SIAM J. MATRIX ANAL. APPL.
Vol. 37, No. 1, pp. 195–214


# STABILITY ANALYSIS OF THE TWO-LEVEL ORTHOGONAL ARNOLDI PROCEDURE*

DING LU†, YANGFENG SU†, AND ZHAOJUN BAI‡

**Abstract.** The second-order Arnoldi (SOAR) procedure is an algorithm for computing an orthonormal basis of the second-order Krylov subspace. It has found applications in solving quadratic eigenvalue problems and model order reduction of second-order dynamical systems among others. Unfortunately, the SOAR procedure can be numerically unstable. The two-level orthogonal Arnoldi (TOAR) procedure has been proposed as an alternative to SOAR to cure the numerical instability. In this paper, we provide a rigorous stability analysis of the TOAR procedure. We prove that under mild assumptions, the TOAR procedure is backward stable in computing an orthonormal basis of the associated linear Krylov subspace. The benefit of the backward stability of TOAR is demonstrated by its high accuracy in structure-preserving model order reduction of second-order dynamical systems.

**Key words.** second-order Krylov subspace, second-order Arnoldi procedure, backward stability, model order reduction, dynamical systems

**AMS subject classifications.** 65F15, 65F30, 65P99

**DOI.** 10.1137/151005142

## 1. Introduction.

The second-order Krylov subspace induced by a pair of matrices and one or two vectors is a generalization of the well-known (linear) Krylov subspace based on a matrix and a vector. An orthonormal basis matrix $Q_k$ of the second-order Krylov subspace can be generated by a second-order Arnoldi (SOAR) procedure [3]. The SOAR procedure has found applications in solving quadratic eigenvalue problems [33, 38] and model order reduction of second-order dynamical systems [2, 6] and structural acoustic analysis [25, 26]. It is implemented in Omega3P for electromagnetic modeling of particle accelerators [18, 19], and in MOR4ANSYS for the model order reduction of ANSYS engineering models [28].

It has been known that SOAR is prone to numerical instability due to the fact that it involves solving potentially ill-conditioned triangular linear systems and implicitly generates a nonorthonormal basis matrix $V_k$ of an associated (linear) Krylov subspace (see examples in section 5). The instability issue has drawn the attention of researchers since the SOAR procedure was proposed. For quadratic eigenvalue problems, Zhu [40] exploited the relations between the second-order Krylov subspace and the associated linear Krylov subspace described in [3] and proposed to represent the basis matrix $V_k$ of the linear Krylov subspace by the product of two orthonormal matrices $Q_k$ and $U_k$, where $Q_k$ is an orthonormal basis of the second-order Krylov subspace and $U_k$ is computed to maintain the orthonormality of $V_k$. The same idea

*Received by the editors January 23, 2015; accepted for publication (in revised form) by T. Stykel December 11, 2015; published electronically February 16, 2016.

http://www.siam.org/journals/simax/37-1/100514.html

†School of Mathematical Sciences, Fudan University, Shanghai 200433, China (dinglu@fudan.edu.cn, yfsu@fudan.edu.cn). Part of this work was done while the first author was visiting the University of California, Davis, supported by China Scholarship Council. The research of the second author was supported in part by the Innovation Program of Shanghai Municipal Education Commission 13zz007, E-Institutes of Shanghai Municipal Education Commission N.E303004, and NSFC key project 91330201.

‡Department of Computer Science and Department of Mathematics, University of California, Davis, CA 95616, USA (zbai@ucdavis.edu). The research of the this author was supported in part by the NSF grants DMS-1522697 and CCF-1527091.

was also independently studied about the same time and presented in [34] for the purposes of curing the numerical instability and finding a memory-efficient representation of the basis matrix $V_k$ in the context of using ARPACK [20] to solve high-order polynomial eigenvalue problems. The term "two-level orthogonal Arnoldi procedure", TOAR in short, was coined in [34]. Recently, the notion of memory-efficient representations of the orthonormal basis matrices of the second-order and higher-order Krylov subspaces was generalized to the model order reduction of time-delay systems [39], solving the linearized eigenvalue problem of matrix polynomials in the Chebyshev basis [17] and implementing a rational Krylov method for solving nonlinear eigenvalue problems [35].

Unlike SOAR, TOAR maintains an orthonormal basis matrix $V_k$ of the associated linear Krylov subspace. Therefore, it is generally believed to be numerically stable. This belief has been adopted in the literature [39, 17, 35]. The motivation of this paper is to provide a rigorous stability analysis of TOAR. We prove that under mild assumptions, TOAR with partial reorthogonalization is backward stable in computing the basis matrix $V_k$ of the associated linear Krylov subspace. In addition, in this paper, we provide a different derivation of the TOAR procedure comparing to the ones in [40, 34] and remove unnecessary normalization steps. A comprehensive TOAR procedure, including the partial reorthogonalization and the treatments of deflation and breakdown, is presented. The advantages of the TOAR procedure are illustrated by an application to the structure-preserving model order reduction of second-order dynamical systems.

The rest of this paper is organized as follows. In section 2, we review the definition and essential properties of the second-order Krylov subspace. In section 3, we derive the TOAR procedure. The backward stability of the TOAR procedure is proven in section 4. Numerical examples for the application of the TOAR procedure in the model order reduction of second-order dynamical systems are presented in section 5. Concluding remarks are in section 6.

*Notations.* Throughout the paper, we use the upper case letter for matrices, lower case letter for vectors, particularly, $I$ for the identity matrix with $e_i$ being the $i$th column, and the dimensions of these matrices and vectors conform with the dimensions used in the context. Following the convention of matrix analysis, we use $\cdot^{\mathrm{T}}$ for transpose, and $\cdot^{\dagger}$ for the pseudo-inverse, $|\cdot|$ for elementwise absolute value, $\|\cdot\|_2$ and $\|\cdot\|_F$ for 2-norm and Frobenius norm, respectively, $\mathrm{span}\{U, v\}$ for the subspace spanned by the columns of matrices $U$ and vector $v$. We also use MATLAB conventions $v(i:j)$ for the $i$th to $j$th entries of vector $v$, and $A(i:j, k:\ell)$ for the submatrix of matrix $A$ by the intersection of row $i$ to $j$ and column $k$ to $\ell$. Other notations will be explained as used.

**2. Second-order Krylov subspace.** Let $A$ and $B$ be $n\times n$ matrices and $r_{-1}$ and $r_0$ be length-$n$ vectors such that $[r_{-1}, r_0] \neq 0$. Then the sequence $r_{-1}, r_0, r_1, r_2, \ldots$ with

$$r_j = Ar_{j-1} + Br_{j-2} \quad \text{for } j \geq 1 \tag{2.1}$$

is called a *second-order Krylov sequence* based on $A$, $B$, $r_{-1}$, and $r_0$. The subspace

$$\mathcal{G}_k(A, B; r_{-1}, r_0) \equiv \mathrm{span}\{r_{-1}, r_0, r_1, \ldots, r_{k-1}\} \tag{2.2}$$

is called a $k$th *second-order Krylov subspace*. If the vector $r_j$ lies in the subspace spanned by the vectors $r_{-1}, \ldots r_{j-1}$, i.e., $\mathcal{G}_{j+1}(A, B; r_{-1}, r_0) = \mathcal{G}_j(A, B; r_{-1}, r_0)$, then

a *deflation* occurs. We should stress that the deflation does not necessarily imply that $r_{j+1}$ and all the following vectors still lie in $\mathcal{G}_j(A, B; r_{-1}, r_0)$. The latter case is referred to as a *breakdown* of the second-order Krylov sequence.

The second-order Krylov subspace $\mathcal{G}_k(A, B; r_{-1}, r_0)$ can be embedded in the linear Krylov subspace

$$\begin{aligned}\mathcal{K}_k(L, v_0) &\equiv \operatorname{span}\{v_0, Lv_0, L^2v_0, \ldots, L^{k-1}v_0\} \\ &= \operatorname{span}\left\{ \begin{bmatrix} r_0 \\ r_{-1} \end{bmatrix}, \begin{bmatrix} r_1 \\ r_0 \end{bmatrix}, \begin{bmatrix} r_2 \\ r_1 \end{bmatrix}, \ldots, \begin{bmatrix} r_{k-1} \\ r_{k-2} \end{bmatrix} \right\},\end{aligned} \tag{2.3}$$

where

$$L = \begin{bmatrix} A & B \\ I & 0 \end{bmatrix} \in \mathbb{R}^{2n\times 2n} \quad \text{and} \quad v_0 = \begin{bmatrix} r_0 \\ r_{-1} \end{bmatrix} \in \mathbb{R}^{2n},$$

and $I$ is an identity matrix of size $n$. Specifically, let $Q_k$ and $V_k$ be basis matrices of the subspaces $\mathcal{G}_k(A, B; r_{-1}, r_0)$ and $\mathcal{K}_k(L, v_0)$, respectively. Then by (2.3), we know that

$$\operatorname{span}\{V_k(1:n,:)\} = \operatorname{span}\{r_0, r_1, \ldots, r_{k-1}\}, \tag{2.4}$$

$$\operatorname{span}\{V_k(n{+}1:2n,:)\} = \operatorname{span}\{r_{-1}, r_0, \ldots, r_{k-2}\}, \tag{2.5}$$

and

$$\operatorname{span}\{Q_k\} = \operatorname{span}\{V_k(1:n,:), V_k(n{+}1:2n,:)\}. \tag{2.6}$$

By (2.4)–(2.6), the basis matrix $V_k$ of the linear Krylov subspace $\mathcal{K}_k(L, v_0)$ can be written in terms of the basis matrix $Q_k$ of the second-order Krylov subspace $\mathcal{G}_k(A, B; r_{-1}, r_0)$:

$$V_k = \begin{bmatrix} V_k(1:n,:) \\ V_k(n{+}1:2n,:) \end{bmatrix} = \begin{bmatrix} Q_k U_{k,1} \\ Q_k U_{k,2} \end{bmatrix} = \begin{bmatrix} Q_k & \\ & Q_k \end{bmatrix} \begin{bmatrix} U_{k,1} \\ U_{k,2} \end{bmatrix} \equiv Q_{[k]} U_k. \tag{2.7}$$

Without loss of generality, we assume the linear Krylov subspace $\mathcal{K}_k(L, v_0)$ is not reduced, i.e., $V_k$ is of dimension $2n \times k$. Otherwise, $\mathcal{K}_k(L, v_0) = \mathcal{K}_j(L, v_0)$ for some $j < k$ with $\mathcal{K}_j(L, v_0)$ being unreduced, and $V_k = V_j$. Note that $Q_k$ is $n \times \eta_k$, $U_{k,1}$ and $U_{k,2}$ are $\eta_k \times k$, where $\eta_k$ is the dimension of $\mathcal{G}_k(A, B; r_{-1}, r_0)$, and $\eta_k \le k + 1$ due to possible deflations. Equation (2.7) indicates a memory-efficient compact representation of the basis matrix $V_k$ since the memory size $2nk$ required to store $V_k$ is reduced to $(n + 2k)\eta_k$ for $Q_k$ and $U_k$.

Equation (2.6) provides a means to compute an orthonormal basis matrix $Q_k$ of $\mathcal{G}_k(A, B; r_{-1}, r_0)$. One can first generate a basis matrix $V_k$ of $\mathcal{K}_k(L, v_0)$, say by the Arnoldi procedure [31, Chap.5], and then extract $Q_k$ by a rank-revealing orthogonalization procedure applied to the columns of $[V_k(1:n,:), V_k(n+1:2n,:)]$. However, this scheme is too expensive. A computationally efficient algorithm exploits the compact representation (2.7) to directly compute $Q_k$ without first generating $V_k$ explicitly. The SOAR procedure [3] is one such algorithm. In SOAR, and similarly in SOAR-like procedures [23, 5, 15], to compute the orthonormal basis matrix $Q_k$, it imposes $U_{k,1} = I$ (assuming no deflation for simplicity) and $U_{k,2}$ is a strict upper triangular matrix. At each iteration of the SOAR procedure, it needs to solve potentially ill-conditioned triangular linear systems; see [3, Algorithm 4] for detail. Moreover, due to the choice of $U_{k,1}$ and $U_{k,2}$, the corresponding basis matrix $V_k$ of the associated linear Krylov subspace of $\mathcal{K}_k(L, v_0)$ is nonorthonormal. Consequently, the SOAR procedure could be numerically unstable. The instability problem of SOAR has been observed in several practical applications and will be illustrated numerically in section 5.

**3. TOAR Procedure.** To cure the potential instability of the SOAR procedure, in this section we study a procedure that computes an orthonormal basis matrix $Q_k$ of the second-order Krylov subspace $\mathcal{G}_k(A, B; r_{-1}, r_0)$, and meanwhile, maintains the basis matrix $V_k$ of the associated linear Krylov subspace $\mathcal{K}_k(L, v_0)$ as orthonormal. The procedure is called as TOAR.

Recall that the Arnoldi procedure for computing an orthonormal basis of the $k$th Kryov subspace $\mathcal{K}_k(L, v_0)$ is governed by the following Arnoldi decomposition of order $k-1$:

$$LV_{k-1} = V_k \underline{H}_k, \tag{3.1}$$

where $V_{k-1}$ consists of the first $k-1$ columns of $V_k$, and $\underline{H}_k$ is a $k \times (k-1)$ unreduced upper Hessenberg matrix; see, for example, [31, Chap. 5]. By the compact representation (2.7), we have the *compact Arnoldi decomposition* of order $k-1$,

$$\begin{bmatrix} A & B \\ I & 0 \end{bmatrix} \begin{bmatrix} Q_{k-1}U_{k-1,1} \\ Q_{k-1}U_{k-1,2} \end{bmatrix} = \begin{bmatrix} Q_k U_{k,1} \\ Q_k U_{k,2} \end{bmatrix} \underline{H}_k. \tag{3.2}$$

We now show how to compute $Q_k$, $U_{k,1}$, and $U_{k,2}$ without explicitly generating $V_k$. We note that the orthonormality of $Q_k$ and $V_k$ implies that $U_k = [U_{k,1}^T, U_{k,2}^T]^T$ is also orthonormal. Let us begin with the following lemma describing the relations between $Q_{k-1}$ and $Q_k$, $U_{k-1,i}$ and $U_{k,i}$ for $i = 1, 2$.

LEMMA 3.1. *For the compact Arnoldi decomposition* (3.2),

$$\operatorname{span}\{Q_k\} = \operatorname{span}\{Q_{k-1}, r\}, \tag{3.3}$$

*where* $r = AQ_{k-1}U_{k-1,1}(:, k-1) + BQ_{k-1}U_{k-1,2}(:, k-1)$. *Furthermore,* (a) *if* $r \in \operatorname{span}\{Q_{k-1}\}$, *then there exist vectors* $x_k$ *and* $y_k$ *such that*

$$Q_k = Q_{k-1}, \quad U_{k,1} = \begin{bmatrix} U_{k-1,1} & x_k \end{bmatrix}, \quad \textit{and} \quad U_{k,2} = \begin{bmatrix} U_{k-1,2} & y_k \end{bmatrix}; \tag{3.4}$$

(b) *otherwise, there exist vectors* $x_k$ *and* $y_k$ *and a scalar* $\beta_k \neq 0$ *such that*

$$Q_k = \begin{bmatrix} Q_{k-1} & q_k \end{bmatrix}, \quad U_{k,1} = \begin{bmatrix} U_{k-1,1} & x_k \\ 0 & \beta_k \end{bmatrix}, \quad \textit{and} \quad U_{k,2} = \begin{bmatrix} U_{k-1,2} & y_k \\ 0 & 0 \end{bmatrix}, \tag{3.5}$$

*where* $q_k = (I - Q_{k-1}Q_{k-1}^{\mathrm{T}})r/\alpha$ *and* $\alpha$ *is a normalization factor such that* $\|q_k\|_2 = 1$.

*Proof.* Note that the $(k-1)$st column of (3.1) is

$$Lv_{k-1} = V_{k-1}\underline{H}_k(1 : k-1, k-1) + \underline{H}_k(k, k-1) \cdot v_k,$$

where $v_{k-1}$ and $v_k$ are the last columns of $V_{k-1}$ and $V_k$, respectively. Therefore, given $V_{k-1}$ we have

$$\operatorname{span}\{V_{k-1}, Lv_{k-1}\} = \operatorname{span}\{V_{k-1}, v_k\} = \operatorname{span}\{V_k\} \equiv \mathcal{K}_k(L, v_0).$$

Since

$$\begin{bmatrix} V_{k-1} & Lv_{k-1} \end{bmatrix} = \begin{bmatrix} Q_{k-1}U_{k-1,1} & r \\ Q_{k-1}U_{k-1,2} & Q_{k-1}U_{k-1,1}(:, k-1) \end{bmatrix},$$

by applying the subspace relation (2.6), we have

$$\operatorname{span}\{Q_k\} = \operatorname{span}\{Q_{k-1}U_{k-1,1},\ r,\ Q_{k-1}U_{k-1,2}\} = \operatorname{span}\{Q_{k-1}, r\}.$$

Hence the relation (3.3) is proven.

To prove (a) and (b), we first note that by (2.4), the top $n$ elements of the $k$th column $v_k$ of $V_k$ satisfy

$$v_k(1:n) \in \text{span}\{r_0, r_1, \ldots, r_{k-1}\} \subset \mathcal{G}_k(A, B; r_{-1}, r_0) = \text{span}\{Q_k\}, \tag{3.6a}$$

and by (2.5), the bottom $n$ elements of $v_k$ satisfy

$$v_k(n+1:2n) \in \text{span}\{r_{-1}, r_0, \ldots, r_{k-2}\} \subset \mathcal{G}_{k-1}(A, B; r_{-1}, r_0) = \text{span}\{Q_{k-1}\}. \tag{3.6b}$$

If $r \in \text{span}\{Q_{k-1}\}$, then (3.3) implies $\text{span}\{Q_k\} = \text{span}\{Q_{k-1}\}$. Furthermore, by (3.6), the vector

$$v_k = \begin{bmatrix} v_k(1:n) \\ v_k(n+1:2n) \end{bmatrix} = \begin{bmatrix} Q_{k-1}x_k \\ Q_{k-1}y_k \end{bmatrix}$$

for some vectors $x_k$ and $y_k$. Therefore,

$$V_k = \begin{bmatrix} V_{k-1} & v_k \end{bmatrix} = \begin{bmatrix} Q_{k-1}U_{k-1,1} & Q_{k-1}x_k \\ Q_{k-1}U_{k-1,2} & Q_{k-1}y_k \end{bmatrix} = \begin{bmatrix} Q_{k-1}[U_{k-1,1}\ x_k] \\ Q_{k-1}[U_{k-1,2}\ y_k] \end{bmatrix}.$$

Thus the result (a) is proven.

If $r \notin \text{span}\{Q_{k-1}\}$, then by orthogonalizing $r$ against $Q_{k-1}$, we have

$$r = Q_{k-1}s + \alpha q_k,$$

where $q_k = (I - Q_{k-1}Q_{k-1}^{\mathrm{T}})r/\alpha$ is a unitary vector orthogonal to the columns of $Q_{k-1}$. Consequently, $Q_k = \begin{bmatrix} Q_{k-1} & q_k \end{bmatrix}$. Furthermore, according to (3.6), the Arnoldi vector

$$v_k = \begin{bmatrix} v_k(1:n) \\ v_k(n+1:2n) \end{bmatrix} = \begin{bmatrix} Q_{k-1}x_k + \beta_k q_k \\ Q_{k-1}y_k \end{bmatrix}$$

for some vectors $x_k$, $y_k$, and scalar $\beta_k$. Therefore, the result (b) is immediately proven by the following equations:

$$V_k = \begin{bmatrix} V_{k-1} & v_k \end{bmatrix} = \begin{bmatrix} Q_{k-1}U_{k-1,1} & Q_{k-1}x_k + \beta_k q_k \\ Q_{k-1}U_{k-1,2} & Q_{k-1}y_k \end{bmatrix}. \qquad \square$$

Now we derive the TOAR procedure to compute the compact Arnoldi decomposition (3.2). With the initial vectors $r_{-1}$ and $r_0$ such that $v_0 = \begin{bmatrix} r_0 \\ r_{-1} \end{bmatrix} \neq 0$, we apply the rank revealing QR decomposition of the $n \times 2$ matrix

$$\begin{bmatrix} r_{-1} & r_0 \end{bmatrix} = Q_1 X,$$

where $Q_1$ is an $n \times \eta_1$ orthonormal matrix, and $X$ is an $\eta_1 \times 2$ matrix. Note that $\eta_1 = 2$ if the starting vectors $r_{-1}$ and $r_0$ are linearly independent, otherwise $\eta_1 = 1$. Then it follows

$$V_1 = \frac{1}{\gamma}\begin{bmatrix} r_0 \\ r_{-1} \end{bmatrix} = \frac{1}{\gamma}\begin{bmatrix} Q_1 X(:,2) \\ Q_1 X(:,1) \end{bmatrix} = \begin{bmatrix} Q_1 & \\ & Q_1 \end{bmatrix}\begin{bmatrix} U_{1,1} \\ U_{1,2} \end{bmatrix} \equiv Q_{[1]}U_1,$$

where $\gamma = \|v_0\|_2$, $U_{1,1} = X(:,2)/\gamma$ and $U_{1,2} = X(:,1)/\gamma$. After the initialization, we have $Q_1$, $U_1$, and an empty $1 \times 0$ Hessenberg matrix $\underline{H}_1 = [\,]$.

Let us assume that the compact Arnoldi decomposition (3.2) has been computed for $k = j$, $j \geq 1$. Next, the TOAR procedure consists of two steps to compute the decomposition (3.2) for $k = j + 1$, namely,

$$\begin{bmatrix} A & B \\ I & 0 \end{bmatrix} \begin{bmatrix} Q_j U_{j,1} \\ Q_j U_{j,2} \end{bmatrix} = \begin{bmatrix} Q_{j+1} U_{j+1,1} \\ Q_{j+1} U_{j+1,2} \end{bmatrix} \underline{H}_{j+1}, \tag{3.7}$$

where $Q_j$, $U_{j,1}$, and $U_{j,2}$ and the first $j-1$ columns of $\underline{H}_{j+1}$ in (3.7) have been computed.

At the first step, we compute the orthonormal matrix $Q_{j+1}$. According to Lemma 3.1, we have $\text{span}\{Q_{j+1}\} = \text{span}\{Q_j, r\}$, where $r = AQ_jU_{j,1}(:,j) + BQ_jU_{j,2}(:,j)$. By orthogonalizing $r$ against $Q_j$, it yields

$$q_{j+1} = (r - Q_j s)/\alpha \quad \text{with} \quad s = Q_j^{\mathrm{T}} r,\ \alpha = \|r - Q_j s\|_2, \tag{3.8}$$

where it is assumed that $\alpha \neq 0$. Subsequently, $Q_{j+1} = \begin{bmatrix} Q_j & q_{j+1} \end{bmatrix}$ and $\eta_{j+1} = \eta_j + 1$. If $\alpha = 0$, then $r \in \text{span}\{Q_j\}$ and $Q_{j+1} = Q_j$ and $\eta_{j+1} = \eta_j$. A deflation occurs and $\mathcal{G}_{j+1}(A, B; r_{-1}, r_0) = \mathcal{G}_j(A, B; r_{-1}, r_0)$.

At the second step, we compute $U_{j+1,1}$ and $U_{j+1,2}$ such that $U_{j+1} = \begin{bmatrix} U_{j+1,1} \\ U_{j+1,2} \end{bmatrix}$ is orthonormal. By left multiplying $Q_{[j+1]}^{\mathrm{T}}$ of (3.7) and taking the $j$th column, we have

$$\begin{bmatrix} Q_{j+1}^{\mathrm{T}} r \\ Q_{j+1}^{\mathrm{T}} Q_j U_{j,1}(:,j) \end{bmatrix} = \begin{bmatrix} U_{j+1,1} \\ U_{j+1,2} \end{bmatrix} \underline{H}_{j+1}(:,j). \tag{3.9}$$

When there is no deflation in the first step, by Lemma 3.1, $U_{j+1}$ is of the form

$$U_{j+1} = \begin{bmatrix} U_{j+1,1} \\ U_{j+1,2} \end{bmatrix} = \begin{bmatrix} U_{j,1} & x_{j+1} \\ 0 & \beta_{j+1} \\ U_{j,2} & y_{j+1} \\ 0 & 0 \end{bmatrix}$$

for some vectors $x_{j+1}$ and $y_{j+1}$, and scalar $\beta_{j+1} \neq 0$. Denote (see (3.8))

$$\begin{matrix} \eta_j \\ 1 \\ \eta_j \\ 1 \end{matrix} \begin{bmatrix} s \\ \alpha \\ u \\ 0 \end{bmatrix} \equiv \begin{matrix} \eta_{j+1} \\ \eta_{j+1} \end{matrix} \begin{bmatrix} Q_{j+1}^T r \\ Q_{j+1}^T Q_j U_{j,1}(:,j) \end{bmatrix},$$

where $u \equiv U_{j,1}(:,j)$. The equation (3.9) can be written as

$$\begin{bmatrix} s \\ \alpha \\ u \\ 0 \end{bmatrix} = \begin{bmatrix} U_{j,1} \\ 0 \\ U_{j,2} \\ 0 \end{bmatrix} h_j + h_{j+1,j} \begin{bmatrix} x_{j+1} \\ \beta_{j+1} \\ y_{j+1} \\ 0 \end{bmatrix},$$

where $h_j = \underline{H}_{j+1}(1:j, j)$ and $h_{j+1,j} = \underline{H}_{j+1}(j+1, j)$. Since $U_{j+1}$ is also imposed to be orthonormal, we have

$$h_j = \begin{bmatrix} U_{j,1} \\ 0 \\ U_{j,2} \\ 0 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} s \\ \alpha \\ u \\ 0 \end{bmatrix} \quad \text{and} \quad h_{j+1,j} = \left\| \begin{bmatrix} s \\ \alpha \\ u \\ 0 \end{bmatrix} - \begin{bmatrix} U_{j,1} \\ 0 \\ U_{j,2} \\ 0 \end{bmatrix} h_j \right\|_2. \tag{3.10}$$

Assume $h_{j+1,j} \neq 0$, the vectors $x_{j+1}$ and $y_{j+1}$ and scalar $\beta_{j+1}$, namely, the $(j+1)$st column of $U_{j+1}$, are then given by

$$\begin{bmatrix} x_{j+1} \\ \beta_{j+1} \\ y_{j+1} \\ 0 \end{bmatrix} = \frac{1}{h_{j+1,j}} \begin{bmatrix} s \\ \alpha \\ u \\ 0 \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} s \\ \alpha \\ u \\ 0 \end{bmatrix} := \begin{bmatrix} s \\ \alpha \\ u \\ 0 \end{bmatrix} - \begin{bmatrix} U_{j,1} \\ 0 \\ U_{j,2} \\ 0 \end{bmatrix} h_j. \tag{3.11}$$

If there is a deflation in the first step, then $Q_{j+1} = Q_j$. By Lemma 3.1, $U_{j+1}$ is of the form

$$U_{j+1} = \begin{bmatrix} U_{j,1} & x_{j+1} \\ U_{j,2} & y_{j+1} \end{bmatrix}$$

for some vectors $x_{j+1}$ and $y_{j+1}$. Denote

$$\begin{matrix} \eta_j \\ \eta_j \end{matrix} \overset{1}{\begin{bmatrix} s \\ u \end{bmatrix}} = \begin{bmatrix} Q_{j+1}^T r \\ Q_{j+1}^T Q_j U_{j,1}(:,j) \end{bmatrix} = \begin{bmatrix} Q_j^T r \\ Q_j^T Q_j U_{j,1}(:,j) \end{bmatrix} = \begin{bmatrix} Q_j^T r \\ U_{j,1}(:,j) \end{bmatrix}.$$

Equation (3.9) can be written as

$$\begin{bmatrix} s \\ u \end{bmatrix} = \begin{bmatrix} U_{j,1} \\ U_{j,2} \end{bmatrix} h_j + h_{j+1,j} \begin{bmatrix} x_{j+1} \\ y_{j+1} \end{bmatrix},$$

where $h_j = \underline{H}_{j+1}(1:j,j)$ and $h_{j+1,j} = \underline{H}_{j+1}(j+1,j)$. Since $U_{j+1}$ is imposed to be orthonormal, we have

$$h_j = \begin{bmatrix} U_{j,1} \\ U_{j,2} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} s \\ u \end{bmatrix}, \quad h_{j+1,j} = \left\| \begin{bmatrix} s \\ u \end{bmatrix} - \begin{bmatrix} U_{j,1} \\ U_{j,2} \end{bmatrix} h_j \right\|_2. \tag{3.12}$$

Assume $h_{j+1,j} \neq 0$, the vectors $x_{j+1}$ and $y_{j+1}$, namely, the $(j+1)$st column of $U_{j+1}$, are then given by

$$\begin{bmatrix} x_{j+1} \\ y_{j+1} \end{bmatrix} = \frac{1}{h_{j+1,j}} \begin{bmatrix} s \\ u \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} s \\ u \end{bmatrix} := \begin{bmatrix} s \\ u \end{bmatrix} - \begin{bmatrix} U_{j,1} \\ U_{j,2} \end{bmatrix} h_j. \tag{3.13}$$

Finally, we note that if $h_{j+1,j} = 0$ in (3.10) or (3.12), then by (3.7), we have

$$\begin{bmatrix} A & B \\ I & 0 \end{bmatrix} \begin{bmatrix} Q_j U_{j,1} \\ Q_j U_{j,2} \end{bmatrix} = \begin{bmatrix} Q_j U_{j,1} \\ Q_j U_{j,2} \end{bmatrix} H_j, \tag{3.14}$$

where $H_j = \underline{H}_{j+1}(1:j, 1:j)$. It implies that $V_j = Q_{[j]}U_j$ spans an invariant subspace of $L$, i.e., $\mathcal{K}_\ell(L, v_0) = \mathcal{K}_j(L, v_0)$ for $\ell \geq j$. Consequently, by (2.3), $\mathcal{G}_\ell(A, B; r_{-1}, r_0) = \mathcal{G}_j(A, B; r_{-1}, r_0)$ for all $\ell \geq j$. This is the case where the breakdown occurs.

A summary of the TOAR procedure is presented in Algorithm 1, where the modified Gram–Schmidt (MGS) process is applied for the orthogonalization of $Q_j$ (lines 5–8) and $U_j$ (lines 11–14). While the MGS is known to be numerically more accurate than the classical Gram–Schmidt process [13], it could still fail to compute a matrix orthonormal with respect to the machine precision. Specifically, by the error analysis in [7, 8], the computed $Q_k$ by the MGS (lines 5–8) satisfies

$$\|I - Q_k^{\mathrm{T}} Q_k\|_2 \leq c\kappa_2(R_k)\varepsilon, \tag{3.15}$$

ALGORITHM 1. Two-level Orthogonal ARnoldi(TOAR) Procedure

**Input:** Matrices $A$, $B$, and initial length-$n$ vectors $r_{-1}$, $r_0$ with $\gamma \equiv \|[r_{-1}, r_0]\|_F \neq 0$, and subspace order $k$.

**Output:** $Q_k \in \mathbb{R}^{n\times\eta_k}$, $U_{k,1}$, $U_{k,2} \in \mathbb{R}^{\eta_k\times k}$ and $\underline{H}_k = \{h_{ij}\} \in \mathbb{R}^{k\times k-1}$.

1: Rank revealing QR: $\begin{bmatrix} r_{-1} & r_0 \end{bmatrix} = QX$ with $\eta_1$ being the rank.
2: Initialize $Q_1 = Q$, $U_{1,1} = X(:,2)/\gamma$ and $U_{1,2} = X(:,1)/\gamma$.
3: **for** $j = 1, 2, \ldots, k-1$ **do**
4: $\quad r = A\left(Q_j U_{j,1}(:,j)\right) + B\left(Q_j U_{j,2}(:,j)\right)$
5: $\quad$ **for** $i = 1, \ldots, \eta_j$ **do**
6: $\quad\quad s_i = q_i^{\mathrm{T}} r$
7: $\quad\quad r = r - s_i q_i$
8: $\quad$ **end for**
9: $\quad \alpha = \|r\|_2$
10: $\quad$ Set $s = [s_1, \ldots, s_{\eta_j}]^{\mathrm{T}}$ and $u = U_{j,1}(:,j)$
11: $\quad$ **for** $i = 1, \ldots, j$ **do**
12: $\quad\quad h_{ij} = U_{j,1}(:,i)^{\mathrm{T}} s + U_{j,2}(:,i)^{\mathrm{T}} u$
13: $\quad\quad s = s - h_{ij} U_{j,1}(:,i)$; $u = u - h_{ij} U_{j,2}(:,i)$
14: $\quad$ **end for**
15: $\quad h_{j+1,j} = (\alpha^2 + \|s\|_2^2 + \|u\|_2^2)^{1/2}$
16: $\quad$ **if** $h_{j+1,j} = 0$ **then**
17: $\quad\quad$ **stop** (breakdown)
18: $\quad$ **end if**
19: $\quad$ **if** $\alpha = 0$ **then**
20: $\quad\quad \eta_{j+1} = \eta_j$ (deflation)
21: $\quad\quad Q_{j+1} = Q_j$; $U_{j+1,1} = \begin{bmatrix} U_{j,1} & s/h_{j+1,j} \end{bmatrix}$; $U_{j+1,2} = \begin{bmatrix} U_{j,2} & u/h_{j+1,j} \end{bmatrix}$
22: $\quad$ **else**
23: $\quad\quad \eta_{j+1} = \eta_j + 1$
24: $\quad\quad Q_{j+1} = \begin{bmatrix} Q_j & r/\alpha \end{bmatrix}$;
$\quad\quad U_{j+1,1} = \begin{bmatrix} U_{j,1} & s/h_{j+1,j} \\ 0 & \alpha/h_{j+1,j} \end{bmatrix}$; $\quad U_{j+1,2} = \begin{bmatrix} U_{j,2} & u/h_{j+1,j} \\ 0 & 0 \end{bmatrix}$
25: $\quad$ **end if**
26: **end for**

where $c$ is a constant only depending on the dimension of $Q_k$, $\kappa_2(R_k)$ is the condition number of the matrix $R_k = [r_1, r_2, \ldots, r_k]$ with $r_j$ being the computed vector $r$ at line 4 in the $j$th iteration. Therefore, the loss of orthonormality occurs when $R_k$ is ill-conditioned. Fortunately, this problem can be cured by applying a partial reorthogonalization [11], which runs the MGS twice on selective columns of $R_k$. Specifically, after line 9, if it holds

$$\alpha \leq \theta \alpha_1, \tag{3.16}$$

where $\alpha_1 = \|AQ_j U_{j,1}(:,j) + BQ_j U_{j,2}(:,j)\|_2$ and $\theta \leq 1$ is a threshold parameter, then we orthogonalize the resulting $r$ against $Q_j$ again, and then update $q_{j+1}$. This is also known as the Kahan–Parlett "twice-is-enough" algorithm [24, page 115] and [12]. To incorporate the partial reorthogonalization into Algorithm 1, we need to insert the following code segments between lines 9 and 10:

**if** $\alpha \leq \theta \alpha_1$
$\quad$ **for** $i = 0, \ldots, \eta_j$

$\widetilde{s}_i = q_i^{\mathrm{T}} r$
$r = r - \widetilde{s}_i q_i$
$s_i = s_i + \widetilde{s}_i$
**end for**
$\alpha = \|r\|_2$
**end if**

where $\alpha_1 = \|AQ_jU_{j,1}(:,j) + BQ_jU_{j,2}(:,j)\|_2$ can be precomputed at line 4 of Algorithm 1. Analogously, we can apply the partial reorthogonalization for the second MGS, lines 11 to 14 in Algorithm 1, to ensure the computed $U_j$ is orthonormal with respect to the machine precision.

To end this section, we note that the idea of using the product of two orthonormal matrices $Q_k$ and $U_k$ to represent an orthonormal basis matrix $V_k$ has appeared in [40, 34]. In this section, we derived the algorithm based on the well-known Arnoldi decomposition and the relations between $Q_{k-1}$, $U_{k-1}$ and $Q_k$, $U_k$. The derivation is more straightforward than the one presented in [40]. In addition, we removed some unnecessary normalization steps for computing $q_j$ as required in [40]. In section 4, we will discuss the treatments of deflation and breakdown in detail.

**4. Backward Error Analysis.** In this section we will provide a backward error analysis of the TOAR procedure in the presence of finite precision arithmetic. By taking into account floating point arithmetic errors, the computed compact Arnoldi decomposition by the TOAR procedure satisfies

$$\begin{bmatrix} A & B \\ I & 0 \end{bmatrix} \begin{bmatrix} \widehat{Q}_{k-1}\widehat{U}_{k-1,1} \\ \widehat{Q}_{k-1}\widehat{U}_{k-1,2} \end{bmatrix} = \begin{bmatrix} \widehat{Q}_k\widehat{U}_{k,1} \\ \widehat{Q}_k\widehat{U}_{k,2} \end{bmatrix} \underline{\widehat{H}}_k + E, \tag{4.1}$$

where $\widehat{Q}_k$, $\widehat{U}_{k,1}$, $\widehat{U}_{k,2}$, and $\underline{\widehat{H}}_k$ are computed matrices of $Q_k$, $U_{k,1}$, $U_{k,2}$, and $\underline{H}_k$ of the exact compact Arnoldi decomposition (3.2), and $E$ is the error matrix. Let $\widehat{V}_k = \begin{bmatrix} \widehat{Q}_k\widehat{U}_{k,1} \\ \widehat{Q}_k\widehat{U}_{k,2} \end{bmatrix}$ and $\widehat{U}_k = \begin{bmatrix} \widehat{U}_{k,1} \\ \widehat{U}_{k,2} \end{bmatrix}$, then (4.1) can be written as

$$(L + \Delta L)\widehat{V}_{k-1} = \widehat{V}_k \underline{\widehat{H}}_k, \tag{4.2}$$

where $\Delta L = E\widehat{V}_{k-1}^{\dagger}$. $\widehat{V}_{k-1}^{\dagger}$ is the pseudoinverse of $\widehat{V}_{k-1}$, and $\widehat{V}_{k-1}^{\dagger}=(\widehat{V}_{k-1}^{\mathrm{T}}\widehat{V}_{k-1})^{-1}\widehat{V}_{k-1}^{\mathrm{T}}$. Here we assume that the computed orthonormal matrices $\widehat{Q}_k$ and $\widehat{U}_k$ are of full column rank, which implies that $\widehat{V}_k$ is also of full column rank. Equation (4.2) indicates that $\widehat{V}_k$ is an exact basis matrix of the Krylov subspace of $L+\Delta L$, a perturbed matrix of $L$. Therefore, the relative backward error $\tau = \|\Delta L\|/\|L\|$ is a measure of the numerical stability of the TOAR procedure. We will show that under mild assumptions, the TOAR procedure is backward stable in the sense that the relative backward error $\tau$ is of the order of machine precision $\varepsilon$.

The proof of the backward stability of the TOAR procedure consists of two parts. In the first part, we derive an upper bound of the error matrix $E$. In the second part, we bound the backward error $\Delta L = E\widehat{V}_{k-1}^{\dagger}$. For the sake of exposition, we assume there is no deflation or breakdown, and no partial reorthogonalization. These cases will be addressed later in this section. In the following analysis, we adopt the standard rounding off error model for the floating point arithmetic of two real scalars $\alpha$ and $\beta$:

$$fl(\alpha \text{ op } \beta) = (\alpha \text{ op } \beta)(1+\delta) \quad \text{with } |\delta| \le \varepsilon \text{ for op} = +, -, *, /, \tag{4.3}$$

where $fl(x)$ denotes the computed quantity and $\varepsilon$ denotes the machine precision. In addition, we recall the following results which will be used repeatedly in this section.

LEMMA 4.1.

(a) *For $x, y \in \mathbb{R}^n$, $fl(x+y) = x + y + f$, where $\|f\|_2 \le (\|x\|_2 + \|y\|_2)\varepsilon$.*

(b) *For $X \in \mathbb{R}^{n\times k}$ and $y \in \mathbb{R}^k$, $fl(Xy) = Xy + w$, where $\|w\|_2 \le k\|X\|_F\|y\|_2\varepsilon + \mathcal{O}(\varepsilon^2)$.*

(c) *For $X \in \mathbb{R}^{n\times k}$, $y \in \mathbb{R}^k$, $b \in \mathbb{R}^n$, and $\beta \in \mathbb{R}$, $\widehat{c} \equiv fl((b - Xy)/\beta)$ satisfies*

$$\beta\widehat{c} = b - Xy + g, \qquad \|g\|_2 \le (k+1)\left\|\begin{bmatrix} X & \widehat{c}\end{bmatrix}\right\|_F \left\|\begin{bmatrix} y \\ \beta\end{bmatrix}\right\|_2 \varepsilon + \mathcal{O}(\varepsilon^2).$$

*Proof.* Result (a) is a direct consequence of the model (4.3) applied elementwisely to the vector $x + y$.

For (b), repeatedly applying the model (4.3) leads to $fl(Xy) = Xy + w$ with $|w| \le k|X||y|\varepsilon + \mathcal{O}(\varepsilon^2)$; see e.g. [14, (3.12), page 78], where $|\cdot|$ denotes elementwise absolute values. By taking 2-norms and using $\|(|X|)\|_2 \le \|X\|_F$, we prove (b).

Result (c) is a direct consequence of Lemma 8.4 in [14, pp.154]. For the sake of completeness, we provide a proof here. We first note that for the scalar operation, we have

$$fl((\alpha - \gamma)/\beta) = (\alpha - \gamma)/\beta(1 + \delta_1)(1 + \delta_2) = (\alpha - \gamma)/\beta + e, \tag{4.4}$$

where $|\delta_i| \le \varepsilon$ for $i = 1, 2$ and $|e| \le 2|fl((\alpha - \gamma)/\beta)|\varepsilon + \mathcal{O}(\varepsilon^2)$. Applying (4.4) elementwisely to the vector $\widehat{c}$ gives rise to

$$\widehat{c} = fl((b - Xy)/\beta) = (b - fl(Xy))/\beta + f, \quad \|f\|_2 \le 2\|\widehat{c}\|_2\varepsilon + \mathcal{O}(\varepsilon^2).$$

Multiplying the equation by $\beta$ and using (b), we obtain

$$\|g\|_2 \equiv \|\beta f - w\|_2 \le (k+1)(|\beta|\|\widehat{c}\|_2 + \|X\|_F\|y\|_2)\varepsilon + \mathcal{O}(\varepsilon^2),$$

where we used $\max\{k, 2\} \le k+1$ for $k \ge 1$. Then by the Cauchy–Schwartz inequality $ab + cd \le (a^2 + c^2)^{1/2}(b^2 + d^2)^{1/2}$ we prove (c). ∎

To derive an upper bound for the error matrix $E$ of (4.1), let us first introduce the following two matrices to represent the floating point errors of matrix-vector multiplications and the orthogonalization processes, respectively,

$$F_{\rm mv} = A(\widehat{Q}_{k-1}\widehat{U}_{k-1,1}) + B(\widehat{Q}_{k-1}\widehat{U}_{k-1,2}) - \widehat{R}_{k-1}, \tag{4.5}$$

$$F = \begin{bmatrix} \widehat{R}_{k-1} \\ \widehat{Q}_{k-1}\widehat{U}_{k-1,1}\end{bmatrix} - \begin{bmatrix} \widehat{Q}_k\widehat{U}_{k,1} \\ \widehat{Q}_k\widehat{U}_{k,2}\end{bmatrix}\underline{\widehat{H}}_k, \tag{4.6}$$

where $\widehat{R}_{k-1} = [\widehat{r}_1, \widehat{r}_2, \ldots, \widehat{r}_{k-1}]$ and $\widehat{r}_j = fl\big(A(\widehat{Q}_j\widehat{U}_{j,1}(:, j)) + B(\widehat{Q}_j\widehat{U}_{j,2}(:, j))\big)$ which is computed at the $j$th iteration of Algorithm 1 (line 4). By (4.1), it is easy to verify that

$$E = \begin{bmatrix} F_{\rm mv} \\ 0\end{bmatrix} + F. \tag{4.7}$$

The following lemma gives an upper bound of $\|F_{\rm mv}\|_F$.

LEMMA 4.2. *Let $\widehat{Q}_{k-1}$ and $\widehat{U}_{k-1}$ be the computed orthonormal matrices by the TOAR procedure (Algorithm 1). Then*

$$\|F_{\rm mv}\|_F \le \varphi\|\widehat{Q}_{k-1}\|_2\|\widehat{U}_{k-1}\|_2\|L\|_F\varepsilon + \mathcal{O}(\varepsilon^2), \tag{4.8}$$

*where $\varphi = 2k(2n + 1)$.*

*Proof.* By the definition (4.5) of $F_{\rm mv}$, the $j$th column of $F_{\rm mv}$ is given by

$$F_{\rm mv}(:,j) = A(\widehat{Q}_j\widehat{U}_{j,1}(:,j)) + B(\widehat{Q}_j\widehat{U}_{j,2}(:,j)) - \widehat{r}_j, \tag{4.9}$$

where $\widehat{r}_j \equiv fl\big(\, A(\widehat{Q}_j\widehat{U}_{j,1}(:,j)) + B(\widehat{Q}_j\widehat{U}_{j,2}(:,j))\,\big)$. Note that here we use the fact that $\widehat{U}_{k-1,1}$ and $\widehat{U}_{k-1,2}$ are upper Hessenberg matrices and $\widehat{Q}_{k-1}\widehat{U}_{k-1,i}(:,j) = \widehat{Q}_j\widehat{U}_{j,i}(:,j)$, for $i = 1, 2$ and $j \le k-1$.

By repeatedly applying Lemmas 4.1(a) and 4.1(b), we have

$$\begin{aligned}\widehat{r}_j &= fl(A\widehat{Q}_j\widehat{U}_{j,1}(:,j)) + fl(B\widehat{Q}_j\widehat{U}_{j,2}(:,j)) + w_j\\ &= A\widehat{Q}_j\widehat{U}_{j,1}(:,j) + B\widehat{Q}_j\widehat{U}_{j,2}(:,j) + w_j^{(1)} + w_j^{(2)} + w_j,\end{aligned}$$

where $w_j^{(1)}$, $w_j^{(2)}$ and $w_j$ are floating point error vectors and satisfy

$$\begin{aligned}\|w_j^{(1)}\|_2 &\le 2n\|A\|_F\|\widehat{Q}_j\|_F\|\widehat{U}_{j,1}(:,j)\|_2\varepsilon + \mathcal{O}(\varepsilon^2),\\ \|w_j^{(2)}\|_2 &\le 2n\|B\|_F\|\widehat{Q}_j\|_F\|\widehat{U}_{j,2}(:,j)\|_2\varepsilon + \mathcal{O}(\varepsilon^2),\\ \|w_j\|_2 &\le (\|fl(A\widehat{Q}_j\widehat{U}_{j,1}(:,j))\|_2 + \|fl(B\widehat{Q}_j\widehat{U}_{j,2}(:,j))\|_2)\varepsilon\\ &\le (\|A\widehat{Q}_j\widehat{U}_{j,1}(:,j)\|_2 + \|B\widehat{Q}_j\widehat{U}_{j,2}(:,j)\|_2)\varepsilon + \mathcal{O}(\varepsilon^2).\end{aligned}$$

Therefore, $F_{\rm mv}(:,j) = -(w_j^{(1)} + w_j^{(2)} + w_j)$. By the upper bounds of $\|w_j^{(1)}\|$, $\|w_j^{(2)}\|$, and $\|w_j\|$, we have

$$\begin{aligned}\|F_{\rm mv}(:,j)\|_2 &\le (2n+1)\Big(\|A\|_F\|\widehat{Q}_j\|_F\|\widehat{U}_{j,1}(\,:\,,j)\|_2 + \|B\|_F\|\widehat{Q}_j\|_F\|\widehat{U}_{j,2}(\,:\,,j)\|_2\Big)\varepsilon + \mathcal{O}(\varepsilon^2)\\ &\le 2(2n+1)\|L\|_F\|\widehat{Q}_{k-1}\|_F\|\widehat{U}_j(\,:\,,j)\|_2\varepsilon + \mathcal{O}(\varepsilon^2),\end{aligned}$$

where for the second inequality we used the fact that $\max\{\,\|A\|_F,\,\|B\|_F\,\} \le \|L\|_F$, $\|\widehat{Q}_j\|_F \le \|\widehat{Q}_{k-1}\|_F$, and $\|\widehat{U}_{j,i}(\,:\,,j)\|_2 \le \|\widehat{U}_j(\,:\,,j)\|_2$ for $i = 1, 2$. In matrix terms, we have

$$\|F_{\rm mv}\|_F = \left(\sum_{j=1}^{k-1}\|F_{\rm mv}(:,j)\|_2^2\right)^{1/2} \le 2(2n+1)\|L\|_F\|\widehat{Q}_{k-1}\|_F\|\widehat{U}_{k-1}\|_F\varepsilon + \mathcal{O}(\varepsilon^2).$$

Since $\widehat{Q}_{k-1}$ and $\widehat{U}_{k-1}$ have at most $k$ columns, so $\|\widehat{Q}_{k-1}\|_F \le \sqrt{k}\|\widehat{Q}_{k-1}\|_2$ and $\|\widehat{U}_{k-1}\|_F \le \sqrt{k}\|\widehat{U}_k\|_2$, we obtain the upper bound (4.8). ∎

The following lemma gives an upper bound of $\|F\|_F$.

LEMMA 4.3. *Let $\widehat{Q}_k$, $\widehat{U}_k$, and $\underline{\widehat{H}}_k$ be computed by the $k$-step TOAR procedure (Algorithm 1). Then*

$$\|F\|_F \le \varphi\|\widehat{Q}_k\|_2\|\widehat{U}_k\|_2\|\underline{\widehat{H}}_k\|_F\varepsilon + \mathcal{O}(\varepsilon^2), \tag{4.10}$$

*where $\varphi = (k+1)(2k+1)$.*

*Proof.* In the $j$th iteration of Algorithm 1, the computed matrix-vector multiplication (line 4) is

$$\widehat{r}_j = fl(A\widehat{Q}_j\widehat{U}_{j,1}(:,j) + B\widehat{Q}_j\widehat{U}_{j,2}(:,j)).$$

For the first orthogonalization process (lines 5 to 9) and normalization (line 24), by applying Lemma 4.1(c), the computed $\eta_{j+1}$th column $\widehat{q}_{j+1}$ of $Q_k$ (which is evaluated by $\widehat{q}_{j+1} = fl((\widehat{r}_j - \widehat{Q}_j\widehat{s})/\widehat{\alpha})$) satisfies

$$\widehat{\alpha}\widehat{q}_{j+1} = \widehat{r}_j - \widehat{Q}_j\widehat{s} - \widetilde{f}_j, \tag{4.11}$$

where $\widehat{s}$ and $\widehat{\alpha}$ are computed in lines 6 and 9, respectively, and $\widetilde{f}_j$ is the error vector bounded by

$$\|\widetilde{f}_j\|_2 \le \varphi_1 \|\widehat{Q}_{j+1}\|_F \left\| \begin{bmatrix} \widehat{s} \\ \widehat{\alpha} \end{bmatrix} \right\|_2 \varepsilon + \mathcal{O}(\varepsilon^2), \tag{4.12}$$

where $\varphi_1 = j + 2$ since $\eta_{j+1} \le j + 2$. Similarly, for the second orthogonalization process (lines 11 to 15) and normalization (line 24), by Lemma 4.1(c), the computed $(j + 1)$st column of $U_{j+1}$ satisfies

$$\widehat{h}_{j+1,j} \begin{bmatrix} \widehat{U}_{j+1,1}(:, j+1) \\ \widehat{U}_{j+1,2}(:, j+1) \end{bmatrix} = \begin{bmatrix} \widehat{s} \\ \widehat{\alpha} \\ \widehat{u} \\ 0 \end{bmatrix} - \begin{bmatrix} \widehat{U}_{j,1} \\ 0 \\ \widehat{U}_{j,2} \\ 0 \end{bmatrix} \widehat{h}_j - g_j, \tag{4.13}$$

where $\widehat{s}$ and $\widehat{u} = \widehat{U}_{j,1}(:, j)$ are computed in line 10, $\widehat{h}_j = \underline{\widehat{H}}_k(1 : j, j)$ and $\widehat{h}_{j+1,j}$ are computed in lines 12 and 15, respectively. $g_j$ is the error vector bounded by

$$\|g_j\|_2 \le \varphi_2 \|\widehat{U}_{j+1}\|_F \|\underline{\widehat{H}}_k(1 : j + 1, j)\|_2 \varepsilon + \mathcal{O}(\varepsilon^2), \tag{4.14}$$

where $\varphi_2 = j + 1$.

With error bounds (4.12) and (4.14), we are ready to estimate the $j$th column of $F$. First, by exploiting the upper-Hessenberg structure of $\underline{\widehat{H}}_k$, the $j$th column of $F$ in (4.6) is given by

$$\begin{aligned}
f_j &= \begin{bmatrix} \widehat{r}_j \\ \widehat{Q}_k\widehat{U}_{k,1}(:, j) \end{bmatrix} - \begin{bmatrix} \widehat{Q}_k\widehat{U}_{k,1}(:, 1 : j) \\ \widehat{Q}_k\widehat{U}_{k,2}(:, 1 : j) \end{bmatrix} \widehat{h}_j - \widehat{h}_{j+1,j} \begin{bmatrix} \widehat{Q}_k\widehat{U}_{k,1}(:, j+1) \\ \widehat{Q}_k\widehat{U}_{k,2}(:, j+1) \end{bmatrix} \\
&= \begin{bmatrix} \widehat{r}_j \\ \widehat{Q}_j\widehat{U}_{j,1}(:, j) \end{bmatrix} - \begin{bmatrix} \widehat{Q}_j\widehat{U}_{j,1} \\ \widehat{Q}_j\widehat{U}_{j,2} \end{bmatrix} \widehat{h}_j - \widehat{h}_{j+1,j} \begin{bmatrix} \widehat{Q}_{j+1}\widehat{U}_{j+1,1}(:, j+1) \\ \widehat{Q}_{j+1}\widehat{U}_{j+1,2}(:, j+1) \end{bmatrix} \\
&= \begin{bmatrix} \widehat{r}_j \\ \widehat{Q}_j\widehat{U}_{j,1}(:, j) \end{bmatrix} - \begin{bmatrix} \widehat{Q}_j\widehat{s} + \widehat{q}_{j+1}\widehat{\alpha} \\ \widehat{Q}_j\widehat{u} \end{bmatrix} + \widehat{Q}_{[j+1]} g_j \\
&= \begin{bmatrix} \widetilde{f}_j \\ 0 \end{bmatrix} + \widehat{Q}_{[j+1]} g_j,
\end{aligned}$$

where, for the third equality, we use the following equation obtained by left multiplying $\widehat{Q}_{[j+1]}$ of (4.13):

$$\begin{bmatrix} \widehat{Q}_j\widehat{U}_{j,1} \\ \widehat{Q}_j\widehat{U}_{j,2} \end{bmatrix} \widehat{h}_j + \widehat{h}_{j+1,j} \begin{bmatrix} \widehat{Q}_{j+1}\widehat{U}_{j+1,1}(:, j+1) \\ \widehat{Q}_{j+1}\widehat{U}_{j+1,2}(:, j+1) \end{bmatrix} = \begin{bmatrix} \widehat{Q}_j\widehat{s} + \widehat{q}_{j+1}\widehat{\alpha} \\ \widehat{Q}_j\widehat{u} \end{bmatrix} - \widehat{Q}_{[j+1]} g_j.$$

Consequently, we have

$$\begin{aligned}
\|f_j\|_2 &\le \|\widetilde{f}_j\|_2 + \|\widehat{Q}_{j+1}\|_2\|g_j\|_2 \\
&\le \left(\varphi_1 \left\| \begin{bmatrix} \widehat{s} \\ \widehat{\alpha} \end{bmatrix} \right\|_2 \varepsilon + \|g_j\|_2\right) \|\widehat{Q}_{j+1}\|_F + \mathcal{O}(\varepsilon^2) \\
&\le \left(\varphi_1 \|\widehat{U}_{j+1}\|_2 \|\underline{\widehat{H}}_k(1:j+1,j)\|_2 \varepsilon + \|g_j\|_2\right) \|\widehat{Q}_{j+1}\|_F + \mathcal{O}(\varepsilon^2) \\
&\le (2j+3)\|\widehat{Q}_{j+1}\|_F \|\widehat{U}_{j+1}\|_F \|\underline{\widehat{H}}_k(1:j+1,j)\|_2 \varepsilon + \mathcal{O}(\varepsilon^2) \\
&\le (2k+1)\|\widehat{Q}_k\|_F \|\widehat{U}_k\|_F \|\underline{\widehat{H}}_k(1:j+1,j)\|_2 \varepsilon + \mathcal{O}(\varepsilon^2), \qquad (4.15)
\end{aligned}$$

where the upper bounds (4.12) and (4.14) of $\|\widetilde{f}_j\|_2$ and of $\|g_j\|_2$ are used for the second and fourth inequalities, respectively. In addition, for the third inequality, we use the following upper bound based on (4.13):

$$\left\| \begin{bmatrix} \widehat{s} \\ \widehat{\alpha} \end{bmatrix} \right\|_2 \le \left\| \begin{bmatrix} \widehat{s} \\ \widehat{\alpha} \\ \widehat{u} \\ 0 \end{bmatrix} \right\|_2 \le \|\widehat{U}_{j+1}\|_2 \cdot \|\underline{\widehat{H}}_k(1:j+1,j)\|_2 + \|g_j\|_2.$$

For the whole orthogonalization error matrix $F$, by (4.15), we have

$$\begin{aligned}
\|F\|_F^2 = \sum_{j=1}^{k-1} \|f_j\|_2^2 &\le (2k+1)^2 \sum_{j=1}^{k-1} \|\widehat{Q}_k\|_F^2 \|\widehat{U}_k\|_F^2 \|\underline{\widehat{H}}_k(1:j+1,j)\|_2^2 \varepsilon^2 + \mathcal{O}(\varepsilon^3) \\
&= (2k+1)^2 \|\widehat{Q}_k\|_F^2 \|\widehat{U}_k\|_F^2 \|\underline{\widehat{H}}_k\|_F^2 \varepsilon^2 + \mathcal{O}(\varepsilon^3).
\end{aligned}$$

The lemma is proven by using the Frobenius norm to 2-norm conversion. ∎

*Remark* 1. If the partial reorthogonalization is applied, Lemma 4.3 still holds subject to a minor change of the coefficient $\varphi$ in the upper bound (4.10). This is due to the fact that the inequalities (4.12) and (4.14) still hold with the coefficients $\varphi_1, \varphi_2 = 2(j+1)c + 1$, where $c$ is a small constant. Specifically, for the first MGS process (lines 5 to 8), by the rounding error analysis (see, e.g., [7, Eq. (5.6)]), the computed $\widehat{\widehat{s}}$ and $\widehat{r}$ ($s$ of line 6 and $r$ of line 7) satisfy

$$\widehat{r} = \widehat{r}_j - \widehat{Q}_j \widehat{\widehat{s}} + f_j^{(1)}, \qquad (4.16)$$

where $\|f_j^{(1)}\|_2 \le c(j+1)\|\widehat{r}_j\|_2 \varepsilon$ and $c$ is a small constant. Now, by running the reorthogonalization of $\widehat{r}$ against $\widehat{Q}_j$ and normalization (line 24), the computed $\eta_{j+1}$th column $\widehat{q}_{j+1}$ of $\widehat{Q}_k$ satisfies

$$\widehat{\alpha}\widehat{q}_{j+1} = \widehat{r} - \widehat{Q}_j \widehat{\widehat{s}} + f_j^{(2)}, \qquad (4.17)$$

where $\|f_j^{(2)}\|_2 \le c(j+1)\|\widehat{r}\|_2 \varepsilon$, and $\widehat{\alpha}_j$ and $\widehat{\widehat{s}}$ are computed quantities after the reorthogonalization. Meanwhile, the updated $s$ is given by

$$\widehat{s} = fl(\,\widehat{\widehat{s}} + \widehat{\widehat{s}}\,) = \widehat{\widehat{s}} + \widehat{\widehat{s}} + e, \qquad (4.18)$$

where $|e| \le |\widehat{s}|\varepsilon + \mathcal{O}(\varepsilon^2)$, due to the model (4.3) and

$$|fl(a+b) - (a+b)| = |fl(a+b)\delta/(1+\delta)| \le |fl(a+b)|\varepsilon + \mathcal{O}(\varepsilon^2).$$

By summing up (4.16) and (4.17) to eliminate $\widehat{r}$ and utilizing (4.18), we have the inequality (4.11) with

$$\begin{aligned}\|\widetilde{f}_j\|_2 \equiv \|f_j^{(1)} + f_j^{(2)} + \widehat{Q}_j e\|_2 \;&\le\; c(j+1)(\|\widehat{r}_j\|_2 + \|\widehat{r}\|_2)\varepsilon + \|\widehat{Q}_j\|_2\|e\|_2 \\ &\le\; 2c(j+1)\|\widehat{r}_j\|_2\varepsilon + \|\widehat{Q}_j\|_2\|\widehat{s}\|_2\varepsilon + \mathcal{O}(\varepsilon^2) \\ &\le (2c(j+1)+1)\|\widehat{Q}_{j+1}\|_2 \left\|[\widehat{s}^{\mathrm{T}}, \widehat{\alpha}]\right\|_2 \varepsilon + 2c(j+1)\|\widetilde{f}_j\|_2\varepsilon + \mathcal{O}(\varepsilon^2),\end{aligned}$$

where for the last inequality we used $\|\widehat{r}_j\|_2 \le \|\widehat{Q}_{j+1}\|_2 \left\|[\widehat{s}^{\mathrm{T}}, \widehat{\alpha}]\right\|_2 + \|\widetilde{f}_j\|_2$ as indicated by (4.11), and for the third inequality we used the partial reorthogonalization condition (3.16), i.e., $\|\widehat{r}\|_2 \le \theta\|\widehat{r}_j\|_2 \le \|\widehat{r}_j\|_2$. Since the term $\|\widetilde{f}_j\|_2\varepsilon$ is of order $\mathcal{O}(\varepsilon^2)$, we immediately have (4.12) with $\varphi_1 = 2(j+1)c + 1$.

Similarly, we can show in the second orthogonalization (lines 11 to 14), the inequality (4.14) holds with the coefficients $\varphi_2 = 2(j+1)c + 1$ when the partial reorthogonalization is applied.

*Remark* 2. In practice, the thresholds for the deflation (line 19) and the breakdown (line 16) in Algorithm 1 can be set to be

$$\alpha \le j\|s_j\|_2\varepsilon \quad \text{and} \quad h_{j+1,j} \le j\|h_j\|_2\varepsilon, \tag{4.19}$$

respectively. We can show that Lemma 4.3 still holds subject to a minor change of the coefficient $\varphi$ in the upper bound (4.10) when the deflation or breakdown occurs. Specifically, for the deflation, since $\|\widehat{q}_{j+1}\|_2 = 1 + \mathcal{O}(\varepsilon)$, we have

$$|\widehat{\alpha}|\|\widehat{q}_{j+1}\|_2 \le j\|\widehat{s}\|_2\varepsilon + \mathcal{O}(\varepsilon^2). \tag{4.20}$$

Therefore, the term $\widehat{\alpha}\widehat{q}_{j+1}$ on the left side of (4.11) is of order $\mathcal{O}(\varepsilon)$, so we can move and absorb it into $\widetilde{f}_j$ on the right side to obtain

$$0 = \widehat{r}_j - \widehat{Q}_j\widehat{s}_j + \widetilde{f}_j, \qquad \widetilde{f}_j := \widetilde{f}_j + \widehat{\alpha}\widehat{q}_{j+1}, \tag{4.21}$$

where by the error bound of $\widetilde{f}_j$ in (4.12), the updated $\widetilde{f}_j$ satisfies

$$\begin{aligned}\|\widetilde{f}_j\|_2 \le\; &(j+2)(\|\widehat{Q}_j\|_F + \|\widehat{q}_{j+1}\|_2)(\|\widehat{s}\|_2 + |\widehat{\alpha}|)\varepsilon + \widehat{\alpha}\|\widehat{q}_{j+1}\|_2 + \mathcal{O}(\varepsilon^2) \\ &\le (3j+4)\|\widehat{Q}_j\|_F\|\widehat{s}\|_2\varepsilon + \mathcal{O}(\varepsilon^2),\end{aligned} \tag{4.22}$$

where for the last inequality we used $\|\widehat{q}_{j+1}\|_2 \le \|\widehat{Q}_j\|_F + \mathcal{O}(\varepsilon)$ and $\widehat{\alpha} \le j\|\widehat{s}\|_2\varepsilon$. Therefore, when the deflation occurs, (4.11) and the bound (4.12) are replaced by (4.21) and the bound of (4.22), respectively. The proof of Lemma 4.3 will be the same by setting $\widehat{\alpha} = 0$ in the case of the deflation.

Analogously, for the breakdown, since the left side of (4.13) is of the order of $\mathcal{O}(\varepsilon)$, we can move and absorb it into $g_j$ on the right. The updated $g_j$ is then bounded by

$$\|g_j\|_2 \le (3j+1)\|\widehat{U}_{j+1}\|_F\|\underline{\widehat{H}}_k(1:j,j)\|_2\varepsilon + \mathcal{O}(\varepsilon^2).$$

Using the bound above to replace (4.14), Lemma 4.3 is proven for $k = j$.

We now present the main theorem on an upper bound of the relative backward error of the TOAR procedure.

THEOREM 4.4. *Suppose that $\widehat{Q}_k$ and $\widehat{U}_k$ have full column rank and*

$$\kappa = \max\{\kappa_2(\widehat{Q}_k), \kappa_2(\widehat{U}_k)\},$$

*where $\kappa_2(X)$ denotes the 2-norm condition number of a matrix $X$. Then the relative backward error of the computed compact Arnoldi decomposition* (4.2) *satisfies*

$$\frac{\|\Delta L\|_F}{\|L\|_F} \le \varphi_1 \kappa^4 \varepsilon + \mathcal{O}(\varepsilon^2), \tag{4.23}$$

*where $\varphi_1 = 4k(2n+1)$. Here we assume $(k+1)(2k+1)\kappa^4\varepsilon < 1$.*

When $\widehat{Q}_k$ and $\widehat{U}_k$ are generated by the MGS process with the partial reorthogonalization, we have $\kappa = 1 + \mathcal{O}(\varepsilon)$. Consequently, $\varphi_1\kappa^4 \approx \varphi_1$. Thus the inequality (4.23) implies that the TOAR procedure is relatively backward stable.

*Proof of Thoerem* 4.4. By (4.2) and (4.7), we have

$$\|\Delta L\|_F = \|E\widehat{V}_{k-1}^{\dagger}\|_F \le \|E\|_F \|\widehat{V}_{k-1}^{\dagger}\|_2 \le (\|F_{\mathrm{mv}}\|_F + \|F\|_F)\|\widehat{V}_{k-1}^{\dagger}\|_2,$$

where the inequality $\|XY\|_F \le \|X\|_2\|Y\|_F$ is used for the first inequality.

For $\|\widehat{V}_{k-1}^{\dagger}\|_2$, by using the property $\sigma_{\min}(XY) \ge \sigma_{\min}(X)\sigma_{\min}(Y)$ for matrices $X$ and $Y$ with full column rank, we derive that

$$\begin{aligned} \|\widehat{V}_{k-1}^{\dagger}\|_2 &\equiv \frac{1}{\sigma_{\min}(\widehat{V}_{k-1})} = \frac{1}{\sigma_{\min}(\widehat{Q}_{[k-1]}\widehat{U}_{k-1})} \\ &\le \frac{1}{\sigma_{\min}(\widehat{Q}_{[k-1]})\sigma_{\min}(\widehat{U}_{k-1})} = \frac{1}{\sigma_{\min}(\widehat{Q}_{k-1})\sigma_{\min}(\widehat{U}_{k-1})}. \end{aligned} \tag{4.24}$$

Together with the upper bounds of $\|F_{\mathrm{mv}}\|_F$ (Lemma 4.2), and $\|X\|_2 \equiv \sigma_{\max}(X)$, we have

$$\begin{aligned} \|F_{\mathrm{mv}}\|_F\|\widehat{V}_{k-1}^{\dagger}\|_2 &\le \widetilde{\varphi}_1 \kappa_2(\widehat{Q}_{k-1})\kappa_2(\widehat{U}_{k-1})\|L\|_F\varepsilon + \mathcal{O}(\varepsilon^2) \\ &\le \widetilde{\varphi}_1 \kappa^2 \|L\|_F \varepsilon + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{4.25}$$

where $\widetilde{\varphi}_1 = 2k(2n+1)$. Note that for the second inequality we utilize the fact $\kappa_2(\widehat{Q}_{k-1}) \le \kappa$ and $\kappa_2(\widehat{U}_{k-1}) \le \kappa$ since $\widehat{Q}_{k-1}$ and $\widehat{U}_{k-1}$ are submatrices of $\widehat{Q}_k$ and $\widehat{U}_k$.

For the term $\|F\|_F\|\widehat{V}_{k-1}^{\dagger}\|_2$, by (4.2), $\underline{\widehat{H}}_k = \widehat{V}_k^{\dagger}(L + \Delta L)\widehat{V}_{k-1}$. Repeatedly applying the inequality $\|XY\|_F \le \|X\|_2\|Y\|_F$ leads to

$$\begin{aligned} \|\underline{\widehat{H}}_k\|_F &\le \|\widehat{V}_{k-1}\|_2\|\widehat{V}_k^{\dagger}\|_2(\|L\|_F + \|\Delta L\|_F) \\ &\le \frac{\|\widehat{Q}_{k-1}\|_2\|\widehat{U}_{k-1}\|_2}{\sigma_{\min}(Q_k)\sigma_{\min}(U_k)}(\|L\|_F + \|\Delta L\|_F), \end{aligned}$$

where for the second inequality we used $\|\widehat{V}_{k-1}\|_2 \equiv \|\widehat{Q}_{[k-1]}\widehat{U}_{k-1}\|_2 \le \|\widehat{Q}_{k-1}\|_2\|\widehat{U}_{k-1}\|_2$, and the upper bound (4.24) with $\widehat{V}_k^{\dagger}$. Therefore, together with the upper bounds of $\|F\|_F$ in Lemma 4.3, we have

$$\begin{aligned} \|F\|_F\|\widehat{V}_{k-1}^{\dagger}\|_2 &\le \varphi_2 \frac{\|\widehat{Q}_k\|_2\|\widehat{U}_k\|_2}{\sigma_{\min}(\widehat{Q}_{k-1})\sigma_{\min}(\widehat{U}_{k-1})} \frac{\|\widehat{Q}_{k-1}\|_2\|\widehat{U}_{k-1}\|_2}{\sigma_{\min}(\widehat{Q}_k)\sigma_{\min}(\widehat{U}_k)}(\|L\|_F + \|\Delta L\|_F)\varepsilon + \mathcal{O}(\varepsilon^2) \\ &= \varphi_2 \kappa_2(\widehat{Q}_k)\kappa_2(\widehat{U}_k)\kappa_2(\widehat{Q}_{k-1})\kappa_2(\widehat{U}_{k-1})(\|L\|_F + \|\Delta L\|_F)\varepsilon + \mathcal{O}(\varepsilon^2) \\ &\le \varphi_2 \kappa^4(\|L\|_F + \|\Delta L\|_F)\varepsilon + \mathcal{O}(\varepsilon^2), \end{aligned} \tag{4.26}$$

where $\varphi_2 = (k+1)(2k+1)$.

Combining (4.25) and (4.26), we have

$$\begin{aligned}\|\Delta L\|_F &\le (\widetilde{\varphi}_1 + \varphi_2\kappa^2)\kappa^2\|L\|_F\varepsilon + \varphi_2\kappa^4\|\Delta L\|_F\varepsilon + \mathcal{O}(\varepsilon^2)\\ &\le 2\widetilde{\varphi}_1\kappa^4\|L\|_F\varepsilon + \varphi_2\kappa^4\|\Delta L\|_F\varepsilon + \mathcal{O}(\varepsilon^2),\end{aligned}$$

where for the second inequality, we have used the fact that $\varphi_2 \le \widetilde{\varphi}_1$ and $1 \le \kappa^2$. Move the term involving $\|\Delta L\|_F$ to the left side to obtain

$$\frac{\|\Delta L\|_F}{\|L\|_F} \le \frac{2\widetilde{\varphi}_1\kappa^4}{1-\varphi_2\kappa^4\varepsilon}\varepsilon + \mathcal{O}(\varepsilon^2),$$

where we assume $\varphi_2\kappa^4\varepsilon < 1$. The theorem is proven by omitting the denominator as it can be covered by the $\mathcal{O}(\varepsilon^2)$ term. ∎

*Remark* 3. In this section, we have assumed that $A$ and $B$ are explicitly given matrices, so that we can apply the standard error bound of the matrix-vector multiplication as in Lemma 4.1. This is the same notion for showing the standard Arnoldi process is backward stable; see [31, Theorem 2.5]. When $A$ and $B$ are linear operators that are implicitly given, the matrix-vector multiplication error will rely on the specific formulation of $A$ and $B$. For example, when $A = K^{-1}D$ and $B = K^{-1}M$ with $M$, $D$, and $K$ being square matrices, the error of the matrix-vector multiplications $Av$ and $Bv$ will have an amplification factor of the condition number $\kappa_2(K)$ of the matrix $K$. The analysis will be the same. In this case, the backward error bound will correspondingly be increased by a factor in $\kappa_2(K)$. Therefore, to ensure small backward error, $K$ should not be severely ill-conditioned.

**5. Numerical Examples.** In this section, we apply the TOAR procedure to the application of the model order reduction of second-order dynamical systems, and illustrate its superior accuracy.

A continuous time-invariant single-input single-output second-order dynamical system in the frequency domain is described by

$$\Sigma_n : \begin{cases} s^2Mx(s) + sDx(s) + Kx(s) = bu(s),\\ \qquad\qquad\qquad\quad y(s) = c^{\mathrm{T}}x(s), \end{cases} \tag{5.1}$$

where $M$, $D$, and $K$ are $n \times n$ system matrices, $x(s)$ is the state vector, $u(s)$ is the external input function with $b \in \mathbb{R}^n$ being the input distribution vector, $y(s)$ is the output response with $c \in \mathbb{R}^n$ being the output distribution vector. $s = \jmath\omega$ with $\jmath = \sqrt{-1}$ and $\omega \ge 0$ being the frequency. By (5.1), the input and output functions $u(s)$ and $y(s)$ satisfy the relation $y(s) = h(s)u(s)$, where $h(s)$ is the transfer function

$$h(s) = c^{\mathrm{T}}(s^2M + sD + K)^{-1}b. \tag{5.2}$$

The second-order systems arise from the study of dynamics of physical systems, such as electrical, mechanical, and structural systems, electromagnetics, and microelectromechanical systems; see, for example, [10, 4, 9, 1, 27, 30, 36, 37].

One of the main objectives of the second-order Krylov subspace-based model order reduction techniques [32, 2, 29, 26] is to construct a reduced second-order system $\Sigma_k$ of the same form (5.1), such that the transfer function

$$h_k(s) = c_k^{\mathrm{T}}(s^2M_k + sD_k + K_k)^{-1}b_k$$

of $\Sigma_k$ is an accurate approximation of $h(s)$ over a wide range of frequency intervals around a prescribed shift $s_0$, where $M_k$, $D_k$, and $K_k$ are $\eta_k \times \eta_k$ system matrices, $b_k$

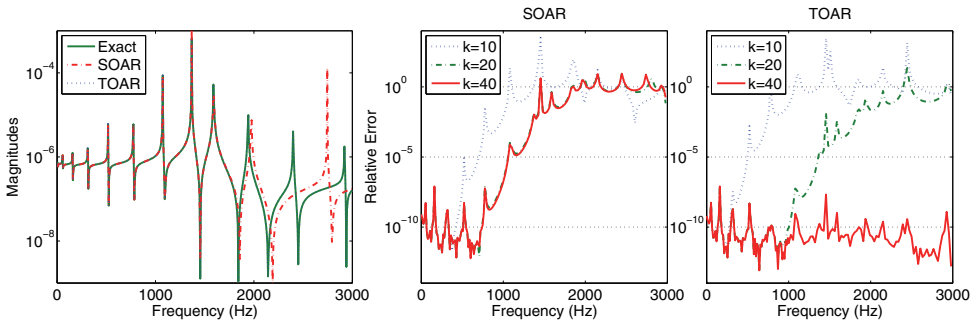


FIG. 1. *Magnitudes of transfer functions $h(s)$ and $h_k(s)$ with $k = 40$ (left). Relative errors $|h(s) - h_k(s)|/|h(s)|$ for $k = 10, 20, 40$ (middle and right).*

and $c_k$ are length-$\eta_k$ vectors, and $\eta_k \le k$. Toward this objective, one first rewrites the transfer function $h(s)$ to include the shift $s_0$:

$$h(s) = c^{\mathrm{T}} \left( (s - s_0)^2 M + (s - s_0)\widetilde{D} + \widetilde{K} \right)^{-1} b,$$

and then computes an orthonormal basis matrix $Q_k \in \mathbb{R}^{n \times \eta_k}$ of the second-order Krylov subspace

$$\mathcal{G}_k(-\widetilde{K}^{-1}\widetilde{D}, -\widetilde{K}^{-1}M; 0, r_0 = \widetilde{K}^{-1}b),$$

where $\widetilde{D} = 2s_0M + D$ and $\widetilde{K} = s_0^2M + s_0D + K$, and $\eta_k \le k$ due to $r_{-1} = 0$. Once $Q_k$ is computed, the system matrices and vectors of the reduced second-order system $\Sigma_k$ are given by $M_k = Q_k^{\mathrm{T}} M Q_k$, $D_k = Q_k^{\mathrm{T}} D Q_k$, $K_k = Q_k^{\mathrm{T}} K Q_k$, $c_k = Q_k^{\mathrm{T}} c$, and $b_k = Q_k^{\mathrm{T}} b$.

We now compare the accuracy of the reduced second-order systems $\Sigma_k$, where the orthonormal basis matrix $Q_k$ are computed by the SOAR procedure [3] and the TOAR procedure (Algorithm 1), respectively. The deflation and breakdown thresholds are set as in (4.19). In addition, we apply the partial reorthogonalization with $\theta = \sqrt{2}/2$ in both SOAR and TOAR to ensure that the computed basis matrix $\widehat{Q}_k$ is orthonormal with respect to the machine precision. All algorithms are implemented in MATLAB and were run on a machine with Intel(R) Core(TM) i7-3632QM CPU@ 2.20 GHz and 6 GB RAM memory.

*Example* 1. This is a second-order system of dimension $n = 400$ from a finite element model of a shaft on bearing supports with a damper in MSC/NASTRAN [16]. The system matrices $M$ and $D$ are symmetric, and $K$ is symmetric positive definite. To approximate the transfer function $h(s)$ over the frequency interval $[0, 3000]$, we select the expansion point $s_0 = 150 \times 2\pi$. The 1-norm condition number of the matrix $\widetilde{K} = s_0^2M + s_0D + K$ is $\mathcal{O}(10^7)$.

The left plot of Figure 1 shows the magnitudes of the transfer functions $h(s)$ of the full-order system $\Sigma_n$, and the transfer functions $h_k(s)$ of the reduced-order systems $\Sigma_k$ generated by SOAR and TOAR with $k = 40$. The relative errors $|h(s) - h_k(s)|/|h(s)|$ are shown in the middle (SOAR) and right (TOAR) plots of Figure 1 with $k = 10, 20, 40$. As we can see the $h_k(s)$ by TOAR is a more accurate approximation than the one by SOAR over the wide range of the interval. Furthermore, when $k$ is increased from 10 to 40, the approximation accuracy of $h_k(s)$ computed by the SOAR

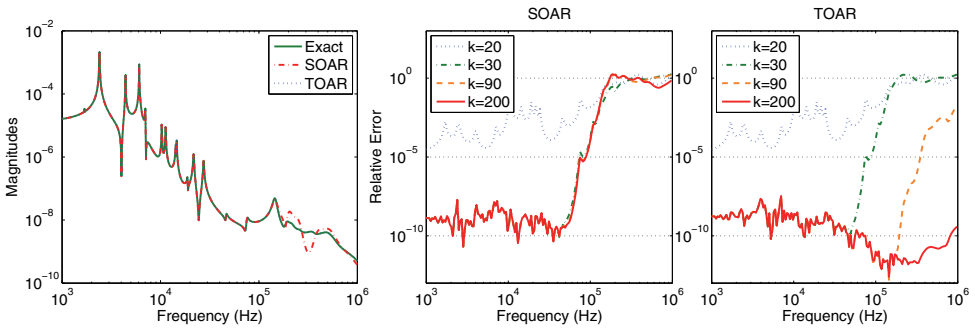


FIG. 2. *Magnitudes of transfer functions $h(s)$ and $h_k(s)$ with $k = 200$ (left). Relative errors $|h(s) - h_k(s)|/|h(s)|$ for $k = 20, 30, 90,$ and 200 (middle and right).*

has stagnated at $k = 20$. In contrast to SOAR, the approximation accuracy of $h_k(s)$ computed by TOAR continues to improve.

The reason for the stagnation of SOAR in accuracy after $k \geq 20$ is due to the fact that the condition numbers of triangular matrices in SOAR grow from 10 to $10^{28}$ and the procedure becomes numerically unstable. There is no such stagnation for the TOAR procedure. When $k = 40$, the condition numbers of the computed $\widehat{Q}_k$ and $\widehat{U}_k$ are $\kappa_2(\widehat{Q}_k) = 1 + \epsilon_Q$ and $\kappa_2(\widehat{U}_k) = 1 + \epsilon_U$, where both $\epsilon_Q$ and $\epsilon_U$ are at the order of machine precision, namely, $\epsilon_Q = 1.33 \times 10^{-15}$ and $\epsilon_U = 8.88 \times 10^{-16}$. The TOAR procedure is numerically stable.

*Example* 2. This is the butterfly gyroscope problem in the Oberwolfach collection [22]. The second-order system $\Sigma_n$ is of the order $n = 17361$ with the proportional Rayleigh damping matrix $D = \alpha M + \beta K$, where the mass and stiffness matrices $M$ and $K$ are symmetric. The second-order system has 1 input vector and 12 output vectors. In the experiment we use the first output vector as the output vector $c$. The same as in [21], we take the damping parameters $\alpha = 0$ and $\beta = 10^{-7}$ and use the expansion point $s_0 = 1.05 \times 10^5$. The condition number of $\widetilde{K}$ is about $\mathcal{O}(10^8)$.

The magnitudes of the transfer functions are shown in the left plot of Figure 2. The relative errors of the reduced transfer function $h_k(s)$ are shown in the middle and right plots of Figure 2. We can clearly observe the advantage of TOAR over SOAR in the frequency range of $10^5 - 10^6$Hz. For $k = 20$ and 30, the reduced transfer functions $h_k(s)$ by SOAR and TOAR are of about the same accuracy. But when $k$ increases to 90 and 200, the accuracy of SOAR stagnates, while the accuracy of TOAR is improved. The TOAR procedure is numerically stable. When $k = 200$, the condition numbers of the computed $\widehat{Q}_k$ and $\widehat{U}_k$ are $\kappa_2(\widehat{Q}_k) = 1+\epsilon_Q$ and $\kappa_2(\widehat{U}_k) = 1+\epsilon_U$, where $\epsilon_Q = 3.11\times10^{-15}$ and $\epsilon_U = 4.66\times10^{-16}$. Furthermore, since $\kappa_2(\widehat{V}_k) \leq \kappa_2(\widehat{Q}_{[k]})\kappa_2(\widehat{U}_k)$, the condition number of $\widehat{V}_k$ is also close to 1.0. Similar to Example 1, the reason for the accuracy stagnation of SOAR is due to the fact that the condition numbers of triangular matrices in SOAR quickly grow to $\mathcal{O}(10^{31})$.

**6. Concluding Remarks.** In this paper we rederived a TOAR procedure for computing an orthonormal basis of the second-order Krylov subspace and presented a rigorous numerical stability analysis of the TOAR procedure. TOAR solves the potential instability encountered by the existing SOAR procedure. We proved TOAR is backward stable in computing an orthonormal basis matrix $V_k$ of the associated Krylov subspace $\mathcal{K}_k(L, v_0)$. Numerical examples show that the basis matrices com-

puted by TOAR are much more accurate than the ones generated by SOAR in the application of dimension reduction of second-order dynamical systems. The stability analysis presented in the paper can be generalized to the two-level orthogonalization procedures for higher-order Krylov subspaces [34, 39, 17, 35].

We should stress that the backward error analysis presented in this paper is for the associated linear Krylov subspace. It is shown that the computed orthonormal basis $\widehat{Q}_k$ of the second-order Krylov subspace satisfies the embedding property $\mathcal{K}_k(L + \Delta L, v_0) \subset \text{span}\{\widehat{Q}_{[k]}\}$. It is still an open problem of whether one is able to provide backward error analysis in terms of the original matrix pair $(A, B)$ of the second-order Krylov subspace $\mathcal{G}_k(A, B; r_{-1}, r_0)$ and show that the computed $\widehat{Q}_k$ is an exact basis matrix of $\mathcal{G}_k(A + \Delta A, B + \Delta B; r_{-1}, r_0)$ with small $\Delta A$ and $\Delta B$.

**Acknowledgment.** The authors would like to express their gratitude to the referees for their numerous valuable comments and suggestions to improve the presentation of the paper.

## REFERENCES

[1] Z. Bai, D. Bindel, J. Clark, J. Demmel, K. S. J. Pister, and N. Zhou, *New numerical techniques and tools in SUGAR for* 3*D MEMS simulation*, in Technical Proceedings of the Fourth International Conference on Modeling and Simulation of Microsystems, Hilton Head Island, USA, Computational Publications, Cambridge, MA, 2001, pp. 31–34.

[2] Z. Bai and Y. Su, *Dimension reduction of large-scale second-order dynamical systems via a second-order Arnoldi method*, SIAM J. Sci. Comput., 26 (2005), pp. 1692–1709.

[3] Z. Bai and Y. Su, SOAR: *A second-order Arnoldi method for the solution of the quadratic eigenvalue problem*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 640–659.

[4] M. J. Balas, *Trends in large space structure control theory: Fondest hopes, wildest dreams*, IEEE Trans. Automat. Control, 27 (1982), pp. 522–535.

[5] L. Bao, Y. Lin, and Y. Wei, *Restarted generalized Krylov subspace methods for solving large-scale polynomial eigenvalue problems*, Numer. Algorithms, 50 (2009), pp. 17–32.

[6] D. S. Bindel, *Structured and Parameter-Dependent Eigensolvers for Simulation-Based Design of Resonant MEMS*, PhD thesis, EECS Department, University of California, Berkeley, CA, 2006.

[7] A. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT, 7 (1967), pp. 1–21.

[8] A. Björck and C. C. Paige, *Loss and recapture of orthogonality in the modified Gram–Schmidt algorithm*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 176–190.

[9] J. V. Clark, N. Zhou, D. Bindel, L. Schenato, W. Wu, J. Demmel, and K. S. J. Pister, 3*D MEMS simulation modeling using modified nodal analysis*, in Proceedings of the Microscale Systems: Mechanics and Measurements Symposium, SEM, Bethel, CT, 2000, pp. 68–75.

[10] R. R. Craig, *Structural Dynamics: An Introduction to Computer Methods*, John Wiley, New York, 1981.

[11] J. W. Daniel, W. B. Gragg, L. Kaufman, and G. W. Stewart, *Reorthogonalization and stable algorithms for updating the Gram-Schmidt factorization*, Math. Comput., 30 (1976), pp. 772–795.

[12] L. Giraud and J. Langou, *When modified Gram–Schmidt generates a well-conditioned set of vectors*, IMA J. Numer. Anal., 22 (2002), pp. 521–528.

[13] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, MD, 1996.

[14] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[15] Z. Jia and Y. Sun, *Implicitly restarted generalized second-order Arnoldi type algorithms for the quadratic eigenvalue problem*, Taiwanese J. Math., 19 (2015), pp. 1–30. Doi: 10.11650/tjm.18.2014.4577.

[16] T. R. Kowalski, *Extracting a Few Eigenpairs of Symmetric Indefinate Matrix Pencils*, PhD thesis, Department of Mathematics, University of Kentucky, Lexington, KY, 2000.

[17] D. Kressner and J. E. Roman, *Memory-efficient Arnoldi algorithms for linearizations of matrix polynomials in Chebyshev basis*, Numer. Linear Algebr., 21 (2014), pp. 569–588.

[18] L.-Q. Lee, L. Ge, Z. Li, C. Ng, G. Schussman, and K. Ko, *Achievements in ISICs/SAPP collaborations for electromagnetic modeling of accelerators*, J. Phys. Conf. Ser., 16 (2005), pp. 205–209.
[19] L.-Q. Lee, Z. Li, C. Ng, and K. Ko, *Omega3P: A parallel finite-element eigenmode analysis code for accelerator cavities*, Technical report SLAC-PUB-13529, Stanford Linear Accelerator Center, Menlo Park, CA, 2009.
[20] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, Software Eniviron. Tools 6, SIAM, Philadelphia, 1998.
[21] Y.-T. Li, Z. Bai, W.-W. Lin, and Y. Su, *A structured quasi-Arnoldi procedure for model order reduction of second-order systems*, Linear Algebra Appl., 436 (2012), pp. 2780–2794.
[22] *Oberwolfach Model Reduction Benchmark Collection*, `http://simulation.uni-freiburg.de/downloads/benchmark` (29 May 2014).
[23] C. Otto, *Arnoldi and Jacobi-Davidson Methods for Quadratic Eigenvalue Problems*, Master's thesis, Institut für Mathematik, Technische Universität Berlin, Berlin, 2004.
[24] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood cliffs, NJ, 1980.
[25] R. S. Puri, *Krylov Subspace Based Direct Projection Techniques for Low Frequency, Fully Coupled, Structural Acoustic Analysis and Optimization*, PhD thesis, Oxford Brookes University, Oxford, 2008.
[26] R. S. Puri and D. Morrey, *A comparison of one-and two-sided Krylov–Arnoldi projection methods for fully coupled, damped structural-acoustic analysis*, J. Comput. Acoust., 21 (2013), 1350004.
[27] D. Ramaswamy and J. White, *Automatic generation of small-signal dynamic macromodels from 3-D simulation*, in Technical Proceedings of the Fourth International Conference on Modeling and Simulation of Microsystems, Computational Publications, Cambridge, MA, 2000, pp. 27–30.
[28] E. B. Rudnyi, *MOR for ANSYS*, in System-Level Modeling of MEMS, Wiley-VCH Verlag, Weinheim,Germany, 2013, pp. 425–438.
[29] B. Salimbahrami and B. Lohmann, *Order reduction of large scale second-order systems using Krylov subspace methods*, Linear Algebra Appl., 415 (2006), pp. 385–405.
[30] R. D. Slone, *Fast Frequency Sweep Model Order Reduction of Polynomial Matrix Equations Resulting from Finite Element Discretizations*, PhD thesis, The Ohio State University, Columbus, OH, 2002.
[31] G. W. Stewart, *Matrix Algorithms, Volumn* II*: Eigensystems*, SIAM, Philadelphia, 2001.
[32] T.-J. Su and R. R. Craig, *Model reduction and control of flexible structures using Krylov vectors*, J. Guidance Control Dynam., 14 (1991), pp. 260–267.
[33] Y. Su, J. Wang, X. Zeng, Z. Bai, C. Chiang, and D. Zhou, *SAPOR: second-order Arnoldi method for passive order reduction of RCS circuits*, in Proceedings of the 2004 IEEE/ACM International Conference on Computer-aided Design, IEEE, Piscataway, NJ, 2004, pp. 74–79.
[34] Y. Su, J. Zhang, and Z. Bai, *A compact Arnoldi algorithm for Polynomial Eigenvalue Problems*, `http://math.cts.nthu.edu.tw/Mathematics/RANMEP%20Slides/Yangfeng%20Su.pdf` (January 2008).
[35] R. Van Beeumen, K. Meerbergen, and W. Michiels, *Compact rational Krylov methods for nonlinear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 820–838.
[36] T. Wittig, I. Munteanu, R. Schuhmann, and T. Weiland, *Two-step Lanczos algorithm for model order reduction*, IEEE Trans. Magn., 38 (2002), pp. 673–676.
[37] H. Wu and A. C. Cangellaris, *Krylov model order reduction of finite element approximations of electromagnetic devices with frequency-dependent material properties*, Int. J. Numer. Model. Electron. Netw. Devices Fields, 20 (2007), pp. 217–235.
[38] C. Yang, *Solving large-scale eigenvalue problems in SciDAC applications*, J. Phys. Conf. Ser., 16 (2005), pp. 425–434.
[39] Y. Zhang and Y. Su, *A memory-efficient model order reduction for time-delay systems*, BIT, 53 (2013), pp. 1047–1073.
[40] J. Zhu, *Structured Eigenvalue Problems and Quadratic Eigenvalue Problems*, PhD thesis, Department of Mathematics, University of California, Berkeley, CA, 2005.