - i|$. Taking into account the possible values that k, ℓ, and i can take in (22), it may be easily checked that this index is smaller than or equal to $n - 2$. The third index is $n - i$ which necessarily is smaller than or equal to $n - 2$, and, hence, the result is true in this case. Finally, consider the case $k = 1$ and $i = n$. There are two kinds on indices of Chebyshev polynomials in (23). The first one is $n - 1 - \ell$, which is equal to $n - 1$ if and only if $\ell = 0$. This gives a contribution $\Gamma_0T_{n-1}(x) = 2I_pT_{n-1}(x)$. The second index is 0, which is smaller than $n - 2$. Therefore, the result is also true in this case.

Now suppose that $i \leq n - k$. We will we prove that $T_{n-i}(x)H_{n-k}(x)$ is of the form (31). Notice that there are four kinds of indices in (24) when $k \geq 2$, namely, $i + k - 2 - \ell$, $|k + \ell - i|$, $|n - i + 1 - \ell - r|$ and $k - 1$, and three kinds on indices in (25) when $k = 1$, namely, $i - 1 - \ell$, $i - 1$ and $|n - i + 1 - \ell - r|$. Taking into account the possible values that k, ℓ, r, and i can take in (31), in both cases ($k \geq 2$ and $k = 2$), it may be checked that these indices do not exceed $n - 2$.

Now, we proceed to prove part (b). Again, we need to distinguish several cases. First, suppose that $i \geq n - k + 1$ and also assume that $k \geq 2$ (the argument when $k = 1$ is similar and simpler, so we omit it), and consider the three kinds of indices of Chebyshev polynomials that appear in (22), namely, $2n - i - k - \ell$, $|k + \ell - i|$, and $n - i$. For $\ell = 0, 1, \ldots, n - k$, a careful look at these indices reveals that if $k + \ell - i \geq 0$, then the three of them are different. Therefore, we can write (22) as

$$\sum_{\ell=0}^{n-1} B_\ell T_\ell(x) + \sum_{\ell=-(1-n)}^{-1} B_\ell T_{-\ell}(x), \tag{32}$$

where B_ℓ is equal to either 0 or Γ_t for some t. It follows that $(M_k)_{ij}$ is equal to either 0, Γ_t for some t, or $\Gamma_{t_1} + \Gamma_{t_2}$ for some t_1, t_2. Finally, recall from (21) that Γ_t is equal to $2I_p + 2A_{n-2} + 2A_{n-4} + \cdots$ if t is even, or to $2A_{n-1} + 2A_{n-3} + \cdots$ if t is odd. Therefore, $(M_k)_{ij} = \sum_{t=0}^{n} \alpha_{t,ijk}A_t$, with $|\alpha_{t,ijk}| \leq 4$.

Then suppose that $i \leq n - k$ and also assume that $k \geq 2$ (again, the argument when $k = 1$ is similar and simpler, so we omit it). First, consider the three kinds of indices of Chebyshev polynomials that appear in the first summand in (24), namely, $i + k - 2 - \ell$, $|k + \ell - i|$, $k - 1$. For $\ell = 0, 1, \ldots, i - 2$, again, it may be checked that if $k + \ell - i \geq 0$, then these three indices are different. Therefore, the first summand in (24) is also of the form (32). Finally, consider the index of the Chebyshev polynomials and the index of the coefficients A_i that appear in the second summand in (24), namely, $|n - i + 1 - \ell - r|$, and $k + 1 + \ell - r$. If

$n-i+1-\ell-r \geq 0$, it may be checked that for any two allowed different pairs (ℓ, r) that realize the same value of $n-i+1-\ell-r$, then the associate indices $k+1+\ell-r$ must be different. Since the same occur when $n-i+1-\ell-r<0$, it follows that (24) is of the form

$$\sum_{\ell=0}^{n-2} C_\ell T_\ell(x) + \sum_{\ell=-(2-n)}^{-1} C_\ell T_{-\ell}(x) - 2\sum_{\ell=0}^{n-2} D_\ell T_\ell(x) - 2\sum_{\ell=2-n}^{-1} D_\ell T_{-\ell}(x) + r_{ik}(x)P(x)$$

where C_ℓ is equal to either 0 or Γ_t for some t, and D_ℓ is equal to $\sum_{t=1}^{q_\ell} A_{i_t}$, where $i_{t_1} \neq i_{t_2}$ whenever $t_1 \neq t_2$. Then, it follows that

$$(M_k)_{ij} = \sum_{\ell=0}^{n} \delta_\ell A_\ell - \sum_{\ell=0}^{n} \rho_\ell A_\ell,$$

where δ_ℓ and ρ_ℓ are equal to either 4, or 2 or 0, therefore $(M_k)_{ij} = \sum_{t=0}^{n} \alpha_{t,ijk} A_t$ with $|\alpha_{t,ijk}| \leq 4$. $\square$

If necessary, explicit expressions of the entries of the Clenshaw matrices $M_1, M_2, \ldots, M_n$ may be obtained from Theorem 4.1. However, since Theorem 4.3 is the only information that we will need about them to prove our main results in the following section, we do not pursue that idea.

4.2 Backward error of the Chebyshev rootfinding method using colleague matrices

In this section we prove Arnold transversality theorem for colleague matrices of monic polynomials in the Chebyshev basis. That is, we show that any matrix $E \in \mathbb{R}^{pn \times pn}$ may be decomposed as

$$E = F_0 + T, \tag{33}$$

where $F_0 \in \mathcal{BFR}_{n,p}$ is a first block row matrix and $T \in T_{C_T}\mathcal{O}(C_T)$, constructing the matrix F_0 explicitly.

As in the case of the monomial basis, generically, $\dim(T_{C_T}\mathcal{O}(C_T)) + \dim(\mathcal{BFR}_{n,p}) = n^2p^2 - np + np^2 \geq n^2p^2$ (see [8, 15]). In words, the tangent space $T_{C_T}\mathcal{O}(C_T)$ and the vector space of first block row matrices $\mathcal{BFR}_{n,p}$ may have a nontrivial intersection when $p > 1$. For this reason, following [8, 15] we choose a particular subspace of the tangent space that will give a unique decomposition (33). This subspace is denoted by Sub $T_{C_T}\mathcal{O}(C_T)$ and it is given by

$$\text{Sub } T_{C_T}\mathcal{O}(C_T) = \{X \in T_{C_T}\mathcal{O}(C_T) \quad \text{such that} \quad X \text{ has 0 first block column}\}.$$

In order to get the decomposition (33) with $T \in \text{Sub } T_{C_T}\mathcal{O}(C_T)$ we will make use of the Clenshaw matrices $M_1, M_2, \ldots, M_n \in \mathbb{R}^{np \times np}$, defined in (26), of the matrix polynomial $P(x)$ in (3). Recall that the Clenshaw matrices satisfy the following recurrence relation (see Proposition 4.2):

$$\begin{aligned}
&M_n = I_n \otimes 2I_p, \\
&M_{n-1} = 2M_n C_T + I_n \otimes 2A_{n-1}, \\
&M_k = 2M_{k-1}C_T - M_{k-2} + I_n \otimes 2A_k, \quad \text{for } k = n-2, \ldots, 3, 2, \quad \text{and} \\
&M_1 = M_2 C_T - M_3/2 + I_n \otimes A_1.
\end{aligned} \tag{34}$$

We illustrate these matrices with a small example ($n = 4$). From (34), it is easy to check that

$$M_4 = \begin{bmatrix} 2I_n & 0 & 0 & 0 \\ 0 & 2I_n & 0 & 0 \\ 0 & 0 & 2I_n & 0 \\ 0 & 0 & 0 & 2I_n \end{bmatrix}, \quad M_3 = \begin{bmatrix} 0 & 2I_n - 2A_2 & -2A_1 & -2A_0 \\ 2I_n & 2A_3 & 2I_n & 0 \\ 0 & 2I_n & 2A_3 & 2I_n \\ 0 & 0 & 4I_n & 2A_3 \end{bmatrix},$$

$$M_2 = \begin{bmatrix} 0 & -2A_1 & 2I_n - 2A_2 - 4A_0 & -2A_1 \\ 0 & 2I_n & 2A_3 - 2A_1 & 2I_n - 2A_0 \\ 2I_n & 2A_3 & 4I_n + 2A_2 & 2A_3 \\ 0 & 4I_n & 4A_3 & 2I_n + 2A_2 \end{bmatrix}, \quad \text{and}$$

$$M_1 = \begin{bmatrix} 0 & -2A_0 & -2A_1 & I_n - A_2 - A_0 \\ 0 & 0 & 2I_n - 2A_0 & A_3 - A_1 \\ 0 & 2I_n & 2A_3 & I_n + A_2 - A_0 \\ 2I_n & 2A_3 & 2I_n + 2A_2 & A_3 + A_1 \end{bmatrix}.$$

The only information that we need about Clenshaw matrices are those stated in Theorem 4.3.

Following [8], we also define the *block trace* of a $np \times np$ block matrix $Z = [Z_{ij}]$, with $Z_{ij} \in \mathbb{R}^{p\times p}$, as the $p \times p$ matrix

$$\operatorname{tr}_p(Z) := \sum_{i=1}^{n} Z_{ii}.$$

The block trace is used in Theorem 4.4, which provides a characterization of the subspace $\operatorname{Sub} T_{C_T}\mathcal{O}(C_T)$, and is a generalization of [8, Theorem 4.1] when the matrix polynomial $P(x)$ is expressed in the Chebyshev basis.

Theorem 4.4. *For any $Z \in \mathbb{R}^{pn\times pn}$,*

$$\operatorname{tr}_p(M_{k+1}Z) = 0, \quad \text{for } k = 0, 1, \ldots, n-1, \tag{35}$$

if and only if

$$Z = C_T X - X C_T \quad \text{for some } X \in \mathbb{R}^{np\times np} \text{ with 0 first block column.} \tag{36}$$

Moreover, either condition determines the first block row of Z uniquely given the remaining block rows.

Proof. From part (a) in Theorem 4.3, the $(n-k, 1)$ block entry of M_{k+1} is equal to $2I_p$, and the $(i, 1)$ block entry of M_{k+1}, with $i \neq n-k$, is equal to 0. Therefore, $Z_{1,n-k}$, for $k = 0, 1, \ldots, n-1$, is uniquely determined from (35). Also, if X has 0 first block column, it may be easily checked that the map from X to the last $n-1$ block rows of $C_T X - X C_T$ has a trivial nullspace. Thus, Z is uniquely determined by (36).

To finish the proof we need to prove that (36) implies (35). That is, we need to show that $\operatorname{tr}_p(M_{k+1}(C_T X - X C_T)) = 0$ for any block matrix X with 0 first block column. In order to do this, first, we show that if X has 0 first block column, then $\operatorname{tr}_p(M_{k+1} X C_T) = \operatorname{tr}_p(X C_T M_{k+1})$. The proof of the previous equation is not completely immediate when $p > 1$ since, in this situation, $\operatorname{tr}_p(AB) = \operatorname{tr}_p(BA)$ does not hold in general. So, consider a block matrix Y that has 0 first block column. Then,

$$\operatorname{tr}_p(C_T Y) = \sum_{i=1}^{p-2} \left(\frac{Y_{i,i+1}}{2} + \frac{Y_{i+2,i+1}}{2} \right) + Y_{p-1,p} = \operatorname{tr}_p(Y C_T).$$

Therefore, if X has 0 first block column, then $\operatorname{tr}_p(M_{k+1} X C_T) = \operatorname{tr}_p(C_T M_{k+1} X)$.

Then, we show that $\operatorname{tr}_p(C_T M_{k+1} X) = \operatorname{tr}_p(M_{k+1} C_T X)$. To do this, note that the Clenshaw matrix M_{k+1} is of the form $2^{n-k} C_T^{n-k-1} + \sum_{t=1}^{n-k-1} (I_n \otimes B_k) C_T^{n-k-1-t}$, for some $B_1, B_2, \ldots, B_{n-k-1} \in \mathbb{R}^{p\times p}$ (this can be verified by induction using (34)). So, we only need to show that $\operatorname{tr}_p(C_T(I_n \otimes B) C_T^j X) = \operatorname{tr}_p((I \otimes B) C_T^j C_T X)$. Indeed, since the matrix $C_T(I_n \otimes B) - (I_n \otimes B)C_T$ is 0 except the first block row, and since $C_T X$ has 0 first block column, it follows that $\operatorname{tr}_p(C_T(I_n \otimes B) C_T^j X - (I \otimes B) C_T^j C_T X) = 0$. Therefore, $\operatorname{tr}_p(C_T M_{k+1} X) = \operatorname{tr}_p(M_{k+1} C_T X)$. Thus, we conclude that $\operatorname{tr}_p(M_{k+1} X C_T) = \operatorname{tr}_p(C_T M_{k+1} X) = \operatorname{tr}_p(M_{k+1} C_T X)$. □

In Theorem 4.5 we present the proof of Arnold transversality theorem for block colleague matrices. Part (a) in Theorem 4.3 will be key here.

Theorem 4.5. *Let $p(x) = I_p T_n(x) + \sum_{k=0}^{n-1} A_k T_k(x)$ be a $p \times p$ monic matrix polynomial in the Chebyshev basis of degree n, and let C_T be its block colleague matrix. Then, any matrix $E \in \mathbb{R}^{np\times np}$ can be expressed as*

$$E = F_0 + T, \tag{37}$$

where $F_0 \in \mathcal{BFR}_{n,p}$ is a first block row matrix, and $T \in \text{Sub } T_{C_T}\mathcal{O}(C_T)$. Moreover, if the first block row of F_0 is written as $\begin{bmatrix} F_0^{(n-1)} & \cdots & F_0^{(1)} & F_0^{(0)} \end{bmatrix}$, then

$$F_0^{(k)} = \frac{1}{2}\text{tr}_p\,(EM_{k+1}), \quad \textit{for } k = 0, 1, \ldots, n-1, \tag{38}$$

where the matrix M_{k+1} is Clenshaw matrix defind in (26)*.*

Proof. Define $F_0^{(k)} = \frac{1}{2}\text{tr}_p\,(EM_{k+1})$, for $k = 0, 1, \ldots, n-1$, and let $F_0 \in \mathcal{BFR}_{n,p}$ be a first block row matrix such that its first block row is $\begin{bmatrix} F_0^{(n-1)} & \cdots & F_0^{(1)} & F_0^{(0)} \end{bmatrix}$. We may write the matrix $T := E - F_0$. Then, we have to check that $T \in \text{Sub } T_{C_T}\mathcal{O}(C_T)$. From Theorem 4.4, we see that it is sufficient to show that $\text{tr}_p\,(TM_{k+1}) = 0$, for $k = 0, 1, \ldots, n-1$. Indeed, using part (a) in Theorem 4.3,

$$\text{tr}_p\,(TM_{k+1}) = \text{tr}_p\,(EM_{k+1}) - \text{tr}_p\,(F_0M_{k+1}) = \text{tr}_p\,(EM_{k+1}) - 2F_0^{(k)} = \text{tr}_p\,(EM_{k+1}) - \text{tr}_p\,(EM_{k+1}) = 0,$$

for $k = 0, 1, \ldots, n-1$. So, we conclude that $T \in \text{Sub } T_{C_T}\mathcal{O}(C_T)$. □

An important consequence of the decomposition in Theorem 4.5 is that if E is a small perturbation of the block colleague matrix C_T, then

$$C_T + E = C_T + F_0 + T = C_T + F_0 + (C_TX - XC_T) = (I + X)^{-1}(C_T + F_0)(I + X) + O(\|E\|_2^2),$$

where we have used that T can be written as $C_TX - XC_T$, for some $X \in \mathbb{R}^{np\times np}$ with 0 first block column. Noticing that $C_T + F_0$ is in turn a block colleague matrix of another matrix polynomial, we deduce that a small perturbation of the block colleague matrix of $P(x)$ is similar, to first order in the norm of the perturbation, to a block colleague matrix of a perturbed polynomial $\widetilde{P}(x)$. This observation allows us to formulate the next corollary.

Corollary 4.6. *Let $P(x) = I_p T_n(x) + \sum_{k=0}^{n-1} A_k T_k(x)$ be a $p \times p$ monic matrix polynomial in the Chebyshev basis of degree n, and let C_T be its block colleague matrix. Assume that the eigenvalues of $P(x)$ are computed as the eigenvalues of C_T with a backward stable algorithm, i.e., an algorithm that computes the exact eigenvalues of some matrix $C_T + E$, with $\|E\|_F = O(u)\|C_T\|_F$, where u is the machine precision. Then, to first order in u, the computed roots are the exact roots of a polynomial $\widetilde{P}(x)$ such that*

$$\frac{\|\widetilde{P} - P\|_F}{\|P\|_F} = O(u)\|P\|_F.$$

Proof. If a backward stable eigensolver is given C_T as an input, the computed eigenvalues are the exact eigenvalues of a matrix $C_T + E$, for some E with $\|E\|_F = \epsilon\|C_T\|_F$, where $\epsilon = uh(n)$, for some low degree polynomial h with moderate coefficients. In other words, the computed eigenvalues are the exact roots of the polynomial $\det(xI - C_T - E)$.

Using Theorem 4.5, we can write $E = F_0 + T$, where $T \in \text{Sub } T_{C_T}\mathcal{O}(C_T)$ and F_0 is a first block row matrix with first block row as in (38). Therefore,

$$C_T + E = C_T + F_0 + C_TX - XC_T = (I + X)^{-1}(C_T + F_0)(I + X) + O(u^2),$$

for some matrix $X \in \mathbb{R}^{np\times np}$ with 0 first block column, that is, the matrix $C_T + E$ is similar, to first order in u, to the colleague matrix $C_T + F_0$. Writing E as a $np \times np$ block matrix $E = [E_{ij}]$, with $E_{ij} \in \mathbb{R}^{p\times p}$, and noticing that $C_T + F_0$ is the colleague matrix of the matrix polynomial $\widetilde{P}(x) = I_pT_n(x) + \sum_{k=0}^{n-1}(A_k -$

$F_0^{(k)})T_k(x)$, we have that, to first order in u, the computed eigenvalues are the exact eigenvalues of a matrix polynomial $\widetilde{P}(x) = I_pT_n(x) + \sum_{k=0}^{n-1} \widetilde{A}_kT_k(x)$, with $\|\widetilde{A}_k - A_k\|_F = \|F_0^{(k)}\|_F = \|\mathrm{tr}_p\,(EM_{k+1})\|_F = \|\sum_{i,j=1}^n E_{ij}(M_{k+1})_{ji}\|_F$. Therefore, for $k = 0, 1, \ldots, n-1$, we have

$$\|\widetilde{A}_k - A_k\|_F \leq \sum_{i,j=1}^n \|E_{ij}\|_F\|(M_{k+1})_{ji}\|_F \leq \sqrt{\sum_{i,j=1}^n \|E_{ij}\|_F^2}\sqrt{\sum_{i,j=1}^n \|(M_{k+1})_{ij}\|_F^2} = \|E\|_F\|M_{k+1}\|_F,$$

where we have used the Cauchy-Schwarz inequality. Then, using part (b) of Theorem 4.3, we have

$$\begin{aligned}\|M_{k+1}\|_F &= \sqrt{\sum_{i,j=1}^n \|(M_{k+1})_{ij}\|_F^2} = \sqrt{\sum_{i,j=1}^n \|\sum_{t=0}^n \alpha_{t,ij,k+1}A_t\|_F^2} \leq \sqrt{\sum_{i,j=1}^n \left(\sum_{t=0}^n \|\alpha_{t,ij,k+1}A_t\|_F\right)^2} \\ &\leq 4\sqrt{\sum_{i,j=1}^n \left(\sum_{t=0}^n \|A_t\|_F\right)^2} \leq 4n^{3/2}\|P\|_F,\end{aligned}$$

where we have used $\sum_{t=0}^n \|A_t\|_2 \leq \sqrt{n}\|P\|_2$. Finally, using that $\|E\|_F = \epsilon\|C_T\|_F$, we get that

$$\begin{aligned}\|E\|_F = \epsilon\|C_T\|_F &\leq \frac{\epsilon}{2}\left\|\begin{bmatrix} -A_{n-1} & -A_{n-1} & \cdots & -A_0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}\right\|_F + \frac{\epsilon}{2}\left\|\begin{bmatrix} & I_p & & & \\ I_p & & I_p & & \\ & \ddots & & \ddots & \\ & & I_p & 0 & I_p \\ & & & 2I_p & \end{bmatrix}\right\|_F \\ &\leq \frac{\epsilon}{2}\left(\sqrt{\sum_{t=0}^{n-1} \|A_t\|_F^2} + \sqrt{2np}\right) \leq \epsilon\frac{\sqrt{2n}}{2}\left(\sqrt{\sum_{t=0}^{n-1} \|A_t\|_F^2} + \sqrt{p}\right) \leq \epsilon\sqrt{2n}\|P\|_F.\end{aligned}$$

Thus, the computed eigenvalues, to first order in u, are the exact eigenvalues of a monic matrix polynomial in the Chebyshev basis $\widetilde{P}(x)$ such that,

$$\|\widetilde{P} - P\|_F = \sqrt{\sum_{k=0}^{n-1} \|\widetilde{A}_k - A_k\|_F^2} \leq \sum_{k=0}^{n-1} \|\widetilde{A}_k - A_k\|_F \leq \sum_{k=0}^{n-1} \|M_{k+1}\|_F\|E\|_F \leq 4n^{5/2}\|P\|_F\|E\|_F \leq \widetilde{\epsilon}\|P\|_F^2,$$

where $\tilde{\epsilon} = u\widehat{h}(n)$, for some low degree polynomial $\widehat{h}$ with moderate coefficients. □

5 Conclusions

In this paper, we have analyzed the backward stability of a Chebyshev-basis polynomial rootfinder (or matrix polynomial eigensolver) based on the solution of the standard eigenvalue problem for the corresponding colleague matrix. More precisely, given a monic scalar polynomial in the Chebyshev basis $p(x)$, we have proved that if the roots of $p(x)$ are computed as the eigenvalues of a colleague matrix using a backward stable eigenvalue algorithm, like the QR algorithm, then the computed roots are the exact roots of a monic polynomial in the Chebyshev basis $\widetilde{p}(x)$ such that

$$\frac{\|\widetilde{p} - p\|_2}{\|p\|_2} = O(u)\|p\|_2,$$

Similarly, if the eigenvalues of a monic matrix polynomial in the Chebyshev basis are computed as the eigenvalues of a block colleague matrix using a backward stable eigenvalue algorithm, then the computed eigenvalues are the exact eigenvalues of a monic matrix polynomial in the Chebyshev basis $\widetilde{P}(x)$ such that

$$\frac{\|\widetilde{P} - P\|_F}{\|P\|_F} = O(u)\|P\|_F,$$

These backward error analysis show that these methods are backward stable when the norms $\|p\|_2$ and $\|P\|_F$ are moderate.

A Proof of Theorem 4.1

In this section we present the proof of Theorem 4.1, that is, given the Clenshaw shift $H_{n-k}(x)$ associated with the matrix polynomial $P(x)$ in (3), and the Chebyshev polynomial $T_{n-i}(x)$, we show that

$$T_{n-i}(x)H_{n-k}(x) = Q_{ik}(x) + r_{ik}(x)P(x), \tag{39}$$

for some scalar polynomial $r_{ik}(x)$, where $Q_{ik}(x)$ is the matrix polynomial of degree less than or equal to $n-1$ in (22)–(25). Moreover, we show that the decomposition (39) is unique.

Along the proof, quite often products of two of Chebyshev polynomials will occur. For this reason, the following formula [1, Chapter 22] is of fundamental importance here:

$$2T_m(x)T_n(x) = T_{m+n}(x) + T_{|m-n|}(x). \tag{40}$$

The first step is to expand the Clenshaw shifts $H_k(x)$, for $k = 0, 1, \ldots, n-1$, in the Chebyshev basis. We will prove

$$H_k(x) = \sum_{\ell=0}^{k-1} 2\Gamma_\ell T_{k-\ell}(x) + \Gamma_k T_0(x), \quad \text{for } k = 0, 1, \ldots, n-2, \quad \text{and} \tag{41}$$

$$H_{n-1}(x) = \sum_{\ell=0}^{n-2} \Gamma_\ell T_{n-1-\ell}(x) + \frac{1}{2}\Gamma_{n-1}T_0(x), \tag{42}$$

where Γ_ℓ is defined in (21). The proof proceeds by induction on k. From (19) we get $H_0(x) = 2I_p = \Gamma_0 T_0(x)$ and $H_1(x) = 4I_p x + 2A_{n-1} = 2\Gamma_0 T_1(x) + \Gamma_1 T_0(x)$, so the result is true for $k = 0$ and $k = 1$. Then, assume that the result is true for $H_0(x), H_1(x), \ldots, H_{k-1}(x)$, with $2 \le k \le n-2$. Using the induction hypothesis, together with (19), we have

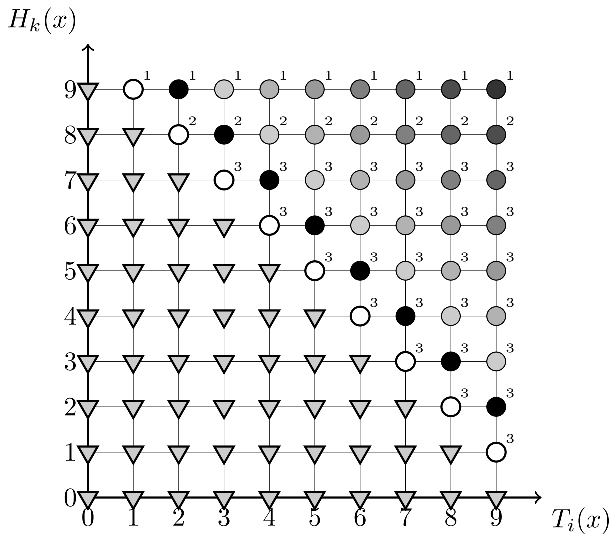
$$\begin{aligned}
H_k(x) &= 2xH_{k-1}(x) - H_{k-2}(x) + 2A_{n-k} \\
&= 2x\left(\sum_{\ell=0}^{k-2} 2\Gamma_l T_{k-1-\ell}(x) + \Gamma_{k-1}T_0(x)\right) - \sum_{\ell=0}^{k-3} 2\Gamma_\ell T_{k-2-\ell}(x) - \Gamma_{k-2}T_0(x) + 2A_{n-k}.
\end{aligned}$$

Using $T_0(x) = 1$, $T_1(x) = x$, and (40) with $m = 1$ and $n = k$, from the previous equation we get

$$\begin{aligned}
H_k(x) &= \sum_{\ell=0}^{k-2} 2\Gamma_\ell\left(T_{k-\ell}(x) + T_{k-2-\ell}(x)\right) + 2\Gamma_{k-1}T_1(x) - \sum_{\ell=0}^{k-3} 2\Gamma_\ell T_{k-2-\ell}(x) - \Gamma_{k-2}T_0(x) + 2A_{n-k}T_0(x) \\
&= \sum_{\ell=0}^{k-2} 2\Gamma_\ell T_{k-\ell}(x) + 2\Gamma_{k-2}T_0(x) + 2\Gamma_{k-1}T_1(x) - \Gamma_{k-2}T_0(x) + 2A_{n-k}T_0(x) \\
&= \sum_{\ell=0}^{k-1} 2\Gamma_\ell T_{k-\ell}(x) + (\Gamma_{k-2} + 2A_{n-k})T_0(x) = \sum_{\ell=0}^{k-1} 2\Gamma_\ell T_{k-\ell}(x) + \Gamma_k T_0(x),
\end{aligned}$$

where in the last equality we have used $\Gamma_{k-2} + 2A_{n-k} = \Gamma_k$. Therefore, the result is also true for $H_k(x)$. Finally, the proof that (42) holds is similar to the previous one, but starting with $H_{n-1}(x) = xH_{n-2}(x) - H_{n-3}(x)/2 + A_1$, so we omit the details.

Now we proceed to show that (39) holds with $Q_{ik}(x)$ as in (22)–(25). In order to do that, we will proceed in certain order. To help the reader to follow the steps, we depict all the possible products $T_{n-i}(x)H_{n-k}(x)$ for $n = 10$ in the following 10×10 grid.



The vertices with triangular shape in the previous grid represent the cases in which the degree of $T_{n-i}(x)H_{n-k}(x)$ does not exceed $n-1$, that is, when $i \geq n-k+1$. In this case, the polynomial $Q_{ik}(x)$ coincide with $T_{n-i}(x)H_{n-k}(x)$, so we just need to expand $T_{n-i}(x)H_{n-k}(x)$ in the Chebyshev basis. Indeed, when $i=n$ and $k=1$, from (42), we have

$$T_0(x)H_{n-1}(x) = H_{n-1}(x) = \sum_{\ell=0}^{n-2} \Gamma_\ell T_{n-1-\ell}(x) + \frac{1}{2}\Gamma_{n-1}T_0(x),$$

and when $n-1 \geq i \geq n-k+1$, from (40) and (41), we have

$$\begin{aligned} T_{n-i}(x)H_{n-k}(x) &= \sum_{\ell=0}^{n-k-1} 2\Gamma_\ell T_{n-i}(x)T_{n-k-\ell}(x) + \Gamma_{n-k}T_{n-i}(x)T_0(x) \\ &= \sum_{\ell=0}^{n-k-1} \Gamma_\ell(T_{2n-i-k-\ell}(x) + T_{|k+\ell-i|}(x)) + \Gamma_{n-k}T_{n-i}(x). \end{aligned}$$

As can be checked, the two previous equations correspond to (22) and (23), respectively.

Next, we consider the products $T_{n-i}(x)H_{n-k}(x)$ with $i < n-k+1$, represented in the grid by vertices with circular shape. This case is much more involved, since the degree of $T_{n-i}(x)H_{n-k}(x)$ is larger than or equal to n. We will prove that (39) holds, with $Q_{ik}(x)$ as in (22)–(25), each diagonal in the grid at a time (from left to right), showing that each product $T_{n-i}(x)H_{n-k}(x)$ can be computed using, at most, a product represented by a vertex in the same diagonal and two products represented by vertices in the diagonal on its left.

The first step is to consider the products $T_k(x)H_{n-k}(x)$, for $k=1,2,\ldots,n-1$, that is, products represented by the diagonal with white circular vertices in the grid. We show that Theorem 4.1 holds for those products from top to bottom. We start with the white circular vertex labeled with 1 in the grid, that is, with the product $T_1(x)H_{n-1}(x)$. From (20) and (41), together with $T_1(x)=x$, we have

$$T_1(x)H_{n-1}(x) = xH_{n-1}(x) = \frac{1}{2}H_{n-2}(x) - A_0T_0(x) + \cdots = \sum_{\ell=0}^{n-3} \Gamma_\ell T_{n-2-\ell}(x) + \frac{1}{2}\Gamma_{n-2}T_0(x) - A_0T_0(x) + \cdots,$$

where the dots correspond to something of the form $r(x)P(x)$, with $r(x)$ a scalar polynomial. As can be easily checked, the previous equation corresponds to (25) with $i=n-1$.

Then, we consider the white circular vertex labeled with 2 in the grid, that is, the product $T_2(x)H_{n-2}(x)$.

From (1) and (19), we have

$$\begin{aligned} H_{n-2}(x)T_2(x) &= H_{n-2}(x)\left(2xT_1(x)-T_0(x)\right) = 2xT_1(x)H_{n-2}(x) - T_0(x)H_{n-2}(x) \\ &= T_1(x)\left(2H_{n-1}(x)+H_{n-3}(x)-2A_1\right) - T_0(x)H_{n-2}(x) \\ &= 2T_1(x)H_{n-1}(x) + T_1(x)H_{n-3}(x) - T_0(x)H_{n-2}(x) - 2A_1T_1(x). \end{aligned}$$

As can be seen from the previous equation, the product $H_{n-2}(x)T_2(x)$ may be computed from products represented by two triangular vertices: $T_1(x)H_{n-3}(x)$ and $T_0(x)H_{n-2}(x)$, and the product $T_1(x)H_{n-1}(x)$. Then, using (41), (42), and the result previously obtained for $T_1(x)H_{n-1}(x)$, we get

$$T_2(x)H_{n-2}(x) = \sum_{\ell=0}^{n-4} \Gamma_\ell (T_{n-2-\ell}(x) + T_{|\ell+4-n|}(x)) + \Gamma_{n-3}T_1(x) - 2A_0T_0(x) - 2A_1T_1(x) + \cdots,$$

where the dots correspond to something of the form $r(x)P(x)$, with $r(x)$ a scalar polynomial. The previous equation corresponds to (24) with $i = n-2$ and $k = 2$.

Finally, we consider the white circular vertices labeled with 3, that is, the products $T_k(x)H_{n-k}(x)$, for $k = 3, 4, \ldots, n$. From (1) and (19), we have

$$\begin{aligned} T_k(x)H_{n-k}(x) &= (2xT_{k-1}(x) - T_{k-2}(x))H_{n-k}(x) = 2xT_{k-1}(x)H_{n-k}(x) - T_{k-2}(x)H_{n-k}(x) \\ &= T_{k-1}(x)(H_{n-k+1}(x) + H_{n-k-1}(x) - 2A_{k-1}) - T_{k-2}(x)H_{n-k}(x) \\ &= T_{k-1}(x)H_{n-k+1}(x) + T_{k-1}(x)H_{n-k-1}(x) - T_{k-2}(x)H_{n-k}(x) - 2A_{k-1}T_{k-1}(x). \end{aligned}$$

As can be seen from the previous equation, $T_k(x)H_{n-k}(x)$ may be computed from $T_{k-1}(x)H_{n-k-1}(x)$ and $T_{k-2}(x)H_{n-k}(x)$, represented in the grid by triangular vertices, and $T_{k-1}(x)H_{n-k+1}(x)$, represented in the grid by the white circular vertex above the white circular vertex corresponding to $T_k(x)H_{n-k}(x)$. Since we have previously seen that Theorem 4.1 holds for $T_1(x)H_{n-1}(x)$ and $T_2(x)H_{n-2}(x)$, and for products represented by triangular vertices, this shows how to prove inductively (from top to bottom) that Theorem 4.1 holds for products represented by white circular vertices labeled with 3. Indeed, assuming that the result holds for $T_{k-1}(x)H_{n-k+1}(x)$ and using (22), we get

$$\begin{aligned} T_k(x)H_{n-k}(x) &= T_{k-1}(x)H_{n-k-1} + \sum_{r=2}^{k}(-2A_{k-r})T_{k-r}(x) - 2A_{k-1}T_{k-1}(x) + \cdots \\ &= \sum_{\ell=0}^{n-k-2} \Gamma_\ell (T_{n-2-\ell}(x) + T_{|\ell+2k-n|}) + \Gamma_{n-k-1}T_{k-1}(x) + \sum_{r=1}^{k}(-2A_{k-r})T_{k-r}(x) + \cdots, \end{aligned}$$

where the dots correspond to something of the form $r(x)P(x)$, with $r(x)$ a scalar polynomial. It is immediate to check that the previous equation corresponds to (24) when $i = n-k$.

The second step is to consider the products $T_{k+1}(x)H_{n-k}(x)$, for $k = 2, 3, \ldots, n-2$, that is, the diagonal with black circular vertices in the grid. This step is very similar to the previous one, so we will only sketch the main ideas. We have to distinguish the cases $k = 2$, $k = 3$ and $k > 3$. When $k = 2$, using (1), (19) and (20), it may be proved

$$T_2(x)H_{n-1}(x) = T_1(x)H_{n-2}(x) - T_0(x)H_{n-1}(x) - 2A_0T_1(x) + \cdots,$$

where the dots correspond to something of the form $r(x)P(x)$, with $r(x)$ a scalar polynomial. The previous equation shows that $T_2(x)H_{n-1}(x)$ may be computed from two products represented by triangular vertices in the grid: $T_1(x)H_{n-2}(x)$ and $T_0(x)H_{n-1}(x)$. Since we have seen that Theorem 4.1 holds for products represented by triangular vertices, it may be proved that (25) holds for $T_2(x)H_{n-1}(x)$.

Then, from (1) and (19), it may be proved that, when $k = 2$,

$$T_3(x)H_{n-2}(x) = 2T_2(x)H_{n-1}(x) + T_2(x)H_{n-3}(x) - 2A_1T_2(x),$$

and, when $k > 3$,

$$T_{k+1}(x)H_{n-k}(x) = T_k(x)H_{n-k+1}(x) + T_k(x)H_{n-k-1}(x) - T_{k-1}(x)H_{n-k}(x) - 2A_{k-1}T_k(x).$$

These two equations show that $T_{k+1}(x)H_{n-k}(x)$ may be computed from two products represented by triangular vertices, and the product represented by the black circular vertex above the black circular vertex corresponding to $T_{k+1}(x)H_{n-k}(x)$. Assuming that Theorem 4.1 holds for $T_2(x)H_{n-1}(x)$, the previous observation shows how to prove inductively (from top to bottom) that Theorem 4.1 holds for products corresponding to black circular vertices labeled with 2 and 3.

Now, we address the products represented by circular vertices colored with different shades of grey, that is, the products $T_{k+r-1}(x)H_{n-k}(x)$, for $r = 3, 4, \dots, n-2$ and $k = 1, 2, \dots, n-1-r$. We will show that Theorem 4.1 holds for products represented by vertices in the same diagonal (same shade of grey) assuming that it holds for products represented by (non-triangular) vertices in the diagonal on its left. Since we have previously proved that Theorem 4.1 holds for products represented by the white and black diagonals, this will imply that Theorem 4.1 holds for all products represented by grey vertices. For each grey diagonal, we have to distinguish the products represented by vertices labeled with 1, 2, and 3.

First, we consider the product $T_r(x)H_{n-1}(x)$, with $r \geq 3$, represented by a grey vertex labeled with 1. From (1) and (20), we get

$$\begin{aligned}T_r(x)H_{n-1}(x) &= (2xT_{r-1}(x) - T_{r-2}(x))H_{n-1}(x) = 2xT_{r-1}(x)H_{n-1}(x) - T_{r-2}(x)H_{n-1}(x) \\ &= T_{r-1}(x)H_{n-2}(x) - T_{r-2}(x)H_{n-1}(x) - 2A_0T_{r-1}(x) + \cdots ,\end{aligned}$$

where the dots correspond to something of the form $r(x)P(x)$, with $r(x)$ a scalar polynomial. The previous equation shows that $T_r(x)H_{n-1}(x)$ may be computed from two products represented by vertices in the diagonal on its left: $T_{r-1}(x)H_{n-2}(x)$ and $T_{r-2}(x)H_{n-1}(x)$. Assuming that (24) and (25) hold for those products, we have

$$T_{r-1}(x)H_{n-2}(x) = \sum_{\ell=0}^{n-r-1} \Gamma_\ell \left(T_{n-r+1-\ell}(x) + T_{|\ell-n+r+1|}(x)\right) + \Gamma_{n-r}T_1(x) - \sum_{\ell=1}^{r-2}\sum_{s=1}^{\ell+1} 2A_{\ell+1-s}T_{|r-\ell-s|}(x) + \cdots ,$$

and

$$T_{r-2}(x)H_{n-1}(x) = \sum_{\ell=0}^{n-r} \Gamma_\ell T_{n-r+1-\ell}(x) + \frac{1}{2}\Gamma_{n-r+1}T_0(x) - \sum_{\ell=1}^{r-2}\sum_{s=1}^{\ell} A_{\ell-s}T_{|r-\ell-s-1|}(x) + \cdots ,$$

where the dots correspond to something of the form $r(x)P(x)$, with $r(x)$ a scalar polynomial. Using

$$\begin{aligned}&\sum_{\ell=0}^{n-r-1} \Gamma_\ell \left(T_{n-r+1-\ell}(x) + T_{|\ell-n+r+1|}(x)\right) + \Gamma_{n-r}T_1(x) - \sum_{\ell=0}^{n-r} \Gamma_\ell T_{n-r+1-\ell}(x) - \frac{1}{2}\Gamma_{n-r+1}T_0(x) \\ &= \sum_{\ell=0}^{n-r-2} \Gamma_\ell T_{n-r-1-\ell}(x) + \frac{1}{2}\Gamma_{n-r-1}T_0(x) - A_{r-1}T_0(x),\end{aligned}$$

where we have used $(\Gamma_{n-r+1} - \Gamma_{n-r-1})/2 = A_{r-1}$, and

$$\begin{aligned}&-\sum_{\ell=1}^{r-2}\sum_{s=1}^{\ell+1} 2A_{\ell+1-s}T_{|r-\ell-s|}(x) + \sum_{\ell=1}^{r-2}\sum_{s=1}^{\ell} A_{\ell-s}T_{|r-\ell-s-1|}(x) = -\sum_{\ell=1}^{r-2}\sum_{s=1}^{\ell+1} A_{\ell+1-s}T_{|r-\ell-s|}(x) - \sum_{\ell=1}^{r-2} A_\ell T_{|r-\ell-1|}(x) \\ &= -\sum_{\ell=1}^{r-2}\sum_{s=1}^{\ell+1} A_{\ell+1-s}T_{|r-\ell-s|}(x) - \sum_{s=1}^{r} A_{r-s}T_{|s-1|}(x) + A_{r-1}T_0(x) + A_0T_{r-1}(x) \\ &= -\sum_{\ell=0}^{r-1}\sum_{s=1}^{\ell+1} A_{\ell+1-s}T_{|r-\ell-s|}(x) + A_{r-1}T_0(x) + 2A_0T_{r-1}(x) \\ &= -\sum_{\ell=1}^{r}\sum_{s=1}^{\ell} A_{\ell-s}T_{|r+1-\ell-s|}(x) + A_{r-1}T_0(x) + 2A_0T_{r-1}(x),\end{aligned}$$

we get

$$T_{r-1}(x)H_{n-2}(x) = \sum_{\ell=0}^{n-r-2} \Gamma_\ell T_{n-r-1-\ell}(x) + \frac{1}{2}\Gamma_{n-r-1}T_0(x) - \sum_{\ell=1}^{r}\sum_{s=1}^{\ell} 2A_{\ell-s}T_{|r+1-\ell-s|}(x) + \cdots ,$$

where the dots correspond to something of the form $r(x)P(x)$, with $r(x)$ a scalar polynomial. As can be checked, the previous equation corresponds to (25) with $k=1$ and $i=n-r$.

The proof that Theorem 4.1 holds for products represented by grey vertices labeled with 2 is very similar to the previous one, so we omit it.

Finally, consider a product $T_{n-i}(x)H_{n-k}(x)$ represented by a grey vertex labeled with 3. From (1) and (19), we have

$$\begin{aligned}T_{n-i}(x)H_{n-k}(x) =&(2xT_{n-i-1}(x)-T_{n-i-2}(x))H_{n-k}(x)\\ =&T_{n-i-1}(x)H_{n-k+1}(x)+T_{n-i-1}(x)H_{n-k-1}(x)-T_{n-i-2}(x)H_{n-k}(x)-2A_{k-1}T_{n-i-1}(x)\end{aligned}$$

The previous equation shows that $T_{n-i}(x)H_{n-k}(x)$ may be computed from two products represented by (non-triangular) vertices in the diagonal on its left: $T_{n-i-1}(x)H_{n-k-1}(x)$ and $T_{n-i-2}(x)H_{n-k}(x)$, and a product represented by a vertex in the same diagonal, above the vertex corresponding to $T_{n-i}(x)H_{n-k}(x)$, that is, the product $T_{n-i-1}(x)H_{n-k+1}(x)$. This observation shows how to prove inductively (from top to bottom) that Theorem 4.1 holds for the grey vertices labeled with 3 in the same diagonal. Assuming that (24) holds for $T_{n-i-1}(x)H_{n-k-1}(x)$, $T_{n-i-2}(x)H_{n-k}(x)$ and $T_{n-i-2}(x)H_{n-k}(x)$, and using $\Gamma_{i+1}-\Gamma_{i-1}=2A_{n-i-1}$,

$$\begin{aligned}&\sum_{\ell=0}^{i-1}\Gamma_\ell(T_{i+k-2-\ell}(x)+T_{|k+\ell-i-2|}(x))+\Gamma_iT_{k-2}(x)+\sum_{\ell=0}^{i-1}\Gamma_\ell(T_{i+k-\ell}(x)+T_{|k+\ell-i|}(x))+\Gamma_iT_k(x)\\ &-\sum_{\ell=0}^{i}\Gamma_\ell(T_{i+k-\ell}(x)+T_{|k+\ell-i-2|}(x))-\Gamma_{i+1}T_{k-1}(x)\\ &=\sum_{\ell=0}^{i-2}\Gamma_\ell(T_{i+k-2-\ell}(x)+T_{|k+\ell-i|}(x))+\Gamma_{i-1}T_{k-1}(x)-2A_{n-i-1}T_{k-1},\end{aligned}$$

and

$$\begin{aligned}&-\sum_{\ell=1}^{n-k+1-i}\sum_{r=1}^{k-2+\ell}2A_{k-2+\ell-r}T_{|n-i-\ell-r|}(x)-\sum_{\ell=1}^{n-k-1-i}\sum_{r=1}^{k+\ell}2A_{k+\ell-r}T_{|n-i-\ell-r|}(x)\\ &+\sum_{\ell=1}^{n-k-1-i}\sum_{r=1}^{k-1+\ell}2A_{k-1+\ell-r}T_{|n-i-1-\ell-r|}(x)=-\sum_{\ell=1}^{n-k+1-i}\sum_{r=1}^{k-1+\ell}2A_{k-1+\ell-r}T_{|n-i+1-\ell-r|}(x)\\ &+2A_{k-1}T_{n-i-1}(x)+2A_{n-i-1}T_{k-1}(x).\end{aligned}$$

we get

$$\begin{aligned}&T_{n-i}(x)H_{n-k}(x)=T_{n-i-1}(x)H_{n-k+1}(x)+T_{n-i-1}(x)H_{n-k-1}(x)-T_{n-i-2}(x)H_{n-k}(x)-2A_{k-1}T_{n-i-1}(x)\\ &=\sum_{\ell=0}^{i-2}\Gamma_\ell(T_{i+k-2-\ell}(x)+T_{|k+\ell-i|}(x))+\Gamma_{i-1}T_{k-1}(x)-\sum_{\ell=1}^{n-k+1-i}\sum_{r=1}^{k-1+\ell}2A_{k-1+\ell-r}T_{|n-i+1-\ell-r|}(x)+\cdots,\end{aligned}$$

where the dots correspond to something of the form $r(x)P(x)$, with $r(x)$ a scalar polynomial, which shows that (24) holds also for $T_{n-i}(x)H_{n-k}(x)$.

The final step of the proof consists in proving the uniqueness of $r_{ik}(x)$ and $Q_{ik}(x)$ in (39). For this purpose, assume that there exist two scalar polynomials $r_{ik}(x)$ and $\widetilde{r}_{ik}(x)$, and two matrix polynomials $Q_{ik}(x)$ and $\widetilde{Q}_{ik}(x)$ of degree at most $n-1$ such that $T_{n-i}(x)H_{n-k}(x)=Q_{ik}(x)+r_{ik}(x)P(x)=\widetilde{Q}_{ik}(x)+\widetilde{r}_{ik}(x)P(x)$. Then, $Q_{ik}(x)-\widetilde{Q}_{ik}(x)=(\widetilde{r}_{ik}(x)-r_{ik}(x))P(x)$ is a matrix polynomial of degree at most $n-1$, but, if $r_{ik}(x)\neq\widetilde{r}_{ik}(x)$, the matrix polynomial $(\widetilde{r}_{ik}(x)-r_{ik}(x))P(x)$ has degree larger than or equal to n, hence $r_{ik}(x)=\widetilde{r}_{ik}(x)$ and $Q_{ik}(x)=\widetilde{Q}_{ik}(x)$.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables.* Number 55. Courier Dover Publications, 1972.

[2] A. Amiraslami, R. M. Corless, P. Lancaster. Linearization of matrix polynomials expressed in polynomial bases. *IMA. J. Numer. Anal.* 29(1), pp. 141–157, 2009.

[3] V. I. Arnold. On matrices depending on parameters. *Russian Matehmatical Surveys*, 26(2), pp. 29–43. 1971

[4] S. Barnett. *Polynomials and Linear Control Systems.* Marcel Dekker Inc., 1983.

[5] C. W. Clenshaw. A note on the summation of Chebyshev series. *Math. Comp.*, 9, pp. 118–120, 1955.

[6] F. De Terán, F. M. Dopico and D. S. Mackey. Fiedler companion linearizations and the recovery of minimal indices. SIAM J. Matrix Anal. Appl., 31(4), pp. 2181–2204,2009/2010.

[7] F. De Terán, F. M. Dopico, J. Pérez. Backward stability of polynomial root-finding using Fiedler companion matrices. *MIMS EPrint* 2014.38, preprint. 2014.

[8] A. Edelman, H. Murakami. Polynomial roots from companion matrix eigenvalues. *Math. Comp.*, 64, pp. 763–776. 1995.

[9] C. Effenberger, D. Kressner. Chebyshev interpolation for nonlinear eigenvalue problems. *BIT*, 52, pp. 933–951, 2012.

[10] I. J. Good. The colleague matrix, a Chebyshev analogue of the companion matrix. *Q. J. Math.*, 12, pp. 61–68, 1961.

[11] P. Lancaster, M. Tismenetsky. *The Theory of Matrices.* Second Ed., Academic Press, San Diego, 1985.

[12] P. W. Lawrence, R. M. Corless, Stability of rootfinding for barycentric Lagrange interpolants. *Numer. Algorithms*, 65, pp. 447–464, 2014.

[13] P. Lawrence, M. Van Barel, P. Van Dooren. Structured backward error analysis of polynomial eigenvalue problems solved by linearizations. Preprint, submitted.

[14] D. Lemmonier, P. Van Dooren. Optimal scaling of companion pencils for the QZ-algorithm. *Proceedings SIAM Appl. Lin. Alg. Conference*, Paper CP7-4. 2003

[15] D. Lemmonier, P. Van Dooren. Optimal scaling of block companion pencils. *Proceedings of the International Symposium on Mathematical Theory of Networks and Systems*, Leuven, Belgium, 2004.

[16] J. Maroulas, S. Barnett. Polynomials with respect to a general basis. I. Theory. *J. Math. Anal. Appl.*, 72(1), pp. 177–194, 1979

[17] Y. Nakatsukasa, V. Noferini. On the stability of computing polynomial roots via confederate linearizations. To appear in *Math. Comp.*

[18] Y. Nakatsukasa, V. Noferini, A. Townsend. Vector spaces of linearizations for matrix polynomials: a bivariate polynomial approach. Preprint, submitted.

[19] V. Noferini, F. Poloni. Duality of matrix pencils, Wong chains and linearizations. *Linear Algebra Appl.*, 471, pp. 730–767, 2015.

[20] B. N. Parlett, C. Reinsch. Balancing a matrix for calculation of eigenvalues and eigenvectors. *Numer. Math.*, 13, pp. 293–304, 1963.

[21] G. Peters, J. H. Wilkinson. Practical problems arising in the solution of polynomial equations. *J. Inst. Maths. Appl.*, 8, pp. 16–35, 1971.

[22] L. N. Trefethen. *Approximation theory and approximation practice.* SIAM, 2013.

[23] L. N. Trefethen et al. Chebfun Version 5. The Chebfun Development Team, 2014. `http://www.maths.ox.ac.uk/chebfun/`.

[24] K. -C. Toh, L. N. Trefethen. Pseudozeros of polynomials and pseudospectra of companion matrices. *Numer. Math.*, 68, pp. 403–425, 1994.

[25] J. H. Wilkinson. *Rounding Errors in Algebraic Processes.* Prentice-Hall, Englewood Cliffs, 1963.

[26] J. H. Wilkinson. *The Algebraic Eigenvalue Problem.* Clarendon Press, Oxford, 1965.