Backward Stability of the Schur Decomposition under Small Perturbations

Anastasiia Minenkova[a], Evelyn Nitch-Griffin[b], Vadim Olshevsky[c]

[a] *University of Hartford, West Hartford, CT, USA. Email: minenkova@hartford.edu*
[b] *Landmark College, Putney, VT, USA. Email: evelynnitchgriffin@landmark.edu*
[c] *University of Connecticut, Storrs, CT, USA. Email: olshevsky@uconn.edu*

Abstract

In the present paper, we show the backward stability of the Schur decomposition for a given matrix under small perturbations. The norm throughout this paper is unitarily invariant.

Keywords: perturbation theory, Schur decomposition, unitary Hessenberg matrices, the Gohberg-Kaashoek numbers, invariant subspaces, gaps between subspaces.
2000 MSC: 15A23, 47A15

1. Introduction

Lipschitz-Hölder stability was investigated for several canonical forms such as Jordan, flipped-orthogonal, flipped-orthogonal conjugate symmetrical, real canonical forms [2, 4, 5, 6, 7, 10], but it was never studied for the Schur decomposition.

We begin with recalling some classical results.

1.1. Eigenvalues' Stability

The first question to consider is what happens to the eigenvalues of a given matrix under small perturbation. The norm throughout this paper is unitarily invariant. In general even the number of distinct eigenvalues can change under perturbations. However, under additional assumptions imposed on the eigenvalues the following result holds true (see [1]).

Proposition 1.1. *Let A_0 be an $n \times n$ matrix and $\{\lambda_1, \dots, \lambda_n\}$ be its eigenvalues, and A being its perturbation with $\|A - A_0\| < \varepsilon$ for sufficiently small ε depending on A_0 and the eigenvalues μ_j. If the number of eigenvalues of*

A_0 is the same as of A, then there is a certain ordering of them such that for some positive $K = K(A_0)$

$$|\mu_i - \lambda_i| \leq K\|A - A_0\|, \quad i = 1, 2, \ldots, |\sigma(A_0)|. \tag{1.1}$$

The bound in (1.1) is called Lipschits stability, and it is valid in the case when A has the sam number of distinct eigenvalues as A_0. The following result shows that when A has a different number of eigenvalues we will have stability, but it is weaker than in (1.1).

For the general case of the eigenvalues stability we have the following result (see [11, Appendix K]).

Proposition 1.2. *Let A_0 be an $n \times n$ matrix and $\{\lambda_1, \ldots, \lambda_n\}$ be its eigenvalues. Then, for every A with $\|A - A_0\| < \varepsilon$ for sufficiently small ε depending on A_0 there is an ordering of its eigenvalues μ_j's and a positive constant $K = K(A_0)$ such that*

$$|\mu_j - \lambda_j| \leq K\|A - A_0\|^{1/n}. \tag{1.2}$$

This type of bounds is called Hölder because of the power $1/n$ for the matrix norm. The following example shows that the power $1/n$ in (1.2) cannot be relaxed in general.

Example 1.3. *Consider the following matrices $A_0, A \in \mathbb{C}^{2\times 2}$.*

$$A_0 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \text{ and } A = \begin{bmatrix} 0 & \epsilon \\ 1 & 0 \end{bmatrix}.$$

Note that $\|A - A_0\| = \epsilon$. Moreover, $\sigma(A_0) = \{0\}$ and $\sigma(A) = \{\pm\sqrt{\epsilon}\}$. It is easy to see that in this case we have

$$|0 \mp \sqrt{\epsilon}| = \epsilon^{1/2} = \|A - A_0\|^{1/2}.$$

This example can be easily modified for $n \times n$-matrices.

1.2. Backward Stability of the Schur Decomposition

Every $n \times n$-matrix A is unitarily similar to an upper triangular matrix T, i.e, $A = UTU^*$ where U is unitary. This triangular matrix T is called a *Schur Triangular form*, and the factorization itself is called the *Schur Decomposition.*

Note that diagonal entries of T are the eigenvalues of A. Hence, the eigenvalues stability results of the Section 2.1 suggest to consider stability of the Schur decomposition.

What kind of stability can we have for the Schur canonical form? We start by considering the following type of a result, that we call forward stability.

Conjecture 1.4 (**Forward Stability**)**.** *Let* $A_0 = U_0 T_0 U_0^* \in \mathbb{C}^{n\times n}$ *be a Schur decomposition, where* U_0 *is unitary and* T_0 *is upper triangular. Then, there exist constants* $K, \epsilon > 0$ *(depending on* A_0 *only) such that for all* A *with* $\|A - A_0\| < \epsilon$ *there exists a Schur decomposition* $A = UTU^*$ *such that*

$$\|U - U_0\| + \|T - T_0\| \leq K\|A - A_0\|^{1/n}$$

As the following example shows this conjecture is not valid in the form stated, and hence the Schur decomposition is not forward stable in general.

Example 1.5. *Consider the following matrix and its perturbation,*

$$A_0 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \text{ and } A = \begin{bmatrix} 2 & 0 \\ \epsilon & 2 \end{bmatrix}.$$

Let us consider the following Schur factorization of A_0

$$A_0 = U_0 T_0 U_0^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and the Schur factorization of A

$$A = UTU^* = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & \epsilon \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Since the first column of U has to be an eigenvector of A, and the latter is essentially unique, the matrix U is essentially unique as well. Hence, the distance $\|U - U_0\|$ is quite large. Since U was the only possible choice, we can conclude that our Conjecture 1.4 above is false in general. Although forward stability results have been obtained for other canonical forms (see [2]), in case of the Schur canonical form we cannot have the stability mentioned in Conjecture 1.4. The next statement shows us why.

Theorem 1.6 (Different Gohberg-Kaashoek Numbers)**.** *Let the matrix* A_0 *and a Schur decomposition* $A_0 = U_0 T_0 U_0^*$ *be fixed. There exists* $K > 0$ *such that in any neighborhood of* A_0*, i.e.* $\{A : \|A - A_0\| < \varepsilon\}$ *for any* $\varepsilon > 0$*,*

$$\sup_{A} \inf_{U,T} \|U - U_0\| + \|T - T_0\| > M > 0, \tag{1.3}$$

where the supremum is taken over all A *in this neighborhood having different Gohberg-Kaashoek numbers from* A_0 *and the infimum is taken over all their Schur factorizations* $A = UTU^*$*.*

Theorem 1.6 uses Gohberg-Kaashoek (GK) numbers. Let us introduce these numbers now (see [3, 8, 9] for details).

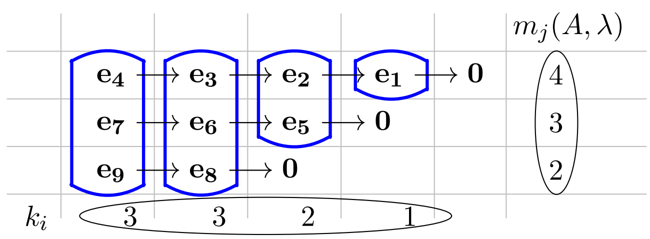
1.2.1. Gohberg-Kaashoek Numbers

Let $A \in \mathbb{C}^{n\times n}$, $\sigma(A)$ be the set of all its eigenvalues, and $m_1(A,\lambda) \le m_2(A,\lambda) \le \cdots \le m_t(A,\lambda)$ be the sizes of all blocks corresponding to $\lambda \in \sigma(A)$ in the Jordan form of A. We set $m_i(A,\lambda) = 0$ $(i = t+1, \ldots, n)$ for convenience. The numbers $m_i(A) = \sum_{\lambda\in\sigma(A)} m_i(A,\lambda)$ are called *the Gohberg-Kaashoek numbers.* The vector $k = [\, k_1 \; k_2 \; \ldots \; k_n \,]^\top$ with $k_i = \max\limits_{1\le l\le n} \{l : m_l \ge i\}$ is called *dual* to m.

In terms of the Gohberg-Kaashoek numbers m_j's it means that if we have

$$A = \left[\begin{array}{cccc|ccc|cc}
\lambda & 1 & 0 & 0 & & & & & \\
0 & \lambda & 1 & 0 & & & & & \\
0 & 0 & \lambda & 1 & & & & & \\
0 & 0 & 0 & \lambda & & & & & \\
\hline
 & & & & \lambda & 1 & 0 & & \\
 & & & & 0 & \lambda & 1 & & \\
 & & & & 0 & 0 & \lambda & & \\
\hline
 & & & & & & & \lambda & 1 \\
 & & & & & & & 0 & \lambda
\end{array}\right]$$

then we can put the Jordan chains corresponding to λ in the following order.



Therefore, $m(A) = [4,3,2,0,0,0,0,0,0]^\top$ and $k(A) = [3,3,2,1,0,0,0,0,0]^\top$. These numbers were introduced in [3], where the problem of complete description for the Jordan structure of a matrix, which is a small perturbation of a given matrix, was posed. This problem was solved independently in [1] and [7].

1.2.2. Backward Stability

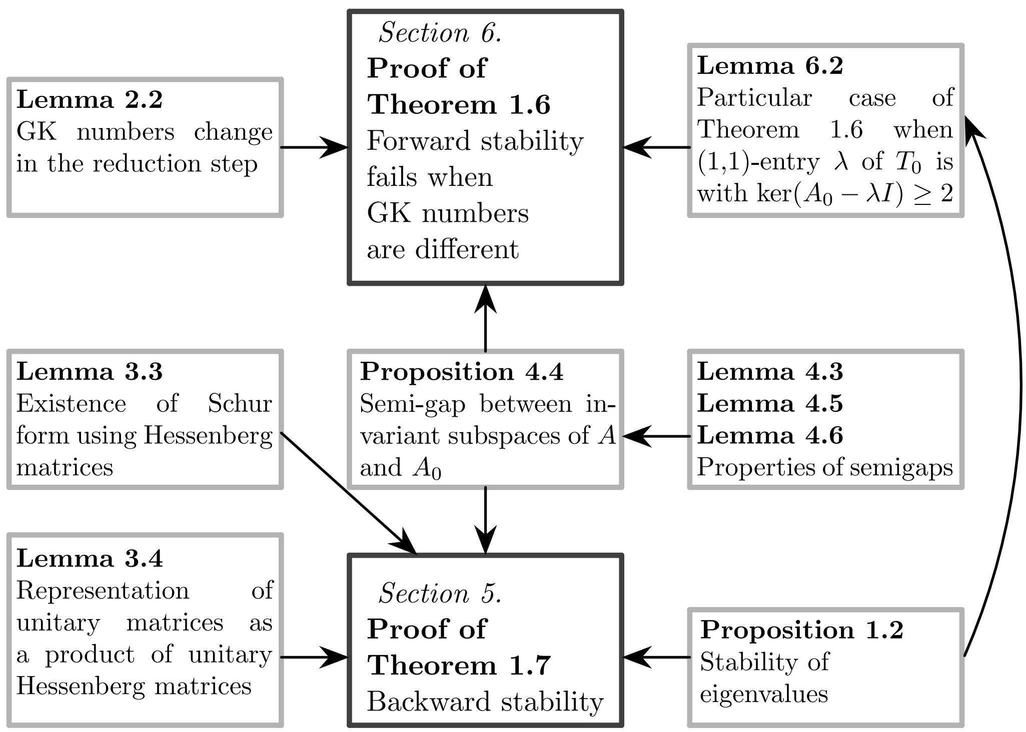
Although a general forward stability result is not valid, we can get the backward stability result.

Theorem 1.7 (Backward Stability)**.** *Let $A_0 \in \mathbb{C}^{n\times n}$ be given. There exist constants $K, \epsilon > 0$ (depending on A_0 only) such that for all A with $\|A - A_0\| < \epsilon$ and for any factorization UTU^* of A (U unitary and T is* *upper triangular) there exist U_0 and T_0 such that $A_0 = U_0 T_0 U_0^*$ is a Schur factorization of A_0 with*

$$\|U - U_0\| + \|T - T_0\| \leq K\|A - A_0\|^{1/n}. \tag{1.4}$$

1.2.3. Organization of the Paper

In Section 2 we consider some auxillary results related to GK numbers after applying a reduction step. In Section 3 we discuss facts related to unitary Hessenberg matrices. In Section 4 we give a short overview of theory of gaps and semigaps. In Section 5 and 6 we present the proofs of Theorem 1.7 and Theorem 1.6 respectively.



2. Auxiliary Results

Before proving Theorem 1.6 we need a couple of technical lemmas. Let us start by introducing the following fact.

Lemma 2.1. *For every eigenvector x of A there is a Jordan basis of A including x.*

Proof. Let us fix a Jordan basis of A, with the Jordan chains corresponding to λ_t ordered by length. Now, given another eigenvector x decompose it in that basis. Look for the last non-zero coefficient, say, α that is corresponding to the eigenvector, say, y. Then the chain for x has the same length as for y and we can replace the chain for y with the chain for x. All it remains to prove is the linear independence of the new set of vectors.

Let Y stands for the matrix whose columns are the Jordan basis we started with. Y is invertible. Denote by X the matrix where the chain for y is replaced by the chain for x. Then $X = YR$, where R is an upper triangular matrix that is invertible, since it has either 1 on its diagonal or α. Note that for the generalized eigenvectors of the chain for x we have the same decompositions with the same coefficients as for x with the vectors from the corresponding chains for the original basis, so we can get R. $\square$

The next result describes the recursion we will use. In particular, we want to figure out what happens to the GK numbers during each step of recursion. Here is the idea behind it:

$$\underbrace{m_1(A,\lambda_t) \geq m_2(A,\lambda_t) \geq m_3(A,\lambda_t) \geq \ldots \geq m_{l-1}(A,\lambda_t)} \geq m_l(A,\lambda_t)$$

The corresponding Jordan chains stay the same.

$m_j(A,\lambda_t) < m_l(A,\lambda_t)$ for $j > l$

So what happens when $m_j(A,\lambda_t) = m_l(A,\lambda_t)$ for some j's greater than l? Let j^* be the maximal such index.

Recursion decreases this chain by one vector.

$$\boxed{m_l(A,\lambda_t)} = m_{l+1}(A,\lambda_t) = \ldots = m_{j^*}(A,\lambda_t) > m_l(A,\lambda_t) - 1$$

Now let us formalize it.

Lemma 2.2. *Consider matrix B with the eigenvalues $\{\lambda_j\}$'s, having the GK numbers $\{m_j(B,\lambda_i)\}$ and e_1 as its eigenvector corresponding to the Jordan chain for λ_t and $m_l(B,\lambda_t)$, i.e.*

$$B = \left[\begin{array}{c|ccc} \lambda_t & \star & \cdots & \star \\ \hline 0 & & & \\ \vdots & & C & \\ 0 & & & \end{array}\right].$$

Then

- $m_j(C, \lambda_i) = m_j(B, \lambda_i)$ *for all* $i \neq t$ *or* $i = t$ *and* $j > l+1$*;*
- $m_l(C, \lambda_t) = m_l(B, \lambda_t) - 1$, $m_{l+1}(C, \lambda_t) = m_{l+1}(B, \lambda_t)$ *if* $m_l(A, \lambda_t) > m_{l+1}(B, \lambda_t)$*;*
- $m_{j^*}(C, \lambda_t) = m_l(B, \lambda_t) - 1$, $m_j(C, \lambda_t) = m_{j+1}(B, \lambda_t)$ *for* $j = l, \ldots, j^* - 1$ *if* $m_l(B, \lambda_t) = m_{l+1}(B, \lambda_t) = \ldots = m_{j^*}(B, \lambda_t)$ *and* j^* *is the maximal such index;*
- $m_j(C, \lambda_t) = m_j(B, \lambda_t)$ *if* $m_j(B, \lambda_1) < m_l(B, \lambda_t)$.

Proof. Note that due to Lemma 2.1 there is a Jordan basis of B containing e_1. Let J be the canonical Jordan form of B where the first block corresponds to the Jordan chain for λ_t that we mentioned. Thus, there is a invertible matrix R containing the Jordan basis $\{f_{i,j}^{(k)}\}_{i,j,k}$ (i is the place in the Jordan chain for $f_{i,j}^{(k)}$ corresponding to $m_j(B, \lambda_k)$) as its columns, where the first $m_l(B, \lambda_t)$ vectors forms the chain of A, having $f_{0,l}^{(t)} = e_1$, i.e. $R = \left[f_{0,l}^{(t)} | f_{1,l}^{(t)} | \ldots | f_{m_l(B,\lambda_t)-1,l}^{(t)} | \ldots \right]$. That is R and R^{-1} are of the following form.

$$R = \left[\begin{array}{c|ccc} 1 & \blacklozenge & \cdots & \blacklozenge \\ \hline 0 & & & \\ \vdots & & R_1 & \\ 0 & & & \end{array}\right] \quad \text{and} \quad R^{-1} = \left[\begin{array}{c|ccc} 1 & \clubsuit & \cdots & \clubsuit \\ \hline 0 & & & \\ \vdots & & R_1^{-1} & \\ 0 & & & \end{array}\right].$$

This argument implies that

$$\left[\begin{array}{c|cccc} \lambda_t & 1 & 0 & \cdots & 0 \\ \hline 0 & & & & \\ \vdots & & J_1 & & \\ 0 & & & & \end{array}\right] = J = R^{-1}BR =$$

$$= \left[\begin{array}{c|ccc} 1 & \blacklozenge & \cdots & \blacklozenge \\ \hline 0 & & & \\ \vdots & & R_1 & \\ 0 & & & \end{array}\right] \left[\begin{array}{c|ccc} \lambda_t & \star & \cdots & \star \\ \hline 0 & & & \\ \vdots & & C & \\ 0 & & & \end{array}\right] \left[\begin{array}{c|ccc} 1 & \clubsuit & \cdots & \clubsuit \\ \hline 0 & & & \\ \vdots & & R_1^{-1} & \\ 0 & & & \end{array}\right].$$

Note that $J_1 = R_1^{-1} C R_1$ is the Jordan form of C.

So what is the difference between J and J_1? The only Jordan chain that is affected is

$$0 \leftarrow f_{0,l}^{(t)} \leftarrow f_{1,l}^{(t)} \leftarrow \ldots \leftarrow f_{m_l(B,\lambda_t)-1,l}^{(t)}.$$

We delete the eigenvector from this chain and truncate the rest of the vectors to get a Jordan chain of length $m_l(B, \lambda_t) - 1$ of C. The length of the rest Jordan chains of C stay the same as they were in B. The conclusion of the lemma follows from this observation. $\square$

3. Unitary Hessenberg Matrices and Schur Canonical Forms

Throughout this paper we are going to use the special type of structured matrices that are called Hessenberg. Let us introduce it to the reader first.

A matrix is called the *upper Hessenberg* if it has zero entries below the first subdiagonal. Similarly, it is called the *lower Hessenberg* if it has zeros above the first super diagonal.

The following is a well-known fact (for example see [12]).

Proposition 3.1. *An $n \times n$ lower unitary Hessenberg matrix can be represented in the following way*

$$U = \begin{bmatrix} -\rho_1 & \mu_1 & 0 & \ldots & 0 \\ -\rho_2\mu_1 & -\rho_2\bar{\rho}_1 & \mu_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \mu_{n-1} \\ -\rho_n\mu_{n-1}\ldots\mu_1 & -\rho_n\mu_{n-1}\ldots\mu_2\bar{\rho}_1 & -\rho_n\mu_{n-1}\ldots\mu_3\bar{\rho}_2 & \ldots & -\rho_n\bar{\rho}_{n-1} \end{bmatrix},$$

where $\mu_j = \sqrt{1-\rho_j^2}$ for all j's.

The following statement is an immediate consequence of the previous proposition.

Corollary 3.2. *If $\rho_j < 1$ for all j, the first column of a lower unitary Hessenberg matrix completely defines the whole matrix.*

Let us consider the following properties of Hessenberg matrices first. Unitary Hessenberg matrices have a number of interesting properties, and are of particular importance in their relationship with the Schur form. In the classical proof of the construction of the Schur form, one builds an orthonormal set using eigenvectors of the matrix A_0, typically through the Gram-Schmidt process. With the above observation, we can derive the Schur form specifically through the use of unitary Hessenberg matrices.

Lemma 3.3. *For any $A \in \mathbb{C}^{n\times n}$ there exist U unitary and T upper triangular with the eigenvalues of A along the diagonal such that $A = UTU^*$.*

Proof. Let $\lambda_1, \dots, \lambda_m$ be the eigenvalues of A and let x be a unit eigenvector of A corresponding to eigenvalue λ_1. Moreover, pick H_1 to be a lower unitary Hessenberg matrix with x as its first column. By Proposition 3.1 this determines H_1 completely.

Then, we have $H_1^* A H_1 e_1 = H_1^* A x = H_1^* \lambda_1 x = \lambda_1 e_1$. In other words,

$$H_1^* A H_1 = \left[\begin{array}{c|ccc} \lambda_1 & \star & \cdots & \star \\ \hline 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{array}\right].$$

By repeating the process of reducing the matrix dimensions, i.e. for each matrix A_k constructing matrix H_k in a way we described, we get a string of matrices $H_1, \dots, H_n$ that are all unitary Hessenberg and

$$\widetilde{H}_{n-1}^* \dots \widetilde{H}_1^* A_0 \widetilde{H}_1 \dots \widetilde{H}_{n-1} = T_0,$$

Where T_0 is an upper triangular matrix, $\widetilde{H}_1 = H_1$, and $\widetilde{H}_k = \begin{bmatrix} I_{k-1} & \mathbf{0} \\ \mathbf{0} & H_k \end{bmatrix}$ for $k = 2, \dots, n-1$ with I_j being the $j \times j$ identity matrix. By taking $U = \widetilde{H}_1 \cdot \ldots \cdot \widetilde{H}_{n-1}$ we get the result. $\square$

Observe that we constructed our unitary matrix U using only unitary Hessenberg matrices H_k, each of whose first column was an eigenvector of the corresponding matrix A_k.

Lemma 3.4. *Every unitary matrix U admits a factorization*

$$U = H_1 \cdot \ldots \cdot H_{n-1},$$

where matrices H_i's are unitary Hessenberg.

Proof. Let H_1 be the unitary Hessenberg matrix whose first column x_1 is the same as U. Note that $x_1^* x_j = \delta_{1,j}$, since U is unitary. Then

$$H_1^* U = \left[\begin{array}{c|ccc} 1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & U_2 & \\ 0 & & & \end{array}\right],$$

where U_2 is unitary. As before we repeat the process until we get

$$H_{n-1}^* \cdot \ldots \cdot H_1^* U = I.$$

The result follows from simply multiplying the both parts of this equality by $H_1 \cdot \ldots \cdot H_{n-1}$. $\square$

4. Gap and Semi-gap

In this section we discuss some topological properties of the set of subspaces in $\mathbb{C}^n$, since in order to prove our main result, we require some facts from the theory of gaps. We begin by stating some definitions.

A matrix $P_{\mathcal{M}}$ is called an *orthogonal projector* onto a subspace $\mathcal{M} \subset \mathbb{C}^n$ if

- $\mathrm{Im} P_{\mathcal{M}} = \mathcal{M}$;
- $P_{\mathcal{M}}^2 = P_{\mathcal{M}}$;
- $P_{\mathcal{M}}^* = P_{\mathcal{M}}$.

The following concept is the key definition.

Let $\mathcal{M}, \mathcal{N}$ be subspaces of $\mathbb{C}^n$, and let $P_{\mathcal{M}}, P_{\mathcal{N}}$ be the orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$ respectively. We define *the gap* $\theta(\mathcal{M}, \mathcal{N})$ between $\mathcal{M}$ and $\mathcal{N}$ as follows

$$\theta(\mathcal{M}, \mathcal{N}) = \|P_{\mathcal{M}} - P_{\mathcal{N}}\|$$

or, equivalently,

$$\theta(\mathcal{M}, \mathcal{N}) = \max \left\{ \sup_{\substack{x \in \mathcal{M} \\ \|x\|=1}} \inf_{y \in \mathcal{N}} \|x - y\|, \sup_{\substack{y \in \mathcal{N} \\ \|y\|=1}} \inf_{x \in \mathcal{M}} \|x - y\| \right\}.$$

It follows immediately from the definition that $\theta(\mathcal{M},\mathcal{N})$ is a metric on the set of all subspaces in $\mathbb{C}^n$. Moreover, $\theta(\mathcal{M},\mathcal{N}) \le 1$.

Note that the Hausdorff distance between sets Inv A and Inv B of all invariant subspaces matrices A and B can be defined as follows

$$\text{dist (Inv } A, \text{Inv } B) = \max\{\sup_{\mathcal{M}\in \text{Inv } A} \theta(\mathcal{M}, \text{Inv } B), \sup_{\mathcal{N}\in \text{Inv } B} \theta(\mathcal{N}, \text{Inv } A)\}.$$

This distance is a metric as well. We are going to use the following property of gaps between subspaces, which can be found in [4].

Proposition 4.1. *For subspaces $\mathcal{M},\mathcal{N} \subset \mathbb{C}^n$, we have*

$$\theta(\mathcal{M},\mathcal{N}) = \theta(\mathcal{N}^{\perp},\mathcal{M}^{\perp}). \tag{4.5}$$

The symmetry of the gap is actually a disadvantage.

Proposition 4.2. *Let $\mathcal{M},\mathcal{N}$ be subspaces of $\mathbb{C}^n$.*

(i) If $\dim(\mathcal{M}) = \dim(\mathcal{N})$ then for any $x \in \mathcal{M}$ there exists a $y \in \mathcal{N}$ such that $\|x - y\| \le \theta(\mathcal{M},\mathcal{N})$.

(ii) If $\dim(\mathcal{M}) \ne \dim(\mathcal{N})$ then $\theta(\mathcal{M},\mathcal{N}) = 1$.

The above result shows that the gap is often not useful to consider when $\dim(\mathcal{M}) \ne \dim(\mathcal{N})$. We would like to find bounds on the kernels of the matrices, however, the dimension of the kernels are, in general, not equal.

The gap provides many useful results in providing a variety of bounds but the usefulness is limited to when the dimensions are equal. The concept of a semi-gap can be helpful when the dimensions are not equal. This advantage is highly useful when considering matrix perturbations.

Let $\mathcal{M},\mathcal{N}$ be subspaces of $\mathbb{C}^n$. The quantity

$$\theta_0(\mathcal{M},\mathcal{N}) = \sup_{\substack{x\in\mathcal{M}\\ \|x\|=1}} \inf_{y\in\mathcal{N}} \|x - y\|$$

is called the *semigap* (or one-sided gap) from $\mathcal{M}$ to $\mathcal{N}$.

We notice some immediate properties of the semi-gap.

Lemma 4.3. *Let $\mathcal{M},\mathcal{N} \subset \mathbb{C}^n$ be two subspaces. Then the following statements hold.*

(i) $\theta(\mathcal{M},\mathcal{N}) = \max\{\theta_0(\mathcal{M},\mathcal{N}), \theta_0(\mathcal{N},\mathcal{M})\}$.

(ii) $\theta_0(\mathcal{M}, \mathcal{N}) = \sup_{\substack{x \in \mathcal{M} \\ \|x\|=1}} \|x - P_{\mathcal{N}} x\|$.

(iii) *If* $\mathcal{N}_1 \subset \mathcal{N}_2$, *then* $\theta_0(\mathcal{M}, \mathcal{N}_2) \leq \theta_0(\mathcal{M}, \mathcal{N}_1)$, $\theta_0(\mathcal{N}_1, \mathcal{M}) \leq \theta_0(\mathcal{N}_2, \mathcal{M})$.

(iv) $\theta_0(\mathcal{M}, \mathcal{N}) \leq 1$.

(v) *If* $\dim \mathcal{M} > \dim \mathcal{N}$, *then* $\theta_0(\mathcal{M}, \mathcal{N}) = 1$.

(vi) $\theta_0(\mathcal{M}, \mathcal{N}) < 1$ *if and only if* $\mathcal{M} \cap \mathcal{N}^\perp = \emptyset$.

These facts are well-known and can be found e.g. in [4, 5].

We derive some new results on gap and semigap that we need.

Proposition 4.4. *Let* A_0 *be fixed. Then, there exist* $\epsilon, K > 0$ *such that for all* A *with* $\|A - A_0\| < \epsilon$, *we have*

$$\theta_0(\ker(A), \ker(A_0)) \leq K\|A - A_0\|. \tag{4.6}$$

Since $\ker(A) \oplus \mathrm{Im}(A^\top) = \mathbb{C}^n$, often times it is easier to prove a result for the image rather than for the kernel. Because of this, the above proposition will follow from the next results.

Lemma 4.5. *Let* A_0 *be fixed. Then, there exist* $\epsilon, K > 0$ *such that for all* A *with* $\|A - A_0\| < \epsilon$, *we have*

$$\theta_0(\mathrm{Im}(A_0), \mathrm{Im}(A)) \leq K\|A - A_0\|.$$

Proof. Let A_0 be an $n \times n$-matrix with $\dim(\mathrm{Im}(A_0)) = k$. Consider an orthonormal basis $g_1, g_2, \ldots, g_k$ of $\mathrm{Im}(A_0)$. That is there are $f_1, f_2, \ldots, f_k$ such that $g_i = A_0 f_i$ for $i = 1, \ldots, k$. Define $h_i = Af_i$, so that $h_i \in \mathrm{Im}(A)$ for all i. Then, we have that

$$\|h_i - g_i\| = \|Af_i - A_0 f_i\| = \|(A - A_0)f_i\| \leq \|f_i\|\|A - A_0\|.$$

Now, let $x \in \mathrm{Im}(A_0)$ and $\|x\| = 1$. This means that $x = \alpha_1 g_1 + \ldots \alpha_k g_k$ for some α_i's. Define $y = \alpha_1 h_1 + \ldots \alpha_k h_k$. Note that $y \in \mathrm{Im}(A)$ and

$$\inf_{z \in \mathrm{Im} A} \|x - z\| \leq \|x - y\| \leq \sum_{i=1}^{k} \alpha_i \|h_i - g_i\|$$

$$\leq \max_j(|\alpha_j|) \sum_{i=1}^{k} \|f_i\|\|A - A_0\| \leq k \max_j(|\alpha_j|) \max_i \|f_i\|\|A - A_0\|.$$

Therefore, by taking the supremum over x we arrive at $\theta_0(\mathrm{Im}(A_0), \mathrm{Im}(A)) \leq K\|A - A_0\|$, where $K = k \max_j(|\alpha_j|) \max_i \|f_i\|$. $\square$

Next, in order to show that Lemma 4.5 implies Proposition 4.4, we need the following result.

Lemma 4.6. *Let $\mathcal{M}, \mathcal{N}$ be subspaces of $\mathbb{C}^n$. Then we have*

$$\theta_0(\mathcal{M}, \mathcal{N}) = \theta_0(\mathcal{N}^\perp, \mathcal{M}^\perp).$$

Proof. First, notice that if $\dim \mathcal{M} > \dim \mathcal{N}$ or $\theta_0(M, N) = 1$, the result follows immediately. Now, we consider the case where $\dim \mathcal{M} \le \dim \mathcal{N}$ and $\theta_0(M, N) < 1$. Define P to be the subspace of all the projection of vectors in $\mathcal{M}$ to $\mathcal{N}$, i.e. $P = \mathrm{proj}_{\mathcal{N}} \mathcal{M}$. Since $\theta_0(M, N) < 1$, we have that $\dim P = \dim \mathcal{M}$. Additionally, we have that $\theta(M, P) = \theta_0(M, N)$. Recall that $P \subset \mathcal{N}$ implies that $\mathcal{N}^\perp \subset P^\perp$. By Proposition 4.1 we get $\theta(\mathcal{M}, P) = \theta(P^\perp, \mathcal{M}^\perp)$. Then, using Lemma 4.3, we have $\theta(P^\perp, \mathcal{M}^\perp) \ge \theta_0(P^\perp, \mathcal{M}^\perp) = \theta(\mathcal{N}^\perp, \mathcal{M}^\perp)$. That is,

$$\theta(\mathcal{N}^\perp, \mathcal{M}^\perp) \le \theta(\mathcal{M}, \mathcal{N}).$$

Now, repeating the argument for $\mathcal{N}^\perp$ and $\mathcal{M}^\perp$ gives us the result. $\square$

Recall that $\ker(A) = \mathrm{Im}(A^*)^\perp$ and $\mathrm{Im}(A) = \ker(A^*)^\perp$. So, by combining Lemma 4.5 and Lemma 4.6, we see that Proposition 4.4 holds true.

5. Backward Stability

In this section we finalize the proof of Theorem 1.7.

Proof of Theorem 1.7. Let A_0 be given. Then, according to Proposition 4.4 there exists $K, \epsilon > 0$ such that (4.6) holds. Let A be a matrix such that $\|A - A_0\| < \epsilon$ and A has Schur decomposition $A = UTU^*$. Let λ_i's denote the eigenvalues of A_0 and μ_j's denote the eigenvalues of A. In the Schur decomposition for A, we have a sequence of unitary Hessenberg matrices $\{V_k\}_1^n$ such that $U = V_1 \cdot \ldots \cdot V_n$. Since the Schur decomposition construction is an iteration of steps as was shown in Lemma 3.3, it is enough to show that we can obtain a bound on the first step. Let μ_1 be an eigenvalue of A and let V_1 be the unitary matrix such that

$$V_1^* A V_1 = \left[\begin{array}{c|ccc} \mu_1 & \star & \cdots & \star \\ \hline 0 & & & \\ \vdots & & A_2 & \\ 0 & & & \end{array}\right]$$

where the first column v_1 of V_1 is a unit eigenvector of A corresponding to eigenvalue μ_1, and the matrix forms an orthogonal basis for $\mathbb{C}^n$.

Thus, by Proposition 4.4 we can find a unit vector $u_1 \in \ker(A_0 - \lambda_1 I)$ and constants K_i's such that $\|v_1 - u_1\| \le K_0\|A - A_0 + (\lambda_1 - \mu_1)I\| \le K_0\|A - A_0\| + K_1|\lambda_1 - \mu_1| \le K_0\|A - A_0\| + \widetilde{K}_1\|A - A_0\|^{\frac{1}{n}} \le K_2\|A - A_0\|^{\frac{1}{n}}$.

Hence, we can find a corresponding orthonormal basis forming U_1 such that $\|v_i - u_i\| \le K_i\|A - A_0\|^{\frac{1}{n}}$. Therefore, we have that $\|V_1 - U_1\| \le M_1\|A - A_0\|^{\frac{1}{n}}$ and by construction, we have that

$$U_1^* A_0 U_1 = \left[\begin{array}{c|ccc} \lambda_1 & \star & \cdots & \star \\ \hline 0 & & & \\ \vdots & & A_{0,2} & \\ 0 & & & \end{array}\right]$$

Repeating the process the same way we did in the proof of Lemma 3.4, we acquire a unitary matrix $U_0 = U_1 \cdot \ldots \cdot U_n$ such that $\|V_i - U_i\| \le M_i\|A_i - A_{0,i}\|^{\frac{1}{n}}$ for all $i = 1, \ldots, n$.

Note that by construction

$$\|A_{i+1} - A_{0,i+1}\| \le \|V_i^* A_i V_i - U_i^* A_{0,i} U_i\| \le \|V_i^* A_i V_i - V_i^* A_{0,i} V_i + V_i^* A_{0,i} V_i -$$

$$- V_i^* A_{0,i} U_i + V_i^* A_{0,i} U_i - U_i^* A_{0,i} U_i\| \le \|A_i - A_{0,i}\| + \|A_{0,i}\|\|V_i -$$

$$- U_{0,i}\| + \|V_i^* - U_i^*\|\|A_{0,i}\| \le \|A_i - A_{0,i}\| + 2M_i\|A_{0,i}\|\|A_i - A_{0,i}\|^{\frac{1}{n}} \le \widetilde{M}_i\|A - A_0\|^{\frac{1}{n}}.$$

It follows by induction that $\|U - U_0\| \le M\|A - A_0\|^{1/n}$ and $M := \sum_i \widetilde{M}_i$ with $\widetilde{M}_i$ depending only on A_0.

Using the argument similar to above, we conclude that

$$\|T - T_0\| = \|U^* A U - U_0^* A_0 U_0\| \le \widehat{M}\|A - A_0\|,$$

where $\widehat{M}$ depends only on A_0.

Hence, we arrive at the conclusion of one of our main results, i.e. formula (1.4) holds true with $K = M + \widehat{M}$. □

6. Different GK numbers and Failure of the Forward Stability of the Schur Decomposition

As it turns out the GK numbers of the original and perturbed matrices give us the information whether the forward stability of the Schur decomposition is impossible. The intuition behind the non-stable case comes from the following fact.

Proposition 6.1 (see [9]). *We have the inequality*

$$\inf \operatorname{dist}(\mathrm{Inv}A, \mathrm{Inv}A_0) > 0$$

where the infimum is taken over all possible pairs of A, A_0, having different GK numbers.

That is why we got the backward stability result and could not get the general result for forward stability.

Lemma 6.2. *Let $A_0 \in \mathbb{C}^{n\times n}$ and $A_0 = U_0 T_0 U_0^*$ its fixed Schur decomposition, where the (1,1)-entry of T_0 is an eigenvalue λ with $\dim \operatorname{Ker}(A_0 - \lambda I) \geq 2$. There exists $M > 0$ such that in any neighborhood of A_0,, i.e. $\{A : \|A - A_0\| < \varepsilon\}$ for any $\varepsilon > 0$,*

$$\sup_{A} \inf_{\substack{U,T \\ \text{Schur Form} \\ \text{of } A}} \|U - U_0\| + \|T - T_0\| > M > 0,$$

where the supremum is taken over all $A \in \mathcal{U}$ not having an eigenvector close to the first column u_1 of U_0, i.e. we have $\|u_1 - v\| > M$ for all v, eigenvectors of A, and the infimum is taken over all their Schur factorizations.

Proof. Let u_1 be the first column of U_0 and λ is the corresponding eigenvalue of A_0. By Proposition 1.2 we know that there are eigenvalues of A that lie relatively close to λ and the difference is equivalent to $\|A - A_0\|^{1/n}$. We will list those eigenvalues as $\mu_1, \ldots, \mu_l$.

Let $\{v_1, \ldots, v_{l_j}\}$ be a basis of $\operatorname{Ker}(A - \mu_j)$ (for $j = 1, \ldots, l$). Moreover, let us denote by $\{u_1, \ldots, u_{k_1(A_0)}\}$ the basis of $\operatorname{Ker}(A_0 - \lambda)$, where u_1 as before and for each v_j there is u_s ($s \neq 1$) such that

$$\|v_j - u_s\| \leq C\|A - A_0\|^{1/n}$$

for some positive number C. We can assume this by using the backward stability result proven above. Then, $\|u_1 - v_j\| \geq \|u_1 - u_s\| - \|u_s - v_j\| \geq \min_{i\neq 1} \|u_1 - u_i\| - C\|A - A_0\|^{1/n}$. We can always choose A close to A_0 so $M := \frac{\min_{i\neq 1}\|u_1 - u_i\|}{2} > C\|A - A_0\|^{1/n}$. Denote by j^* the index that minimize the left-hand side of the above inequalities. Therefore, when supremum and infimum is taken under the conditions of this lemma

$$\sup_{\widetilde{A}} \inf_{U,T} \|U - U_0\| + \|T - T_0\| \geq \inf_{U,T:A=UTU^*} \|U - U_0\| \geq \|u_1 - v_{j^*}\| > M.$$

$\square$

Remark 6.3. *The assumption that we can find such A in every neighborhood of A_0 not having an eigenvector close to the first column of U_0 is based on the following fact. Consider the Jordan form of $A_0 = P_0 J_0 P_0^{-1}$ such that the first two blocks correspond to the eigenvalue λ with the second one having u_1 as the eigenvector. For any $\varepsilon > 0$ take $A = P_0(J_0 + J_\varepsilon)P_0^{-1}$, where J_ε has the only non-zero entry equal to $\frac{\varepsilon}{\|P_0\|\|P_0^{-1}\|}$ on the $(j, j+1)$ spot ($j \times j$) is the size of the first Jordan block in J_0 and the second block corresponds to the eigenvector u_1). Hence, u_1 is not an eigenvector of A and $\|A - A_0\| = \|P_0 J_\varepsilon P_0^{-1}\| \leq \varepsilon$. So the set of A that we are taking supremum over in Lemma 6.2 is not empty.*

Now let us show that the statement of Theorem 1.6 is valid.

Proof of Theorem 1.6. First, note that having different GK numbers for A and A_0 implies that A_0 is derogatory, i.e. there is an eigenvalue λ of A_0 such that $\dim \operatorname{Ker}(A_0 - \lambda I) \geq 2$. Moreover, A_0 and T_0 are similar so they have the same GK numbers and $\dim \operatorname{Ker}(T_0 - \lambda I) \geq 2$ as well. Therefore, we can show the equivalent fact instead, i.e.

$$\sup_{\widetilde{B}} \inf_{U,T} \|U - I\| + \|T - T_0\| \geq \inf_{U,T: B = UTU^*} \|U - I\| > M.$$

If λ is the (1,1) entry of T_0 the we use Lemma 6.2 to get the desired result. If it is not then we are required to perform an extra step.

$$T_0 = U_0^* A_0 U_0 = \left[\begin{array}{ccccc|ccc} \star & \star & \cdots & \star & \star & \star & \cdots & \star \\ 0 & \star & \cdots & \star & \star & \star & \cdots & \star \\ 0 & 0 & \cdots & \star & \star & \star & \cdots & \star \\ \vdots & \ddots & \ddots & \ddots & \vdots & \star & \cdots & \star \\ & & & & & \star & \cdots & \star \\ 0 & 0 & \cdots & 0 & \star & \star & \cdots & \star \\ \hline & & 0 & & & & T_1 & \end{array}\right],$$

where the first say j rows do not have λ on its main diagonal and the next row of T_0 is the first time we meet λ. In addition, the (1,1)-entry of T_1 is λ. According to Lemma 2.2, it means that we have the same GK numbers related to λ for T_1 as for A_0. Recall that the truncation was using eigenvectors not corresponding to λ, so T_1 will have the same number and length of Jordan chains for λ as A_0 has. Thus, $\dim \operatorname{Ker}(T_1 - \lambda I) \geq 2$. Now, we use Lemma 6.2 to get the result for $\widetilde{A}_0$, V_0, and $\widetilde{T}_0$ by constructing B_1

(note that B is not upper triangular, since e_1 is not its eigenvector) for T_1 as described by the lemma. Then define

$$A = U_0 \left[\begin{array}{c|c} \begin{matrix} \star & \star & \cdots & \star & \star \\ 0 & \star & \cdots & \star & \star \\ 0 & 0 & \cdots & \star & \star \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \star \end{matrix} & \begin{matrix} \star & \cdots & \star \\ \star & \cdots & \star \\ \star & \cdots & \star \\ \star & \cdots & \star \\ \star & \cdots & \star \\ \star & \cdots & \star \end{matrix} \\ \hline \Large 0 & \Large B_1 \end{array}\right] U_0^*,$$

where the first j rows marked with stars coincide with T_0. By the construction, we can see that (1.3) holds true and hence finishing the proof of Theorem 1.6. $\square$

To summarize, we have showed that the Schur decomposition is backward stable and why it fails to be forward stable.

References

[1] H. Den Boer, G.Ph. Thijsse, *Semi-stability of sums of partial multiplicities under additive perturbations*, Integral Equations and Operator Theory, **3**, 1980, 23–42.

[2] S. Dogruer Akgul, A. Minenkova, V. Olshevsky, *Lipschitz stability of some canonical Jordan bases of real H-selfadjoint matrices under small perturbations*, arXiv:2204.04639.

[3] I. Gohberg, M. A. Kaashoek, *Unsolved problems in matrix and operator theory*, Integral Eqs. Oper. Theory, **1**, 1978, 278–283.

[4] I. Gohberg, P. Lancaster, L. Rodman, *Invariant Subspaces of Matrices with Applications*, Canadian Mathematical Society Series of Monographs and Advanced Texts. A Wiley-Interscience Publication, New York, 1986, xviii+692 pp.

[5] T. Kato, *Perturbation theory for linear operators*, Die Grundlehren der mathematischen Wissenschaften, Band 132 Springer-Verlag New York, Inc., New York, 1966, xix+592 pp.

[6] M. Konstantinov, D. Gu, V. Mehrmann, P. Petkov, *Perturbation Theory for Matrix Equations Studies in Computational Mathematics*, **9**, North Holland Publishing Co., 2003, 429 pp.

[7] A. Markus, E. Parilis, *The change of the Jordan structure of a matrix under small perturbations*, Mat. Issled., **54** (1980), 98–109 (*in Russian*), *English translation*: Linear Algebra Appl., **54**, 1983, 139–152.

[8] V. Matsaev, V. Olshevsky, *Cyclic dimensions, kernel multiplicities, and Gohberg-Kaashoek numbers*, Linear Algebra Appl., **239**, 1996, 161–174.

[9] V. Olshevsky, *A condition for the nearness of sets of invariant subspaces of near matrices in terms of their Jordan structures.*, (in Russian), Siberian Math. Journal, **30**, 4, 1989, 102–110.

[10] J. Moro, J. Burke, M. Overton, *On the Lidskii-Lyusternik-Vishik Perturbation Theory for Eigenvalues with Arbitrary Jordan Structure*, SIAM J.Matrix Anal. Appl., **18**, 1997, 793–817.

[11] A.M. Ostrowski, *Solution of equations in Euclidean and Banach spaces. Third edition of Solution of equations and systems of equations. Pure and Applied Mathematics*, Vol. 9. Academic Press, New York-London, 1973. xx+412 pp.

[12] B. Simon, *Orthogonal Polynomials on the Unit Circle, Part 1: Classical Theory*, AMS Colloquium Series, American Mathematical Society, Providence, RI, 2005.