SIAM J. Matrix Anal. Appl.
Vol. 39, No. 1, pp. 148–177

REDUCTION OF MATRIX POLYNOMIALS TO SIMPLER FORMS*

YUJI NAKATSUKASA†, LEO TASLAMAN‡, FRANÇOISE TISSEUR‡, AND ION ZABALLA§

Abstract. A square matrix can be reduced to simpler form via similarity transformations. Here "simpler form" may refer to diagonal (when possible), triangular (Schur), or Hessenberg form. Similar reductions exist for matrix pencils if we consider general equivalence transformations instead of similarity transformations. For both matrices and matrix pencils, well-established algorithms are available for each reduction, which are useful in various applications. For matrix polynomials, unimodular transformations can be used to achieve the reduced forms but we do not have a practical way to compute them. In this work we introduce a practical means to reduce a matrix polynomial with nonsingular leading coefficient to a simpler (diagonal, triangular, Hessenberg) form while preserving the degree and the eigenstructure. The key to our approach is to work with structure preserving similarity transformations applied to a linearization of the matrix polynomial instead of unimodular transformations applied directly to the matrix polynomial. As an application, we illustrate how to use these reduced forms to solve parameterized linear systems.

Key words. triangularization, matrix polynomial, quasi-triangular, diagonalization, Hessenberg form, companion linearization, controller form linearization, equivalence, quadratic eigenvalue problem, Schur form, parameterized linear systems

AMS subject classifications. 15A18, 15A22, 65F15

DOI. 10.1137/17M1125182

1. Introduction.

Almost all matrices in $\mathbb{C}^{n\times n}$ can be reduced to diagonal form via a similarity transformation. (The exceptions constitute the measure-zero set of defective matrices.) Furthermore, all matrices in $\mathbb{C}^{n\times n}$ can be reduced to triangular and upper Hessenberg form via unitary similarity transformations. For matrices in $\mathbb{R}^{n\times n}$, we have similar results with the difference that we now have quasi-diagonal and quasi-triangular forms instead of diagonal and triangular forms. Here the prefix "quasi" means that all diagonal blocks are either of size 1×1 or 2×2. Now, consider matrix polynomials with a nonsingular leading coefficient

$$P(\lambda) = \lambda^\ell P_\ell + \cdots + \lambda P_1 + P_0 \quad \text{with} \quad \det(P_\ell) \neq 0, \tag{1}$$

over $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$. Is it possible to reduce such matrix polynomials to the simpler forms mentioned above while preserving the degree and the eigenstructure, that is, the eigenvalues and their partial multiplicities? If we use only similarity transformations, the answer is, in general, no. Even if we use the broader class of strict equivalence transformations, that is, multiplication by nonsingular matrices from left and right, it

*Received by the editors April 11, 2017; accepted for publication (in revised form) October 27, 2017; published electronically January 30, 2018.
http://www.siam.org/journals/simax/39-1/M112518.html
Funding: The work of the first author was supported by JSPS as an Overseas Research Fellow. The work of the second author was supported by EPSRC grant EP/I005293. The work of the third author was supported by EPSRC grant EP/I005293 and by a Royal Society-Wolfson Research Merit Award. The work of the fourth author was supported by the Dirección General de Investigación Proyecto de Investigación MTM2013-40960-P and Red de Excelencia MTM2015-68805-REDT and Subvención a Grupos UPV-EHU GIU16/42.
†Mathematical Institute, University of Oxford, Oxford, OX2 6GG, UK (Yuji.Nakatsukasa@maths.ox.ac.uk).
‡School of Mathematics, The University of Manchester, Manchester, M13 9PL, UK (leotaslaman@gmail.com, francoise.tisseur@manchester.ac.uk).
§Departamento de Matemática Aplicada y EIO, Euskal Herriko Univertsitatea (UPV/EHU), Apdo. Correos 644, 48080 Bilbao, Spain (ion.zaballa@ehu.eus).

is, in general, not possible. Indeed, if there were to exist nonsingular matrices E and F such that $EP(\lambda)F = T(\lambda)$ is triangular, say, of degree $\ell > 1$ and with $\det P_\ell \neq 0$ then the family of matrices $(P_\ell^{-1}P_{\ell-1}, \ldots, P_\ell^{-1}P_1, P_\ell^{-1}P_0)$ would be simultaneously triangularizable by similarity. This would imply (see, for example, [7, Thms. 2.4.8.6 and 2.4.8.7]) that for all $i \neq j$, $i, j = 0, 1, \ldots, \ell-1$, the eigenvalues of $P_\ell^{-1}P_iP_\ell^{-1}P_j - P_\ell^{-1}P_jP_\ell^{-1}P_i$ are all equal to zero. This is a very restrictive condition.

A type of transformation that gives us a sufficient amount of freedom while preserving the eigenstructure is multiplication by unimodular matrix polynomials. A matrix polynomial $U(\lambda) \in \mathbb{F}[\lambda]^{n\times n}$ is said to be unimodular if $\det U(\lambda) \in \mathbb{F} \setminus \{0\}$, and two matrix polynomials that differ only by multiplication by unimodular matrix polynomials (from the left and the right) are said to be *equivalent*. It was shown in [16] and [17] that unimodular transformations are enough to reduce any square matrix polynomial to triangular form over $\mathbb{C}$ and quasi-triangular form over $\mathbb{R}$, while preserving the degree. Of course, this includes the case of Hessenberg form since (quasi-) triangular matrices are also Hessenberg. Further, it is a straightforward exercise to show that any complex/real matrix polynomial with semisimple eigenstructure is equivalent to a diagonal/quasi-diagonal matrix polynomial of the same degree.

The reduction to diagonal form has applications in structural engineering, where it has been used to decouple systems of second-order differential equations (see, for example, [3] and [12]). In applications where parametrized linear systems of the form $P(\omega)x = b(\omega)$ with P as in (1) need to be solved for many values of ω over a large range, it may be useful to first reduce P to simpler form before solving the linear systems (see section 5).

How can we compute these simpler forms in practice? The approach taken in the recent paper [2] is to define a pseudoinner product on the vector space $\mathbb{F}(\lambda)^n$ where $\mathbb{F}(\lambda)$ is the field of rational functions. Then a Krylov-like subspace method is applied to any matrix polynomial to reduce it to Hessenberg form. In general, the entries in this Hessenberg matrix are rational functions. On the other hand, the discussions in [4, Thm. 1.7], [16] are based on applying unimodular transformations to the Smith form, and its numerical implementation is nontrivial. To avoid working with unimodular transformations, which, in general, affect the degree, we use linearizations. Recall that a pencil $\lambda I - A$ is a monic linearization of the matrix polynomial $P(\lambda) \in \mathbb{F}[\lambda]^{n\times n}$ in (1) if $A \in \mathbb{F}^{\ell n\times \ell n}$ and $\lambda I - A$ has the same elementary divisors as $P(\lambda)$. Suppose $P(\lambda)$ has the same eigenstructure as the monic matrix polynomial $R(\lambda) = \lambda^\ell I + \sum_{j=0}^{\ell-1} \lambda^j R_j$ and take any monic linearization $\lambda I - A$ of $P(\lambda)$. Note that $\lambda I - A$ is also a linearization of $R(\lambda)$. The Gohberg, Lancaster, Rodman theory [4, sect. 1.10] tells us that there is an $\ell n \times n$ matrix X such that (A, X) is a left standard pair for $R(\lambda)$, that is, the $\ell n \times \ell n$ matrix

$$S = [X \;\; AX \;\; \ldots \;\; A^{\ell-1}X] \tag{2}$$

is nonsingular and

$$A^\ell X + A^{\ell-1}XR_{\ell-1} + \cdots + AXR_1 + XR_0 = 0. \tag{3}$$

Taken together, (2) and (3) can be rewritten as

$$S^{-1}AS = \begin{bmatrix} & & & -R_0 \\ I & & & -R_1 \\ & \ddots & & \vdots \\ & & I & -R_{\ell-1} \end{bmatrix} =: C_L(R) \tag{4}$$

```
n = 5; deg = 3;  % size and degree
P0 = randn(n); P1 = randn(n); P2 = randn(n); % coefficient matrices
C_P = [ zeros(n) zeros(n) -P0 % left companion form
        eye(n)   zeros(n) -P1
        zeros(n) eye(n)   -P2 ]
[U,~] = schur(C_P,'complex');
X = U*kron(eye(n), ones(deg,1));
S = [X C_P*X C_P^(deg-1)*X];
C_R = (S\C_P)*S;
spy(abs(C_R)>1e-12)
```

FIG. 1. *Basic MATLAB M-file that generates a random monic cubic matrix polynomial, then computes the left companion form of an equivalent triangular matrix polynomial, and plots its (numerical) nonzero pattern.*

showing that A is similar to the left companion matrix associated with $R(\lambda)$. Actually, for any given monic linearization $\lambda I - A$ of $P(\lambda)$ and any nonsingular matrix S of the form (2), $S^{-1}AS$ will always be the left companion matrix of some matrix polynomial, as in (4). This matrix polynomial $R(\lambda) = \lambda^\ell I + \lambda^{\ell-1}R_{\ell-1} + \cdots + \lambda R_1 + R_0$ will have the same degree and eigenstructure as $P(\lambda)$. The above discussion suggests that in order to reduce $P(\lambda)$ in (1) to a simpler form, it is enough to find an $n\ell \times n$ matrix X such that S in (2) is nonsingular and $S^{-1}AS$ has the desired zero pattern in the coefficient matrices (in the last block column), where A can be any matrix such that $\lambda I - A$ is a linearization of $P(\lambda)$. One of the main contributions in this paper is to give a characterization of such a matrix X in terms of block Krylov subspaces (see section 2).

In the generic case, when all the eigenvalues are distinct, it turns out to be surprisingly easy to find X such that $S^{-1}AS$ is the left companion matrix of a matrix polynomial in triangular, diagonal, or Hessenberg form. We illustrate this with a snippet of MATLAB code in Figure 1. If we replace `schur(C_P,'complex')` by `eig(C_P)`, then `C_R` becomes the companion matrix of an equivalent diagonal matrix polynomial, and if we replace `schur(C_P,'complex')` by `hess(C_P)` and `ones(deg,1)` by `eye(deg,1)`, then `C_R` becomes the companion matrix of an equivalent matrix polynomial in Hessenberg form. The code can be generalized to any degree and works as long as the block Krylov matrix `S` on line 8 is nonsingular, which it is for almost all coefficient matrices, as we will see in section 3.2. A colored spy plot from one execution of the MATLAB code in Figure 1 is shown on the left of Figure 2. The other plots correspond to the diagonal reduction (middle plot) and the Hessenberg reduction (right plot). We remark that the reduction to Hessenberg form requires no iterative process (such as computing the eigenvalues) and uses a fixed number of arithmetic operations. Our reduction gives a Hessenberg matrix polynomial with all but the second leading coefficient being triangular.

The code in Figure 1 is not meant to be a numerically efficient or stable algorithm. Although the threshold specified in the last line of the M-file works for all tried monic matrix polynomials with randomly chosen coefficients, the choice of X is by no means unique and likely improvable. Also, we need only the last block column of $C_L(R)$ and there may be more efficient ways to compute it. Nevertheless, the code suggests a possible practical procedure to reduce $P(\lambda)$ in (1) to triangular form while preserving its degree and eigenstructure when all eigenvalues are distinct. In this paper we

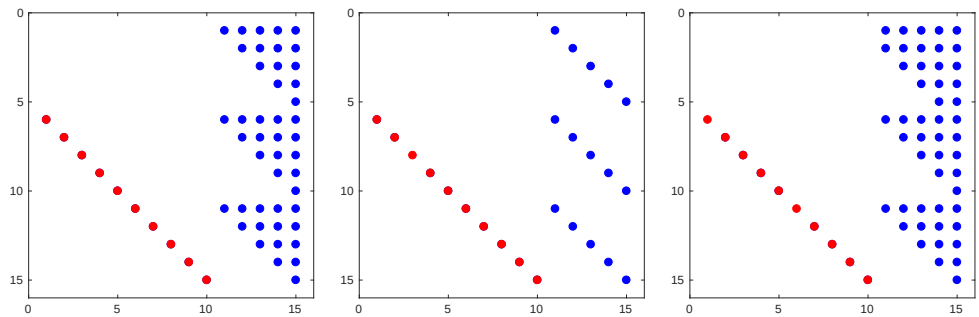

FIG. 2. *Colored spy plots of the left companion linearization of the reduced matrix polynomials (size $n = 5$, degree $\ell = 3$) obtained by the MATLAB code in Figure 1 (or its modification as explained in the text): triangular (left), diagonal (middle), and Hessenberg (right). The red dots are 1s, the blue dots are nonzero entries of the coefficient matrices. (Figure is in color online.)*

discuss why and when this procedure works. We also want to allow the monic matrix polynomials to have multiple eigenvalues and to use real arithmetic if the given matrix polynomial is real. In these cases, additional computations to those described in the code of Figure 1 may be needed. To be precise, one of the main goals of this work is to give a practical procedure to reduce any $P(\lambda)$ in (1) to triangular or quasi-triangular form according as $\mathbb{F} = \mathbb{C}$ or $\mathbb{R}$, while preserving its degree and eigenstructure. The proposed procedure consists of the following steps:

1. Choose a monic linearization $\lambda I - A$ of $P(\lambda)$.
2. Compute a real or complex Schur form, T_0, of A according as $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$.
3. Reorder the diagonal entries of T_0 and, in the real case, the 2×2 blocks along its diagonal to produce a new Schur form $T = U^* AU$ that can be split into blocks that are suited to construct the matrix X of the next step.
4. Use U and the diagonal blocks of T to produce a matrix $X \in \mathbb{F}^{\ell n \times n}$ of full column rank such that S in (2) is nonsingular and $S^{-1}AS$ is the left companion matrix of a monic upper triangular matrix polynomial.
5. Compute $S^{-1}A^\ell X$, i.e., the last block column of $C_L(R)$ in (4), and extract the blocks R_j, $j = 0, \ldots, \ell - 1$ defining $R(\lambda) = \lambda^\ell I + \lambda^{\ell-1} R_{\ell-1} + \cdots + \lambda R_1 + R_0$.

The matrix polynomial $R(\lambda)$ will be upper (quasi-) triangular and have the same eigenstructure as $P(\lambda)$. We remark that the structure of A and S can be exploited to compute $S^{-1}A^\ell X$ at a reduced cost in step 5, but this is outside the scope of this work. Notice also that we could replace A in steps 4 and 5 by the Schur form T obtained in step 3. Nevertheless, the analysis of how to implement steps 4 and 5 in a numerically reliable and efficient way will remain as an open problem.

We will show in section 3 how to implement step 3 in a numerically stable manner when all the eigenvalues of $P(\lambda)$ have algebraic multiplicity not greater than n (the size of $P(\lambda)$). The matrices X that are used to implement step 4 are characterized in section 2; a method to obtain them explicitly is provided in section 3. It will also be shown in that section that X can be constructed to have orthogonal columns. As mentioned above, it is left as an open problem how to obtain it in such a way that step 5 can be computed in a numerically stable manner. The quadratic case ($\ell = 2$) is fully examined in section 4, where a stable way of implementing step 3 is given that works independently of the algebraic multiplicity of the eigenvalues of $P(\lambda)$.

To be slightly more general, we will also study how to construct matrices X to reduce $P(\lambda)$ to one of the following forms:

- block-diagonal form:

$$D(\lambda) = D_1(\lambda) \oplus D_2(\lambda) \oplus \cdots \oplus D_k(\lambda) \in \mathbb{F}[\lambda]^{n\times n} \tag{5}$$

monic of degree ℓ with $D_i(\lambda) \in \mathbb{F}[\lambda]^{s_i\times s_i}$, $1 \le i \le k$ and $s_1 + \cdots + s_k = n$,

- block-triangular form:

$$T(\lambda) = \begin{bmatrix} T_{11}(\lambda) & T_{12}(\lambda) & \dots & T_{1k}(\lambda) \\ & T_{22}(\lambda) & & \vdots \\ & & \ddots & \vdots \\ & & & T_{kk}(\lambda) \end{bmatrix} \in \mathbb{F}[\lambda]^{n\times n}, \tag{6}$$

monic of degree ℓ with $T_{jj}(\lambda) \in \mathbb{F}[\lambda]^{s_j\times s_j}$, $1 \le j \le k$, and $s_1 + \cdots + s_k = n$, and

- Hessenberg form:

$$H(\lambda) = \lambda^\ell I + \lambda^{\ell-1}H_{\ell-1} + \cdots + \lambda H_1 + H_0 \in \mathbb{F}[\lambda]^{n\times n}, \tag{7}$$

with coefficient matrices H_i, $i = 0, \dots, \ell-1$, in Hessenberg form.

We will discuss in section 5 how to use the simpler forms to solve parameterized linear systems $P(\omega)x = b(\omega)$, where x is to be computed for many values of the parameter ω.

2. Conditions for reduction to simpler forms. For matrices $A \in \mathbb{F}^{m\times m}$ and $V \in \mathbb{F}^{m\times j}$ we define the block Krylov matrix

$$K_\ell(A, V) = [V \ AV \ \cdots \ A^{\ell-1}V] \in \mathbb{F}^{m\times \ell j}$$

and the block Krylov subspace

$$\mathcal{K}_\ell(A, V) = \text{range} K_\ell(A, V).$$

For a subspace $\mathcal{X}$ of $\mathbb{F}^m$ and a matrix A operating on that subspace, we define $A\mathcal{X} = \{Ax : \ x \in \mathcal{X}\}$.

Assume that $P(\lambda)$ is given by (1), and let $\lambda I - A$ be any monic linearization of $P(\lambda)$, for example, the left companion linearization of $P_\ell^{-1}P(\lambda)$. Recall that we are looking for a matrix $X \in \mathbb{F}^{\ell n\times n}$ such that

(i) $S := K_\ell(A, X) = [X \ AX \ \cdots \ A^{\ell-1}X]$ is nonsingular, and

(ii) $\lambda I - S^{-1}AS$ is the left companion linearization of one of the reduced forms in (5)–(7).

If (i) holds, then $S^{-1}AS$ is the left companion matrix of a monic matrix polynomial, say $R(\lambda) = \lambda^\ell I + \cdots + \lambda R_1 + R_0$, and

$$S^{-1}A^\ell X = S^{-1}AS(e_\ell \otimes I_n) = -\begin{bmatrix} R_0 \\ R_1 \\ \vdots \\ R_{\ell-1} \end{bmatrix}, \tag{8}$$

where e_j denotes the jth column of the identity matrix and $\otimes$ denotes the Kronecker product. Then we see that the (i, j) entry of $R(\lambda)$ with $i \ne j$ is zero if and only if the vector $S^{-1}A^\ell Xe_j$ has zeros in the entries $i, i+n, \dots, i+(\ell-1)n$. This means that $S^{-1}A^\ell Xe_j$ is in the span of the columns of the submatrix of $I_{n\ell}$ obtained

by deleting the columns $i, i+n, \ldots, i+(\ell-1)n$. Thus, taking into account that $S^{-1}[X\ AX\ \cdots\ A^{\ell-1}X] = I$ and (8), it follows that

$$[R(\lambda)]_{ij} \equiv 0,\ i \neq j, \quad \Longleftrightarrow \quad A^{\ell}x_j \in \mathcal{K}_{\ell}(A, [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n]), \tag{9}$$

where x_j denotes the jth column of X.

We are now ready to state our main theorem, but before we do so we introduce some new notation. For the block reductions (5) and (6), it is convenient to partition X as

$$X = [X_1\ X_2\ \cdots\ X_k],$$

where $X_j \in \mathbb{F}^{\ell n \times s_j}$ and $s_1 + \cdots + s_k = n$. Also, we let $x_{1:i}$ and $X_{1:i}$ denote the matrices $[x_1\ x_2\ \cdots\ x_i] \in \mathbb{F}^{\ell n \times i}$ and $[X_1\ X_2\ \cdots\ X_i] \in \mathbb{F}^{\ell n \times \sigma_i}$, respectively, where $\sigma_i := s_1 + \cdots + s_i$. Finally, we define $\sigma_0 := 0$.

THEOREM 1. *Let $P(\lambda) \in \mathbb{F}[\lambda]^{n\times n}$ be of degree ℓ with nonsingular leading matrix coefficient and let $\lambda I - A$ be any monic linearization of $P(\lambda)$. Then $P(\lambda)$ is equivalent to a monic matrix polynomial $R(\lambda)$ of degree ℓ having one of the reduced forms* (5)–(7) *if and only if there exists a full rank matrix $X \in \mathbb{F}^{\ell n \times n}$ such that*

(i) *the matrix $[X\ AX\ \ldots\ A^{\ell-1}X] \in \mathbb{F}^{\ell n \times \ell n}$ is nonsingular, and*

(ii) (a) *$\mathcal{K}_{\ell}(A, X_i)$ for $1 \le i \le k$ is A-invariant for block-diagonal form as in* (5),

(b) *$\mathcal{K}_{\ell}(A, X_{1:i})$ for $1 \le i \le k$ is A-invariant for block-triangular form as in* (6),

(c) *$A^{\ell}x_i \in \mathcal{K}_{\ell}(A, x_{1:i+1})$, $1 \le i \le n-1$, for Hessenberg form as in* (7).

Proof. ($\Rightarrow$) Suppose that $P(\lambda)$ is equivalent to $R(\lambda)$. Then $\lambda I - A$ is also a monic linearization of $R(\lambda)$ and as explained in the introduction (recall (2)–(4)), there is a matrix X such that (A, X) is a left standard pair for $R(\lambda)$, which implies (i) and $AS = SC_L(R)$, where $S = [X\ AX\ \ldots\ A^{\ell-1}X] = K_{\ell}(A, X)$.

(ii)(a): Suppose that $R(\lambda)$ has the block-diagonal form of $D(\lambda)$ in (5). Define

$$\Pi_i = [e_{\sigma_{i-1}+1}\ \cdots\ e_{\sigma_i}\ e_{n+\sigma_{i-1}+1}\ \cdots\ e_{n+\sigma_i}\ \cdots\ e_{(\ell-1)n+\sigma_{i-1}+1}\ \cdots\ e_{(\ell-1)n+\sigma_i}],$$

where e_i is the ith column of $I_{\ell n}$. Then $S\Pi_i = K_{\ell}(A, X)\Pi_i = K_{\ell}(A, X_i)$ and $C_L(D)\Pi_i = \Pi_i C_L(D_i)$. It follows from $AS = SC_L(D)$ that

$$AK_{\ell}(A, X_i) = AS\Pi_i = SC_L(D)\Pi_i = S\Pi_i C_L(D_i) = K_{\ell}(A, X_i)C_L(D_i),$$

which proves (ii)(a).

(ii)(b): Suppose that $R(\lambda)$ has the block-triangular form of $T(\lambda)$ in (6). Let

$$\Pi_{1:i} = [\, e_1 \quad \cdots \quad e_{\sigma_i} \quad e_{n+1} \quad \cdots \quad e_{n+\sigma_i} \quad \cdots \quad e_{(\ell-1)n+1} \quad \cdots \quad e_{(\ell-1)n+\sigma_i}\,].$$

Then $K_{\ell}(A, X_{1:i}) = K_{\ell}(A, X)\Pi_{1:i} = S\Pi_{1:i}$ and $C_L(T)\Pi_{1:i} = \Pi_{1:i}C_L(T_i)$, where $T_i(\lambda)$ is the leading $\sigma_i \times \sigma_i$ principal submatrix of $T(\lambda)$. Then from $AS = SC_L(T)$ we obtain

$$AK_{\ell}(A, X_{1:i}) = AS\Pi_{1:i} = SC_L(T)\Pi_{1:i} = S\Pi_{1:i}C_L(T_i) = K_{\ell}(A, X_{1:i})C_L(T_i).$$

(ii)(c): Suppose that $R(\lambda)$ has the Hessenberg form of $H(\lambda)$ in (7). From $AS = SC_L(H)$ and (9), we see that $A^{\ell}x_i$ lies in the span of $K_{\ell}(A, x_{1:i+1})$.

($\Leftarrow$) Suppose that there exists X such that $S = [X\ AX\ \cdots\ A^{\ell-1}X]$ is nonsingular. Then the matrix $S^{-1}AS$ is the left companion form of a monic matrix polynomial of degree ℓ, say $R(\lambda)$, equivalent to $P(\lambda)$.

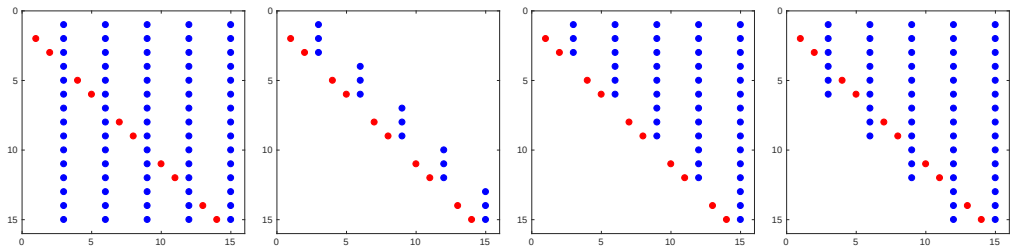

FIG. 3. *Spy plots for the controller form of the left companion matrices for cubic* 5×5 *matrix polynomials with (from left to right) dense, diagonal, triangular, and Hessenberg matrix coefficients, respectively. The red dots are* 1*'s, the blue dots are nonzero entries. (Figure is in color online.)*

Now, $AS = SC_L(R)$, (ii)(a), and (9) imply that the $n \times n$ blocks $R_0, \ldots, R_{\ell-1}$ in the last block column of $C_L(R)$ (see (8)) are block-diagonal with k diagonal blocks, the ith diagonal block being $s_i \times s_i$, where s_i is the number of columns of X_i, $i = 1\colon k$. The proofs for (ii)(b) and (ii)(c) are similar. ∎

3. Construction of the matrix X. In this section we discuss a process to construct the matrix X in Theorem 1 such that properties (i) and (ii) hold.

3.1. Auxiliary results. We start by proving some technical results that will be needed for the triangularization. Let $\lambda I - C_L$ be the left companion matrix of a monic matrix polynomial $P(\lambda)$ of size $n \times n$ and degree ℓ, and let Π denote the permutation matrix

$$\Pi = [\,\pi_1 \quad \pi_2 \quad \cdots \quad \pi_n\,], \qquad \pi_i = [\,e_i \quad e_{n+i} \quad \cdots \quad e_{(\ell-1)n+i}\,] \quad \text{for} \quad i = 1, \ldots, n.$$

Then the permuted linearization $\lambda I - \Pi^T C_L \Pi$ will be called the *left companion linearization of* $P(\lambda)$ *in controller form.* This is not a common term in the context of linearizations. The name comes from the theory of linear control systems, where controllable systems whose state matrices have the form of $\Pi^T C_L \Pi$ are said to be in *controller form* [8]. If we view this linearization as an $n \times n$ block pencil, then the zero-block structure of the pencil is the same as the zero structure of $P(\lambda)$. Furthermore, the diagonal $\ell \times \ell$ blocks are the companion matrices of the corresponding scalar polynomials on the diagonal of $P(\lambda)$. To illustrate the controller form, Figure 3 shows the spy plots of the left companion matrix for $P(\lambda)$ in dense (no structure), diagonal, triangular, and Hessenberg forms.

The controller form is useful in the proofs of the following theorems. In these theorems we will work with matrices having eigenvalues of geometric multiplicity at most n. The rationale behind this is that if $\lambda I - A$ is a linearization of an $n \times n$ matrix polynomial $P(\lambda)$ as in (1), then, by [4, Thm. 1.7], the geometric multiplicity of the eigenvalues of A cannot be greater than n.

3.1.1. Existence of Schur form for triangular reduction. Recall that a matrix is called *nonderogatory* if every eigenvalue has geometric multiplicity one.

THEOREM 2 (Schur form with nonderogatory blocks, complex version). *Let* $A \in \mathbb{C}^{\ell n \times \ell n}$ *be a matrix whose eigenvalues have geometric multiplicity at most* n*. Then* A

has a Schur decomposition

$$A = Q \begin{bmatrix} T_{11} & * & * & * \\ & T_{22} & * & * \\ & & \ddots & * \\ & & & T_{nn} \end{bmatrix} Q^H,$$

where the diagonal blocks $T_{ii} \in \mathbb{C}^{\ell\times\ell}$, $i = 1, \ldots, n$, are upper triangular and nonderogatory.

Proof. Since A has no eigenvalue with geometric multiplicity greater than n, it follows from [4, Proof of Thm. 1.7] that $\lambda I - A$ is a linearization of an $n \times n$ upper triangular monic matrix polynomial $R(\lambda)$ of degree ℓ. This matrix polynomial has a left companion linearization in controller form, which itself must be monic. Denote this linearization by $\lambda I - B$. Then $A = SBS^{-1}$ for some nonsingular S. Furthermore, B is block upper triangular, with blocks of size $\ell \times \ell$, and all diagonal blocks must be nonderogatory (since they are companion matrices). Let $U_iT_iU_i^H$ be a Schur decomposition of the ith diagonal block and set $U = U_1 \oplus U_2 \oplus \cdots \oplus U_n$. Then

$$B = UTU^H, \quad \text{with} \quad T = \begin{bmatrix} T_1 & * & * & * \\ & T_2 & * & * \\ & & \ddots & * \\ & & & T_n \end{bmatrix},$$

is a Schur decomposition. Finally, let $SU = QR$ be a QR factorization of SU and note that since R is upper triangular and nonsingular, $A = Q(RTR^{-1})Q^H$ is a Schur decomposition of A. The next theorem follows from the fact that the ith diagonal $\ell \times \ell$ block of RTR^{-1} is similar to T_i. ∎

We now prove the real analog of Theorem 2.

THEOREM 3 (Schur form with nonderogatory blocks, real version). *Let $A \in \mathbb{R}^{\ell n\times\ell n}$ be a matrix whose eigenvalues have geometric multiplicity at most n. Then A has a real Schur decomposition*

$$A = Q \begin{bmatrix} T_{11} & * & * & * \\ & T_{22} & * & * \\ & & \ddots & * \\ & & & T_{rr} \end{bmatrix} Q^T, \tag{10}$$

where each T_{ii} is either of size $\ell \times \ell$ and nonderogatory or of size $2\ell \times 2\ell$ and such that all eigenvalues have geometric multiplicity one or two.

Proof. Since all eigenvalues of A have geometric multiplicity at most n, it follows that $\lambda I - A$ has a real Smith form $D(\lambda) \oplus I_{(\ell-1)n}$ with $\deg\det D(\lambda) = n\ell$. By [16, Theorem 4.1] $D(\lambda)$ is equivalent to some real quasi-triangular matrix polynomial $T(\lambda)$ of degree ℓ, which may be assumed to be monic. It follows that

$$\lambda I - A \sim \begin{bmatrix} D(\lambda) & \\ & I_{(\ell-1)n} \end{bmatrix} \sim \begin{bmatrix} T(\lambda) & \\ & I_{(\ell-1)n} \end{bmatrix},$$

where $\sim$ denotes the equivalence relation for matrix polynomials. In other words, A is a linearization of some monic quasi-triangular matrix polynomial of degree ℓ. If

B denotes the constant matrix of the left companion linearization of $T(\lambda)$ in controller form, then the rest of the proof is essentially the same as the last part of the proof of Theorem 2, with the only difference being that we consider the real Schur decomposition instead of the complex one. ∎

3.1.2. Numerically stable construction of a Schur form for (block-) triangular reduction. The above theorems are key stones in the process of constructing the matrix X of Theorem 1 for the (block-) triangular reduction of $P(\lambda)$ in (1). To be numerically useful we need to overcome the drawback that the linearization $\lambda I - B$ in the proofs of Theorems 2 and 3 is obtained from $\lambda I - A$ via unimodular transformations. In what follows we propose a numerically stable procedure to construct the desired Schur form of A in Theorem 2 or Theorem 3 out of any of its Schur forms. This procedure works as long as all eigenvalues of A have algebraic multiplicity at most n. This will be our assumption.

We will proceed by induction on n. If $n = 1$, then $A \in \mathbb{F}^{\ell\times\ell}$ is a nonderogatory matrix because, by assumption, all its eigenvalues have algebraic multiplicity $n = 1$. Thus any Schur form of A (real or complex) is nonderogatory. Assume $n > 1$ and that any matrix $A \in \mathbb{F}^{m\ell\times m\ell}$, with $m \leq n-1$ and all its eigenvalues of algebraic multiplicity at most m, admits a Schur form satisfying the conditions of Theorems 2 or 3 according as $\mathbb{F} = \mathbb{C}$ or $\mathbb{F} = \mathbb{R}$, respectively.

First, compute any (real or complex) Schur decomposition of A. Then reorder the diagonal entries/blocks using the procedure in Bai and Demmel [1] according to the rules described below. We discuss the real and complex cases separately.

(I) Complex case. Suppose there are k distinct eigenvalues of algebraic multiplicity n and s distinct eigenvalues of algebraic multiplicity less than n. Note that $k \leq \ell$ and $s = 0$ or $s > \ell - k$ according as $k = \ell$ or $k < \ell$, respectively. Reorder the Schur form such that the leading $k \times k$ principal submatrix has one instance of each eigenvalue of algebraic multiplicity n. If there are $k < \ell$ such eigenvalues, pick any $\ell - k (< s)$ distinct eigenvalues of algebraic multiplicity less than n and reorder the diagonal such that these appear after the k eigenvalues of algebraic multiplicity n. The leading $\ell \times \ell$ submatrix obtained in this way has simple eigenvalues and is thus nonderogatory. We can use the induction hypothesis on the lower right $(n-1)\ell \times (n-1)\ell$ part of the matrix because all eigenvalues of A with algebraic multiplicity n have been used.

(II) Real case. The procedure over $\mathbb{R}$ is more involved because we need to move nonreal eigenvalues in complex conjugate pairs in order to keep the decomposition real. In addition, $2\ell \times 2\ell$ diagonal blocks may appear with eigenvalues of geometric multiplicity two. An example that illustrates the main features of the procedure that follows is given in Example 5.

At this point it is important for us to recall that when applying the Bai–Demmel algorithm [1] to block triangular matrices of size 3×3 (one element and one 2×2 block in the diagonal) and 4×4 (two blocks of size 2×2 in the diagonal), the blocks of size 2×2 before and after applying the algorithm are similar but not necessarily identical. In what follows we will very often use phrases like "reordering" or "moving the diagonal blocks" to mean that consecutive diagonal elements and blocks are swapped to place them in desired diagonal positions. As in the complex case, our goal is to move (if needed) diagonal elements and blocks in order to obtain a real Schur form T of A whose $\ell\times\ell$ diagonal blocks have distinct eigenvalues and the eigenvalues of the $2\ell\times 2\ell$ diagonal blocks are of algebraic multiplicity at most 2. This matrix would, of course, satisfy the conditions of Theorem 3.

Let us assume that the matrix A has the following:

- k_r distinct real eigenvalues of algebraic multiplicity n;
- k_c distinct pairs of nonreal complex conjugate eigenvalues of algebraic multiplicity n;
- s_i distinct real eigenvalues of algebraic multiplicity $i < n$;
- q_i distinct pairs of nonreal complex conjugate eigenvalues of algebraic multiplicity $i < n$.

Define $s := s_1 + s_2 + \cdots + s_{n-1}$, $q := q_1 + q_2 + \cdots + q_{n-1}$ and $k := k_r + 2k_c$. So, s and q are the number of distinct real and distinct pairs of nonreal complex conjugate eigenvalues, respectively, of multiplicity smaller than n. We need some inequalities that will be useful to ensure that the inductive process can be completed. First, $k \leq \ell$ and $n(\ell - k) = s_1 + 2s_2 + \cdots + (n-1)s_{n-1} + 2q_1 + 4q_2 + \cdots + 2(n-1)q_{n-1}$. Then,

$$n = 2 \Longrightarrow 2(\ell - k) = s_1 + 2q_1 = s + 2q \tag{11}$$

and

$$n(\ell - k) \leq (n-1)s + 2q_1 + 4q_2 + \cdots + 2(n-1)q_{n-1}. \tag{12}$$

We also claim that

$$n \geq 3 \quad \text{and} \quad \ell - k > 0 \Longrightarrow s + 2q - q_1 > \ell - k. \tag{13}$$

In fact, first we have $(s+2q-q_1)n = (s+q_1+2q_2+\cdots+2q_{n-1})n$. Define $h := s+(n-2)q_1+(2n-4)q_2+(2n-6)q_3+\cdots+(2n-2n+2)q_{n-1}$. If $\ell-k > 0$, then $s > 0$ or $q_i > 0$ for some i. And if $n \geq 3$, then $n-2 > 0$ and $2n-2i > 0$ for $2 \leq i \leq n-1$. Thus if $\ell-k > 0$ and $n \geq 3$, then $h > 0$. But $(s+2q-q_1)n-h = (n-1)s+2q_1+4q_2+\cdots 2(n-1)q_{n-1}$. Thus, if $\ell-k > 0$ and $n \geq 3$, then $(s+2q-q_1)n > (n-1)s+2q_1+4q_2+\cdots 2(n-1)q_{n-1}$ and (13) follows from (12).

Start by reordering the Schur form such that one instance of each of the k_r real eigenvalues and one instance of the 2×2 blocks corresponding to the k_c pairs of nonreal complex conjugate eigenvalues of algebraic multiplicity n appear in the leading $k \times k$ principal submatrix. Let $T_{11}^{(k)}$ denote this submatrix. If $\ell - k = 0$, then $T_{11} := T_{11}^{(k)}$ is nonderogatory and we can apply the induction hypothesis to the $(n-1)\ell \times (n-1)\ell$ lower right part of the matrix. If $\ell - k > 0$, this positive integer is either even or odd. Let us assume first that it is even. If there is at least $(\ell - k)/2$ diagonal blocks of size 2×2 corresponding to distinct nonreal complex conjugate eigenvalues of algebraic multiplicity smaller than n (i.e., if $q \geq (\ell-k)/2$), then we can move $(\ell-k)/2$ diagonal blocks of size 2×2 to appear after $T_{11}^{(k)}$ and the obtained leading $\ell\times\ell$ submatrix would be nonderogatory. If, on the contrary, $q < (\ell - k)/2$, then we will have to move all diagonal blocks of nonreal complex conjugate eigenvalues of multiplicity less than n and $\ell - k - 2q$ distinct real eigenvalues to appear after $T_{11}^{(k)}$. The question is: Do we have $\ell - k - 2q$ distinct real eigenvalues? The answer is in the affirmative because by either (11) or (13) $s + 2q \geq \ell - k$. In summary:

(i) If $\ell-k$ is even, then let $k_1 = \min\{q, \frac{\ell-k}{2}\}$. Choose k_1 2×2 blocks corresponding to distinct nonreal complex conjugate eigenvalues of algebraic multiplicity less than n and move them so that they appear directly after $T_{11}^{(k)}$. Denote the new submatrix $T_{11}^{(k+2k_1)}$.

(i$_1$) If $k_1 = \frac{\ell-k}{2}$ (that is, $\ell = k+2k_1$), then $T_{11} := T_{11}^{(k+2k_1)}$ is nonderogatory.

(i$_2$) If $k_1 = q < \frac{\ell-k}{2}$, then it follows from either (11) or (13) that $s > \ell - k - 2q = \ell - k - 2k_1$. Move (if necessary) $\ell - k - 2k_1$ distinct real

eigenvalues of algebraic multiplicity less than n so that they appear after $T_{11}^{(k+2k_1)}$. The leading $\ell \times \ell$ submatrix is nonderogatory.

Apply the induction hypothesis to the $(n-1)\ell \times (n-1)\ell$ lower right part of the matrix as above.

Let us assume now that $\ell - k$ is odd. In order to complete $T_{11}^{(k)}$ up to an $\ell \times \ell$ upper (block-) triangular matrix, we need at least one real eigenvalue of algebraic multiplicity less than n; i.e., $s > 0$. If this is the case, then we can proceed as in the case when $\ell - k$ is even but replacing $\ell - k$ by $\ell - k - 1$ and moving one available real eigenvalue to the position (ℓ, ℓ). On the other hand, if $s = 0$, then we can try to produce a $2\ell \times 2\ell$ diagonal block with eigenvalues of algebraic multiplicity at most 2. We will see that this is always possible.

(ii) If $\ell - k$ is odd and $s > 0$, then let $k_1 = \min\{q, \frac{\ell-k-1}{2}\}$. As in the case when $\ell - k$ is even, choose k_1 2×2 blocks corresponding to distinct nonreal complex conjugate eigenvalues of algebraic multiplicity less than n and move them so that they appear directly after $T_{11}^{(k)}$. Denote the new leading submatrix $T_{11}^{(k+2k_1)}$.

(ii$_1$) If $k_1 = \frac{\ell-k-1}{2} < q$, then $\ell = k + 2k_1 + 1$. Since $s > 0$, one real eigenvalue of algebraic multiplicity less than n can be placed after $T_{11}^{(k+2k_1)}$ so that the $\ell \times \ell$ principal submatrix is nonderogatory.

(ii$_2$) If $k_1 = q$, then as in case (i$_2$), $s > \ell - k - 2k_1$. Move (if necessary) $\ell - k - 2k_1$ distinct real eigenvalues of algebraic multiplicity less than n so that they appear after $T_{11}^{(k+2k_1)}$ and the resulting $\ell \times \ell$ principal submatrix is nonderogatory.

Apply the induction hypothesis as above.

(iii) If $\ell - k$ is odd and $s = 0$, we aim to form a $2\ell \times 2\ell$ block with eigenvalues of geometric multiplicity at most two. Recall that we already have one instance of each of the k_r real eigenvalues and one instance of the 2×2 blocks corresponding to the k_c pairs of nonreal complex conjugate eigenvalues of algebraic multiplicity $n \geq 2$ in $T_{11}^{(k)}$. Next, move another instance of each of the k_r real eigenvalues and another instance of the 2×2 blocks corresponding to the k_c pairs of nonreal complex conjugate eigenvalues of algebraic multiplicity n, so that they appear just after $T_{11}^{(k)}$. Let $T_{11}^{(2k)}$ denote this matrix. These eigenvalues may have geometric multiplicity two in $T_{11}^{(2k)}$.

(iii$_1$) If $n = 2$, then, from (11), $2(\ell - k) = 2q_1 + s = 2q_1$. This means that there are $\ell - k$ 2×2 diagonal blocks corresponding to pairs of nonreal complex conjugate eigenvalues of algebraic multiplicity one. Reorder (if necessary) the diagonal blocks so that they appear just after $T_{11}^{(2k)}$. The resulting $2\ell \times 2\ell$ submatrix has all its eigenvalues of geometric multiplicity 2 at the most.

(iii$_2$) If $n \geq 3$, then we have two possibilities: either $\ell - k \leq q$ or $\ell - k > q$. If $\ell - k \leq q$, then we have $\ell - k$ 2×2 blocks corresponding to distinct nonreal complex conjugate eigenvalues of algebraic multiplicity less than n that can be moved so that they appear directly after $T_{11}^{(2k)}$. Then all the eigenvalues of the obtained $2\ell \times 2\ell$ submatrix have geometric multiplicity one or two. If $\ell - k > q$, the process is a little more involved. First, we move q 2×2 blocks corresponding to distinct nonreal complex conjugate eigenvalues of algebraic multiplicity less than n so that they appear directly after $T_{11}^{(2k)}$. Let $T_{11}^{(2k+2q)}$ be the obtained submatrix. We

```
A = [ 1 0 0 1; 0 0 -1 0; 0 1 0 1; 0 0 0 1];
E = ordeig(A);
[Q, T] = ordschur(eye(4),A,imag(E)==0);
```

FIG. 4. *MATLAB M-file that implements the Bai–Demmel algorithm to swap diagonal block* $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ *and diagonal element* 1 *in position* $(4,4)$ *of the real Schur matrix* A *of Example* 4.

need to move $\ell - k - q$ additional 2×2 blocks corresponding to distinct nonreal complex conjugate eigenvalues of algebraic multiplicity less than n. Notice that, since we have moved q such blocks to form $T_{11}^{(2k+2q)}$, we have already used all blocks corresponding to nonreal complex conjugate eigenvalues of algebraic multiplicity 1. So we are left with $q - q_1$ 2×2 blocks corresponding to distinct nonreal complex conjugate eigenvalues of algebraic multiplicity less than n. But it follows from (13) (recall that $\ell - k > 0$ and $n \geq 3$) that $\ell - k < 2q - q_1 + s = 2q - q_1$ and this means that $\ell - k - q < q - q_1$. Therefore, another copy of $\ell - k - q$ 2×2 blocks corresponding to nonreal complex conjugate eigenvalues of algebraic multiplicities between 2 and $n - 1$ can be moved to appear directly after $T_{11}^{(2k+2q)}$. The eigenvalues of the resulting $\ell \times \ell$ matrix have algebraic multiplicity at most 2.

We can now apply the induction hypothesis to the $(n-2)\ell \times (n-2)\ell$ lower right part of the matrix.

We note that when $\ell - k$ is odd and $s = 0$ (case (iii)), the constructed $2\ell \times 2\ell$ may or may not be further split into two $\ell \times \ell$ nonderogatory blocks by moving the eigenvalues and blocks along the diagonal. The following example illustrates the two possibilities.

Example 4. (a) Let $n = 2$, $\ell = 2$, and

$$A = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

This matrix is in real Schur form and satisfies the requirements of Theorem 3: one $2\ell \times 2\ell$ block with eigenvalues of geometric multiplicity at most 2. However, we can swap the diagonal block $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ and the last diagonal element of A to obtain another Schur form that also satisfies the conditions of Theorem 3. The MATLAB code in Figure 4 implements the Bai–Demmel algorithm [1] to perform the swapping. The returned matrices Q and T are

$$Q = \begin{bmatrix} 1.0000 & 0 & 0 & 0 \\ 0 & -0.4082 & 0.7071 & 0.5774 \\ 0 & 0.4082 & 0.7071 & -0.5774 \\ 0 & 0.8165 & -0.000 & 0.5774 \end{bmatrix}, \; T = \left[\begin{array}{rr|rr} 1.0000 & 0.8165 & -0.0000 & 0.5774 \\ 0 & 1.0000 & 0.5774 & 0.7071 \\ \hline 0 & 0 & -0.0000 & 1.2247 \\ 0 & 0 & -0.8165 & 0.0000 \end{array}\right].$$

The 2×2 submatrix in the lower right corner of T is (approximately) similar to $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ and T is another (approximate) real Schur form of A with two nonderogatory diagonal blocks. Notice that if we replace $a_{14} = 1$ by $a_{14} = 0$ in A, then the eigenvalue 1 of the new matrix would have geometric multiplicity two and there would not be a Schur form with two nonderogatory blocks of size 2×2 in the diagonal.

(b) Let $n = 2$, $\ell = 3$, and

$$A = \operatorname{diag}\left(\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -3 \\ 3 & 0 \end{bmatrix}\right).$$

Despite the eigenvalues of A being simple, there is no real Schur form of A with two nonderogatory diagonal blocks of size 3×3. If we apply the procedure of item **(II)** to A, then $k_r = k_c = k = 0$, $q_1 = q = 3$, and $s = 0$. Since $\ell - k = 3$ is odd and $s = 0$, we use item (iii$_1$). In fact, $2(\ell - k) = 2\ell = 6 = 2q_1$ and we must put together three diagonal blocks of size 2×2. This means that A is itself the desired matrix.

The following example clarifies the main features of the procedure for the real case (item **(II)**) to bring a matrix in real Schur form to another one satisfying the requirements in Theorem 3.

Example 5. Let $n = 4$, $\ell = 2$, and let $A \in \mathbb{R}^{8\times 8}$ be a matrix in real Schur form with the following diagonal blocks:

$$B = \operatorname{diag}\left(1, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, 1, 2, 1\right).$$

We can write $A = B+T$ where T is a strict block-upper triangular matrix (block-upper triangular with zero blocks in the diagonal). We are going to apply the procedure of item **(II)** to A to find an orthogonal matrix Q such that $Q^T AQ$ is a real Schur form satisfying the requirements in Theorem 3.

Step 1. For A we have $k_r = k_c = 0$, $s = 2$, and $q_1 = 1$. Then, $k = k_r + k_c = 0$ and $\ell - k = 2$. Thus $\ell - k$ is even and $k_1 = \min\left\{q, \frac{\ell-k}{2}\right\} = \frac{\ell-k}{2} = 1$. We apply (i$_1$): use the Bai–Demmel algorithm to move the first $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ block to place it in the upper left corner: there is an orthogonal matrix Q_1 such that $A_1 = Q_1^T AQ = B_1 + T_1$ with T_1 strict block-upper triangular and $B_1 = \operatorname{diag}\left(B_{11}, 1, \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right], 1, 2, 1\right)$, where B_{11} is similar to $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$.

We remove the two first rows and columns of A_1 and work with the lower right 6×6 matrix $\widehat{A}_1 = \widehat{B}_1 + \widehat{T}_1$ where $\widehat{B}_1 = \operatorname{diag}\left(1, \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right], 1, 2, 1\right)$. Now $n = 3$ and $\ell = 2$.

Step 2. For $\widehat{A}_1$ we have $k_r = 1$, $k_c = 0$, $s = 1$, and $q_1 = 1$. Since $k_r = 1$ for eigenvalue 1, first of all, we must move it to position $(1, 1)$. In this case no action is needed because it is already there. Now, $k = k_r + k_c = 1$, $\ell - k = 1$ is odd, $s > 0$, and $k_1 = \min\left\{q, \frac{\ell-k-1}{2}\right\} = \min\{1, 0\} = 0$. Hence we use (ii$_2$): move a real eigenvalue of multiplicity less than $n = 3$ to position $(2, 2)$. There is only one choice: use the Bai–Demmel algorithm to exchange the block $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ and the entry in position $(5, 5)$ (actually, we must swap first diagonal entries 1 and 2 and then swap 2 and block $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$). So, there is an orthogonal matrix $\widehat{Q}_2$ such that $\widehat{A}_2 = \widehat{Q}_2^T \widehat{A}_1 \widehat{Q}_2 = \widehat{B}_2 + \widehat{T}_2$ with $\widehat{T}_2$ strict block-upper triangular and $\widehat{B}_2 = \operatorname{diag}(1, 2, B_{21}, 1, 1)$, where B_{21} is similar to $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$.

We remove again the two first rows and columns of $\widehat{A}_2$ and pay attention to $\widetilde{A}_2 = \widetilde{B}_2 + \widetilde{T}_2$ with $\widetilde{B}_2 = \operatorname{diag}(B_{21}, 1, 1)$. Now $n = 2$ and $\ell = 2$.

Step 3. For $\widetilde{A}_2$ we have $k_r = 1$, $k_c = 0$, $s = 0$, and $q_1 = 1$. Again $k_r = 1$ and we must place the eigenvalue of algebraic multiplicity 2 in position $(1, 1)$. We use the Bai–Demmel algorithm to swap the diagonal block B_{21} and the diagonal entry $(3, 3)$. Let $\widehat{B}_{21}$ be the resulting 2×2 block.

Now $k = k_r + k_c = 1$, $\ell - k = 1$ is odd and $s = 0$ so case (iii) applies: move another copy of the eigenvalues of algebraic multiplicity $n = 2$ to position $(2, 2)$. We use the

Bai–Demmel algorithm to exchange the diagonal block $\widehat{B}_{21}$ and the entry in position $(4,4)$. Let B_{21} be the obtained block. We observe that $n = 2$ and $2(\ell - k) = 2 = 2q_1$. So we proceed as indicated in item (iii$_1$): move the block B_{21} to place it right after the two repeated eigenvalues to get a diagonal block of size 4. In this case, no action is needed. Thus there is an orthogonal $\widetilde{Q}_3$ such that $\widetilde{A}_3 = \widetilde{Q}_3^T \widetilde{A}_2 \widetilde{Q}_3 = \widetilde{B}_3 + \widetilde{T}_3$ with $\widetilde{T}_3$ strict block-upper triangular, $\widetilde{B}_3 = \operatorname{diag}(1, 1, B_{21})$, and B_{21} similar to $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$.

If we define $Q = Q_1 \operatorname{diag}(I_2, \widehat{Q}_2) \operatorname{diag}(I_4, \widetilde{Q}_3)$, then Q is an orthogonal matrix and

$$Q^T A Q = \left[\begin{array}{c|cc|ccc} B_{11} & * & * & * & * & * \\ \hline & 1 & * & * & * & * \\ & & 2 & * & * & * \\ \hline & & & 1 & * & * \\ & & & & 1 & * \\ \hline & & & & & B_{21} \end{array}\right]$$

is a matrix in real Schur form. Blocks B_{11} and B_{21} are both similar to $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$. The 4×4 block in the lower-right corner will be nonderogatory if its $(1,2)$ entry is not zero; otherwise, the geometric multiplicity of 1 in that block would be 2.

Matrix A in Example 5 has repeated eigenvalues, but even in the generic case of real matrices with simple eigenvalues, the diagonal blocks of a computed real Schur form might need to be rearranged in order to satisfy the requirements of Theorem 3. In addition, as part (b) of Example 4 shows, the diagonal blocks in the Schur form of Theorem 3 for matrices with simple eigenvalues may need to be of size $2\ell \times 2\ell$.

If one eigenvalue has algebraic multiplicity greater than n, the problem of computing the desired Schur forms in a stable manner, using unitary/orthogonal transformations, becomes significantly more complicated. We devote section 4 to this problem for quadratic matrix polynomials ($\ell = 2$). The higher-degree case $\ell > 2$ is left as an open problem. There is a process to obtain a desired form by manipulating Jordan forms, but we omit the details as it is an unstable process.

3.1.3. Sufficient conditions for nonsingular $K_\ell(A, X)$. Theorems 2 and 3 will be used in combination with the following lemmas. They show a nice connection with the following known result in the theory of linear control systems: the minimum number of inputs needed to control a linear time-invariant system is the geometric multiplicity of the eigenvalues with highest geometric multiplicity (see, for example, [19]).

LEMMA 6. *If $B \in \mathbb{F}^{\ell \times \ell}$ is nonderogatory, then there exists $x \in \mathbb{F}^\ell$ such that the Krylov matrix $K_\ell(B, x)$ is nonsingular.*

Proof. Since B is nonderogatory it is similar to the left companion matrix C_L of its characteristic polynomial [7, Thm. 3.3.15], that is, $B = SC_LS^{-1}$ for some nonsingular matrix S. It is now easy to see that $K_\ell(C_L, e_1) = I$. Hence letting $x = Se_1$ yields the desired result. ∎

The next lemma is the real counterpart of Lemma 6.

LEMMA 7. *Let $B \in \mathbb{R}^{2\ell \times 2\ell}$ have eigenvalues with geometric multiplicity at most two. Then there exist two real vectors x and y such that $K_\ell(B, [x\ y])$ is nonsingular.*

Proof. We can rearrange the real Jordan decomposition [18, sect. 2.4] of B so that

$$S^{-1}BS = \begin{array}{c} \\ m_1 \\ m_2 \end{array} \begin{array}{c} \begin{array}{cc} m_1 & m_2 \end{array} \\ \begin{bmatrix} J_1 & \\ & J_2 \end{bmatrix} \end{array} \in \mathbb{R}^{2\ell \times 2\ell}, \quad m_1 \geq m_2 \geq 0,$$

with J_1 and J_2 nonderogatory. Note that the matrix B is allowed to be nonderogatory: in this case $S^{-1}BS = J_1$, $m_2 = 0$, and J_2 is empty. Since J_1 and J_2 are nonderogatory matrices, they are similar (via real arithmetic) to the left companion matrices $C_1 \in \mathbb{R}^{m_1 \times m_1}$ and $C_2 \in \mathbb{R}^{m_2 \times m_2}$ of their characteristic polynomials, respectively. Hence there exists a nonsingular $W \in \mathbb{R}^{2\ell \times 2\ell}$ such that

$$W^{-1}BW = C_1 \oplus C_2 =: C.$$

If $m_2 = 0$, then $C := C_1$. It suffices to prove that there exist $u, v \in \mathbb{R}^{2\ell}$ such that $M = [K_\ell(C, u)\ K_\ell(C, v)]$ is nonsingular because we then get the desired result by taking $x = Wu$ and $y = Wv$.

If $m_1 = m_2$ or $m_2 = 0$, then $u = e_1$ and $v = e_{\ell+1}$ yield $M = I_{2\ell}$ and we are done. If $m_1 > m_2 > 0$, we let $u = e_1$ and $v = e_{\ell-m_2+1} + e_{m_1+1}$. Then direct calculations show that

$$M = \begin{array}{r} \\ \ell-m_2 \\ m_2 \\ \ell-m_2 \\ m_2 \end{array} \begin{array}{c} \begin{array}{cccc} \ell-m_2 & m_2 & m_2 & \ell-m_2 \end{array} \\ \begin{bmatrix} I & & & \\ & I & I & \\ & & & I \\ & & I & * \end{bmatrix} \end{array},$$

where $*$ is some $m_2 \times (\ell - m_2)$ matrix. It is now easy to see that M has full column rank, and thus is nonsingular. ∎

Finally, we provide a lemma that can be seen as a block generalization of Lemmas 6 and 7.

LEMMA 8. *If all eigenvalues of $A \in \mathbb{F}^{k\ell \times k\ell}$ have geometric multiplicity at most k, then there exists $X \in \mathbb{F}^{k\ell \times k}$ such that $K_\ell(A, X)$ is nonsingular.*

Proof. We will handle the real and complex case simultaneously. Let $A = ZTZ^{-1}$ be the decomposition from Theorem 2 or Theorem 3 and denote the diagonal blocks by T_{ii}, $i = 1\colon r$. For each T_{ii} we define W_i in the following way. If T_{ii} is of size $\ell \times \ell$, take W_i to be the $\ell \times 1$ vector in Lemma 6 such that $K_\ell(T_{ii}, W_i)$ is nonsingular, and if T_{ii} is of size $2\ell \times 2\ell$, take W_i to be the $2\ell \times 2$ matrix whose columns are the two real vectors in Lemma 7. Letting $W = W_1 \oplus W_2 \oplus \cdots \oplus W_r$ and $X = ZW$ yields $K_\ell(A, X) = ZK_\ell(T, W)$, which is of full rank. ∎

3.2. Reduced forms. For a given matrix polynomial with nonsingular leading matrix coefficient and monic linearization $\lambda I - A$, we now discuss how to construct a matrix X such that properties (i) and (ii) in Theorem 1 hold.

3.2.1. Block-triangular form. For the reduction to (block-) triangular form we have the following result.

PROPOSITION 9. *Let $s_1, \ldots, s_k$ be positive integers such that $s_1 + \cdots + s_k = n$, and let*

$$T = \begin{bmatrix} T_{11} & * & * & * \\ & T_{22} & * & * \\ & & \ddots & * \\ & & & T_{kk} \end{bmatrix} \in \mathbb{F}^{\ell n \times \ell n},$$

where $T_{ii} \in \mathbb{F}^{\ell s_i \times \ell s_i}$ has no eigenvalues of geometric multiplicity more than s_i for $i = 1, \ldots, k$. If $A \in \mathbb{F}^{\ell n \times \ell n}$ is similar to T, then there exists $X = [X_1\ X_2\ \cdots\ X_k]$ with $X_i \in \mathbb{F}^{n\ell \times s_i}$ such that $S = K_\ell(A, X)$ is nonsingular and $\mathcal{K}_\ell(A, X_{1:i})$ is A-invariant for $i = 1, \ldots, k$.

Proof. By Lemma 8, we can for each T_{ii} pick a $V_i \in \mathbb{F}^{\ell s_i \times s_i}$ such that $K_\ell(T_{ii}, V_i)$ is nonsingular. Thus, if $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$, then $K_\ell(T, V)$ is nonsingular. Let Z be a nonsingular matrix such that $Z^{-1}AZ = T$ and put $X = ZV$. Then $S = K_\ell(A, X) = ZK_\ell(T, V)$ is nonsingular. In addition, if $\sigma_i = s_1 + \cdots + s_i$ and

$$W_i = \begin{bmatrix} V_1 \oplus \cdots \oplus V_i \\ 0 \end{bmatrix} \in \mathbb{F}^{\ell n \times \sigma_i}, \quad i = 1, \ldots, k,$$

then $\mathcal{K}_\ell(A, X_{1:i})$ is A-invariant if and only if $\mathcal{K}_\ell(T, W_i)$ is T-invariant. Since the columns of $T^j W_i$ are also columns of $K_\ell(T, W_i)$ for $j < \ell$, we only have to show that there is a matrix R such that $T^\ell W_i = K_\ell(T, W_i)R$. If T_i is the submatrix of T formed by its $\ell\sigma_i$ first rows and columns and $\widehat{V}_i = V_1 \oplus \cdots \oplus V_i$, then

$$K_\ell(T, W_i) = \begin{bmatrix} K_\ell(T_i, \widehat{V}_i) \\ 0 \end{bmatrix}.$$

Since $K_\ell(T_i, \widehat{V}_i)$ is nonsingular, there is a matrix $R \in \mathbb{F}^{\ell n \times \sigma_i}$ such that

$$T^\ell W_i = \begin{bmatrix} T_i^\ell \widehat{V}_i \\ 0 \end{bmatrix} = \begin{bmatrix} K_\ell(T_i, \widehat{V}_i)R \\ 0 \end{bmatrix} = K_\ell(T, W_i)R,$$

as desired. ∎

Remark 10. The proof of Proposition 9 provides a practical means to construct X. From the proof we see that the columns of $K_\ell(A, X_{1:i})$ must be a basis for the invariant subspace of A corresponding to the eigenvalues of $T_{11}, T_{22}, \ldots, T_{ii}$.

We now explain why the MATLAB M-file in Figure 1 successfully reduced $P(\lambda)$ to triangular form (see the left plot of Figure 2). Since the coefficients are generated randomly, the eigenvalues are all distinct with probability one. Therefore, MATLAB's `schur` function computes a Schur decomposition $C_P = ZTZ^H$, where $Z^H = Z^{-1}$ and the $\ell \times \ell$ diagonal blocks are all nonderogatory. Thus each $K_\ell(T_{ii}, V_i) \in \mathbb{R}^{\ell \times \ell}$ becomes nonsingular by taking each $V_i \in \mathbb{F}^{\ell \times 1}$ to be a vector of ones (almost any random vector would do). Hence, $X := Z(V_1 \oplus V_2 \oplus \cdots \oplus V_k)$ is as in Proposition 9, and so the conditions (i) and (ii)(b) in Theorem 1 are fulfilled.

3.2.2. Block-diagonal form. For the reduction to block-diagonal form we have the following result.

PROPOSITION 11. *Let $A \in \mathbb{F}^{\ell n \times \ell n}$ and assume that for some nonsingular Z*

$$Z^{-1}AZ = \begin{bmatrix} D_{11} & & & \\ & D_{22} & & \\ & & \ddots & \\ & & & D_{kk} \end{bmatrix} \in \mathbb{F}^{\ell n \times \ell n}, \tag{14}$$

where $D_{ii} \in \mathbb{F}^{s_i \ell \times s_i \ell}$ has eigenvalues of geometric multiplicity at most $s_i \in \mathbb{N}$, $i = 1, \ldots, k$, with $s_1 + \cdots + s_k = n$. Then there exists $X = [X_1 \ X_2 \ \ldots \ X_k]$ with $X_i \in \mathbb{F}^{n\ell \times s_i}$ such that $S = K_\ell(A, X)$ is nonsingular and $\mathcal{K}_\ell(A, X_i)$ is A-invariant for $i = 1, \ldots, k$.

The proof is similar to that of Proposition 9 and is omitted. We have the following analog to Remark 10.

Remark 12. With the notation of Proposition 11, the columns of $K_\ell(A, X_i)$ are a basis for the invariant subspace of A corresponding to the eigenvalues of D_{ii}.

We now explain how the diagonalization corresponding to the middle plot of Figure 2 was achieved. The eigenvalues are again all distinct (with probability one), and the `eig` function computes Λ, Z such that $C_P = Z\Lambda Z^{-1}$ is an eigenvalue decomposition with Λ diagonal. Thus by taking $V_i \in \mathbb{F}^{\ell\times 1}$ to be vectors of ones and letting $X := Z(V_1 \oplus V_2 \oplus \cdots \oplus V_k)$, the conditions (i), (ii)(a) in Theorem 1 are satisfied.

Clearly the number of blocks in the decomposition (14) of Proposition 11 is not arbitrary. Indeed, the linear matrix polynomial $\lambda I - J_\alpha$, where

$$J_\alpha = \begin{bmatrix} \alpha & 1 & & & \\ & \alpha & 1 & & \\ & & \alpha & \ddots & \\ & & & \ddots & 1 \\ & & & & \alpha \end{bmatrix}$$

is of size $2\ell \times 2\ell$, cannot be reduced to a block-diagonal structure with smaller block sizes. Further, since $\lambda I - J_\alpha$ is a linearization of

$$P(\lambda) = \begin{bmatrix} (\lambda-\alpha)^\ell & 1 \\ & (\lambda-\alpha)^\ell \end{bmatrix},$$

it may also be the case that for matrix polynomials of degree $\ell > 1$, the block sizes of a block-diagonal form cannot be reduced.

Let $\lambda I - A$ be a linearization of $P(\lambda) \in \mathbb{F}[\lambda]^{n\times n}$ in (1). From Theorem 1 and Proposition 11, we see that a reduction to diagonal form is possible if we can partition the Jordan blocks associated with A into n sets such that

(a) each set has at most one Jordan block of each eigenvalue, and
(b) the sizes of all Jordan blocks in each set sum up to ℓ.

The result also holds in the opposite direction, that is, it is possible to reduce $P(\lambda)$ to diagonal form, *only if* we can partition the Jordan blocks of A such that (a) and (b) hold. To see this, we simply note that any diagonal monic matrix polynomial $D(\lambda) = d_1(\lambda) \oplus d_2(\lambda) \oplus \cdots \oplus d_n(\lambda)$ has left companion linearization in controller form:

$$\lambda I - (C_L(d_1) \oplus C_L(d_2) \oplus \cdots \oplus C_L(d_n)).$$

The following question arises: When is it possible to partition the Jordan blocks such that (a) and (b) are satisfied? This problem was solved by Lancaster and Zaballa [10] for the special case of quadratic matrix polynomials with nonsingular leading matrix coefficient, and by Zúñiga Anaya [20] for general regular quadratics. For matrix polynomials of higher degree the problem is still open.

3.2.3. Hessenberg form. For the reduction to Hessenberg form we have the following result.

PROPOSITION 13. *Let $A \in \mathbb{F}^{\ell n\times \ell n}$, and let Z be a nonsingular matrix such that*

$$Z^{-1}AZ = H, \tag{15}$$

where H is upper Hessenberg and partitioned in $\ell \times \ell$ blocks. Assume that the $\ell \times \ell$ diagonal blocks are unreduced, that is, $H_{i+1,i} \neq 0$ for all i. If we let $V = [e_1\ e_{\ell+1}\ \cdots\ e_{(n-1)\ell+1}]$ and $X = ZV \in \mathbb{F}^{\ell n\times n}$, then $K_\ell(A, X)$ is nonsingular and $A^\ell x_i \in \mathcal{K}_\ell(A, x_{1:i+1})$ for $i = 1, \ldots, n-1$.

Proof. We have $K_\ell(A, X) = ZK_\ell(H, V)$, which is obviously nonsingular. Furthermore, if v_i and x_i are the ith columns of V and X, respectively, then

$$A^\ell x_i = ZH^\ell v_i = ZH^\ell e_{(i-1)\ell+1} \in Z\mathcal{K}_\ell(H, v_{1:i+1}) = \mathcal{K}_\ell(A, x_{1:i+1}),$$

completing the proof. ∎

In practice we are interested in Hessenberg decompositions $A = UHU^H$, where U is unitary or real orthogonal, depending on whether we work over $\mathbb{C}$ or $\mathbb{R}$. By the implicit Q-theorem [5, Thm. 7.4.2], the Hessenberg matrix H is uniquely defined, up to products by real or complex numbers of absolute value 1, by the first column of U. Hence a random Hessenberg matrix similar to A via unitary/real orthogonal transformations can be constructed using, e.g., the Arnoldi algorithm with a random starting vector (or equivalently, the standard Hessenberg reduction step [5, sect. 7.4.2] applied to QAQ^H, where Q is a random orthogonal matrix). If a matrix has distinct eigenvalues, the resulting Hessenberg matrix will be unreduced with probability one. Since this is the generic case for matrix polynomials, Proposition 11 may be used to reduce almost all matrix polynomials to Hessenberg form without further care. This is how the right plot of Figure 2 was obtained.

If a matrix, on the other hand, has an eigenvalue of geometric multiplicity greater than one, then any similar Hessenberg matrix is necessarily reduced. Now, according to Proposition 13 the reduction of $P(\lambda)$ to Hessenberg form is still valid if H is reduced, as long as the diagonal $\ell \times \ell$ blocks are unreduced. This means that all zeros on the subdiagonal are in some of the positions $(\ell+1, \ell), (2\ell+1, 2\ell), \ldots, ((n-1)\ell + 1, (n-1)\ell)$. If H has a zero in any other position on the subdiagonal (that is, if some Hessenberg diagonal block is reduced), $K_\ell(A, X)$ becomes singular and the reduction will fail with the matrix X selected in the statement of Proposition 13. This raises the following question: Is it possible to move zeros on the subdiagonal, from unwanted to wanted positions, using a finite number of Givens rotations or Householder reflectors? Intuitively, this should not be possible since if moving a zero is possible, then we can change the number of "deflated" eigenvalues; a rigorous argument can be found in [15, pp. 104–105].

PROPOSITION 14. *Matrices X of Propositions* 9 *and* 13 *can be taken to have orthonormal columns.*

Proof. Let X be the matrix constructed in the proof of either Proposition 9 or that of Proposition 13. Let $X = QR$ be a QR factorization of X. Since X has full column rank, R is nonsingular and $Q = XR^{-1}$. Thus

$$K_\ell(A, X) = K_\ell(A, Q)(R \oplus R \oplus \cdots \oplus R),$$

and so $K_\ell(A, Q)$ is nonsingular.

Now, write $Q = [\,Q_1 \quad Q_2 \quad \cdots \quad Q_k\,]$ with $Q_i \in \mathbb{F}^{\ell n \times s_i}$ and $s_1 + \cdots + s_k = n$. Define $Q_{1:i} = [\,Q_1 \quad \cdots \quad Q_i\,] \in \mathbb{F}^{\ell n \times \sigma_i}$ with $\sigma_i = s_1 + \cdots + s_i$, and let R_i denote the $i \times i$ upper left principal submatrix of R. Then R_i is nonsingular, $X_{1:i} = Q_{1:i} R_{\sigma_i}$ and

$$K_\ell(A, X_{1:i}) = K_\ell(A, Q_{1:i})(R_{\sigma_i} \oplus R_{\sigma_i} \oplus \cdots \oplus R_{\sigma_i}).$$

Therefore, $\mathcal{K}_\ell(A, X_{1:i}) = \mathcal{K}_\ell(A, Q_{1:i})$, $i = 1, \ldots, k$, and so if X is the matrix of Proposition 9, then $\mathcal{K}_\ell(A, Q_{1:i})$ is A-invariant.

Similarly, the ith column of X and Q are $x_i = q_{1:i} r_i$ and $q_i = x_{1:i} u_i$, respectively, where r_i and u_i are the last columns of R_i and R_i^{-1}, respectively. A simple induction argument shows that $A^\ell x_i \in \mathcal{K}_\ell(A, x_{1:i+1})$ if and only if $A^\ell q_i \in \mathcal{K}_\ell(A, q_{1:i+1})$. ∎

In practice, using a matrix X with orthonormal columns to construct $S = K_\ell(A, X)$ may result in a more reliable way of computing $S^{-1}AS$ to obtain the left companion matrix of a block-triangular or Hessenberg matrix polynomial equivalent to $P(\lambda)$. We note that while we have discussed (when possible) how to compute X in a stable manner, finding the reduced matrix polynomial $R(\lambda)$ appears to require further computing $S^{-1}AS$. As mentioned in the introduction, we leave the stable computation of $R(\lambda)$ as an open problem.

4. Stable computation of a special Schur form for quadratic matrix polynomials. A procedure was exhibited in section 3.1.2 to compute the Schur decompositions in Theorems 2 and 3 in a numerically stable manner when all eigenvalues have algebraic multiplicity at most n. In this section we aim to complete the study to cover the case of eigenvalues of arbitrary algebraic multiplicity for quadratic matrix polynomials, that is, when $\ell = 2$. Recall that the eigenvalues of any linearization of $P(\lambda)$ in (1) have geometric multiplicity at most n [4, Thm. 1.7], but they may have algebraic multiplicity greater than n. We will assume in this section that one eigenvalue (and only one because $\ell = 2$) has algebraic multiplicity greater than n. It must be real (again because $\ell = 2$) and will be denoted by α.

We collect key tools in three lemmas, which all have algorithmic proofs. These proofs rely on the possibility of computing the geometric multiplicity of α; equivalently, computing rank $(A - \alpha\, I)$. Since we are dealing with a very ungeneric case, this is not an unreasonable assumption. Our goal is to follow a procedure similar to that of section 3.1.2: we start with a computed Schur form of a linearization of $P(\lambda)$ and use the Bai–Demmel algorithm [1] to move the eigenvalues (or 2×2 blocks with complex conjugate eigenvalues if $\mathbb{F} = \mathbb{R}$) along the main diagonal. Now we have one eigenvalue, α, whose algebraic multiplicity is $n + t$ with $t > 0$. When $\mathbb{F} = \mathbb{C}$ we will pair a copy of α with an eigenvalue different from α. The corresponding 2×2 diagonal block will be nonderogatory. Once the $n - t$ eigenvalues different from α have been used and the rows and columns associated to the corresponding 2×2 blocks have been constructed, we are left with a $2t \times 2t$ triangular matrix, T_1 say, whose only eigenvalue is α. We also need the 2×2 diagonal blocks of T_1 to be nonderogatory. Hence α, as an eigenvalue of T_1, must have geometric multiplicity at most t. So, our strategy will be to move the eigenvalues along the diagonal in order to pair a copy of α with an eigenvalue different from α in such a way that, when removing the rows and columns of the corresponding 2×2 diagonal block, the geometric multiplicity of α, as an eigenvalue of the resulting submatrix, is at most $n - 1$. This is Lemma 15. In this way, α as an eigenvalue of T_1 will have geometric multiplicity at most t. Then we show that any $2n \times 2n$ matrix with α as the only eigenvalue and geometric multiplicity at most n admits a Schur form with 2×2 nonderogatory diagonal blocks (Lemma 16). When $\mathbb{F} = \mathbb{R}$ the Schur form may have 2×2 diagonal blocks associated to pairs of complex conjugate eigenvalues in addition to real eigenvalues. In this case, we first construct 4×4 blocks with two copies of α so that, after removing the first four rows and columns, the geometric multiplicity of α as an eigenvalue of the obtained submatrix is at most $n - 2$. This is Lemma 17. Examples are provided at the end of the section.

Here and below, MATLAB notation is used to denote the submatrices of a given matrix. For instance, $X(i_1 : i_2, j_1 : j_2)$ is the submatrix of $X \in \mathbb{F}^{m\times n}$ formed with the i_1 through i_2 rows and the j_1 through j_2 columns and $X(:, j_1 : j_2) = X(1 : m, j_1 : j_2)$.

In some parts of the proofs we will use "bottom-up" QR factorizations of matrices. For a matrix $M \in \mathbb{F}^{m\times n}$, we say that $M = QR$ is a bottom-up QR factorization of

M if $Q \in \mathbb{F}^{m\times m}$ is a unitary matrix (orthogonal in the real case) and $R(m:-1:1,:)$ is upper triangular. If $M = Q_f R_f$ is a (full) QR factorization of M and

$$P = \begin{bmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{bmatrix},$$

then a bottom-up QR factorization of M is $M = QR$ where $Q = Q_f P^H$ and $R = PR_f$. The reason for using the bottom-up QR factorization is that if $m > n$, then the last $m-n$ rows of R are zero rows in the "usual" QR factorization of M. However, we will need a QR factorization of M where the first $m-n$ rows of R are zero.

We discuss the complex and real cases in different subsections.

4.1. The complex case. We start by proving two lemmas.

LEMMA 15. *Assume that $A \in \mathbb{C}^{2n\times 2n}$ $(n \geq 2)$ has at least two distinct eigenvalues α and β with α of geometric multiplicity at most n. Then there exists a Schur form of A, $A = UTU^H$, such that*

(i) $T(1\colon 2, 1\colon 2) = \begin{bmatrix} \beta & * \\ 0 & \alpha \end{bmatrix}$.

(ii) *α is an eigenvalue of $T(3\colon 2n, 3\colon 2n)$ with geometric multiplicity at most $n-1$.*

Proof. Let $\widehat{T}$ be a Schur form of A. By using, if necessary, the Bai–Demmel algorithm [1], we can assume that the blocks in the diagonal of $\widehat{T}$ are so that $\widehat{T}(1\colon 2, 1\colon 2)$ is as in (i). Then the condition (ii) necessarily holds if the geometric multiplicity of α as an eigenvalue of A is less than n. Hence below we suppose that it is equal to n.

Let $m_1 \geq m_2 \geq \cdots \geq m_n$ be the partial multiplicities of α as an eigenvalue of A (that is, the sizes of the Jordan blocks) and $s = m_1 + \cdots + m_n$. Since A has at least two distinct eigenvalues, $n \leq s < 2n$ and, given that the geometric multiplicity of α is n and $s < 2n$, we have $m_n = 1$. We aim to detect one eigenvalue α in the diagonal of $\widehat{T}$ associated with a Jordan block of size 1. This will be the copy of α to be placed in $T(1:2, 1:2)$.

We use again the Bai–Demmel algorithm [1] to reorder the diagonal of $\widehat{T}(2\colon 2n, 2\colon 2n)$ so that in the new matrix T_0 the s copies of the eigenvalue α appear in the submatrix $T_1 = T_0(2\colon s+1, 2\colon s+1)$. Observe that T_0 is still a Schur form of A.

Thus α is the only eigenvalue of T_1 and recall that we are assuming that its geometric multiplicity is n. Let $Q_1 \in \mathbb{C}^{s\times n}$ be a matrix whose columns are an orthonormal basis of $\operatorname{Ker}(T_1 - \alpha I_s)$ and complete Q_1 up to a unitary matrix $\widetilde{Q} = [Q_1\ \widetilde{Q}_1] \in \mathbb{C}^{s\times s}$ (using a full QR factorization of Q_1, for example). Then we have

$$\widetilde{Q}^H T_1 \widetilde{Q} = \begin{bmatrix} \alpha I_n & C_1 \\ 0 & B \end{bmatrix} \tag{16}$$

for some $B \in \mathbb{F}^{(n-s)\times(n-s)}$ and $C_1 \in \mathbb{F}^{n\times(n-s)}$. Let $C = \widetilde{Q}_2 R$ be a bottom-up QR factorization of C. Since $s < 2n$, C has more rows than columns and so the entries of the first row of R are all zero. Hence, if $\widehat{Q} = \operatorname{diag}(\widetilde{Q}_2, I_{s-n})\widetilde{Q}$, then

$$\widehat{Q}^H T_1 \widehat{Q} = \begin{bmatrix} \alpha I_n & R \\ 0 & B \end{bmatrix}.$$

Next, if $Q_2^H B Q_2 = T_B$ is a Schur decomposition of B and

$$Q = \begin{bmatrix} 1 & & \\ & \widehat{Q} & \\ & & I_{2n-s-1} \end{bmatrix} \begin{bmatrix} I_{n+1} & & \\ & Q_2 & \\ & & I_{2n-s-1} \end{bmatrix},$$

then

$$T_2 = Q^H T_0 Q = \begin{bmatrix} \beta & * & * & * \\ & \alpha I_n & RQ_2 & * \\ & & T_B & * \\ & & & T_D \end{bmatrix}$$

is a Schur decomposition of A and α is not an eigenvalue of T_D. Notice that the first row of RQ_2 is still zero and so the first row of $T_2(2:s+1, 2:s+1) - \alpha I_s$ is also zero. Therefore, $\operatorname{rank}(T_2(2:s+1, 2:s+1) - \alpha I_s) = \operatorname{rank}(T_2(3:s+1, 3:s+1) - \alpha I_{s-1})$. Since T_0 and T_2 are similar, the geometric multiplicity of α as an eigenvalue of $T_2(2:s+1, 2:s+1)$ is n and so $s - n = \operatorname{rank}(T_2(2:s+1, 2:s+1) - \alpha I_s) = \operatorname{rank}(T_2(3:s+1, 3:s+1) - \alpha I_{s-1})$. This means that $\operatorname{null}(T_2(3:s+1, 3:s+1) - \alpha I_{s-1}) = n-1$. That is to say, the geometric multiplicity of α as eigenvalue of $T_2(3:s+1, 3:s+1)$ is $n-1$ and T_2 satisfies conditions (i) and (ii). ∎

We note that the structure in (16) is the first step of a proof of the Jordan canonical form (e.g., [18, sect. 2.4]), and a further reduction of B establishes the Weyr characteristics [13], leading to the Weyr canonical form.

The next lemma is needed for dealing with a matrix with only one real eigenvalue.

LEMMA 16. *Let $A \in \mathbb{F}^{2n\times 2n}$ ($\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$) be upper triangular with zero diagonal entries and assume that the geometric multiplicity of the zero eigenvalue is at most n. Then there exists a unitary U (orthogonal if $\mathbb{F} = \mathbb{R}$) such that $U^H AU$ is upper triangular with nonderogatory 2×2 diagonal blocks.*

Proof. Notice that the hypothesis about the geometric multiplicity of zero as an eigenvalue of A is equivalent to $\operatorname{rank}(A) \geq n$. We will assume $\mathbb{F} = \mathbb{C}$ but the proof for the real case is the same changing unitary matrices by orthogonal matrices.

We use induction on n. For $n = 1$, $\operatorname{rank}(A) \geq n$ implies that $A = \left[\begin{smallmatrix} 0 & a_{12} \\ 0 & 0 \end{smallmatrix}\right]$ with $a_{12} \neq 0$, that is, A is nonderogatory. Suppose the result holds for $n-1$. Let $A \in \mathbb{C}^{2n\times 2n}$ be upper triangular with zero diagonal and $\operatorname{rank}(A) \geq n$. If $a_{12} = 0$, then we can unitarily transform A so that its $(1,2)$ entry becomes nonzero as follows. Use a sequence of Givens rotations G to transform the first nonzero column of A, say Ae_m, $m \geq 2$, to a multiple of e_1. Then $G^H AG$ is still upper triangular with zero diagonal, first $m-1$ columns equal to zero, and the mth column equal to a multiple of e_1. Then we move the $(1,m)$ nonzero entry to the $(1,2)$ position with a permutation $P_{2,m}$, where $P_{2,m}$ swaps the second and mth row/column of $G^H AG$. The resulting matrix $P_{2,m}^H G^H AGP_{2,m}$ is still upper triangular. Hence below we assume that $a_{12} \neq 0$ in A.

Write $A = \left[\begin{smallmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{smallmatrix}\right]$, where $A_{11} = A(1\colon 2, 1\colon 2) = \left[\begin{smallmatrix} 0 & a_{12} \\ 0 & 0 \end{smallmatrix}\right]$. To use the induction hypothesis, we need to make sure that A_{22} is upper triangular with $\operatorname{rank}(A_{22}) \geq n-1$, that is, the geometric multiplicity of the eigenvalue zero is at most $n-1$. Since it cannot be greater than n, $\operatorname{rank}(A_{22}) \geq n-2$ and so care is needed only when $\operatorname{rank}(A_{22}) = n-2$. Notice first that, in this case, $n \geq \operatorname{null}(A) \geq \operatorname{null}(A_{22}) = n$ and so $\operatorname{rank}(A) = \operatorname{null}(A) = n$. Hence the geometric multiplicity of the eigenvalue zero in A is n. Now, $\operatorname{rank}(A) = n$ and $\operatorname{rank}(A_{22}) = n-2$ imply that the second row of A cannot be zero. In order to unitarily transform A so that $\operatorname{rank}(A_{22}) = n-1$, we can use the same technique as that in the proof of Lemma 15: since the geometric multiplicity of the eigenvalue zero is n, we can unitarily transform $\widehat{A} = Q^H AQ$ with $Q = [\,1\,] \oplus \widehat{Q}_1$ so that $\widehat{A}$ is upper triangular and the first n columns of $\widehat{A}(2:2n, 2:2n)$

become zero:

$$\widehat{A} = \left[\begin{array}{c|cc|c} 0 & a_{12} & * & * \\ \hline 0 & \multicolumn{2}{c|}{0_{n\times n}} & C \\ \hline 0 & \multicolumn{2}{c|}{0_{(n-1)\times n}} & B \end{array}\right].$$

Notice that the first column of $A(2:2n, 2:2n)$ is zero and so e_1 can be taken as the first column of $\widehat{Q}_1$. In other words, Q can be chosen to have the form $Q = I_2 \oplus Q_1$ with Q_1 a unitary matrix of order $2n-2$. Bearing in mind that the second row of A is not zero, we conclude that the second row of $\widehat{A}$ (and so the first row of C) is not zero.

It follows from $\operatorname{rank}(\widehat{A}) = n$ that $\operatorname{rank}\left(\left[\begin{smallmatrix} C \\ B \end{smallmatrix}\right]\right) = n-1$. If $\operatorname{rank}\left(\left[\begin{smallmatrix} C(2:n,:) \\ B \end{smallmatrix}\right]\right) = n-1$, then $\operatorname{rank}(\widehat{A}(3:2n, 3:2n)) = n-1$ and the induction hypothesis can be applied to $\widehat{A}$.

Assume now that $\operatorname{rank}\left(\left[\begin{smallmatrix} C(2:n,:) \\ B \end{smallmatrix}\right]\right) = n-2$. Since the size of C is $n \times (n-1)$, its rank is not greater than $n-1$, so there exists a row, say k, that linearly depends on the remaining rows of C. But the first row of C is not zero. Hence k can be chosen such that $1 < k \le n+1$. We can use a Givens rotation G applied to C in the planes $(1,k)$ to change the kth row of C in such a way that if $\widetilde{C} = GCG^H$, then

- $\operatorname{rank}\left(\left[\begin{smallmatrix} \widetilde{C}(2:n,:) \\ B \end{smallmatrix}\right]\right) = n-1$ and
- $\left[\begin{smallmatrix} a_{12} & * \end{smallmatrix}\right] G^H = \left[\begin{smallmatrix} \widetilde{a}_{12} & * \end{smallmatrix}\right]$ with $\widetilde{a}_{12} \neq 0$.

(Notice that "almost all" Givens rotations will have this property.) Now, if $Q_2 = \left[\begin{smallmatrix} 1 \end{smallmatrix}\right] \oplus G \oplus I_{n-1}$ and $\widetilde{A} = Q_2 \widehat{A} Q_2^H$, then $\widetilde{A}$ is upper triangular with zero diagonal, $\widetilde{A}(1:2, 1:2) = \left[\begin{smallmatrix} 0 & \widetilde{a}_{12} \\ 0 & 0 \end{smallmatrix}\right]$, and $\operatorname{rank}(\widetilde{A}(3:2n, 3:2n)) = n-1$. Hence the induction hypothesis can be applied to $\widetilde{A}$. ∎

We are now ready to describe an algorithm that stably computes the Schur form in Theorem 2 when $\ell = 2$. Let $A = U^H T U$ be any computed Schur decomposition of the matrix A in Theorem 2, and suppose that some of the 2×2 diagonal blocks of T are derogatory.

If all eigenvalues have algebraic multiplicity at most n, then we can reorder the diagonal entries of T using the Bai–Demmel algorithm [1], as was discussed in section 3.1.2. Thus assume that the eigenvalue α has algebraic multiplicity $n+t$ with $1 \le t \le n$. If $t = n$, we use the procedure described in the proof of Lemma 16 to further unitarily reduce $T - \alpha I$ to an upper triangular matrix T_1 with 2×2 nonderogatory diagonal blocks. $T_1 + \alpha I$ is the desired Schur form of A.

If $t < n$, note that all other eigenvalues must have algebraic multiplicity less than n. By using the Bai–Demmel algorithm [1] we pair as many α as possible with eigenvalues other than α thereby forming nonderogatory blocks in the top-left corner of T. In doing so, we use Lemma 15 to ensure that the resulting $2t \times 2t$ bottom-right corner of T has eigenvalue α with geometric multiplicity no larger than t. Thus we are left with

$$T = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \tag{17}$$

where $T_{22} \in \mathbb{C}^{2t \times 2t}$ contains 2×2 diagonal blocks with eigenvalue α and $\operatorname{rank}(T_{22} - \alpha I) \ge t$. Lemma 16 is then applied to $T_{22} - \alpha I$ as above to obtain a unitarily similar upper triangular matrix with nonderogatory 2×2 diagonal blocks.

4.2. The real case. To describe an algorithm that works in real arithmetic and computes the Schur decomposition in Theorem 3, we need a real version of Lemma 15.

LEMMA 17. *Let $A \in \mathbb{R}^{2n\times 2n}$ $(n \geq 2)$ and suppose that the spectrum of A contains a pair of nonreal complex eigenvalues $a \pm ib$ and a real eigenvalue α of geometric multiplicity at most n and algebraic multiplicity greater than n. Then there exists a Schur form T of A, such that*

(i) $T(1\colon 4, 1\colon 4) = \begin{bmatrix} a & b & * & * \\ -b & a & * & * \\ & & \alpha & * \\ & & & \alpha \end{bmatrix}$ *and*

(ii) *α is an eigenvalue of $T(5\colon 2n, 5\colon 2n)$ with geometric multiplicity at most $n-2$.*

Proof. Let $s \leq 2n-2$ and $k \leq n$ be the algebraic and geometric multiplicities of α as an eigenvalue of A. Consider an arbitrary real Schur form of A and reorder the diagonal blocks, using the Bai–Demmel algorithm [1] so as to obtain a Schur form T of A such that the leading 2×2 block is as in (i) and α appears on the diagonal of $\widehat{X} = T(3\colon s+2, 3\colon s+2)$. Thus

$$T = \begin{bmatrix} a & b & * & * \\ -b & a & * & * \\ & & \widehat{X} & * \\ & & & \widehat{Y} \end{bmatrix}$$

and α is not eigenvalue of $\widehat{Y}$. Then T satisfies condition (ii) of the lemma if and only if the geometric multiplicity of α as an eigenvalue of $\widehat{X}(3:s, 3:s)$ is at most $n-2$. Hence it is enough to show how to construct a Schur form, T_1, of $\widehat{X}$ such that the geometric multiplicity of α as an eigenvalue of $T_1(3:s, 3:s)$ is at most $n-2$. In fact, if Q_1 is an orthogonal matrix such that $Q_1^T\widehat{X}Q = T_1$ and $Q = I_2 \oplus Q_1 \oplus I_{2n-s-2}$, then Q^TTQ satisfies conditions (i) and (ii). In what follows we will show how to obtain the desired Schur form T_1 of $\widehat{X}$.

First, the geometric multiplicity of α as an eigenvalue of $\widehat{X}$ is the same as that of α as an eigenvalue of A and we are assuming that this is k. Then $\dim \operatorname{Ker}(\widehat{X} - \alpha\, I_s) = k \leq n$, and if $k \leq n-2$, then $\dim \operatorname{Ker}(\widehat{X}(3:s, 3:s) - \alpha\, I_s) \leq k \leq n-2$. This means that if $T_1 = \widehat{X}$, then the geometric multiplicity of $T_1(3:s, 3:s)$ is not greater than $n-2$. Hence we only have to analyze the cases $k = n$ and $k = n-1$. In both cases, we proceed as in Lemma 15. We find an orthogonal $\widehat{Q}_0$ such that (see (16))

$$T_0 = \widehat{Q}_0^T\widehat{X}\widehat{Q}_0 = \begin{bmatrix} \alpha I_k & C \\ 0 & T_B \end{bmatrix}$$

is upper triangular and compute a "bottom-up" QR factorization of C, $C = \widehat{Q}_1 R$. Define $Q_1 = \widehat{Q}_1 \oplus I_{s-k}$. Then $T_1 = \widehat{Q}_1^T T_0 \widehat{Q}_1$ is a real Schur form of $\widehat{X}$ and

$$T_1 = \begin{bmatrix} \alpha I_k & R \\ 0 & T_B \end{bmatrix}.$$

Recall that $s \leq 2n-2$ and for $k \leq n-2$ the lemma has been already proved. We split the remaining possibilities into three different cases. Each case requires a different proof.

(i) $k = n$, (ii) $k = n-1$ and $s < 2n-2$, (iii) $k = n-1$ and $s = 2n-2$.

(i) If $k = n$, then $\operatorname{null}(T_1 - \alpha I_s) = n$ and so $\operatorname{rank}(T_1 - \alpha I_s) = s - n$. Also, the size of R is $n \times (s-n)$ and $s - n \leq 2n-2-n = n-2$. Hence the two first rows

of R are zero and $s-n = \operatorname{rank}(T_1 - \alpha I_s) = \operatorname{rank}\left(\left[\begin{smallmatrix} R(3:n,:) \\ T_B - \alpha I_{s-n} \end{smallmatrix}\right]\right) = \operatorname{rank}(T_1(3:s,3:s) - \alpha I_{s-2})$. Thus, $\operatorname{null}(T_1(3:s,3:s) - \alpha I_{s-2}) = n-2$ meaning that the geometric multiplicity of α as an eigenvalue of $T_1(3:s,3:s)$ is $n-2$ as desired.

(ii) If $k = n-1$ and $s < 2n-2$, then $\operatorname{null}(X_1 - \alpha I_s) = n-1$, the size of R is $(n-1)\times(s-n+1)$, and $s-n+1 < 2n-2-n+1 = n-1$. Then the first row of R is zero and since $s-n+1 = \operatorname{rank}(T_1 - \alpha I_s) = \operatorname{rank}\left(\left[\begin{smallmatrix} R(2:n-1,:) \\ T_B - \alpha I_{s-n+1} \end{smallmatrix}\right]\right) = \operatorname{rank}(T_1(2:s,2:s) - \alpha I_{s-1})$, we get $\operatorname{rank}(T_1(3:s,3:s) - \alpha I_{s-2}) \geq s-n$ (with equality if the second row of R is not a linear combination of the remaining rows in $\left[\begin{smallmatrix} R \\ T_B - \alpha I_{s-n+1} \end{smallmatrix}\right]$). Thus $\operatorname{null}(T_1(3:s,3:s) - \alpha I_{s-2}) \leq (s-2)-(s-n) = n-2$. Again T_1 satisfies the desired conditions.

(ii) If $k = n-1$ and $s = 2n-2$, then $\widehat{X} - \alpha I_s$ fulfils the hypothesis of Lemma 16 because this matrix is nilpotent, its size is $2(n-1)$, and the geometric multiplicity of α is $n-1$. Following the procedure designed in the proof of that lemma, an orthogonal matrix U can be obtained such that $U^T(\widehat{X} - \alpha I_{2n-2})U$ is upper triangular with nonderogatory 2×2 diagonal blocks. Therefore, $T_1 = U^T \widehat{X} U$ is a Schur form of $\widehat{X}$ such that the geometric multiplicity of α as an eigenvalue of $T_1(3:s,3:s)$ is at most $n-2$ as desired. ∎

We now have the artillery to describe a stable algorithm that computes the Schur form of A in Theorem 3 for $\ell = 2$. Suppose a real Schur form of A is given. Any 2×2 block on the diagonal associated to a pair of nonreal complex conjugate eigenvalues is obviously nonderogatory, so we need only take care of the real eigenvalues. The case when all eigenvalues have algebraic multiplicities at most n was discussed in section 3. In the real case with $\ell = 2$ there cannot be nonreal complex eigenvalues of algebraic multiplicity greater than n. Hence, we only have to deal with the case when exactly one real eigenvalue α has algebraic multiplicity greater than n. We first use Lemma 17 as many times as possible, that is, we pair two copies of α with as many pairs of nonreal complex conjugate eigenvalues as possible. After doing this we are left with real eigenvalues only. Henceforth, Lemmas 15 and 16 can be used as in the complex case to get a Schur form of A with all its diagonal blocks either nonderogatory of size 2×2 or of size 4×4 with eigenvalues whose geometric multiplicity is at most two.

We illustrate this process in the following long but complete example.

Example 18. Let $n = 4$, $\ell = 2$, and let A be the following matrix in real Schur form:

$$A = \begin{bmatrix} 1 & * & * & * & * & * & * & * \\ & 0 & -1 & * & * & * & * & * \\ & 1 & 0 & * & * & * & * & * \\ & & & 1 & * & * & * & * \\ & & & & 1 & * & * & * \\ & & & & & 1 & * & * \\ & & & & & & a & * \\ & & & & & & & 1 \end{bmatrix},$$

where a is either 1 or 2. Thus the distinct eigenvalues of A are 1, i, and $-i$ when $a = 1$ and 1, 2, i, and $-i$ when $a = 2$. In addition, the algebraic multiplicity of 1 as an eigenvalue of A is 5 or 6 according as $a = 2$ or $a = 1$. In both cases it is greater than n which is 4. Let us also assume that the geometric multiplicity of the eigenvalue 1 is 4. Under these conditions (see [4, Thm. 1.7]) A is a linearization of a 4×4 quadratic matrix polynomial $P(\lambda)$ with nonsingular leading coefficient. Now, by [17, Thm. 3.6], $P(\lambda)$ is not triangularizable over $\mathbb{R}[\lambda]$. Hence there is no real Schur form of A with four nonderogatory blocks of size 2×2 in the diagonal. In other words, any real Schur

form of A with the properties of Theorem 3 must have at least one block of size 4×4 with 1 as an eigenvalue of geometric multiplicity 2. This is consistent with Theorem 4.1 of [17] and must be revealed by the algorithmic process.

Our goal is to find an orthogonal matrix $Q \in \mathbb{R}^{8\times 8}$ such that $Q^T AQ$ is a real Schur form with two nonderogatory blocks of size 2×2 and one block of size 4×4 with only one real eigenvalue whose geometric multiplicity is 2. We use the procedure developed in Lemmas 15–17 as follows.

Step 1. We are under the hypothesis of Lemma 17. Use the Bai–Demmel algorithm to exchange entry $(1, 1)$ and block $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ to obtain a new real Schur form $T_1 = Q_1^T AQ_1$ with diagonal $\operatorname{diag}(B_1, 1, 1, 1, 1, a, 1)$ where B_1 is similar to $\left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$.

Let us assume now that $a = 2$. The case $a = 1$ will be dealt with later on.

Step 2. Use the Bai–Demmel algorithm to exchange entries $(7, 7)$ and $(8, 8)$ so that all equal eigenvalues appear together in the diagonal. We get a new real Schur form

$$T_2 = Q_2^T T_1 Q_2 = \left[\begin{array}{ccc|cccc} B_1 & * & * & * & * & * & * \\ & 1 & * & * & * & * & * \\ & & 1 & * & * & * & * \\ \hline & & & 1 & * & * & * \\ & & & & 1 & * & * \\ & & & & & 1 & * \\ & & & & & & 2 \end{array}\right].$$

T_2 satisfies condition (i) of Lemma 17. We proceed as indicated in the proof of that lemma to produce a real Schur form which also satisfies condition (ii).

Step 3. Extract the following submatrix $T_2(3 : 7, 3 : 7)$:

$$X = \begin{bmatrix} 1 & * & * & * & * \\ & 1 & * & * & * \\ & & 1 & * & * \\ & & & 1 & * \\ & & & & 1 \end{bmatrix},$$

and obtain an orthonormal basis $\widehat{Q}_1$ of $\operatorname{Ker}(X - I_5)$ (using the singular value decomposition, for instance). By hypothesis, the geometric multiplicity of the eigenvalue 1 is 4. Thus $\dim \operatorname{Ker}(X - I_5) = 4$ and so the size of $\widehat{Q}_1$ is 5×4. Complete $\widehat{Q}_1$, as in Lemma 15, up to an orthogonal matrix $\widetilde{Q}_1$ (using, for example, the QR factorization of $\widehat{Q}_1$). Then

$$\widetilde{Q}_1^T X \widetilde{Q}_1 = \begin{bmatrix} I_4 & c \\ 0 & 1 \end{bmatrix},$$

where c is not the zero vector. Compute a bottom-up QR factorization of c. In this case we can use a Householder reflection, $\widehat{Q}_2^T (= \widehat{Q}_2)$ so that $\widehat{Q}_2^T c = [\,0 \quad 0 \quad 0 \quad d\,]^T$ with $d \neq 0$. Define $\widetilde{Q}_2 = \operatorname{diag}(\widehat{Q}_2, 1)$ and $Q_3 = \operatorname{diag}(I_2, \widetilde{Q}_1 \widetilde{Q}_2, 1)$. Then

$$T_3 = Q_3^T T_2 Q_3 = \left[\begin{array}{ccc|cccc} B & * & * & * & * & * & * \\ & 1 & 0 & 0 & 0 & 0 & * \\ & & 1 & 0 & 0 & 0 & * \\ \hline & & & 1 & 0 & 0 & * \\ & & & & 1 & d & * \\ & & & & & 1 & * \\ & & & & & & 2 \end{array}\right].$$

Since $d \neq 0$, the geometric multiplicity of 1 as an eigenvalue of $T_3(5 : 8, 5 : 8)$ is 2.

Step 4. We could successively permute the first and second rows and columns and then the second and third rows and columns of $T_3(5 : 8, 5 : 8)$ to get the desired

matrix. However, $T_3(5:8,5:8)$ satisfies the hypothesis of Lemma 15, and we will proceed as indicated in its proof: use the Bai–Demmel algorithm to swap block $\left[\begin{smallmatrix}1 & d\\ & 1\end{smallmatrix}\right]$ and the entry 2 in position $(8,8)$. The resulting matrix is

$$T_4 = \left[\begin{array}{ccc|cc|c} B & * & * & * & * & * \\ & 1 & 0 & 0 & * & * \\ & & 1 & 0 & * & * \\ \hline & & & 1 & * & * \\ & & & & 2 & * \\ \hline & & & & & C \end{array}\right],$$

where C is similar to $\left[\begin{smallmatrix}1 & d\\ & 1\end{smallmatrix}\right]$, so it is nonderogatory. If C is itself upper triangular, then T_4 is the desired matrix. Otherwise,

Step 5. Reduce C to upper triangular form by orthogonal similarity and apply it to the last two rows and columns of T_4 to obtain

$$T_5 = \left[\begin{array}{ccc|cc|cc} B & * & * & * & * & * & * \\ & 1 & 0 & 0 & * & * & * \\ & & 1 & 0 & * & * & * \\ \hline & & & 1 & * & * & * \\ & & & & 2 & * & * \\ \hline & & & & & 1 & g \\ & & & & & & 1 \end{array}\right].$$

T_5 is a real Schur form of A with two nonderogatory blocks of size 2×2 and one block of size 4×4 in the diagonal. The geometric multiplicity of 1 as an eigenvalue of $T_5(1:4,1:4)$ is two; thus we have a desired Schur form for $a=2$.

Assume now that $a=1$. Since, in this case, the entries in positions $(7,7)$ and $(8,8)$ are both equal to 1 there is no need to exchange these diagonal elements. We put $T_2 = T_1$ and go straight ahead to the next step.

Step 3. Let $X = T_2(3:8,3:8)$ and the columns of $\widehat{Q}_1$ be an orthonormal basis of $\operatorname{Ker}(X - I_6)$. This is a 6×4 matrix with orthonormal columns. We can complete it to a 6×6 orthogonal matrix $\widetilde{Q}_1$ such that

$$\widetilde{Q}_1^T X \widetilde{Q}_1 = \begin{bmatrix} I_4 & C \\ 0 & T_C \end{bmatrix},$$

where $T_C = \left[\begin{smallmatrix}1 & r_{32}\\ & 1\end{smallmatrix}\right]$. Compute a bottom-up QR factorization of $C = \widehat{Q}_2 R$ and define $Q_3 = \operatorname{diag}(I_2, \widetilde{Q}_1\widetilde{Q}_2)$ with $\widetilde{Q}_2 = \operatorname{diag}(\widehat{Q}_2, I_2)$. Then

$$T_3 = Q_3^T T_2 Q_3 = \left[\begin{array}{c|cccccc} B & * & * & * & * & * & * \\ \hline & 1 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 & r_{12} \\ & & & & 1 & r_{21} & r_{22} \\ & & & & & 1 & r_{32} \\ & & & & & & 1 \end{array}\right],$$

where $r_{21}r_{12} \neq 0$ or $r_{21}r_{32} \neq 0$ (that is, $r_{21} \neq 0$ and at least one of r_{12} or r_{32} is not zero) because otherwise $\operatorname{null}(T_3 - I) > 4$. Thus T_3 is a real Schur form of A which satisfies conditions (i) and (ii) of Lemma 17.

Step 4. Deflate $T_3(1:4,1:4)$ and pay attention to $Y = T_3(5:8,5:8)$. This is a 4×4 real matrix with all eigenvalues 1. Its algebraic multiplicity is 4 and its geometric multiplicity is 2 because $r_{21}r_{12} \neq 0$ or $r_{21}r_{32} \neq 0$. We use the proof of Lemma 16 to get a real Schur form with two nonderogatory blocks of size 2×2 in the diagonal.

First, we define $Z = Y - I_4$, which is a nilpotent matrix, and notice that if S is a real Schur form of Z, then $S + I_4$ is a real Schur form of Y. Now we apply the method proposed in the proof of Lemma 16: observe that the first nonzero column of Z is the third one, so we use a Givens rotation in order to replace that column by a multiple of e_1. In the present case a permutation of the first and second rows and columns suffices:

$$P_1^T Z P_1 = \begin{bmatrix} 0 & 0 & r_{21} & r_{22} \\ 0 & 0 & 0 & r_{12} \\ 0 & 0 & 0 & r_{32} \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Next, we permute the second and third rows and columns to get

$$Z_1 = P_2^T P_1^T Z P_1 P_2 = \begin{bmatrix} 0 & r_{21} & 0 & r_{22} \\ 0 & 0 & 0 & r_{32} \\ 0 & 0 & 0 & r_{12} \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

With the notation of Lemma 15, $A_{22} = \left[\begin{smallmatrix} 0 & r_{12} \\ 0 & 0 \end{smallmatrix}\right]$ and $n = 2$. Thus, if $r_{12} \neq 0$, then $\operatorname{rank}(A_{22}) = 1 = n - 1$ and no further transformation is needed on Z_1 because it is upper triangular with 2×2 nonderogatory diagonal blocks. But if $r_{12} = 0$, then $\operatorname{rank}(A_{22}) = 0 = n - 2$ and one additional transformation is needed. In fact, as shown above and in the proof of Lemma 16, $r_{32} \neq 0$, and we can perform a Givens rotation on rows and columns two and three to place simultaneously nonzero elements in $(1, 2)$ and $(3, 4)$ of Z_1.

Summarizing, there is an orthogonal matrix $Q = Q_1 Q_3 Q_4$ with $Q_4 = I_4 \oplus P_1 P_2 G$ where G is an appropriate Givens rotation (the identity if $r_{12} \neq 0$) such that

$$T = Q^T A Q = \left[\begin{array}{ccc|cc|cc} B & * & * & * & * & * & * \\ & 1 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 & 0 \\ \hline & & & 1 & s_{12} & s_{13} & s_{14} \\ & & & & 1 & s_{23} & s_{24} \\ \hline & & & & & 1 & s_{34} \\ & & & & & & 1 \end{array}\right],$$

and $s_{12} \neq 0$ and $s_{34} \neq 0$. $Q^T A Q$ is a real Schur form of A with two nonderogatory blocks of size 2×2 and one block of size 4×4 in the diagonal. Again, the geometric multiplicity of 1 as an eigenvalue of $T(1:4, 1:4)$ is two.

5. Parameterized linear systems. We consider parameterized linear systems of the form

$$P(\omega)x = b(\omega), \quad x = x(\omega). \tag{18}$$

These types of systems appear when computing numerical solutions of differential equations which arise in areas including electromagnetic scattering, wave propagation in porous media, or structural dynamics (see, for example, [6, 9, 14] and the references therein). The coefficient matrix in (18) is the matrix polynomial in (1) and b may be constant [11, 14] or a (in general, nonlinear) function of the parameter ω [6, 9]. For quadratic matrix polynomials ω is either real or pure imaginary with $|\omega| \in \mathcal{I} = [\omega_\ell, \omega_h]$, $\omega_\ell \ll \omega_h$ [6, 9, 11, 14], and the solution of (18) is to be computed for many values of the parameter ω. In particular, in [9] $b(\omega)$ is supposed to be analytic in $\mathcal{I}$ except at points ω where $\det P(\omega) = 0$; the solution $x(\omega)$ then inherits the same property. Whether we are interested in analytic solutions of (18) or in solutions for

finitely many values of ω, reduced forms $R(\omega)$ of $P(\omega)$ can be used to convert system (18) into a simpler equivalent one

$$R(\omega)y = c(\omega), \quad y = y(\omega). \tag{19}$$

We have shown in the previous sections a procedure to compute $R(\omega)$ from $P(\omega)$ without using unimodular transformations. We must show how to compute $c(\omega)$ so that systems (18) and (19) are equivalent. We will show a little more: how to obtain $c(\omega)$ from $b(\omega)$ so that the solution of (18) can be given explicitly in terms of $b(\omega)$ and the solution of (19). As a result we will give an explicit expression of the solution of (18) in terms of $y(\omega)$ and $R(\omega)$ for every $\omega \in \mathbb{C}$ which is not an eigenvalue of P. For simplicity we will consider the case where $R(\omega)$ is triangular.

Let $C_L(P)$ be the left companion matrix of $P(\omega)$. In computing $R(\omega)$ we first use the algorithm of section 3.1.2 (or, in the quadratic case, those of section 4 if needed) to compute a Schur form of $C_L(P)$ satisfying the properties of Theorem 2 (i.e., $T_{ii} \in \mathbb{C}^{\ell\times\ell}$ is upper triangular and nonderogatory). Let $T = Q^H C_L(P)Q$ be such a Schur form. Then we use Proposition 9 to obtain an $\ell n \times n$ matrix $X = [\, x_1 \quad x_2 \quad \cdots \quad x_n \,]$ such that $V = [\, X \quad TX \quad \cdots \quad T^{\ell-1}X \,]$ is nonsingular and $\mathcal{K}_\ell(A, [\, x_1 \quad x_2 \quad \cdots \quad x_i \,])$ is A-invariant for $1 \le i \le n-1$ (notice that V is block-triangular because, following the proof of Proposition 9, X is of the form $X = v_1 \oplus v_2 \oplus \cdots \oplus v_n$ with v_i of size $\ell \times 1$, $i = 1, \ldots, n$). Then it follows from Theorem 1 that $V^{-1}TV = C_L(R)$ is the left companion matrix of a triangular matrix polynomial of degree ℓ. Thus, if $S = QV$, then $S^{-1}C_L(P)S = C_L(R)$ for some upper triangular matrix polynomial. This is the matrix polynomial $R(\omega)$ of system (19).

Now we are going to find $c(\omega)$ so that the solution $x(\omega)$ of system (18) can be explicitly given in terms of $R(\omega)$ and the solution of (19) for that $c(\omega)$.

Using [4, Prop. 1.2] we have

$$\begin{aligned} P(\omega)^{-1} &= (e_\ell^T \otimes I_n)(\omega I - C_L(P))^{-1}(e_1 \otimes I_n) \\ &= (e_\ell^T \otimes I_n)S\big(\omega I - C_L(R)\big)^{-1}S^{-1}(e_1 \otimes I_n) \\ &= S(\ell(n-1)+1 : \ell n, :)\big(\omega I - C_L(R)\big)^{-1}S^{-1}(:, 1:n). \end{aligned} \tag{20}$$

A direct computation shows that

$$\big(\omega I - C_L(R)\big)^{-1} = E(\omega)\begin{bmatrix} R(\omega)^{-1} & 0 \\ 0 & I \end{bmatrix} F(\omega),$$

where

$$E(\omega) = \begin{bmatrix} B_{\ell-1}(\omega) & -I & 0 & \cdots & 0 \\ B_{\ell-2}(\omega) & 0 & -I & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & & \ddots & -I \\ B_0(\omega) & 0 & \cdots & \cdots & 0 \end{bmatrix}, \quad F(\omega) = \begin{bmatrix} I & \omega I & \cdots & \omega^{\ell-2}I & \omega^{\ell-1}I \\ 0 & \ddots & \ddots & & \omega^{\ell-2}I \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \omega I \\ 0 & \cdots & \cdots & 0 & I \end{bmatrix}$$

with $B_0(\omega) = I$ and $B_j(\omega) = \omega B_{j-1}(\omega) + R_{\ell-j}$ for $j = 1, \ldots, \ell-1$.

Let $Y = [\, Y_1^T \quad \cdots \quad Y_\ell^T \,]^T$ and $Z = [\, Z_1 \quad \cdots \quad Z_\ell \,]$ be the first n columns of S^{-1} and the last n columns of S, respectively. That is, $Z = S(\ell(n-1)+1 : \ell n, :)$ and $Y = S^{-1}(:, 1:n)$. On substituting $\big(\omega I - C_L(R)\big)^{-1}$, Y, and Z in (20) we get

$$P(\omega)^{-1} = \left(\sum_{i=1}^{\ell} Z_i B_{\ell-i}(\omega)\right) R(\omega)^{-1} \left(\sum_{i=1}^{\ell} \omega^{i-1} Y_i\right) - \sum_{j=1}^{\ell-1} Z_j \left(\sum_{i=j+1}^{\ell} \omega^{i-(j+1)} Y_i\right).$$

If we let $c(\omega) = \left(\sum_{i=1}^{\ell} \omega^{i-1} Y_i\right) b(\omega)$ and solve (19) for $y(\omega)$, then for the solution $x(\omega)$ to the parameterized linear system (18) we have

$$x(\omega) = \sum_{i=1}^{\ell} Z_i B_{\ell-i}(\omega) y(\omega) - \sum_{j=1}^{\ell-1} Z_j \left(\sum_{i=j+1}^{\ell} \omega^{i-(j+1)} Y_i b(\omega) \right).$$

The structure of S and that of the left companion matrix can be exploited to construct the last n rows of S and the first n columns of S^{-1}. For each value of ω, $x(\omega)$ can be computed in $O(n\ell^2 + n^2\ell)$ operations, by precomputing $Y_i b(\omega)$ for every i and reusing them to obtain the second term in $x(\omega)$.

6. Conclusions. All matrix polynomials with nonsingular leading coefficients can be reduced to triangular form while keeping the size, degree, and eigenstructure of the original matrix polynomial by means of unimodular transformations. We do not have a practical way to compute the unimodular transformations, so instead, we have proposed a practical procedure that, starting from a Schur form of any linearization $\lambda I - A$ of a given $n \times n$ matrix polynomial of degree ℓ, consists of three steps:

1. Moving the diagonal elements (and the 2×2 diagonal blocks, in the real case) of the Schur form so as to obtain a new Schur form satisfying the properties of Theorems 2 or 3.
2. Using the obtained Schur form to construct a full column rank matrix X satisfying the conditions of Theorem 1 (X may be taken to have orthonormal columns).
3. Performing a structure preserving similarity transformation $S = K_\ell(A, X)$ as in (2) so that $S^{-1}AS$ is the left companion matrix of a monic triangular matrix polynomial of degree ℓ (only the last n columns of $S^{-1}AS$ are needed).

We showed how to implement step 1 in a stable way so that the procedure reduces any quadratic matrix polynomials to triangular form. For $\ell > 2$, however, we only discussed how to succeed with step 1 in the case when no eigenvalue has algebraic multiplicity larger than n.

Reduction to other simple forms like block-diagonal, block-triangular, or Hessenberg forms was also considered. In particular, it was shown that if a Hessenberg form of a linearization, when partitioned in $\ell \times \ell$ blocks, has unreduced diagonal blocks, then the matrix polynomial can be brought to Hessenberg form using steps 2 and 3 above (with the obvious substitutions "Schur form" by "Hessenberg form" and "triangular matrix" by "Hessenberg matrix").

Acknowledgments. The authors wish to thank the anonymous referees for several comments and suggestions which led to improvements in this paper.

REFERENCES

[1] Z. Bai and J. W. Demmel, *On swapping diagonal blocks in real Schur form*, Linear Algebra Appl., 186 (1993), pp. 73–95, https://doi.org/10.1016/0024-3795(93)90286-W.

[2] T. R. Cameron, *On the reduction of matrix polynomials to Hessenberg form*, Electron. J. Linear Algebra, 31 (2016), pp. 321–334, https://doi.org/10.13001/1081-3810.3011.

[3] S. D. Garvey, M. I. Friswell, and U. Prells, *Co-ordinate transformations for second order systems. Part* I*: General transformations*, J. Sound Vib., 258 (2002), pp. 885–909, https://doi.org/10.1006/jsvi.2002.5165.

[4] I. Gohberg, P. Lancaster, and L. Rodman, *Matrix Polynomials*, Classics Appl. Math. 58, SIAM, Philadelphia, 2009. Unabridged republication of book first published by Academic Press in 1982, https://doi.org/10.1137/1.9780898719024.

[5] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed., The Johns Hopkins University Press, Baltimore, 2012.

[6] G.-D. Gu and V. Simoncini, *Numerical solution of parameter-dependent linear systems*, Numer. Linear Algebra Appl., 12 (2005), pp. 923–940, https://doi.org/10.1002/nla.442, https://doi.org/10.1002/nla.442.

[7] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, UK, 2013, https://doi.org/10.1017/9781139020411.

[8] T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[9] M. Kuzuoglu and R. Mittra, *Finite element solution of electromagnetic problems over a wide frequency range via the Padé approximation*, Comput. Methods Appl. Mech. Engrg., 169 (1999), pp. 263–277, https://doi.org/10.1016/S0045-7825(98)00157-1.

[10] P. Lancaster and I. Zaballa, *Diagonalizable quadratic eigenvalue problems*, Mech. Systems Signal Process, 23 (2009), pp. 1134–1144, https://doi.org/10.1016/j.ymssp.2008.11.007.

[11] K. Meerbergen, *Fast frequency response computation for Rayleigh damping*, Internat. J. Numer. Methods Eng., 73 (2008), pp. 96–106, https://doi.org/10.1002/nme.2058.

[12] M. Morzfeld, F. Ma, and B. N. Parlett, *The transformation of second-order linear systems into independent equations*, SIAM J. Appl. Math., 71 (2011), pp. 1026–1043, https://doi.org/10.1137/100818637.

[13] H. Shapiro, *The Weyr characteristic*, Amer. Math. Monthly, 106 (1999), pp. 919–929, https://doi.org/10.2307/2589746.

[14] V. Simoncini and F. Perotti, *On the numerical solution of* $(\lambda^2 A + \lambda B + C)x = b$ *and application to structural dynamics*, SIAM J. Sci. Comput., 23 (2002), pp. 1875–1897, https://doi.org/10.1137/S1064827501383373.

[15] L. Taslaman, *Algorithms and Theory for Polynomial Eigenproblems*, Ph.D. thesis, The University of Manchester, Manchester, UK, 2014. Available as MIMS EPrint 2015.4: http://eprints.ma.man.ac.uk/2237.

[16] L. Taslaman, F. Tisseur, and I. Zaballa, *Triangularizing matrix polynomials*, Linear Algebra Appl., 439 (2013), pp. 1679–1699, https://doi.org/10.1016/j.laa.2013.05.006.

[17] F. Tisseur and I. Zaballa, *Triangularizing quadratic matrix polynomials*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 312–337, https://doi.org/10.1137/120867640.

[18] D. Watkins, *The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods*, SIAM, Philadelphia, 2007, https://doi.org/10.1137/1.9780898717808.

[19] I. Zaballa, *Interlacing inequalities and control theory*, Linear Algebra Appl., 101 (1988), pp. 9–31, https://doi.org/10.1016/0024-3795(88)90140-1.

[20] J. C. Zúñiga Anaya, *Diagonalization of quadratic matrix polynomials*, Systems Control Lett., 59 (2010), pp. 105–113, https://doi.org/10.1016/j.sysconle.2009.12.005.