Noferini, Vanni; Robol, Leonardo; Vandebril, Raf

Structured backward errors in linearizations

Published in:
Electronic Transactions on Numerical Analysis

DOI:
10.1553/ETNA_VOL54S420

Published: 01/01/2020

Document Version
Publisher's PDF, also known as Version of record

Electronic Transactions on Numerical Analysis.
Volume 54, pp. 420–442, 2021.

ISSN 1068–9613.
DOI: 10.1553/etna_vol54s420

ETNA
Kent State University and
Johann Radon Institute (RICAM)

STRUCTURED BACKWARD ERRORS IN LINEARIZATIONS*

VANNI NOFERINI†, LEONARDO ROBOL‡§, AND RAF VANDEBRIL¶

Abstract. A standard approach to calculate the roots of a univariate polynomial is to compute the eigenvalues of an associated *confederate* matrix instead, such as, for instance, the companion or comrade matrix. The eigenvalues of the confederate matrix can be computed by Francis's QR algorithm. Unfortunately, even though the QR algorithm is provably backward stable, mapping the errors back to the original polynomial coefficients can still lead to huge errors. However, the latter statement assumes the use of a non-structure-exploiting QR algorithm. In [J. L. Aurentz et al., *Fast and backward stable computation of roots of polynomials*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 942–973] it was shown that a structure-exploiting QR algorithm for companion matrices leads to a structured backward error in the companion matrix. The proof relied on decomposing the error into two parts: a part related to the recurrence coefficients of the basis (a monomial basis in that case) and a part linked to the coefficients of the original polynomial. In this article we prove that the analysis can be extended to other classes of comrade matrices. We first provide an alternative backward stability proof in the monomial basis using structured QR algorithms; our new point of view shows more explicitly how a structured, decoupled error in the confederate matrix gets mapped to the associated polynomial coefficients. This insight reveals which properties have to be preserved by a structure-exploiting QR algorithm to end up with a backward stable algorithm. We will show that the previously formulated companion analysis fits into this framework, and we analyze in more detail Jacobi polynomials (comrade matrices) and Chebyshev polynomials (colleague matrices).

Key words. backward error, structured QR, linearization, comrade matrix, colleague matrix, companion matrix

AMS subject classifications. 65H04, 65F15, 65G50

1. Introduction.

A standard approach to find the solutions of a univariate polynomial equation is to convert the problem into an equivalent one where the eigenvalues of a matrix are computed instead. The algebraic technique used to construct such a matrix is called a linearization, and, albeit ultracentenarian, it is still the most popular initial step of modern rootfinding algorithms, at least if all the polynomial roots are sought. For example, this is what MATLAB's `roots` function does [9] for polynomials expressed in the monomial basis, and it is at the heart of `chebfun/roots` for polynomials expressed in the Chebyshev basis [23].

In the landmark paper [9] Edelman and Murakami cast a shadow on this strategy. They showed that, even if the matrix eigenvalue problem is solved with a backward stable algorithm such as QR [25], the whole approach can (depending on the specific linearized polynomial) be catastrophically unstable. More recently, De Terán, Dopico, and Pérez [7] argued that, if Fiedler linearizations [11] are used instead of the classical companion linearization [9], then the potential misfortunes can be even more pronounced. Fortunately, Van Dooren and Dewilde [24] showed that this problem could be circumvented by solving a generalized eigenproblem instead; the disadvantage, however, is that this is significantly less efficient.

While De Terán, Dopico, and Pérez [7] and Edelman and Murakami [9] focused only on polynomials expressed in the monomial basis, Nakatsukasa and Noferini [18] proved that

*Received December 19, 2019. Accepted April 15, 2021. Published online on June 7, 2021. Recommended by Qiang Ye. The work of Vanni Noferini was supported by an Academy of Finland grant (Suomen Akatemian päätös 331240); the work of Leonardo Robol was supported by an INdAM/GNCS research grant "Metodi low-rank per problemi di algebra lineare con struttura data-sparse".

†Department of Mathematics and Systems Analysis, Aalto University, PL 11000, 00076 Aalto, Finland (vanni.noferini@aalto.fi).

‡Department of Mathematics, University of Pisa, Largo Bruno Pontecorvo 5, 56127 Pisa, Italy (leonardo.robol@unipi.it). The author is a member of the INdAM research group GNCS.

§Institute of Information Science and Technologies "A. Faedo", ISTI-CNR, Pisa, Italy (leonardo.robol@isti.cnr.it).

¶Department of Computer Science, KU Leuven, Celestijnenlaan 200A, 3001 Heverlee, Belgium (raf.vandebril@cs.kuleuven.be).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

analogous results on the dangers of always trusting the linearize-and-use-QR philosophy can be stated for any degree-graded basis. That is, the beautiful idea of constructing the so-called *confederate* matrix and then finding its eigenvalues by QR is potentially, depending on the polynomial input, unstable. Again, for many bases of practical importance, switching to a pencil and employing the QZ algorithm provably avoids any instabilities [15, 16, 18]. Less clear than for the monomial basis is the question under which conditions on the input polynomial the QR-based approach is stable; Noferini and Pérez [20] gave a complete answer for the Chebyshev basis, but we are not aware of any progress for other bases.

This story has recently seen a sudden twist towards positive news. All the aforementioned results rely on the assumption that a *general* eigensolver, e.g., an *unstructured* QR algorithm, is applied to the linearizing matrix. However, confederate matrices are typically highly structured. Algorithms specifically designed to preserve and utilize this structure result in two advantages: reduced computational and storage costs, and a structured backward error. Several such algorithms can be found in the literature; see, for example, [3, 5, 14] for companion matrices and the references therein for the case of unitary (fellow) and symmetric-plus-low-rank (comrade) matrices.

Consider, for instance, the companion algorithm presented by Aurentz et al. [5]. There, the authors proved that the structured QR algorithm has a backward error in the companion matrix of the order of $\|p\|_2^2\epsilon_m$ for the rank-one part and of the order of ϵ_m for the unitary part (with ϵ_m denoting the machine precision). This implies that *as an eigensolver* this particular algorithm is not stable, and a blind application of the results of Edelman and Murakami [9], merging both errors, would yield a backward error for the polynomial of the size $\|p\|_2^3\epsilon_m$: an apparent disaster as this is even worse than what the unstructured QR obtains: $\|p\|_2^2\epsilon_m$. In the numerical experiments, however, only an error of the form $\|p\|_2^2\epsilon_m$ was observed, insinuating that something peculiar was happening with the errors.

Two years later, Aurentz et al. [2, 3], were able to improve their companion code to get an error of the order of $\|p\|_2\epsilon_m$ for the rank-one part. According to the results of Edelman and Murakami, this should have implied an error of size about $\|p\|_2^2\epsilon_m$ for the polynomial coefficients. However, by considering a mixed backward error analysis, it was demonstrated that the specific structure of the backward error in the companion matrix implies that *as a rootfinder*, considering the backward error in the polynomial, the algorithm is backward stable with a backward error bounded by $\|p\|_2\epsilon_m$! This was the first time that a rootfinder based on linearization and (structured) QR was proved to be stable in this stronger sense.

In the current paper we extend the backward error results of Aurentz et al. [3] to other confederate matrices. As a first step, we present an alternative derivation of the same result of [3], which is less coupled with the underlying algorithm and thus easier to generalize to other bases. We examine how to cleverly map the structured backward error in the confederate matrix back to the polynomial coefficients. As an example of particular interest we analyze the case of Chebyshev polynomials (colleague matrices) in detail, see how the companion results [3] fit in, and later discuss the extension to more general Jacobi polynomials (comrade matrices).

More specifically, we address the following problem. We assume we are given a confederate pencil, that is, a structured-plus-rank-one pencil that linearizes a polynomial p expressed in a degree-graded basis. The pencil is of the form $M(x) + ab^T$, where $M(x)$ is independent of p and links to the polynomial basis, and the rank-one addend ab^T encodes the coefficients of p. This is precisely the scenario encountered for polynomials expressed in a broad class of orthogonal polyomial bases, including monomials, Chebyshev, Legendre, ultraspherical, and other Jacobi polynomials. Next, we assume that a *structured* eigensolver is used to compute the eigenvalues of the structured pencil such that backward errors of different form can be

ETNA
Kent State University and
Johann Radon Institute (RICAM)

attached to $M(x)$ and to ab^T. The question of interest is to map the error back to p and to characterize it, thereby assessing the overall stability of the rootfinding algorithm. We show that under minimal assumptions on the pencil $M(x)$ (satisfied in practice by most linearization schemes), only the backward error in $M(x)$ increases when mapping it back to the polynomial.

Our result thus clarifies the direction that should be followed in the development of stable structured QR algorithms for polynomial rootfinding: one has to ensure that the backward error in the "basis part" of the pencil, the addend $M(x)$, is small and independent of the polynomial under consideration.

The paper is structured as follows. In Section 2 we introduce confederate matrices, prove properties essential for the article and refine them to comrade matrices. Section 3 discusses the basic principles for the mixed backward error analysis; it is shown how the structured error can be mapped back to the polynomials. In Section 4 we illustrate the main idea by reconsidering the companion matrix and providing an alternative and simpler derivation of the results of Aurentz et al. [5]. In Section 5 we provide specific bounds for polynomials in the Chebyshev basis (colleague matrices) and come up with a conjecture for Jacobi polynomials. We conclude with Section 6.

2. Confederate matrices.

First we discuss general confederate matrices. Then we refine the results to companion and comrade matrices and discuss the special case of colleague matrices.

2.1. Definition and properties of confederate matrices.

Let ϕ_j be any degree-graded (i.e., $\deg \phi_j = j$) polynomial basis such that, for all $j = 0, \dots, n$, ϕ_j has a leading coefficient $\nu_j \neq 0$ when expressed in the monomial basis. Let p be a polynomial of degree n, monic in the basis $\{\phi_j\}$. Denoting

$$\Phi(x) = \begin{bmatrix} \phi_{n-1}(x) \\ \vdots \\ \phi_1(x) \\ \phi_0(x) \end{bmatrix},$$

we can write $p(x) = \phi_n(x) + c^T\Phi(x)$ for a unique coefficients vector c. Following [6, 18], we now introduce the confederate matrix of $p(x)$.

DEFINITION 2.1. *The confederate matrix of* $p(x) = \phi_n(x) + c^T\Phi(x)$ *is the unique matrix* C *satisfying*

$$C\Phi(x) = x\Phi(x) - \chi^{-1}p(x)e_1,$$

where $\chi = \nu_n\nu_{n-1}^{-1}$.

In the following theorem, the second item is classical [6]. The first item also dates back to [6], although in a weaker form; it was stated in this form (without proof) in [18]. The third item may be new in this general form although some special cases can be deducted from other published results; for example, if $\{\phi_j\}$ is the monomial basis, then it is a consequence of [7], and for the Chebyshev basis it can be proved using the analysis of [20].

ETNA
Kent State University and
Johann Radon Institute (RICAM)

THEOREM 2.2. *The following properties hold:*

1. *C is a (strong) linearization of $p(x)$ (implying $\det(xI - C) = \frac{p(x)}{\nu_n}$);*
2. *C can be written as*

$$C = H - \chi^{-1} e_1 c^T,$$

where H is Hessenberg and only depends on the basis $\{\phi_j\}$;

3. $\operatorname{adj}(xI - C)e_1 = \nu_{n-1}^{-1}\Phi(x)$.

Proof. We prove the three points separately.

1. It can be easily verified that $xI - C$ belongs to the vector space $\mathbb{L}_1$ for the basis $\{\phi_j\}$ [17, 19]. Since it is manifestly a nonsingular pencil, it is a strong linearization for p by [19, Theorem 2.1] (or [17, Theorem 4.3] for the monomial basis). Since $\det(xI - C)$ is monic in the monomial basis and p is monic in the basis $\{\phi_j\}$, the equality $\det(xI - C) = p(x)/\nu_n$ follows.
2. Let H be the matrix that satisfies $H\Phi(x) = x\Phi(x) - e_1\chi^{-1}\phi_n(x)$. Since $x\phi_k(x)$ has degree $k+1$, for all $k = 0, \ldots, n-2$, it follows that H is Hessenberg by the degree-gradedness of $\{\phi_j\}$. Now,

$$\chi(H - C)\Phi = e_1(p(x) - \phi_n(x)) = e_1 c^T \Phi(x).$$

Since this relation holds over $\mathbb{R}(x)$, a fortiori it is still true as a relation over $\mathbb{R}$ after evaluating $\Phi(x)$ at any point. Thus, for any Vandermonde matrix V, we obtain

$$\chi(H - C)V = e_1 c^T V \Rightarrow \chi(H - C) = e_1 c^T,$$

as desired.

3. By definition of C we have

$$\chi(xI - C)\Phi = p(x)e_1.$$

As $xI - C$ is regular, it is invertible over $\mathbb{R}(x)$. Hence we can premultiply by its inverse (using $\det(xI - C) = p(x)/\nu_n$), to obtain

$$\nu_{n-1}^{-1}\Phi(x) = \operatorname{adj}(xI - C)e_1. \qquad \square$$

REMARK 2.3. The matrix H is a square submatrix of the multiplication matrix M in [19, Section 2], and it represents the multiplication-by-x operator in the quotient space $\mathbb{R}[x]/\langle\chi^{-1}\phi_n\rangle$.

EXAMPLE 2.4 (Companion matrix). Consider the monomial basis $\{\phi_j\}$ with $\phi_j(x) = x^j$. We have $\phi_j(x) = x\phi_{j-1}(x)$. As a consequence $x\Phi(x) = H\Phi(x) + x^n e_1$, where H is the downshift matrix, i.e., the matrix with only ones on the subdiagonal and zeroes elsewhere.

2.2. Comrade matrices. When the $\{\phi_j\}$ are orthonormal on a closed real interval and have positive leading coefficients, the three-term recurrence

$$\phi_j(x) = (\alpha_j x + \beta_j)\phi_{j-1}(x) - \gamma_j\phi_{j-2}(x)$$

holds for all j and for some $\beta_j \in \mathbb{R}$, $\alpha_j = \nu_j\nu_{j-1}^{-1} > 0$, $\gamma_k = \nu_j\nu_{j-2}\nu_{j-1}^{-2} > 0$ [22, Theorem 3.2.1]. As a consequence, multiplication-by-x is encoded by

$$x\phi_{j-1}(x) = \frac{1}{\alpha_j}\phi_j(x) - \frac{\beta_j}{\alpha_j}\phi_{j-1}(x) + \frac{\gamma_j}{\alpha_j}\phi_{j-2}(x),$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

which immediately implies that, in this case,

$$x\Phi(x) = H\Phi(x) + \alpha_n^{-1}\phi_n(x)e_1,$$

where H is tridiagonal and has positive subdiagonal/superdiagonal elements. In the case of an orthogonal basis, the confederate matrix is also known as the comrade matrix of p. Note, moreover, that, as displayed above, in this setting $\chi = \alpha_n$ so that for $p(x) = \phi_n(x) + c^T\Phi(x)$ it holds that

$$C = H - \alpha_n^{-1}e_1c^T.$$

The matrix H has the following form:

$$H := \begin{bmatrix} -\frac{\beta_n}{\alpha_n} & \frac{\gamma_n}{\alpha_n} & & & \\ \frac{1}{\alpha_{n-1}} & -\frac{\beta_{n-1}}{\alpha_{n-1}} & \frac{\gamma_{n-1}}{\alpha_{n-1}} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{1}{\alpha_2} & -\frac{\beta_2}{\alpha_2} & \frac{\gamma_2}{\alpha_2} \\ & & & \frac{1}{\alpha_1} & -\frac{\beta_1}{\alpha_1} \end{bmatrix}.$$

We remark that for polynomials represented in the Chebyshev basis, the matrix C is called the colleague matrix.

In addition, we note that, since α_j and γ_j are positive, it is possible to perform a diagonal scaling to the matrix H that makes it symmetric. Indeed, we can consider the matrix $D^{-1}HD$, where D is any diagonal matrix with entries

$$d_k := \sqrt{\frac{\alpha_1}{\alpha_{n-k+1}}\prod_{i=2}^{n-k+1}\gamma_i}.$$

Observe that, in particular, $d_n = 1$. This corresponds to choosing the orthogonal basis $\tilde{\phi}_j(x) := d_{n-j}^{-1}\phi_j(x)$ having formally set $d_0 := 1$. The scaled matrices are as follows:

$$D^{-1}HD = \begin{bmatrix} -\frac{\beta_n}{\alpha_n} & \sqrt{\frac{\gamma_n}{\alpha_n\alpha_{n-1}}} & & & \\ \sqrt{\frac{\gamma_n}{\alpha_n\alpha_{n-1}}} & -\frac{\beta_{n-1}}{\alpha_{n-1}} & \sqrt{\frac{\gamma_{n-1}}{\alpha_{n-1}\alpha_{n-2}}} & & \\ & \ddots & \ddots & \ddots & \\ & & \sqrt{\frac{\gamma_3}{\alpha_3\alpha_2}} & -\frac{\beta_2}{\alpha_2} & \sqrt{\frac{\gamma_2}{\alpha_2\alpha_1}} \\ & & & \sqrt{\frac{\gamma_2}{\alpha_2\alpha_1}} & -\frac{\beta_1}{\alpha_1} \end{bmatrix},$$

$$D^{-1}CD = D^{-1}HD - \tilde{\chi}^{-1}e_1\tilde{c}^T,$$

where $\tilde{c}$ is the coefficient vector of p expressed in the scaled basis $\{\phi_0, \tilde{\phi}_1, \ldots, \tilde{\phi}_{n-1}, \phi_n\}$ and $\tilde{\chi} = \sqrt{\alpha_1\alpha_n\gamma_2\gamma_3 s\gamma_n}$; the coefficient $\tilde{\chi}$ is the analogue of χ for the rescaled basis $\{\phi_0, \tilde{\phi}_1, \ldots, \tilde{\phi}_{n-1}, \phi_n\}$. From now on we work in this symmetrized setting, and we only consider $D^{-1}CD$. From the viewpoint of developing structured algorithms, this is particularly relevant. If $A = H + uv^T$, with H real symmetric or Hermitian and uv^T of rank one, then all the matrices obtained through the iteration of a QR method that can be written as $A_k := Q_kAQ_k^H$, with Q_k orthogonal or unitary, have the same property.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

This observation is key in the development of fast algorithms; in the monomial case, the companion matrix can be similarly decomposed as the sum of a unitary and a rank-one part; this property is also preserved by the QR iterations.

Fast algorithms for these classes of matrices often work on the structured (either Hermitian or unitary) and rank-one part separately. Therefore, it is reasonable to assume that these parts might be contaminated, throughout the iterations, by backward errors of different magnitudes. Classical backward error analysis does not take this property into account, so we present a more general backward error formulation in the next section.

3. Mixed backward error analysis. We are now ready to study the behavior of the linearized polynomial $p(x)$ under perturbations of the pencil $xI - C$. More precisely, we consider $xI - (C + \delta C)$, where $C + \delta C$ has a mixed backward error of the following form:

$$C + \delta C = H + \delta H + (e_1 + \delta e_1)(c + \delta c)^T. \tag{3.1}$$

We note that, for any δC, it is always possible to find a decomposition as the one in (3.1). Indeed, for any choice of δe_1 and δc, it suffices to choose $\delta H := \delta C - \delta e_1 c^T - e_1 \delta c^T - \delta e_1 \delta c^T$. Our analysis will show that these terms provide different contributions to the backward error in the polynomial with different amplification factors. In particular, it will show that errors in δH may be amplified much higher when projecting the error back to the polynomial, whereas the backward errors δe_1 and δc are relatively less harmful.

By Theorem 2.2 the perturbed matrix C linearizes the polynomial

$$p(x) + \delta p(x) := \nu_n \det(xI - C - \delta C). \tag{3.2}$$

Our aim is now to examine the size of $\delta p(x)$ under the assumption that for the various actors in (3.1) a bound is known. In particular, we assume to know appropriate positive $\epsilon_H, \epsilon_1, \epsilon_c$ such that

$$\|\delta H\|_2 \leq \epsilon_H < 1, \qquad \|\delta c\|_2 \leq \epsilon_c, \qquad \|\delta e_1\|_2 \leq \epsilon_1. \tag{3.3}$$

Observe that, as discussed above, for given δC, infinitely many choices exist for $\delta H, \delta c, \delta e_1$. Thus, we may without any loss of generality assume to have picked one that is optimal (in the sense of making our results as sharp as possible) for the values of ϵ_H, ϵ_c, ϵ_1. Of course, it may also happen that in practice an algorithm suggests a particular choice for which bounds for ϵ_H, ϵ_c, ϵ_1 are "naturally" obtained (see, e.g., [3]).

In this section we will discuss the general setting, which holds for all confederate matrices. In Sections 4 and 5 we will specialize to companion and comrade matrices, that is, corresponding either to the monomial basis or a basis of polynomials orthogonal on a real interval.

We will make the following assumptions in our analysis:

Assumption 1. The matrix H is *normal*; for instance, we will consider the case where H is unitary (the monomial basis) and symmetric (the Chebyshev case and in general the case of orthogonal polynomials on the real line).

Assumption 2. The backward errors in H, e_1, c may be of very *different magnitude*. In particular, to obtain strong backward stability on the polynomial, we will need ϵ_H and ϵ_1 to be bounded *independently* of the norm of the vector of polynomial coefficients; on the contrary, ϵ_C may depend linearly on this value.

We stress that the second assumption, in particular the bound for ϵ_H, is the one which is hard to obtain in practice when designing a structured algorithm. It does not hold for the unstructured QR or QZ, and obtaining it was key in developing a backward stable algorithm for

ETNA
Kent State University and
Johann Radon Institute (RICAM)

the monomial case in [3, 5]; we hope that our analysis will help to devise similar algorithms in other bases, in particular for the Chebyshev one.

From now on, we use the notation $\doteq$ to indicate an equality that holds up to second order in the terms $\epsilon_H, \epsilon_1, \epsilon_c$. Based on the expressions above, we can rewrite the perturbed polynomial.

THEOREM 3.1. *With the notation of* (3.1), (3.2), *and* (3.3), *the following first-order expansion in* $\epsilon_H, \epsilon_c, \epsilon_1$ *holds:*

$$\begin{aligned}(p+\delta p)(x) \doteq p(x) &+ \nu_n \left(\det(xI - H - \delta H) - \det(xI - H)\right) \\ &+ \chi \delta c^T \Phi(x) + \nu_n c^T \operatorname{adj}(xI - H)\delta e_1 + \\ &+ \nu_n c^T \left(\operatorname{adj}(xI - H - \delta H) - \operatorname{adj}(xI - H)\right) e_1.\end{aligned}$$

Proof. By (3.2) we have $(p+\delta p)(x) = \nu_n \det(xI - (C + \delta C))$, which by Theorem 2.2 is equal to

$$(p+\delta p)(x) = \nu_n \det(xI - H - \delta H) + \nu_n (c + \delta c)^T \operatorname{adj}(xI - H - \delta H)(e_1 + \delta e_1).$$

To obtain the statement, we first add $p(x)$ and subtract its expansion obtained by Theorem 2.2. Next, we discard higher-order terms and use the equalities $\nu_{n-1} \operatorname{adj}(xI - H)e_1 = \Phi(x)$ and $\nu_{n-1}\chi = \nu_n$. ∎

Theorem 3.1 reveals that, in order to provide bounds for the perturbation $\delta p(x)$, it is essential to do a perturbation analysis related to determinants and adjugates. To this aim, we provide a few results that will be useful in later proofs.

LEMMA 3.2 (Jacobi's formula). *Let* X *be any square matrix and* δX *a small perturbation. Then,*

$$\det(X + \delta X) = \det(X) + \operatorname{tr}(\operatorname{adj}(X)\delta X) + O(\|\delta X\|^2).$$

A similar result can be given for the adjugate as well, which characterizes the effect of small perturbations.

LEMMA 3.3. *Let* δX *be a small perturbation (*$\|\delta X\| < 1$*). Then,*

$$\operatorname{adj}(I + \delta X) = (I - \delta X)(1 + \operatorname{tr}(\delta X)) + O(\|\delta X\|^2). \tag{3.4}$$

Proof. Since $\|\delta X\| < 1$, $I + \delta X$ is invertible, and therefore we can write

$$\operatorname{adj}(I + \delta X) = (I + \delta X)^{-1} \det(I + \delta X).$$

We shall make a first-order approximation of both terms involved in the above equality. Concerning the first one, we have that $(I + \delta X)^{-1} \doteq I - \delta X$. To bound the change in the determinant, we use Lemma 3.2 and obtain

$$\det(I + \delta X) = 1 + \operatorname{tr}(\delta X) + O(\|\delta X\|^2),$$

which provides the sought first-order expansion (3.4). ∎

LEMMA 3.4. *Let* A, B *be two* $n \times n$ *matrices, and assume that* A *is normal with eigenvalues* $\lambda_1, \ldots, \lambda_n$. *Then,*

$$|\operatorname{tr}(AB)| \leq \|B\|_2 \sum_{j=1}^{n} |\lambda_j|.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Proof. Let $A = QDQ^H$ be an eigendecomposition of A with Q unitary. Then, if we denote by q_j the columns of Q, we can write:

$$|\operatorname{tr}(AB)| = |\operatorname{tr}(Q^H AQQ^H BQ)| = |\operatorname{tr}(DQ^H BQ)| \leq |\sum_{j=1}^{n} \lambda_j q_j^H Bq_j|.$$

Since $|q_j^H Bq_j| \leq \|B\|_2$, the result follows. ∎

With these tools at hand, we are now able to bound the pointwise perturbation $\delta p(\xi)$, i.e., the evaluation of the perturbation at any point $\xi \in \mathbb{C}$. Later on, when going to comrade and companion matrices, we will need to specify these points ξ to retrieve tight bounds for the polynomial coefficients.

LEMMA 3.5. *Let $(p+\delta p)(x) = \det(xI - C - \delta C)$ be the perturbed polynomial. Then, for every $\xi \in \mathbb{C}$, we get the following first-order bound:*

$$|\delta p(\xi)| \leq \Gamma_1(\xi)\epsilon_1 + \Gamma_c(\xi)\epsilon_c + \Gamma_H(\xi)\epsilon_H + \mathcal{O}(\epsilon_H^2 + \epsilon_1^2 + \epsilon_c^2),$$

where

$$\Gamma_1(\xi) := M(\xi)|\phi_n(\xi)|\|c\|_2, \qquad \Gamma_c(\xi) := \chi\|\Phi(\xi)\|_2,$$
$$\Gamma_H(\xi) := S(\xi)|\phi_n(\xi)| + \Gamma_c(\xi)\left(M(\xi) + S(\xi)\right)\|c\|_2,$$

having defined

$$S(\xi) := \sum_{j=1}^{n} \frac{1}{|\xi - r_j|}, \qquad M(\xi) := \max_{j=1,\ldots,n} \frac{1}{|\xi - r_j|}.$$

In the expressions for $S(\xi)$ and $M(\xi)$ above, r_j denote the roots of the polynomial ϕ_n of degree n.

Proof. Let us first note that, since $\phi_n(x) = \nu_n \det(xI - H)$ and since $\xi I - H$ is normal, we have

$$\|(\xi I - H)^{-1}\|_2 = \max_{j=1,\ldots,n} \frac{1}{|\xi - r_j|}.$$

By Theorem 3.1 we have the following first-order approximation for $\delta p(\xi)$:

$$\delta p(\xi) \doteq \nu_n \left(\det(\xi I - H - \delta H) - \det(\xi I - H)\right) \tag{3.5}$$
$$+ \chi \delta c^T \Phi(\xi) + \nu_n c^T \operatorname{adj}(\xi I - H)\delta e_1 \tag{3.6}$$
$$+ \nu_n c^T \left(\operatorname{adj}(\xi I - H - \delta H) - \operatorname{adj}(\xi I - H)\right) e_1. \tag{3.7}$$

We bound all the terms separately.

- We consider (3.5) first. By Lemma 3.2 we can write

$$\det(\xi I - H - \delta H) - \det(\xi I - H) \doteq \operatorname{tr}(\operatorname{adj}(\xi I - H)\delta H).$$

Since $\xi I - H$ is a normal matrix, we can use Lemma 3.4 to obtain

$$\nu_n |\det(\xi I - H - \delta H) - \det(\xi I - H)| \doteq \nu_n |\operatorname{tr}(\operatorname{adj}(\xi I - H)\delta H)| \leq \epsilon_H \sum_{j=1}^{n} \frac{\phi_n(\xi)}{|\xi - r_j|}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

- To bound the second term in (3.6), we use

$$\|\mathrm{adj}(\xi I - H)\|_2 = |\det(\xi I - H)|\|(\xi I - H)^{-1}\|_2 = \frac{|\phi_n(\xi)|}{\nu_n} \max_j \frac{1}{|\xi - r_j|}.$$

  Bounding the first term just requires taking norms of all the factors involved.
- To bound (3.7), assuming $\|(\xi I - H)^{-1}\delta H\| \leq 1$ and using Lemma 3.3, we write

$$\begin{aligned}\mathrm{adj}(\xi I - H - \delta H) &= \mathrm{adj}\left((\xi I - H)(I - (\xi I - H)^{-1}\delta H)\right)\\ &= \mathrm{adj}(I - (\xi I - H)^{-1}\delta H)\,\mathrm{adj}(\xi I - H)\\ &\doteq (I + (\xi I - H)^{-1}\delta H)(1 - \mathrm{tr}((\xi I - H)^{-1}\delta H))\,\mathrm{adj}(\xi I - H).\end{aligned}$$

  Then, using the fact that $\nu_{n-1}\,\mathrm{adj}(\xi I - H)e_1 = \Phi(\xi)$, we have

$$\begin{aligned}&\nu_n c^T\left(\mathrm{adj}(\xi I - H - \delta H) - \mathrm{adj}(\xi I - H)\right)e_1\\ &\quad\doteq \chi c^T\left((\xi I - H)^{-1}\delta H - \mathrm{tr}((\xi I - H)^{-1}\delta H)I\right)\Phi(\xi).\end{aligned}$$

Taking norms and combining all the results yields the desired bound. ∎

The results we have proved are valid for any class of polynomials under the assumption that H is normal[1]. We will use this idea to generalize the pointwise bound to a bound for the coefficients in the case of the monomial basis and of orthogonal polynomials on a real interval. These are the subjects of the next sections.

4. Companion matrix. In this section we reconsider the error analysis of Aurentz et al. [3] in view of this new theory. The derivation of [3] is based on running the Faddev-Leverrier algorithm to compute the coefficients of the adjugates, and this is used to provide bounds for its norm. This approach is not easily generalizable, despite the existence of a Faddev-Leverrier scheme for non-monomial bases. Our new point of view yields a simple and clean derivation of the results therein, based instead on an interpolation argument.

To analyze the backward error of an algorithm running on the companion matrix, we have to rewrite the companion matrix slightly. Example 2.4 revealed that the Hessenberg matrix H is the downshift matrix, and the eigenvalues can be retrieved from $C = H - e_1 c^T$, i.e., the downshift matrix plus a rank-one part. Structure-exploiting algorithms, however, rely on the unitary-plus-low-rank structure, and rewriting $C = \tilde{H} - e_1\tilde{c}^T$, with $\tilde{H} = H - e_1 e_n^T$ and $\tilde{c} = c + e_n^T$, is clearly of unitary-plus-low-rank form.

This has some impact on the backward error since we are now working with the basis $1, x, \ldots, x^{n-1}, x^n + 1$, instead of the classical monomial basis. Moreover, also the trailing coefficient of our polynomial p has changed. For simplicity, we will therefore, from now on, assume to be working in the basis $1, x, \ldots, x^{n-1}, x^n + 1$.

Eventually, we will use the Fast Fourier transform to retrieve the coefficients of δp. To do so, we need to bound δp evaluated at the n-th roots of unity ξ_j, for $j = 0, \ldots, n-1$. Lemma 3.5 provides

$$|\delta p(\xi_j)| \leq \Gamma_1(\xi_j)\epsilon_1 + \Gamma_c(\xi_j)\epsilon_c + \Gamma_H(\xi_j)\epsilon_H.$$

Using the formulas of Lemma 3.5 yields $\Gamma_1(\xi_j) = 2M$ and $\Gamma_H \leq 2S + \sqrt{n+3}(M+S)\|c\|_2$. Recalling that $\phi_n(x) = x^n + 1$, we have that $r_j = e^{\frac{(2j+1)\pi}{n}}$. Exploiting that $|\phi_i(\xi_j)| = 1$, for $i < n$, and $\chi = 1$, we obtain $\Gamma_c(\xi_j) = \|\Phi(\xi_j)\|_2 \leq \sqrt{n}$. The quantity M is the inverse of the

[1] We note that these perturbations might make $H + \delta H$ non-normal, but this is not an issue for our analysis as we always deal with expansions centered at H.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

distance between the n-th roots of the unity and r_j, which can be bounded by $M \leq \frac{n}{2}$. To bound $S = \sum_{k=1}^{n} \frac{1}{|\xi_j - r_k|}$, we use

$$S = \sum_{k=1}^{n} \frac{1}{|\xi_j - r_k|} = \sum_{k=1}^{n} \frac{1}{|1 - e^{\frac{2\pi}{2n}(2j+1)}|} = \sum_{k=1}^{n} \frac{1}{\left|2\sin\left(\frac{\pi}{2n}(2j+1)\right)\right|}$$
$$= 2\sum_{k=1}^{n/2} \frac{1}{2\sin\left(\frac{\pi}{2n}(2j+1)\right)}$$

and the fact that $\sin x > \frac{2}{\pi}x$, for $x \in [0, \frac{\pi}{2}]$. As a result we obtain

$$S \leq \frac{n}{2}\log\left(\frac{n}{2} + \frac{1}{2}\right).$$

Combining all of this leads to

$$|\delta p(\xi_j)| \leq n\|c\|_2\epsilon_1 + \sqrt{n}\,\epsilon_c + n\log\left(\frac{n}{2}\right)\epsilon_H + \frac{n\sqrt{n}}{2}\left(1 + \log\left(\frac{n}{2} + \frac{1}{2}\right)\right)\|c\|_2\epsilon_H + \mathcal{O}(\epsilon_H^2 + \epsilon_1^2 + \epsilon_c^2). \tag{4.1}$$

As a result, we get for the Euclidean norm of the vector of coefficients of $\delta p(x)$, denoted as $\|\delta p\|_2$,

$$\|\delta p\|_2 = \left\|\frac{1}{\sqrt{n}}F^* q\right\|_2 \leq \|q\|_\infty + \mathcal{O}(\epsilon_H^2 + \epsilon_1^2 + \epsilon_c^2),$$

where $q = [\delta p(\xi_0), \ldots, \delta p(\xi_{n-1})]^T$ and F is the matrix of the discrete Fourier transform. The last factor can be bounded by (4.1).

Reconsidering the algorithm of Aurentz et al. [3], we have that $\epsilon_1 = 0$, $\epsilon_H = \epsilon_m$, and $\epsilon_c = \|c\|_2\epsilon_m$, where ϵ_m is the machine precision. Clearly we end up with the same bound proposed by Aurentz et al., namely a linear dependency on $\|c\|_2$.

Before moving to orthogonal bases on real intervals and in particular Chebyshev and Jacobi polynomials, we emphasize that the main ingredients playing a role in the bound are related to the eigenvalues of the structured matrix H, namely their separation as measured by the constants M, S of Lemma 3.5, and their good properties as interpolation points for the chosen basis. These two quantities will play an important role in the analysis of the following sections as well.

5. Orthogonal polynomials on a real interval. In this section, we consider a class of degree-graded polynomials $\phi_i(x)$, for $j \geq 0$, that are orthogonal on $[-1, 1]$ with respect to a positive measure $w(x)$.

Our aim is to leverage Lemma 3.5 to provide a bound for the coefficients of the perturbed polynomial $\delta p(x)$. To this aim, we provide the following result, which holds for any polynomial family orthogonal on $[-1, 1]$; since this bound is not very explicit, we will then specialize it to a few particular families of polynomials for which we can be more precise, namely Chebyshev and later on all Jacobi polynomials.

THEOREM 5.1. *In the notation of* (3.1) *and* (3.3)*, let* $\{\phi_i\}$ *be a basis of orthogonal polynomials on* $[-1, 1]$ *such that* H *is real and symmetric. Let* $\{\rho_j\}_{j=0}^n$ *be distinct points in* $[-1, 1]$*, and* $\{r_j\}_{j=1}^n$ *the roots of* $\phi_n(x)$*. Let* $\{\ell_j(x)\}_{j=0}^n$ *be the Lagrange polynomials defined by the nodes* $\rho_0, \ldots, \rho_n$*, and consider the matrix* L *such that* L_{ij} *contains the* i*-th coefficient*

ETNA
Kent State University and
Johann Radon Institute (RICAM)

of $\ell_j(x)$ with respect to the basis $\{\phi_i\}$. Then, the norm of the vector of coefficients of $\delta p(x)$ can be bounded by

$$\|\delta p\|_\infty \leq \|\hat{L}\|_\infty \left(\max_{j=0,\dots,n} \Gamma_1(\rho_j)\epsilon_1 + \Gamma_c(\rho_j)\epsilon_c + \Gamma_H(\rho_j)\epsilon_H \right) + \mathcal{O}(\epsilon_H^2 + \epsilon_1^2 + \epsilon_c^2),$$

where $\hat{L}$ is the matrix with the first n rows of L and $\Gamma_1, \Gamma_c, \Gamma_H$ are defined as in Lemma 3.5.

Proof. We note that $\delta p(x) = \sum_{j=0}^{n-1} \delta p_j \phi_j(x)$ is a polynomial of degree $n-1$. Its coefficients can be recovered by interpolation at the points $\{\rho_0, \dots, \rho_n\}$. Notice that these are $n+1$ points, one more than actually required. Let V_n be the $(n+1) \times (n+1)$ generalized Vandermonde matrix interpolating at these nodes in the prescribed basis. Hence, we have

$$\begin{bmatrix} \delta p_0 \\ \vdots \\ \delta p_{n-1} \\ 0 \end{bmatrix} = V_n^{-1} \begin{bmatrix} \delta p(\rho_0) \\ \delta p(\rho_1) \\ \vdots \\ \delta p(\rho_n) \end{bmatrix}.$$

Note that $L = V_n^{-1}$. Indeed, the entries of the inverse of a Vandermonde matrix are the cofficients of the Lagrange polynomials with nodes $\rho_0, \dots, \rho_n$. Therefore, we have, for $0 \leq i \leq n-1$,

$$|\delta p_i| \leq \sum_{j=1}^{n+1} |L_{i+1,j}||\delta p(\rho_j)| \leq \|\hat{L}\|_\infty \max_{0\leq j\leq n} |\delta p(\rho_j)|,$$

where, by $\hat{L}$, we denote the first n rows of L. The statement then follows by applying Lemma 3.5. ∎

5.1. Chebyshev polynomials. Chebyshev polynomials of the first kind play a special role among orthogonal polynomials on $[-1, 1]$, in particular thanks to their nice approximation properties. For instance, they are the basis of the `chebfun` MATLAB toolbox [8] that aims at making computing with functions as accessible as computing with matrices and vectors.

Their orthogonality measure is defined by the weight function $w(x) = (1-x^2)^{-\frac{1}{2}}$, and they can be obtained through the recursive relations

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), \qquad T_0(x) = 1, \qquad T_1(x) = x.$$

We denote by $U_k(z)$ the Chebyshev polynomials of the second kind, which can be obtained by replacing the degree-1 polynomial by $2x$ and keeping the rest of the recursion unchanged. The latter are orthogonal with respect to the weight $\sqrt{1-x^2}$. Moreover, $T_n'(x) = nU_{n-1}(x)$, and therefore the extrema of $T_n(x)$ are the roots of $U_{n-1}(x)$.

Our aim in this section is to apply Theorem 5.1 to Chebyshev polynomials of the first kind, making all the involved constants explicit or functions of the degree. To this aim, we need to choose the interpolation nodes, and in this case we select $\rho_j = \cos(j\pi/n)$, for $j = 0, \dots, n$, which are the roots of $U_{n-1}(x)$ (with, additionally, the points ± 1) and therefore the extrema of $T_n(x)$ on $[-1, 1]$.

LEMMA 5.2. *Let $\hat{L}$ be the matrix defined as in Theorem 5.1 choosing as $\{\phi_j\}$ the Chebyshev polynomials of the first kind and as nodes $\rho_j = \cos(j\pi/n)$, for $j = 0, \dots, n$. Then, $\|\hat{L}\|_\infty \leq 2$.*

Proof. We prove the result by showing that, for $1 \leq i \leq n$, we have $|\hat{L}_{ij}| \leq \frac{2}{n}$, if $2 \leq j \leq n$, and $|\hat{L}_{ij}| \leq \frac{1}{n}$, for $j \in \{1, n+1\}$. It immediately follows that the row sums of $|\hat{L}|$ are bounded by 2, and thus the claim holds.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

For any i, j, since $\hat{L}_{ij}$ is the Chebyshev coefficient corresponding to T_{i-1} of $\ell_{j-1}(x)$, we can recover it by writing

$$\|T_{i-1}(x)\|^2 \hat{L}_{ij} = \int_{-1}^{1} \frac{\ell_{j-1}(x) T_{i-1}(x)}{\sqrt{1-x^2}}\, dx, \qquad i, j = 1, \dots, n+1.$$

Here $\|T_{i-1}(x)\|$ denotes the norm induced by the scalar product considered above. We note that, if $2 \le j \le n$, then $\ell_{j-1}(x)$ is divisible by $(1-x)^2$, since it vanishes at ± 1. Therefore, for $1 \le j \le n-1$, we can define the degree-$(n-2)$ polynomial $q_j(x) := \ell_j(x)/(1-x^2)$ and rewrite the formula as follows:

$$\hat{L}_{ij} = \frac{1}{\|T_{i-1}(x)\|^2} \int_{-1}^{1} q_{j-1}(x) T_{i-1}(x) \sqrt{1-x^2}\, dx, \qquad 2 \le j \le n.$$

Since $\deg(q_{j-1}(x)T_{i-1}(x)) = n + i - 3 \le 2n-3$, because we are assuming $i \le n$, we can integrate the above identity exactly using a Chebyshev-Gauss quadrature formula with Chebyshev polynomials of the second kind of degree $n-1$, which yields

$$\|T_{i-1}(x)\|^2 \hat{L}_{ij} = \sum_{s=1}^{n-1} \frac{w_s}{1-x_s^2} \ell_{j-1}(x_s) T_{i-1}(x_s) = \frac{w_{j-1}}{1-x_{j-1}^2} T_{i-1}(x_{j-1}).$$

For the Chebyshev-Gauss quadrature of the second kind, the w_s are known explicitly and are $w_s = \frac{\pi}{n}(1-x_s^2)$; this, combined with $\|T_{i-1}(x)\|^2 \ge \frac{\pi}{2}$ and $|T_{i-1}(x_{j-1})| \le 1$, yields $|\hat{L}_{ij}| \le \frac{2}{n}$.

It remains to consider the case $j \in \{1, n+1\}$. Without loss of generality we can consider $j = 1$, which is associated with $\ell_0(x)$. Since $\ell_0(x)$ has as roots the zeros of $U_{n-1}(x)$ and -1, we can write it as $\ell_0(x) = \gamma(1+x)U_{n-1}(x)$ up to a scaling factor γ. The latter can be determined by imposing $\ell_0(\rho_0) = \ell_0(1) = 1$, which yields $\gamma = (2n)^{-1}$ since $U_{n-1}(\pm 1) = n$. Similarly, we can show that $\ell_n(x) = (2n)^{-1}(1-x)U_{n-1}(x)$. In addition, we may write

$$(1+x)U_{n-1}(x) = \sum_{j=0}^{n} f_j T_j(x), \qquad (1-x)U_{n-1}(x) = \sum_{j=0}^{n} (-1)^{n-j+1} f_j T_j(x),$$

where $f_j = 2$ if $1 \le j \le n-1$ and 1 if $j \in \{1, n\}$. These equalities can be easily verified using [1, (22.5.8), p. 778]. Hence, we can conclude that $|\hat{L}_{i1}| = |\hat{L}_{i,n+1}| \le \frac{1}{n}$, and therefore $\|\hat{L}\|_\infty \le (n-1)\frac{2}{n} + \frac{1}{n} + \frac{1}{n} = 2$. □

To apply Theorem 5.1 we need to obtain bounds for the constants Γ_1, Γ_c, and Γ_H, which in turn requires to bound the quantities M and S as defined in Lemma 3.5.

LEMMA 5.3. *For Chebyshev polynomials, with the notation of Lemma 3.5 and $\xi = \rho_j$ as defined in Theorem 3.1, we have*

$$M \le 3n^2, \qquad S \le 5n^2.$$

The above result is somewhat tedious to prove, so we delay the proof to Section 5.2; it allows us to state the following corollary for the case of Chebyshev polynomials. Recall that, given a monic polynomial $p(x) = \sum_{j=0}^{n} p_j T_j(x)$, the (scaled) colleague matrix is given by:

$$C = H - \frac{1}{2} e_1 c^T = \begin{bmatrix} 0 & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \ddots & & \\ & \ddots & \ddots & \frac{1}{2} & \\ & & \frac{1}{2} & 0 & \frac{\sqrt{2}}{2} \\ & & & \frac{\sqrt{2}}{2} & 0 \end{bmatrix} - \frac{1}{2} e_1 \left[p_{n-1}, \dots, p_1, \sqrt{2} p_0 \right], \tag{5.1}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

as described in Section 2.2, since for Chebyshev polynomials of the first kind we have $\alpha_n = 2, \beta_n = 0, \gamma_n - 1$, with the only exception of $\alpha_1 = 1$.

COROLLARY 5.4. *Let $C = H - \chi^{-1} e_1 c^T$ be the scaled linearization for a polynomial $p(x)$ expressed in the Chebyshev basis given by (5.1). Consider perturbations $\|\delta H\|_2 \leq \epsilon_H$, $\|\delta e_1\| \leq \epsilon_1$, and $\|\delta c\| \leq \epsilon_c$. Then, the matrix $C + \delta C := H + \delta H - \chi^{-1}(e_1 + \delta e_1)(c + \delta c)^T$ linearizes the polynomial*

$$p(x) + \delta p(x) := \sum_{j=0}^{n} (p_j + \delta p_j) T_j(x),$$

where $|\delta p_j| \leq \left(6\|c\|_2 \epsilon_1 + 2\sqrt{n}\epsilon_c + (5 + 16\sqrt{n}\|c\|_2)\epsilon_H\right) n^2 + \mathcal{O}(\epsilon_H^2 + \epsilon_1^2 + \epsilon_c^2)$.

Proof. This result follows by combining Lemma 3.5 with Theorem 3.1 and Lemma 5.3. More precisely, the bound is obtained for the coefficients of the polynomial

$$p + \delta p(x) = \sum_{j=0}^{n} (q + \delta q_j) \tilde{T}_j(x),$$

where $T_0(x) = (\sqrt{2})^{-1} T_0(x)$ and $\tilde{T}_j(x) = T_j(x)$ otherwise. Therefore, we have $\delta p_j = \delta q_j$ and $\delta p_0 = \sqrt{2}(\sqrt{2})^{-1} \delta q_0$, so in particular $|\delta p_0| \leq |\delta q_0|$. ∎

The previous result tells us that a structured QR algorithm working on the Hermitian and rank-one part separately and ensuring a low relative backward error for these two components would give a backward stable rootfinding algorithm. Indeed, in that case we would have

$$\epsilon_H \lesssim \|H\|_2 \epsilon_m, \qquad \epsilon_1 \lesssim \epsilon_m, \qquad \epsilon_c \lesssim \|c\|_2 \epsilon_m,$$

where $\lesssim$ is used to denote the first-order inequality up to a constant and a low-degree polynomial in the degree. Combining this fact with the result of Corollary 5.4 would guarantee that the backward error in the polynomial is bounded by $\|\delta p\| \lesssim (1 + \|p\|)\epsilon_m$.

Before providing the details of the proof, we check experimentally the results of Corollary 5.4 by generating polynomials expressed in the Chebyshev basis and measuring the impact of perturbing H, e_1, and c in the (scaled) colleague linearization. More precisely, we have generated polynomials $p(x) = \sum_{j=0}^{n} p_j T_j(x)$ with $n = 5$; the p_j have been specifically designed to be relatively unbalanced, a configuration that often triggers worst case behaviors in QR-based rootfinders. More specifically, we have set:

$$p_j = \begin{cases} \gamma_j \, 3^{5.5\,\eta_j} & j < n \\ 1 & j = n \end{cases}, \qquad \gamma_j, \eta_j \sim N(0,1),$$

and we have perturbed the terms H, u, v with perturbations of relative norm 10^{-6}. Our motivation for this choice of the coefficients' distribution and the perturbation norm is that we wanted to explore difficult examples and check if yet we could retrieve meaningful results in floating point arithmetic. The backward error has been computed in higher precision starting from the eigenvalues relying on the MPFR library [12]. The results, showing the actual backward errors and the bounds are reported in Figure 5.1. In each experiment we have perturbed only one of the input data H, u, v.

The bounds from Corollary 5.4 are rather descriptive for the impact of the perturbations. However, we find that for larger degrees, the quadratic terms in n tend to be pessimistic and are rarely visible in practice. On the contrary, the dependency on c is encountered in generic cases.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

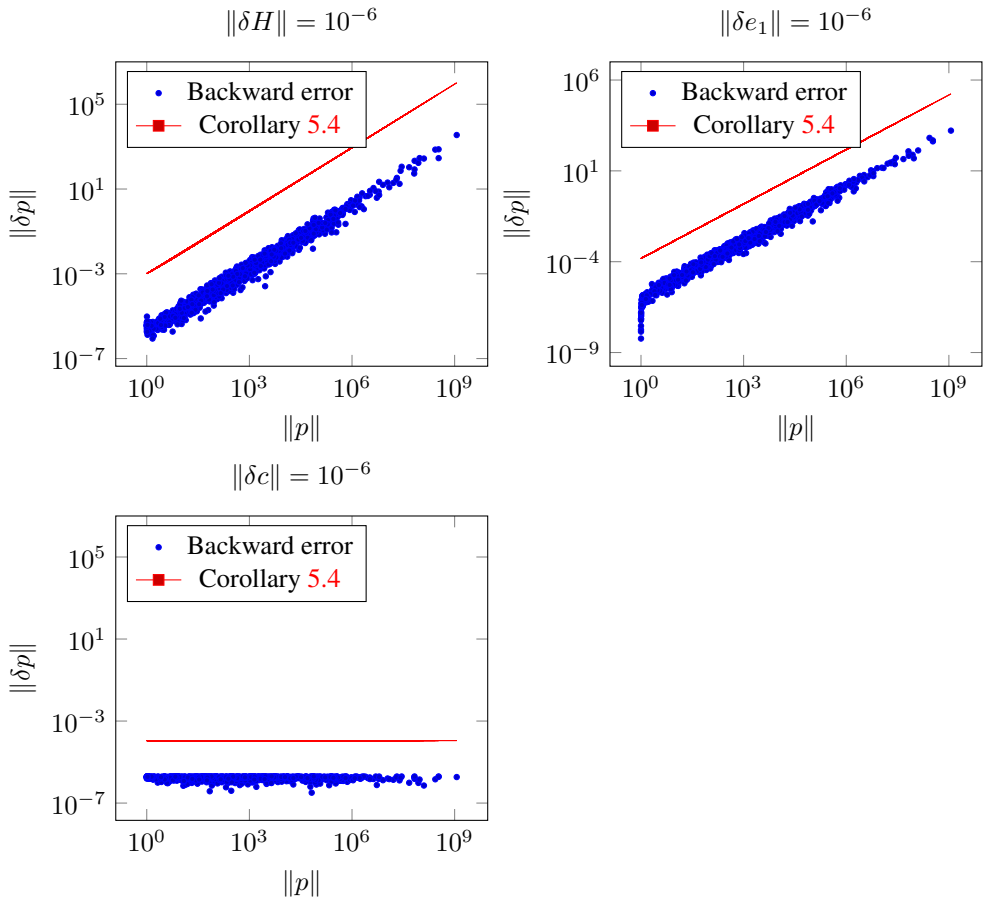

FIG. 5.1. *Experimental validation of the bounds from Corollary 5.4 for random Chebyshev polynomials with unbalanced coefficients and degree* 5. *The dependency of the error on* δc *for perturbations* δH *and* δe_1 *is clearly visible, whereas the perturbations in* c *are not influenced by the norm of the polynomial coefficients.*

5.2. Proof of Lemma 5.3. Bounding the constant M in Lemma 5.3 requires providing a lower bound for the pairwise distance between the roots of the Chebyshev polynomial of the first kind of degree n, denoted by $r_1, \dots, r_n$, and the ones of the second kind of degree $n-1$, denoted by $\rho_1, \dots, \rho_{n-1}$ extended with ± 1 as ρ_0 and ρ_n. In addition, bounding S requires an upper bound for the sum of their inverses. To obtain such results, we exploit the fact that these quantities are explicitly known:

$$r_j = \cos\left(\frac{(2j+1)\pi}{2n}\right), \quad j = 0, \dots, n-1, \qquad \rho_j = \cos\left(\frac{j\pi}{n}\right), \quad j = 0, \dots, n. \tag{5.2}$$

Before stating the main result, we need to establish a few inequalities that will be key in the proof.

LEMMA 5.5. *Let* x, y *be two positive numbers such that* $0 \le x \le \frac{\pi}{2}$ *and* $0 \le x \le y \le \pi$. *Then,*

$$\cos(x) - \cos(y) \ge \frac{4}{3\pi^2}(y^2 - x^2).$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Proof. Let us consider two separate cases; if $y \leq \frac{\pi}{2}$, then we can rewrite $\cos(x) - \cos(y)$ as

$$\cos(x) - \cos(y) = 2\sin\left(\frac{x+y}{2}\right)\sin\left(\frac{y-x}{2}\right) \geq \frac{2}{\pi^2}(y^2 - x^2),$$

where we used that $\sin(z) \geq \frac{2}{\pi}z$, for $z \in [0, \pi/2]$, and the fact that both $\frac{x+y}{2}$ and $\frac{y-x}{2}$ lie in this interval. Then, we may consider $\frac{\pi}{2} \leq y \leq \pi$. In this case, the condition $y \geq x$ is trivially satisfied, so it can be ignored. Then, we note that $\cos(z) \geq 1 - \frac{2}{\pi}z$, for $z \in [0, \frac{\pi}{2}]$, and $\cos(z) \leq 1 - \frac{2}{\pi}z$, if $z \in [\frac{\pi}{2}, \pi]$. Hence,

$$\cos(x) - \cos(y) \geq \left(1 - \frac{2}{\pi}x\right) - \left(1 - \frac{2}{\pi}y\right) = \frac{2}{\pi}(y - x), \qquad \begin{cases} 0 \leq x \leq \frac{\pi}{2}, \\ \frac{\pi}{2} \leq y \leq \pi. \end{cases}$$

Under these assumptions, we also have $(y + x) \leq \frac{3}{2}\pi$, so we can conclude that

$$\cos(x) - \cos(y) \geq \frac{2}{\pi}(y - x) = \frac{2}{\pi}\frac{y^2 - x^2}{y + x} \geq \frac{4}{3\pi^2}(y^2 - x^2).$$

Combining the inequalities obtained in the different parts of the domain yields the final result. □

LEMMA 5.6. *Let $m \geq 1, n \geq 0$ be positive integers. Then,*

$$S_1(m) := \sum_{j=1}^{m-1} \frac{1}{m^2 - j^2} \leq \frac{1}{3}, \qquad S_2(m) := \sum_{j=m+1}^{n} \frac{1}{j^2 - m^2} \leq \frac{3}{4}.$$

Proof. The inequality for $S_2(m)$ can be obtained by extending the summation to infinity and then performing a change of variable:

$$\begin{aligned} \sum_{j=m+1}^{n} \frac{1}{j^2 - m^2} \leq \sum_{j=m+1}^{\infty} \frac{1}{j^2 - m^2} &= \sum_{j=1}^{\infty} \frac{1}{(j+m)^2 - m^2} \\ &= \sum_{j=1}^{\infty} \frac{1}{j^2 + 2mj} \leq \sum_{j=1}^{\infty} \frac{1}{j^2 + 2j} = \frac{3}{4}, \end{aligned}$$

where the last equality can be obtained by proving, e.g., by induction, that the partial sums up to N of the above series are equal to $(3N^2 + 5N)/(4N^2 + 12N + 8)$. Taking the limit for $N \to \infty$ yields the desired result.

For the first inequality, we note that the summand is an increasing function in j, and therefore we can bound the summation by the integral[2]

$$\begin{aligned} \sum_{j=1}^{m-1} \frac{1}{m^2 - j^2} &= \frac{1}{2m-1} + \sum_{j=1}^{m-2} \frac{1}{m^2 - j^2} \leq \frac{1}{2m-1} + \int_0^{m-1} \frac{dx}{m^2 - x^2} \\ &= \frac{1}{2m-1} + \frac{\log(2m-1)}{2m} =: F(m). \end{aligned}$$

[2]The explicit form of the integral can be obtained using the known primitive of $\frac{1}{m^2 - x^2}$ in terms of the hyperbolic arctangent, and then using the expression of the latter by means of logarithms. The derivation is elementary but tedious, so it has been omitted.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Note that the term $\frac{1}{2m-1}$ has been removed from the integral to avoid the singularity at $x = m$. We now show that $F(m)$ is decreasing, and therefore it is sufficient to evaluate it at a certain m to obtain bounds for all $m' > m$. To this aim, we compute

$$F'(m) = \frac{-2}{(2m-1)^2} + \frac{1}{m^2}\left(\frac{m}{2m-1} - \frac{\log(2m-1)}{2}\right) = -\frac{1}{(2m-1)^2 m} - \frac{\log(2m-1)}{2m^2},$$

and it is immediate to verify that $F'(m) < 0$ for $m \geq 1$. We then substitute[3] $m = 6$, and we have

$$\sum_{j=1}^{m-1} \frac{1}{m^2 - j^2} \leq \frac{1}{11} + \frac{\log(11)}{12} \leq 0.3, \qquad m \geq 6.$$

A direct inspection shows that $S_1(2) = \frac{1}{3}$ and $S_1(m) \leq \frac{1}{3}$, for $m \in \{1, 3, 4, 5\}$. Therefore, we conclude that $S_1(m) \leq \frac{1}{3}$ for any $m \geq 1$. ∎

LEMMA 5.7. *Let $r_1, \ldots, r_n$ be the roots of $T_n(x)$, and ρ_j defined as in (5.2), and assume $n \geq 2$. Then, defining the function*

$$f_m(x) = \frac{1}{|x - r_m|},$$

we have that $f_m(\rho_j) \leq 3n^2$, for any $j = 0, \ldots, n$.

Proof. Recall that, in view of (5.2), $\rho_{j+1} \leq r_j \leq \rho_j$. Therefore, we only need to test the bound for $j \in \{m, m+1\}$. Let us consider $j = m$ first. We have:

$$f_m(\rho_m) = \frac{1}{\rho_m - r_m} = \frac{1}{\cos\left(\frac{2m}{2n}\pi\right) - \cos\left(\frac{2m+1}{2n}\pi\right)}.$$

Assume that $\frac{2m}{2n}\pi \leq \pi/2$. This is not restrictive thanks to the symmetry of the problem. Indeed, one can use the change of variable $\theta \mapsto \pi - \theta$, and reduce to the cases considered below.

Then, using Lemma 5.5 to establish a lower bound for the denominator, we obtain

$$f_m(\rho_m) \leq \frac{3\pi^2}{4\left(\frac{(2m+1)^2\pi^2}{4n^2} - \frac{(2m)^2\pi^2}{4n^2}\right)} = \frac{3n^2}{4m+1} \leq 3n^2,$$

since $m \geq 0$. The case $j = m + 1$ is completely analogous. ∎

The previous result provides a bound for the quantity M of Lemma 5.3. It is now necessary to consider the summation of $\frac{1}{|r_m - \rho_j|}$ in order to bound S.

LEMMA 5.8. *Let $r_1, \ldots, r_n$ be the roots of $T_n(x)$, ρ_j defined as in (5.2). If we define the function*

$$g(x) = \sum_{j=1}^{n} \frac{1}{|x - r_j|},$$

then $g(\rho_m) \leq 5n^2$, for any $m = 0, \ldots, n$.

Proof. As a preliminary reduction, note that it is sufficient to prove the claim under the assumption that $\rho_m \in [0, 1]$, which is equivalent to $\frac{2m}{2n}\pi \leq \frac{\pi}{2}$. Indeed, both the sets of r_m and

[3] The choice of $m = 6$ is motivated by the fact that the bound is not sharp for small values of m, so we only use it for the elements $m \geq 6$, and we verify the others by a direct computation.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

ρ_m are symmetric with respect to the imaginary axis, and therefore $g(\rho_m) = g(-\rho_m)$. We now rewrite the summation to remove the absolute values, recalling that $r_{m+1} \leq \rho_m \leq r_m$:

$$g(\rho_m) = \underbrace{\sum_{j=1}^{m-1} \frac{1}{r_j - \rho_m}}_{g_1(m)} + \underbrace{\sum_{j=m+1}^{n} \frac{1}{\rho_m - r_j}}_{g_2(m)} + f_m(\rho_m),$$

where $f_m(x)$ is defined according to Lemma 5.7. The last term can be bounded by $3n^2$. Let us consider $g_1(m)$, for which we can write, using the same arguments of Lemma 5.7 and noting that $r_j = \cos(\frac{2j+1}{2n}\pi)$ are such that $\frac{2j+1}{2n}\pi \leq \frac{2m}{2n}\pi \leq \frac{\pi}{2}$,

$$\begin{aligned} g_1(m) = \sum_{j=1}^{m-1} \frac{1}{r_j - \rho_m} &= \sum_{j=1}^{m-1} \frac{1}{\cos\left(\frac{(2j+1)\pi}{2n}\right) - \cos\left(\frac{2m\pi}{2n}\right)} \leq \sum_{j=1}^{m-1} \frac{3\pi^2}{4\left(\frac{4m^2\pi^2}{4n^2} - \frac{(2j+1)^2\pi^2}{4n^2}\right)} \\ &\leq 3n^2 \sum_{j=1}^{m-1} \frac{1}{4m^2 - (2j+1)^2} \leq 3n^2 \sum_{j=1}^{2m-1} \frac{1}{(2m)^2 - j^2} \leq n^2. \end{aligned}$$

The result concerning $g_2(m)$ can be proven by following similar steps:

$$\begin{aligned} g_2(m) = \sum_{j=m+1}^{n} \frac{1}{\rho_m - r_j} &= \sum_{j=m+1}^{n} \frac{1}{\cos(\frac{2m\pi}{2n}) - \cos(\frac{2j+1}{2n}\pi)} \\ &\leq \frac{3n^2}{4} \sum_{j=m+1}^{n} \frac{1}{j^2 - m^2} \leq \frac{9}{16} n^2, \end{aligned}$$

where once again we used Lemma 5.5, since $\frac{m}{n}\pi \leq \frac{\pi}{2}$, and then applied Lemma 5.6. Combining the bounds yields $g(m) \leq (3 + 1 + \frac{9}{16})n^2 \leq 5n^2$. ∎

5.3. The case of Jacobi polynomials. A natural extension of the approach described in Section 5.1 is to provide explicit constants for Theorem 5.1 for the Jacobi polynomials $P_k^{(\alpha,\beta)}(x)$, which are orthogonal with respect to the scalar product:

$$\langle p, q \rangle := \int_{-1}^{-1} p(x)q(x)(1-x)^\alpha(1+x)^\beta \, dx.$$

The usual normalization for Jacobi polynomials is to impose that

$$P_k^{(\alpha,\beta)}(1) = \binom{k+\alpha}{k}.$$

Note that this choice in case of $\alpha = \beta = -\frac{1}{2}$ provides a scaled version of the Chebyshev polynomials of the first kind and, when $\alpha = \beta = \frac{1}{2}$, of the ones of the second kind. In particular, Jacobi polynomials with this scaling are orthogonal but not orthonormal, and we have

$$\begin{aligned} \|P_k^{(\alpha,\beta)}\|^2 &= \int_{-1}^{1} P_k^{(\alpha,\beta)}(x)^2(1-x)^\alpha(1+x)^\beta \, dx \\ &= \frac{2^{\alpha+\beta+1}}{2k+\alpha+\beta+1} \frac{\Gamma(k+\alpha+1)\Gamma(k+\beta+1)}{\Gamma(k+\alpha+\beta+1)\Gamma(k+1)}. \end{aligned}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The recursion coefficients for Jacobi polynomials are given by (see [1, Section 22]):

$$\alpha_k = \frac{(2k+\alpha+\beta)(2k+\alpha+\beta-1)}{2k(k+\alpha+\beta)}, \qquad \beta_k = \frac{(\alpha^2-\beta^2)(2k+\alpha+\beta-1)}{2k(k+\alpha+\beta)(2k+\alpha+\beta-2)},$$
$$\gamma_k = \frac{(k+\alpha-1)(k+\beta-1)(2k+\alpha+\beta)}{k(k+\alpha+\beta)(2k+\alpha+\beta-2)}.$$

Hence, using the construction and the symmetrization procedure as in Section 2.2, we have that

$$C = \begin{bmatrix} b_n & c_{n-1} & & \\ c_{n-1} & b_{n-2} & \ddots & \\ & \ddots & \ddots & c_1 \\ & & c_1 & b_1 \end{bmatrix} - \tilde{\chi}^{-1} e_1 \left[d_1 p_{n-1}, \ldots, d_n p_0 \right],$$

where

$$b_k = \frac{\beta^2-\alpha^2}{(2k+\alpha+\beta)(2k+\alpha+\beta-2)},$$
$$c_k = \frac{2}{2k+\alpha+\beta}\sqrt{\frac{k(k+\alpha)(k+\beta)(k+\alpha+\beta)}{(2k+\alpha+\beta+1)(2k+\alpha+\beta-1)}},$$

and

$$d_k = \sqrt{\frac{\Gamma(\alpha+k)\Gamma(\beta+k)\Gamma(\alpha+\beta+2)}{(k-1)!(2k+\alpha+\beta+1)\Gamma(\alpha+1)\Gamma(\beta+1)\Gamma(\alpha+\beta+k)}}, \tag{5.3}$$

and we set $d_0 = 1$ as described in Section 2.2. We observe that $d_k = \mathcal{O}(k^{-\frac{1}{2}})$ for large k; if one was to perform the scaling of the basis numerically, this would yield the asymptotic conditioning of the task. For the degrees that are typically of practical interest, this behaviour is mild, and the scaling of the problem to get a structured matrix can be used without significantly altering the conditioning of the problem.

The following lemma will be used in the proof of Lemma 5.10, which provides the analogue result of Lemma 5.3 for Jacobi polynomials.

LEMMA 5.9. *Let $P_{n-1}^{(\alpha+1,\beta+1)}(x)$ be the Jacobi polynomial of degree $n-1$, with $\alpha, \beta \geq \frac{1}{2}$. If the coefficients f_j satisfy*

$$(1 \pm x) P_{n-1}(x)^{(\alpha+1,\beta+1)} = \sum_{j=0}^{n} f_j P_j^{(\alpha,\beta)}(x),$$

then $|f_j| \leq 6$.

Proof. We first consider the case with $(1+x)P_{n-1}^{(\alpha+1,\beta+1)}(x)$. We report the following relations among Jacobi polynomials, which can be found in [1, Section 22.7]. We have:

$$(1+x)P_{n-1}^{(\alpha+1,\beta+1)}(x) = a_n P_{n-1}^{(\alpha+1,\beta)} + b_n P_n^{(\alpha+1,\beta)}, \tag{5.4}$$
$$P_n^{(\alpha+1,\beta)}(x) = c_n P_n^{(\alpha,\beta)}(x) + d_n P_{n-1}^{(\alpha+1,\beta)}(x), \tag{5.5}$$

where $a_n = \frac{2(n+\beta)}{2n+\alpha+\beta+1}$, $b_n = \frac{2n}{2n+\alpha+\beta+1}$, $c_n = \frac{2n+\alpha+\beta+1}{n+\alpha+\beta+1}$, and $d_n = \frac{n+\beta}{n+\alpha+\beta+1}$. We note that the repeated application of (5.5) yields the following:

$$P_n^{(\alpha+1,\beta)}(x) = c_n P_n^{(\alpha,\beta)}(x) + d_n c_{n-1} P_{n-1}^{(\alpha,\beta)}(x) + d_n d_{n-1} c_{n-2} P_{n-2}^{(\alpha,\beta)}(x) + \ldots + d_n s d_1 c_0.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Combining this observation with (5.4) finally yields

$$f_j := \begin{cases} b_n c_n, & \text{if } j = n, \\ (b_n d_n + a_n) c_j \prod_{s=j+1}^{n-1} d_s, & 0 \le j \le n-1. \end{cases}$$

Thanks to our assumption that $\alpha, \beta \ge \frac{1}{2}$, we have that $|d_j| \le 1$, and in particular this implies that $f_j \le c_j(|a_j| + b_j)$. Since $1 \le c_j \le 2$, $b_j \le 1$, and $|a_j| \le 2$, we conclude that $|f_j| \le 6$. The proof for $(1-x)P_{n-1}^{(\alpha+1,\beta+1)}(x)$ is similar so we omit it. ∎

LEMMA 5.10. *Consider the nodes $\rho_0 = -1, \rho_n = 1$ and the roots ρ_j of $P_{n-1}^{(\alpha+1,\beta+1)}$ for $j = 1, \dots, n-1$. Moreover, let $\hat{L}$ be the matrix defined as in Theorem 5.1, choose the nodes as above, and let $\{\phi_j\}$ be the Jacobi polynomials $P_n^{(\alpha,\beta)}$. Then,*

$$\|\hat{L}\|_\infty \le C_n^{(\alpha,\beta)} := 12 + (n-1) \max_j \left| \frac{w_{j-1}}{1 - x_{j-1}^2} \right| \frac{2n + \alpha + \beta + 1}{2^{\alpha+\beta+1}} \binom{\alpha + \beta + n - 1}{\max\{\alpha, \beta\}},$$

where w_j and x_j are the integration weights and nodes associated with the orthogonal polynomial $P_{n-1}^{(\alpha+1,\beta+1)}(x)$.

Proof. The proof follows the same strategy and uses the same notation of the one given for Chebyshev polynomials of the first kind. We have that

$$\|P_{i-1}^{(\alpha,\beta)}(x)\|^2 \hat{L}_{ij} = \int_{-1}^{1} \ell_{j-1}(x) P_{i-1}^{(\alpha,\beta)}(x)(1-x)^\alpha (1+x)^\beta \, dx, \qquad i, j = 1, \dots, n+1.$$

If $2 \le j \le n$, then $\ell_{j-1}(x)$ is divisible by $(1-x)^2$ since it vanishes at ± 1. Therefore, for $1 \le j \le n-1$, we can define the degree-$(n-2)$ polynomial $q_j(x) := \ell_j(x)/(1-x^2)$ and rewrite the formula as follows:

$$\hat{L}_{ij} = \frac{1}{\|P_{i-1}^{(\alpha,\beta)}(x)\|^2} \int_{-1}^{1} q_{j-1}(x) P_{i-1}^{(\alpha,\beta)}(x)(1-x)^{\alpha+1}(1+x)^{\beta+1} \, dx, \qquad 2 \le j \le n.$$

Since $\deg(q_{j-1}(x)P_{i-1}^{(\alpha,\beta)}(x)) = n + i - 3 \le 2n - 3$, because we are assuming $i \le n$, we can integrate the above exactly using the Jacobi-Gauss quadrature formula associated with the orthogonal polynomials $P_n^{(\alpha+1,\beta+1)}$, which yields

$$\|P_{i-1}^{(\alpha,\beta)}(x)\|^2 \hat{L}_{ij} = \sum_{s=1}^{n-1} \frac{w_s}{1 - x_s^2} \ell_{j-1}(x_s) P_{i-1}^{(\alpha,\beta)}(x_s) = \frac{w_{j-1}}{1 - x_{j-1}^2} P_{i-1}^{(\alpha,\beta)}(x_{j-1}).$$

Hence, we have that

$$\begin{aligned} |\hat{L}_{ij}| &\le \max_j \left| \frac{w_{j-1}}{1 - x_{j-1}^2} \right| \binom{\max\{\alpha, \beta\} + i - 1}{i - 1} \frac{2i + \alpha + \beta - 1}{2^{\alpha+\beta+1}} \frac{\Gamma(i + \alpha + \beta)\Gamma(i)}{\Gamma(i + \alpha)\Gamma(i + \beta)} \\ &= \max_j \left| \frac{w_{j-1}}{1 - x_{j-1}^2} \right| \frac{2i + \alpha + \beta + 1}{2^{\alpha+\beta+1}} \binom{\alpha + \beta + i - 1}{\max\{\alpha, \beta\}}. \end{aligned}$$

It remains to consider the case $j \in \{1, n+1\}$. We can consider $j = n+1$ first, which is associated with $\ell_n(x)$. Since $\ell_n(x)$ has as roots the zeros of $P_{n-1}^{(\alpha+1,\beta+1)}(x)$ and -1, we can write it as $\ell_n(x) = \gamma(1+x)P_{n-1}^{(\alpha+1,\beta+1)}(x)$ up to a scaling factor γ. The latter

ETNA
Kent State University and
Johann Radon Institute (RICAM)

can be determined by imposing $\ell_n(\rho_n) = \ell_n(1) = 1$, which yields $\gamma = \frac{\Gamma(\alpha+1)\Gamma(n)}{2\Gamma(\alpha+n)}$ since $P_{n-1}^{(\alpha+1,\beta+1)}(1) = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha+1)\Gamma(n)}$. Similarly, $\ell_0(x) = (-1)^n \frac{\Gamma(\beta+1)\Gamma(n)}{2\Gamma(\beta+n)}(1-x)P_{n-1}^{(\alpha+1,\beta+1)}(x)$. In addition, we may write

$$(1+x)P_{n-1}^{(\alpha+1,\beta+1)}(x) = \sum_{j=0}^{n} f_j P_j^{(\alpha,\beta)}(x), \qquad (1-x)P_{n-1}^{(\alpha+1,\beta+1)}(x) = \sum_{j=0}^{n} g_j P_j^{(\alpha,\beta)}(x),$$

where $|f_j|, |g_j| \le 6$ in view of Lemma 5.9. Hence, we can conclude that $|\hat{L}_{i1}| + |\hat{L}_{i,n+1}| \le 12$, and therefore

$$\|\hat{L}\|_\infty \le 12 + (n-1)\max_j \left| \frac{w_{j-1}}{1-x_{j-1}^2} \right| \frac{2n+\alpha+\beta+1}{2^{\alpha+\beta+1}} \binom{\alpha+\beta+n-1}{\max\{\alpha,\beta\}}. \qquad \square$$

In fact, we cannot directly use Lemma 5.10 as we are working with the scaled basis $d_{n-i+1}^{-1} P_{i-1}^{(\alpha,\beta)}$. In other words, we actually need a bound for $\|D\hat{L}\|_\infty$, D being the diagonal scaling matrix $D = \operatorname{diag}(d_1, \ldots, d_n)$. This is readily obtained as $\|D\hat{L}\|_\infty \le \|D\|\|\hat{L}\|_\infty$, with $\|D\| = \max_{1\le i\le n} d_i$.

REMARK 5.11. We note that $C_n^{(\alpha,\beta)}$ involves the quantity $\mu_n^{(\alpha,\beta)} := \max_j \left| \frac{w_{j-1}}{1-x_{j-1}^2} \right|$. Observe that $\mu_n^{(-\frac{1}{2},-\frac{1}{2})} = \frac{\pi}{n}$, and this fact is used in the proof of Lemma 5.2. For other Jacobi polynomials, numerical experiments suggest that, at least if $\alpha = \beta$, then $\mu_n^{(\alpha,\beta)} \approx \frac{\pi}{n+\alpha+\frac{1}{2}}$. We are not aware of a proof of this conjecture; some asymptotic results in this direction can be found in [21].

In order to provide the final result for Jacobi polynomials, we need the analogue of Lemma 5.3 that is stated for Chebyshev polynomials.

LEMMA 5.12. *For Jacobi polynomials $P_n^{(\alpha,\beta)}$, with the notation of Lemma 3.5 and $\xi = \rho_j$ as defined in Theorem 3.1, there exist two moderate constants η_M and η_S, depending on α, β, such that*

$$M \le \eta_M n^2, \qquad S \le \eta_S n^3.$$

Proof. In view of the Frenzen-Wong formula [13] we may write the roots of $P_n^{(\alpha,\beta)}$ as $\cos(\theta_{n,k})$, with

$$\theta_{n,k} = t_k + \frac{1}{N^2}\left(\left(\alpha^2 - \frac{1}{4}\right)\frac{1 - t_k\cot(t_k)}{2t_k} - \frac{\alpha^2-\beta^2}{4}\tan\left(\frac{t_k}{2}\right)\right) + \mathcal{O}(n^{-3}),$$

where $t_k := \frac{j_{\alpha,k}}{N}$, $N := n + \frac{\alpha+\beta+1}{2}$, and $j_{\alpha,k}$ are the positive roots of the Bessel function $J_\alpha(x)$. We now estimate the distance between consecutive roots by writing

$$\theta_{n,k+1} - \theta_{n,k} = t_{k+1} - t_k + \frac{1}{N^2}h(t_k, t_{k+1}) + \mathcal{O}(n^{-3}),$$

where $h(\cdot,\cdot)$ collects the terms in front of $\frac{1}{N^2}$ in the difference. It is known that the roots $j_{\alpha,k}$ of the Bessel function $J_\alpha(x)$ are simple and asymptotically distributed in such a way that $j_{\alpha,k+1} - j_{\alpha,k} \sim \pi$ for $k \to \infty$, and the smallest root is strictly positive. Hence, we observe that we can give an inclusion for t_k of the form

$$t_k \in \left[\frac{C_{\min}}{n}, \pi - \frac{C_{\max}}{n}\right], \qquad C_{\min}, C_{\max} > 0.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The constants above only depend on α, β and not on n. In addition, since the roots are well separated for $k \to \infty$, we may set $\gamma_\alpha := \inf_k |j_{\alpha,k+1} - j_{\alpha,k}| > 0$.

We now note that, to bound the separation, it is sufficient to consider $k = 1, \dots, \lceil \frac{n}{2} \rceil$; for the other roots, we may just swap the role of α, β and apply the same argument. For such k, t_k belongs to (assuming $n \geq 2$) the interval $[\frac{C_{\min}}{n}, \frac{3\pi}{4}]$, and therefore it is immediate to verify that the coefficient of the $\frac{1}{N^2}$-term in the Frenzen-Wong formula is uniformly bounded. Hence, for $n \to \infty$, we have

$$\theta_{n,k+1} - \theta_{n,k} \geq \frac{2\max\{\gamma_\alpha, \gamma_\beta\}}{2n + \alpha + \beta + 1} + \mathcal{O}(n^{-2}).$$

Evaluating the cosine at these angles yields that, for large n,

$$|\cos(\theta_{n,k+1}) - \cos(\theta_{n,k})|^{-1} \sim \mathcal{O}(n^2).$$

Hence, there exists a constant η_M that uniformly bounds the above quantity for all n, which allows us to derive the first bound of the Lemma. For the second, it is sufficient to sum all these bounds over all $k' \neq k$. Note that, by construction, the constants η_M, η_S do not depend on n but only on α, β. ∎

We remark that we doubt that this bound is optimal for η_S: we conjecture that a clever analysis of the bounds would lead, using similar techniques as the ones in Lemma 5.3, to control the growth of S quadratically in n. We leave the analysis of this conjecture as an open problem.

Combining Lemma 5.10 with Lemma 5.12 yields the following result.

COROLLARY 5.13. *Let $C = H - \chi^{-1} e_1 c^T$ be the linearization for a polynomial $p(x)$ expressed in the scaled Jacobi basis $d_{n-j}^{-1} P_j^{(\alpha,\beta)}$, for $j = 0, \dots, n$, where d_j are defined in (5.3). Consider perturbations $\|\delta H\|_2 \leq \epsilon_H$, $\|\delta e_1\| \leq \epsilon_1$, and $\|\delta c\| \leq \epsilon_c$. Then, the matrix $C + \delta C := H + \delta H - \tilde{\chi}^{-1}(e_1 + \delta e_1)(c + \delta c)^T$ linearizes the polynomial*

$$p(x) + \delta p(x) := \sum_{j=0}^{n} (p_j + \delta p_j) d_{n-j}^{-1} P_j^{(\alpha,\beta)}(x),$$

where

$$|\delta p_j| \leq \eta_D \hat{C}_n^{(\alpha,\beta)} \left(\eta_M \|c\|_2 \epsilon_1 n^2 + \tilde{\chi} \epsilon_c n^{\frac{5}{2}} + (\eta_S n^3 + (\eta_M + \eta_S)\tilde{\chi} n^{\frac{7}{2}}) \|c\|_2 \epsilon_H \right) + \mathcal{O}(\epsilon_H^2 + \epsilon_1^2 + \epsilon_c^2),$$

$\hat{C}_n^{\alpha,\beta} = C_n^{(\alpha,\beta)} \binom{\max\{\alpha,\beta\}+n}{n}$, with $C_n^{(\alpha,\beta)}$ defined as in Lemma 5.10 and $\eta_D := \frac{\max_j d_j}{\min_j d_j}$.

Proof. The result follows by applying Theorem 5.1 together with Lemmas 5.10 and 5.12 and using the fact that

$$|d_{n-j}^{-1} P_j^{(\alpha,\beta)}(x)| \leq \frac{\binom{\max\{\alpha,\beta\}+n}{n}}{\min_j d_j}.$$ ∎

6. Conclusions. We have presented a backward error analysis applicable to computing roots of polynomials through structured QR solvers. The results cover the cases where the error has the same normal-plus-rank-one structure of the confederate matrix, and the backward errors on the various parts have different magnitudes.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

This often happens in practice when the structure is exploited, as in the algorithm presented in [5] for the monomial case. We have provided an alternative derivation that recovers the results of the stability analysis in [5].

These results have then been extended to the Chebyshev and Jacobi bases with explicit bounds provided. This indicates the requirements that a QR-based rootfinder in these bases needs to have in order to obtain a stable rootfinding algorithm.

Some related topics might be subject to future investigation. For instance, an algorithm for symmetric-plus-rank-one matrices arising from polynomial rootfinding, satisfying the proposed stability constraints, does not exist yet. Our hope is that this paper suggests research directions to develop one.

Another research line stemming from this analysis is extending the results to the case of matrix polynomials. Polynomial eigenvalue problems can be solved using unitary-plus-low-rank solvers in the monomial basis [4] or symmetric-plus-low-rank ones for more general bases [10]. However, the use of the determinant to recover the linearized polynomial is not applicable in the matrix polynomial setting, and other more involved questions, such as the accurate (stable) computation of the eigenvectors, are of interest as well.

REFERENCES

[1] M. ABRAMOWITZ AND I. A. STEGUN, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, vol. 55 of National Bureau of Standards Applied Mathematics Series, U.S. Government Printing Office, Washington, D.C., 1964.

[2] J. L. AURENTZ, T. MACH, L. ROBOL, R. VANDEBRIL, AND D. S. WATKINS, *Core-chasing Algorithms for the Eigenvalue Problem*, SIAM, Philadelphia, 2018.

[3] ———, *Fast and backward stable computation of roots of polynomials, part II: backward error analysis; companion matrix and companion pencil*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 1245–1269.

[4] ———, *Fast and backward stable computation of the eigenvalues of matrix polynomials*, Math. Comp., 88 (2010), pp. 313–347.

[5] J. L. AURENTZ, T. MACH, R. VANDEBRIL, AND D. S. WATKINS, *Fast and backward stable computation of roots of polynomials*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 942–973.

[6] S. BARNETT, *Polynomials and Linear Control Systems*, Marcel Dekker, New York, 1983.

[7] F. DE TERÁN, F. M. DOPICO, AND J. PÉREZ, *Backward stability of polynomial root-finding using Fiedler companion matrices*, IMA J. Numer. Anal., 36 (2016), pp. 133–173.

[8] T. A. DRISCOLL, N. HALE, AND L. N. TREFETHEN, *Chebfun Guide*, Pafnuty Publications, Oxford, 2014.

[9] A. EDELMAN AND H. MURAKAMI, *Polynomial roots from companion matrix eigenvalues*, Math. Comp., 64 (1995), pp. 763–776.

[10] Y. EIDELMAN, L. GEMIGNANI, AND I. GOHBERG, *Efficient eigenvalue computation for quasiseparable Hermitian matrices under low rank perturbations*, Numer. Algorithms, 47 (2008), pp. 253–273.

[11] M. FIEDLER, *A note on companion matrices*, Linear Algebra Appl., 372 (2003), pp. 325–331.

[12] L. FOUSSE, G. HANROT, V. LEFÈVRE, P. PÉLISSIER, AND P. ZIMMERMANN, *MPFR: a multiple-precision binary floating-point library with correct rounding*, ACM Trans. Math. Software, 33 (2007), Art. No. 13, 15 pages.

[13] L. GATTESCHI, *On the zeros of Jacobi polynomials and Bessel functions*, in International Conference on Special Functions: Theory and Computation (Turin, 1984), Rend. Sem. Mat. Univ. Politec. Torino, 1985, Special Issue, Università e Politecnico di Torino, Turin, 1985, pp. 149–177.

[14] L. GEMIGNANI AND L. ROBOL, *Fast Hessenberg reduction of some rank structured matrices*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 574–598.

[15] P. W. LAWRENCE AND R. M. CORLESS, *Stability of rootfinding for barycentric Lagrange interpolants*, Numer. Algorithms, 65 (2014), pp. 447–464.

[16] P. W. LAWRENCE, M. VAN BAREL, AND P. VAN DOOREN, *Backward error analysis of polynomial eigenvalue problems solved by linearization*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 123–144.

[17] D. S. MACKEY, N. MACKEY, C. MEHL, AND V. MEHRMANN, *Vector spaces of linearizations for matrix polynomials*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 971–1004.

[18] Y. NAKATSUKASA AND V. NOFERINI, *On the stability of computing polynomial roots via confederate linearizations*, Math. Comp., 85 (2016), pp. 2391–2425.

[19] Y. NAKATSUKASA, V. NOFERINI, AND A. TOWNSEND, *Vector spaces of linearizations for matrix polynomials: a bivariate polynomial approach*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 1–29.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[20] V. NOFERINI AND J. PÉREZ, *Chebyshev rootfinding via computing eigenvalues of colleague matrices: when is it stable?*, Math. Comp., 86 (2017), pp. 1741–1767.
[21] P. OPSOMER, *Asymptotics for Orthogonal Polynomials and High-Frequency Scattering Problems*, PhD. Thesis, Department of Computer Science, KU Leuven, Leuven, 2018.
[22] G. SZEGÖ, *Orthogonal Polynomials*, AMS, Providence, 1975.
[23] L. N. TREFETHEN AND ET AL., *Chebfun version 6*, 2017.
[24] P. VAN DOOREN AND P. DEWILDE, *The eigenstructure of an arbitrary polynomial matrix: computational aspects*, Linear Algebra Appl., 50 (1983), pp. 545–579.
[25] D. S. WATKINS, *The Matrix Eigenvalue Problem*, SIAM, Philadelphia, 2007.