Optimal Scaling of Companion Pencils for the QZ-Algorithm*†

D. Lemonnier‡, *P. Van Dooren*‡

1 Introduction

Computing roots of a monic polynomial may be done by computing the eigenvalues of the corresponding companion matrix using for instance the well-known QR-algorithm. We know this algorithm to be backward stable since it computes exact eigenvalues of a slightly modified matrix. But it may yield very poor backward errors in the coefficients of the polynomial. In this paper we investigate what can be done to improve these errors, using a geometric approach. We will see that preconditioning the companion matrix using some carefully chosen similarity may achieve this goal. In particular, we will give a geometric interpretation of what balancing the companion matrix does. We then naturally extend these results for the non-monic polynomial case where the algorithm we deal with is now the QZ-algorithm acting on companion pencils instead of companion matrices.

The article is divided into two parts: in the first one we examine the monic scalar polynomial case and in the second one the general non-monic scalar case. In each part, we begin by explaining the problem in terms of error analysis, and then we look at the problem from a geometric point of view.

*This paper presents research supported by the Belgian Programme on Inter-university Poles of Attraction, initiated by the Belgian State, Prime Minister's Office for Science, Technology and Culture. This work was also supported by the National Science Foundation under Grant No. CCR-20003050.

†We thank W.-W. Lin and H. Fan (Tsing Hua University, Taiwan) for some useful discussions regarding this paper.

‡Department of Mathematical Engineering, Université Catholique de Louvain, Belgium.

2 The monic scalar case

2.1 Problem statement

We consider the nth-order polynomial $p : \mathbb{C} \to \mathbb{C} : \lambda \to p(\lambda) \doteq \lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0$ with complex coefficients a_k . We can associate with this polynomial the companion matrix

$$C = \begin{bmatrix} 0 & & & & -a_0 \\ 1 & 0 & & & -a_1 \\ & 1 & & & -a_2 \\ & & \ddots & & \vdots \\ & & & 1 & -a_{n-1} \end{bmatrix} \tag{1}$$

whose characteristic polynomial $\chi_C(\lambda) \doteq \det(\lambda I - C) = p(\lambda)$. We then compute the roots of $p(\lambda)$ using any eigenvalue algorithm. If we use the standard QR-algorithm, the computed eigenvalues are exactly those of a matrix $C + \Delta$ for some dense backward error matrix Δ. We can state that this backward error satisfies

$$\|\Delta\| \;=\; O(\epsilon)\|C\| \tag{2}$$

where ϵ denotes the machine precision, and $\|\cdot\|$ the Frobenius norm for example. However we are interested not by the matrix itself but by $p(\lambda)$, its characteristic polynomial. So, we would like to require that the computed eigenvalues are precisely the roots of a polynomial $p(\lambda) + \delta(\lambda)$ where $\delta(\lambda) = \delta a_{n-1}\lambda^{n-1} + \ldots + \delta a_1\lambda + \delta a_0$ satisfies $\|\delta(.)\| = O(\epsilon)\|p(.)\|$ for some polynomial norm, which would mean backward stability for our root-seeking problem. But a more interesting requirement is the so-called *componentwise backward stability* :

$$\max_k \frac{|\delta a_k|}{|a_k|} = O(\epsilon) \; . \tag{3}$$

If (3) holds, we know that the computed roots are those of a polynomial that is ϵ-close to the original polynomial in a relative componentwise sense.

2.2 A geometric approach

Let us consider the euclidian matrix space $\mathbb{C}^{n\times n}$ with the usual Frobenius inner product :

$$\langle A \; , \; B \rangle \;\doteq\; \mathrm{tr}(AB^*) \; .$$

In this space, we consider the manifold

$$\mathbf{Orb} \;\doteq\; \mathrm{orbit}(C) \;=\; \{T^{-1}CT : \det(T) \neq 0\} \; .$$

The dimension of this manifold is $n^2 - n$. A first order calculation shows that the tangent space to this manifold at C, written **Tan**, is the set of *additive commutators* $\{XC - CX : X \in \mathbb{C}^{n\times n}\}$, and the normal space **Nor** at the same point C (of

dimension n) is the *centralizer* of C^* : $\{X \in \mathbb{C}^{n\times n} : XC^* = C^*X\}$, i.e. the set of matrices that commute with C^*.
It has been already shown in [6] how a dense perturbation Δ to the matrix C leads to first order perturbations in the coefficients a_k. To do so, one considers the often called Sylvester space **Syl**, namely the set of companion matrices. It is a n-dimensional affine space that goes through C. One easily shows (see for example [1]) that every matrix may then be decomposed as a linear combination of a companion matrix and a matrix in the tangent space. On the other side, one can prove that perturbations in the tangent space do not affect the coefficients of $\chi_C(\lambda)$ to first order. Indeed, for any small X we may write *up to first order*

$$\begin{aligned}\det((\lambda I - (C + (XC - CX))) &= \det(\lambda I - (I + X)C(I + X)^{-1}) \\ &= \det(I + X)\det(\lambda I - C)\det(I - X) \\ &= (1 + \mathrm{tr}(X))\det(\lambda I - C)(1 - \mathrm{tr}(X)) \\ &= \det(\lambda I - C)\ .\end{aligned}$$

So only the perturbations that lie in **Syl** play a role here up to first order. Decomposing Δ in **Syl** $\bigoplus$ **Tan**, expressing that **Tan** is orthogonal to **Nor**, and taking the special structure of the elements of **Nor** into account, it is then shown in [6] that :

$$|\delta a_k| = |\ \langle \Delta\ ,\ M_{k+1}^*\ \rangle\ | \tag{4}$$

where the M_k^* 's ($k = 1$ to n) defined as

$$M_k = \overbrace{\begin{bmatrix} a_k & & & \\ a_{k+1} & a_k & & \\ \vdots & a_{k+1} & \ddots & \\ a_n = 1 & \vdots & \ddots & a_k \\ & a_n = 1 & \ddots & a_{k+1} \\ & & \ddots & \vdots \\ & & & a_n = 1 \end{bmatrix}}^{k} \overbrace{\begin{bmatrix} -a_0 & & \\ -a_1 & \ddots & \\ \vdots & \ddots & -a_0 \\ -a_{k-1} & \ddots & \vdots \\ & \ddots & \vdots \\ & & -a_{k-1} \end{bmatrix}}^{n-k} \tag{5}$$

span the normal space **Nor**. Notice that the relation (4) only holds *to first order*. The matrices M_k are actually the polynomial coefficients of the adjoint matrix of C. Defining $M_0 = M_{n+1} = 0$, one can show [7] that these matrices satisfy

$$M_k \ = \ CM_{k+1} + a_k I \qquad \text{for } k = 0\ldots, n\ . \tag{6}$$

.

How far are we from getting (3) ? To answer this, we use the Cauchy-Schwarz inequality and the backward stability property of the QR-algorithm (2) which yields

$$|\delta a_k| = O(\epsilon)\ \|C\|\ \|M_{k+1}\|$$

where $\|\cdot\|$ denotes the Frobenius norm. Assuming no a_k 's are zero, we then obtain

$$\max_k \frac{|\delta a_k|}{|a_k|} = O(\epsilon) \max_k \frac{\|C\| \ \|M_{k+1}\|}{|a_k|} \ . \tag{7}$$

One could call this componentwise backward stability because $\max_k \frac{\|C\| \ \|M_{k+1}\|}{|a_k|}$ is a constant. But this can be quite bad if the $|a_k|$'s are very different in size ! An upper bound obtained from (5) for it is given by $\frac{\sqrt{n} \ \|C\|^2}{\min_k \ |a_k|}$.

We now show how to modify the problem in order to get improved componentwise backward errors. All matrices on the manifold **Orb** have the same characteristic polynomial. So we could apply the QR-algorithm to any point of **Orb**, and get a new backward error. To obtain the $|\delta a_k|$'s, we just have to transform this error back to C and project it on **Nor**. If the new point has been well chosen, the new $|\delta a_k|$'s will be smaller. Suppose the new point is $\widehat{C} = T^{-1}CT$. Applying the QR-algorithm to it gives us a new backward error $\|\widehat{\Delta}\| = O(\epsilon) \ \|\widehat{C}\|$. We transform it back to C where it becomes $T\widehat{\Delta}T^{-1}$. Then using (4), we get

$$\begin{aligned} |\widehat{\delta a_k}| &= | \ \langle T\widehat{\Delta}T^{-1} \ , \ M_{k+1}^* \rangle \ | \\ &= | \ \langle \widehat{\Delta} \ , \ T^* M_{k+1}^* T^{-*} \rangle \ | \\ &= | \ \langle \widehat{\Delta} \ , \ \widehat{M}_{k+1}^* \rangle \ | \ . \end{aligned} \tag{8}$$

But it is not difficult to see that the $\widehat{M}_k{}^*$'s span the normal space $\widehat{\mathbf{Nor}}$ at the point $\widehat{C}$. So we keep the same interpretation of what the δa_k 's are. Proceeding as before, we can thus write from (8) that

$$\max_k \frac{|\widehat{\delta a_k}|}{|a_k|} = O(\epsilon) \max_k \frac{\|\widehat{C}\| \ \|\widehat{M}_{k+1}\|}{|a_k|} \ .$$

It looks like (7) but the big difference is that we now can affect the constant $\max_k \frac{\|\widehat{C}\| \ \|\widehat{M}_{k+1}\|}{|a_k|}$ by carefully choosing a good similarity. More precisely, we want to solve :

$$\min_T \max_k \frac{\|T^{-1}CT\| \ \|T^{-1}M_{k+1}T\|}{|a_k|} \ . \tag{9}$$

But it is a well-known fact that under similarity transformations the Frobenius norm of a matrix is minimal for the similarity that makes it diagonal. This can be seen after putting the matrix in its Schur form. If the matrix is not diagonalizable, the result is the same except that the minimum cannot be reached : it becomes an infimum [9]. Moreover, as we have just seen, the similarity T_d that makes C diagonal simultaneously makes the M_k 's diagonal. So the similarity that solve the problem (9) is given by $T = T_d$. where the $T_d^{-1}M_{k+1}T_d$'s are diagonal. Notice that if we replace T_d by T_dU where U is unitary, does not change anything. The optimal points on **Orb** are thus of the form $(T_dU)^{-1} \ C \ (T_dU)$. These points correspond also to normal matrices for they are solutions of

$$(T^{-1}CT) \ (T^{-1}CT)^* = (T^{-1}CT)^* \ (T^{-1}CT) \ . \tag{10}$$

And conversely all solutions of (10) are of this form. Hence, denoting $\mathcal{N} \subset \mathbb{C}^{n\times n}$ the set of normal matrices, we may write :

$$T^{-1}CT \text{ is optimal} \iff T^{-1}CT \in \mathcal{N} \cap \mathbf{Orb}. \tag{11}$$

We will in general not be able to reach these points for several reasons. First of all, C has possibly multiple roots and consequently Λ does not necessarily belong to **Orb** (but it belongs to the closure of **Orb** [9]). Secondly, even if C is diagonalizable, the similarity that makes it diagonal is precisely what we want to compute and is therefore unknown before running the algorithm : it can thus not help to precondition the companion matrix C. Thirdly we want to compute the preconditioning similarity exactly, without introducing new numerical errors, which is e.g. possible with real diagonal similarities whose entries are powers of two [8]. For all these reasons we will limit ourselves to *real diagonal* similarities. Then one quickly sees that we can no longer be optimal since for example (10) becomes : find $T = D$, where D stands for some diagonal similarity, such that

$$D^{-1}CDDC^*D^{-1} = DC^*D^{-1}D^{-1}CD \tag{12}$$

which has no solution, because the zero entries of CC^* do not correspond to those of C^*C. Therefore the optimal scaling D solves

$$\min_D \max_k \frac{\|D^{-1}CD\| \, \|D^{-1}M_{k+1}D\|}{|a_k|} . \tag{13}$$

It can be proven that taking the logarithm and changing the variables leads to minimizing a *convex* function[1]. (The proof being too long, we do not give it in this paper.) Hence there are good algorithms that solve it efficiently.

Let us now end this section by giving a geometric interpretation of what the traditionally used scaling transformation does. Knowing that the optimal points on **Orb** correspond to normal matrices, we would be tempted by rather defining this optimal scaling problem :

$$\inf_{D,U} \|D^{-1}CD - U^*\Lambda U\| \tag{14}$$

which boils down to minimizing the Frobenius distance between $\mathbf{Orb}_D \doteq \{D^{-1}CD : \det(D) \neq 0\}$ and $\mathbf{Orb}_N \doteq \mathcal{N} \cap \mathbf{Orb}$. But as we said before the problem is that we have no knowledge of Λ before running the algorithm. But if we choose a normal matrix that is diagonal, Λ will not appear in the minimization problem (14). And one easily sees that the only diagonal matrix in $\mathbf{Orb}_N$ (and even in **Orb**) is Λ.

[1]Special thanks to Yurii Nesterov (Center for operations research and econometrics (CORE), Université Catholique de Louvain) for having given a proof in a personal communication.

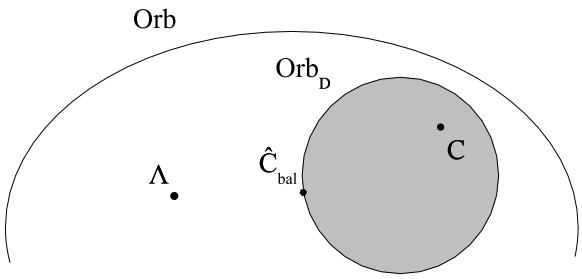

Figure 1. *The balancing similarity minimizes the Frobenius norm between* $\boldsymbol{Orb}_D$ *and* Λ.

This is exactly what the the diagonal similarity that *balances*[2] $\widehat{C}$ does : it minimizes the Frobenius norm of $\widehat{C}$ or equivalently the Frobenius norm of $(\widehat{C} - \Lambda)$. It is thus suboptimal for the problem (14). Figure 1 gives a schematic view of what the balancing similarity does.

3 The general non-monic scalar case

3.1 Problem statement

Here we deal with the nth-order complex polynomial $p(\lambda) \doteq a_n\lambda^n + a_{n-1}\lambda^{n-1} + \ldots + a_1\lambda + a_0$. Because we make no assumption on a_n, we could take the companion matrix (1) into account where we divide each a_k ($k \leq n-1$) by a_n, which can lead to a very large backward error (see (2)) if a_n is very small. So we build the matrix pencil

$$\lambda\, \bar{I} - C = \lambda \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & \ddots & \\ & & & a_n \end{bmatrix} - \begin{bmatrix} 0 & & & -a_0 \\ 1 & 0 & & -a_1 \\ & \ddots & & \vdots \\ & & 1 & -a_{n-1} \end{bmatrix} \tag{15}$$

whose determinant is $p(\lambda)$. Its eigenvalues satisfy $\det(\lambda\, \bar{I} - C) = 0$. Computing the roots of $p(\lambda)$ reduces once again to an eigenvalue problem. If we use the standard QZ-algorithm, the computed eigenvalues are exactly those of a matrix pencil $\lambda\,(\bar{I} + \Delta_B) - (C + \Delta_A)$ for some dense backward error matrix pencil $\lambda\,\Delta_B - \Delta_A$ with

$$\|\Delta_A\| \;=\; O(\epsilon)\|C\| \quad , \quad \|\Delta_B\| \;=\; O(\epsilon)\|\bar{I}\| \tag{16}$$

where $\|.\|$ denotes e.g. the Frobenius norm. Recall that we want the componentwise backward stability (3) to be satisfied (now also for $k = n$).

[2]A matrix is balanced if, for every row, the norm of the row is equal to the norm of the corresponding column.

3.2 A geometric approach

We work inside the $2n^2$-dimensional space of $n \times n$ complex matrix pencils with Frobenius inner product

$$\langle \lambda\ B_1 - A_1\ ,\ \lambda\ B_2 - A_2 \rangle\ \doteq\ \mathrm{tr}(A_1 A_2^* + B_1 B_2^*)\ .$$

In this pencil space we consider the manifold

$$\mathbf{Orb}\ \doteq\ \mathrm{orbit}(\lambda\ \bar{I} - C)\ =\ \{P^{-1}(\lambda\ \bar{I} - C)Q : \det(P)\det(Q) \neq 0\}\ .$$

One can check by performing a first order calculation that the tangent space to this manifold at the point $\lambda\ \bar{I} - C$ consists of matrix pencils of the form

$$\mathbf{Tan}\ =\ \{\lambda\ T_B - T_A = \lambda\ (X\bar{I} - \bar{I}Y) - (XC - CY)\ :\ X,\ Y\ \in\ \mathbb{C}^{n\times n}\}\ .$$

Indeed by definition it holds

$$\lambda\ T_B - T_A\ =\ [(I + \delta P)^{-1}(\lambda\ \bar{I} - C)(I + \delta Q) - (\lambda\ \bar{I} - C)]_{\text{1st order}}\ . \qquad (17)$$

Hence denoting $-\delta P$ by X and $-\delta Q$ by Y, we get the desired expression for the tangent space. Furthermore it is shown in [5] that the normal space at the same point can be seen as

$$\mathbf{Nor}\ =\ \{\lambda\ N_B - N_A\ :\ N_A C^* + N_B \bar{I}^* = 0\ \text{ and }\ C^* N_A + \bar{I}^* N_B = 0\}\ . \qquad (18)$$

Remark that letting $a_n\ =\ 1$ brings us back to the centralizer of C^* for N_A. The dimension of **Nor** is called the *codimension* of the orbit, denoted as $\mathrm{cod}(\lambda\ \bar{I} - C)$. We know from [3] that this number is the sum of different contributions depending on the structure of the Kronecker Canonical Form (KCF) of the pencil (15). Because this pencil is regular and has no eigenvalues at infinity (a_n is supposed not to be zero), the KCF contains only Jordan blocks. If p is the number of distinct eigenvalues, p_i the number of Jordan Blocks for λ_i and $q_j(\lambda_i)$ the size of the j-th block associated with λ_i, taking into account that the companion matrix C is non-derogatory, we find

$$\mathrm{cod}(\lambda\ \bar{I} - C)\ =\ \sum_{i=1}^{p}\sum_{j=1}^{p_i}(2j-1)q_j(\lambda_i)\ =\ n\ .$$

The dimension of **Orb** is thus $2n^2 - n$.

We will now try to proceed in the same way as [6] to find out how a dense perturbation $\lambda\ \Delta_B - \Delta_A$ to the matrix pencil (15) leads to first order perturbations in the coefficients a_k.
We first have to find a *transversal* space through $\lambda\ \bar{I} - C$. The natural space we think of is the set of 'companion pencils' of the form (15). Like in section 2, we will call it **Syl**. This space has dimension $n + 1$. So we can no longer decompose uniquely a perturbation of (15) in one component along **Syl** and another along **Tan**

because the sum of the dimensions of these two subspaces exceeds $2n^2$. Where does this come from ? The answer is given by looking carefully at the tangent space. In the monic case, a perturbation to C that lies in the tangent space does not modify the Jordan Canonical Form (JCF) to first order, i.e. preserve $\chi_C(\lambda)$. Since C is non-derogatory implies there is indeed a one-to-one correspondence between the JCF and the characteristic polynomial. So the totality of the first order coefficient perturbation lies in **Syl**. In the non-monic case, a tangential perturbation to $\lambda\,\bar{I} - C$ does not change the KCF to first order, but does modify the determinant! Our decomposition can thus not work anymore. The idea is then to consider only the points of **Orb** that have the same determinant $p(\lambda)$, such that any tangential perturbations in this subspace

$$\widetilde{\mathbf{Orb}} \doteq \{P^{-1}(\lambda\,\bar{I} - C)Q : \det(P) = \det(Q) \neq 0\}\ .$$

preserve the determinant to first order. The tangent space to $\widetilde{\mathbf{Orb}}$ at the point $\lambda\,\bar{I} - C$ consists of matrix pencils of the form

$$\widetilde{\mathbf{Tan}} = \{\lambda\,\widetilde{T}_B - \widetilde{T}_A = \lambda\,(X\bar{I} - \bar{I}Y) - (XC - CY) \ :\ \mathrm{tr}(X) = \mathrm{tr}(Y)\ \ X,\, Y \in \mathbb{C}^{n\times n}\}\,.$$

The trace condition comes up because in (17) we now impose $\det(I + \delta P)$ to be to first order equal to $\det(I + \delta Q)$ which yields $\mathrm{tr}(\delta P) = \mathrm{tr}(\delta Q)$. It has thus dimension $2n^2 - n - 1$. It follows that the normal space $\widetilde{\mathbf{Nor}}$ will have dimension $n + 1$. We can define this normal space in a similar way to (18). Expressing that each vector of $\widetilde{\mathbf{Nor}}$ must be orthogonal to $\widetilde{\mathbf{Tan}}$, and using appropriate choices of X , Y, we find

$$\widetilde{\mathbf{Nor}} = \{\lambda\,\widetilde{N}_B - \widetilde{N}_A \ :\ \widetilde{N}_A C^* + \widetilde{N}_B \bar{I}^* = \alpha I = C^*\widetilde{N}_A + \bar{I}^*\widetilde{N}_B\ ,\ \ \alpha \in \mathbb{C}\}\ . \quad (19)$$

To make our error decomposition, we need a basis for $\widetilde{\mathbf{Nor}}$. We define first the $\overline{M}_k$'s to be a slightly modified version of the M_k 's we had in section 2 :

$$\overline{M}_k = \overbrace{\begin{bmatrix} a_k & & & \\ a_{k+1} & a_k & & \\ \vdots & a_{k+1} & \ddots & \\ a_n & \vdots & \ddots & a_k \\ & a_n & \ddots & a_{k+1} \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}}^{k} \overbrace{\begin{bmatrix} -a_0 & & \\ -a_1 & \ddots & \\ \vdots & \ddots & -a_0 \\ -a_{k-1} & \ddots & \vdots \\ & \ddots & \vdots \\ & & -a_{k-1} \end{bmatrix}}^{n-k} . \quad (20)$$

After some manipulations one obtains the relations

$$\left\{ \begin{array}{rcl} \bar{I}\,\overline{M}_k & = & C\,\overline{M}_{k+1} + a_k I \\ \overline{M}_k\,\bar{I} & = & \overline{M}_{k+1}\,C + a_k I \end{array} \right. \quad (21)$$

which have to be compared with (6). Hence with the convention $\overline{M}_0 = \overline{M}_{n+1} = 0$, the $n + 1$ independent vectors $\lambda\,\widetilde{N}_B - \widetilde{N}_A \doteq \lambda\,\overline{M}_k^{\,*} + \overline{M}_{k+1}^{\,*}$ for $k = 0, \ldots, n$

satisfy (19). They thus form a basis for the normal space $\widetilde{\mathbf{Nor}}$.
We now have all we need. Writing $\Delta = \lambda\ \Delta_B - \Delta_A$ as $\Delta = \Delta^{\widetilde{\mathbf{Tan}}} + \Delta^{\mathbf{Syl}}$, expressing that $\Delta^{\widetilde{\mathbf{Tan}}}$ is perpendicular to $\widetilde{\mathbf{Nor}}$, and keeping in mind the form of $\Delta^{\mathbf{Syl}}$, we come to the desired equation :

$$|\delta a_k| = |\ \langle \lambda\ \Delta_B - \Delta_A\ ,\ \lambda\ \overline{M}_k^{\ *} + \overline{M}_{k+1}^{\ *} \rangle\ |$$

for all k 's between 0 and n. In particular, we see that $|\delta a_0|$ only depends on Δ_A, and $|\delta a_n|$ only on Δ_B, which corresponds to our intuition.

The same question appears. How far are we from getting (3) ?
To answer this, we follow the same procedure as in section 2. Using (16) yields

$$\begin{aligned} |\delta a_k| &\leq \|\lambda\ \Delta_B - \Delta_A\|\ \|\lambda\ \overline{M}_k + \overline{M}_{k+1}\| \\ &= O(\epsilon)\ \|\lambda\ \bar{I} - C\|\ \|\lambda\ \overline{M}_k + \overline{M}_{k+1}\| \end{aligned}$$

where $\|.\|$ denotes the Frobenius norm, such that if no a_k 's are zero we have

$$\max_k\ \frac{|\delta a_k|}{|a_k|} = O(\epsilon)\ \max_k\ \frac{\|\lambda\ \bar{I} - C\|\ \|\lambda\ \overline{M}_k + \overline{M}_{k+1}\|}{|a_k|}\ .$$

Like before, this maximum is constant but can be very large if the order of magnitude of the $|a_k|$'s varies a lot. An upper bound for it obtained from (20) is $\frac{\sqrt{2n}\ \|\lambda\ \bar{I} - C\|^2}{\min_k\ |a_k|}$.

Would it be possible to improve this bound using left and right preconditioning transformations ? We know we can apply the QZ-algorithm to another point $T_2^{-1}(\lambda\ \bar{I} - C)T_1$ of the manifold $\widetilde{\mathbf{Orb}}$ since all matrix pencils on it have the same eigenvalues, and the idea is to take the point in such a way that the projection on $\widetilde{\mathbf{Nor}}$ of the new backward error $\lambda\ \widehat{\Delta}_B - \widehat{\Delta}_A$ transformed back to $\lambda\ \bar{I} - C$ is smaller than the projection without preconditioning. In view of this, we have

$$\begin{aligned} |\widehat{\delta a_k}| &= |\ \langle T_2(\lambda\ \widehat{\Delta}_B - \widehat{\Delta}_A)T_1^{-1}\ ,\ \lambda\ \overline{M}_k^{\ *} + \overline{M}_{k+1}^{\ *} \rangle\ | \\ &= |\ \langle \lambda\ \widehat{\Delta}_B - \widehat{\Delta}_A\ ,\ {T_2}^*(\lambda\ \overline{M}_k^{\ *} + \overline{M}_{k+1}^{\ *}){T_1}^{-*} \rangle\ | \end{aligned}$$

where the ${T_2}^*(\lambda\ \overline{M}_k^{\ *} + \overline{M}_{k+1}^{\ *}){T_1}^{-*}$'s span the normal space ${T_2}^*\ \widetilde{\mathbf{Nor}}\ {T_1}^{-*}$ at the new point. We can thus write

$$\max_k\ \frac{|\widehat{\delta a_k}|}{|a_k|} = O(\epsilon)\ \max_k\ \frac{\|T_2^{-1}(\lambda\ \bar{I} - C)T_1\|\ \|T_1^{-1}(\lambda\ \overline{M}_k + \overline{M}_{k+1})T_2\|}{|a_k|}\ .$$

Now the problem we want to solve is the following :

$$\min_{\det(T_1)=\det(T_2)}\ \max_k\ \frac{\|T_2^{-1}(\lambda\ \bar{I} - C)T_1\|\ \|T_1^{-1}(\lambda\ \overline{M}_k + \overline{M}_{k+1})T_2\|}{|a_k|}\ . \qquad (22)$$

Since the Frobenius norm of a pencil is invariant under unitary transformations, we can assume both pencil be in generalized Schur form. Indeed one easily sees from (21) that the transformations $Q_1,\ Q_2$ that make it upper triangular make the pencil $\lambda\ \overline{M}_k + \overline{M}_{k+1}$ upper triangular too (for every k). We will minimize this product it in two times. First we take in account only transformations that preserve the diagonal elements, and additionally we consider a diagonal scaling transformation. Assuming that the pencil $\lambda\ \bar{I} - C$ is diagonalizable, i.e. that the roots of the polynomial are distinct, the first part of the job is done by unit upper triangular left and right transformations R_1, R_2 that kill the elements above the diagonal without changing the diagonal self [4]. From (21) also, it is clear that these transformations act simultaneously on both pencils. Clearly this minimizes the norm of each pencil under this class of transformations, and hence minimizes the product of the norms. Then it remains to find the diagonal transformation D (that verifies $\det D = 1$ such that $(Q_2R_2)^{-1}(\lambda\ \bar{I} - C)(Q_1R_1D)$ is still on $\widetilde{\mathbf{Orb}}$) that scale them in such a way that the product is minimum.
If the pencil is not diagonalizable, the result is the same except that the minimum cannot be reached : it becomes an infimum [9]. So the optimal transformations that solve (22) are given by $T_1 = Q_1R_1D$ and $T_2 = Q_2R_2$. Notice here that replacing T_1 by T_1U_1 and T_2 by T_2U_2 where U_1 and U_2 are unitary, does not modify anything. The optimal points on $\widetilde{\mathbf{Orb}}$ are thus of the form $(Q_2R_2U_2)^{-1}\ (\lambda\ \bar{I} - C)\ (Q_1R_1DU_1) = U_2^*(\lambda\ \Lambda_B - \Lambda_A)DU_1$. These points are normal pencils. Indeed one defines in [2] a pencil $\lambda\ B - A$ to be normal if there exist unitary transformations U_1 and U_2 such that $U_2(\lambda\ B - A)U_1^*$ is diagonal. Hence denoting $\widetilde{\mathcal{N}} \subset \mathbb{C}^{n\times n}\times\mathbb{C}^{n\times n}$ as the set of normal pencils 'generated' by the diagonal pencil $(\lambda\ \Lambda_A - \Lambda_B)D$, we may write :

$$T_2^{-1}(\lambda\ \bar{I} - C)T_1 \ \text{ is optimal } \iff T_2^{-1}(\lambda\ \bar{I} - C)T_1 \ \in\ \widetilde{\mathcal{N}} \cap \widetilde{\mathbf{Orb}}\ . \qquad (23)$$

Notice that in the monic case, it was *necessary and sufficient* to be normal for being optimal (11). Here it is only *necessary*. This comes from the fact that there are several diagonal pencils on $\widetilde{\mathbf{Orb}}$ contrarily to the monic case where there is only one diagonal matrix on the manifold.
We will generally not be able to reach these points for the reasons we explained at the end of section 2. We will limit ourselves to *real diagonal* transformations $D_1,\ D_2$. Then one quickly sees that we can no longer be optimal since diagonal transformations do not change the pencil structure. Therefore the optimal scaling problem is defined to be

$$\min_{\det(D_1)=\det(D_2)}\ \max_k\ \frac{\|D_2^{-1}(\lambda\ \bar{I} - C)D_1\|\ \|D_1^{-1}(\lambda\ \overline{M}_k + \overline{M}_{k+1})D_2\|}{|a_k|}\ . \qquad (24)$$

Writing a minimum makes sense since taking the logarithm and changing the variables leads to minimizing a *convex* function on a *convex* constraints set. So the scaling transformations can be computed efficiently.

Bibliography

[1] V. I. Arnol′d. Matrices depending on parameters. *Uspehi Mat. Nauk*, 26(2(158)):101–114, 1971.

[2] J.-P. Charlier and P. Van Dooren. A Jacobi-like algorithm for computing the generalized Schur form of a regular pencil. *J. Comput. Appl. Math.*, 27(1-2):17–36, 1989. Reprinted in *Parallel algorithms for numerical linear algebra*, 17–36, North-Holland, Amsterdam, 1990.

[3] James W. Demmel and Alan Edelman. The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms. *Linear Algebra Appl.*, 230:61–87, 1995.

[4] James Weldon Demmel and Bo Kgstrm. Computing stable eigendecompositions of matrix pencils. *Linear Algebra Appl.*, 88/89:139–186, 1987.

[5] Alan Edelman, Erik Elmroth, and Bo Kgstrm. A geometric approach to perturbation theory of matrices and matrix pencils. II: A stratification-enhanced staircase algorithm. *SIAM J. Matrix Anal. Appl.*, 20(3):667–699, 1999.

[6] Alan Edelman and H. Murakami. Polynomial roots from companion matrix eigenvalues. *Math. Comput.*, 64(210):763–776, 1995.

[7] F.R. Gantmacher. *Theory of Matrices*, volume 1. Chelsea Publishing Company, Chelsea, New York, 1959.

[8] B.N. Parlett and C. Reinsch. Balancing a matrix for calculation of eigenvalues and eigenvectors. *Numer. Math.*, 13:293–304, 1969.

[9] Andrzej Pokrzywa. On perturbations and the equivalence orbit of a matrix pencil. *Linear Algebra Appl.*, 82:99–121, 1986.