On IOMs(q): Incomplete Orthogonalization Methods for Large Unsymmetric Linear Systems*

Zhongxiao Jia†

Abstract

The truncated version, Incomplete Orthogonalization Methods (IOMs(q)), of Full Orthogonalization Methods (FOM), proposed by Saad has been used a long time for solving large unsymmetric linear systems. However, no convergence analysis has been made. In this paper, we analyze in detail IOMs(q) from a theoretical point of view, and establish a convergence theory of the methods. We derive a number of important results of how the unsymmetry of the matrix A to affect the basis vectors generated by IOMs(q), and establish the relationships between the residuals for IOMs(q) and the full orthogonalization method (FOM). The results show that IOMs(q) behave much like FOM once the basis vectors generated by IOMs(q) are well conditioned. However, it is shown that IOMs(q) may generate an ill-conditioned basis for any unsymmetric matrix and the parameter q such that IOMs(q) may fail to converge or at least cannot behave like FOM. Due to the mathematical equivalence between IOMs(q) and the truncated ORTHORES(q) developed by Young and Jea, essentially we give insights into the convergence of ORTHORES(q). As for practical considerations, we suggest one strategy of choosing the parameter q. Numerical experiments are reported to show convergence behaviors of IOMs(q).

Key words. unsymmetric, Krylov subspace, IOMs(q), FOM, Arnoldi, truncated, basis vector, convergence, restarted

AMS(MOS) subject classifications. 65F10

*This work was supported by the Graduiertenkolleg Mathematik at the University of Bielefeld. The results in this paper have been announced on the 15th Biennial Conference on Numerical Analysis, 29 June–2 July, 1993, Dundee and the Lanczos International Centenary Conference, Dec. 12–17, 1993, Raleigh

†Fakultät für Mathematik, Universität Bielefeld, email: jia@mathematik.uni-bielefeld.de, Postfach 10 03 01, 33501 Bielefeld, Germany

1 Introduction

Consider the solution of linear system

$$Ax = b, \tag{1}$$

where $A \in C^{N\times N}$ is nonsingular, large and unsymmetric (non-Hermitian), and b an $N \times 1$ column vector. A major class of methods for solving (1) is Krylov subspace or conjugate gradient type methods. So far, much work has been done; see, e.g. [12, 2, 3, 4, 42, 31, 32, 33, 35, 8, 9, 13, 19, 20, 29, 24, 25, 27] and many others. For survey papers, refer to [17, 36, 18, 14]; for a numerically comparative study of some of them, see [37, 38]; for a synthesis and general description of available conjugate gradient type methods, see [34].

Of these methods, one kind of the most successful schemes is based on the orthogonal projection, typified by GCR or ORTHOMIN [39, 42, 9], ORTHODIR and ORTHORES [42], FOM or Arnoldi's method [31] and GMRES [35]. There are rich references concerning these methods, as mentioned previously. Here we want to stress that there is an intimate connection between FOM and GMRES. Brown [7] has shown that a singular upper Hessenberg matrix H_m in FOM corresponds to the "stagnation" of approximate solutions in GMRES, while the occurrence of the "stagnation" of approximate solutions in GMRES is due to a singular upper Hessenberg matrix H_m. If successful, GMRES performs better than FOM since the residual norms of the former are characterized by a minimization property over Krylov subspace generated by A. In contrast to those algorithms based on the Lanczos like process [26, 12, 13], GCR, ORTHODIR, ORTHORES, FOM and GMRES use long recurrences, which means that work and storage per step grow as the number of steps increases. As a consequence, one has to restart these algorithms in practice, which may result in very slow convergence for some problems. Except for those matrices of very special forms, however, several authors have shown that long recurrences cannot be simplified [10, 11, 23].

Even so, noticing the disadvantage of long recurrences, some researchers have proposed truncated strategies, which use only a few rather than all previously computed vectors in recurrences and can thus be significantly less expensive. For example, Vinsome [39] suggested the truncated version ORTHOMIN(q) of ORTHOMIN or GCR, Young et al. [42] presented the corresponding truncated versions ORTHODIR(q) and ORTHORES(q) of ORTHODIR and ORTHORES, Saad [31] proposed the truncated version IOMs(q) of FOM, and Brown et al. [5, 6] suggested the truncated version IGMRES(q) of GMRES for solving (1). At about the same time, Saad applied IOMs(q) for solving unsymmetric eigenvalue problems [30]. Later, Saad [32, 33, 35], Brown et al. [5, 6] and Gear et al. [15] discussed and used IOMs(q), and some others described or mentioned IOMs(q), e.g. [8, 18, 27]. Besides, there are some relationships between various truncated versions, for example, Saad [33] has shown that IOMs(q) are equivalent to ORTHORES(q) in the mathematical sense.

A convergence analysis of the truncated ORTHOMIN(q) has been made by [8]. For all the other truncated versions mentioned above, however, no convergence theory has been established and how to make convergence analysis is still an open problem.

We devote this paper to the study of IOMs(q), and expose its important features. Due to the mathematical equivalence between IOMs(q) and ORTHORES(q), we essentially give insights into the convergence of ORTHORES(q) in the mathematical sense. We will present two main results: The first is to derive a number of important relationships of how the unsymmetry of A to affect the basis vectors of Krylov subspace generated by IOMs(q), which are crucial in understanding how IOMs(q) converge; the second is to establish the relationships between the residuals for IOMs(q) and FOM. By virtue of these results and the convergence analysis of FOM [31, 21], we prove that IOMs(q) can behave much like FOM once the basis vectors generated by IOMs(q) are well conditioned. However, it is shown that IOMs(q) may generate a very ill-conditioned basis for any unsymmetric matrix such that they may fail to converge or at least cannot behave like FOM.

In Section 2 we briefly review FOM and IOMs(q); in Section 3 we establish a convergence theory of IOMs(q), and prove our main results; as for practical considerations, we suggest in Section 4 one strategy of choosing the parameter q; in Section 5, we report numerical experiments to show convergence behaviors of IOMs(q) and their restarted versions, followed by concluding remarks in Section 6.

2 FOM and IOMs(q)

Let x^* be the solution of (1), $x^{(0)}$ an initial guess to x^* and $r^{(0)} = b - Ax^{(0)}$ the initial residual, and let the matrix $V_m = (v_1, v_2, \ldots, v_m)$ whose columns are linearly independent and the Krylov subspace $\mathcal{K}_m(r^{(0)}, A) = span\{r^{(0)}, Ar^{(0)}, \ldots, A^{m-1}r^{(0)}\} = span\{V_m\}$. Throughout the paper, the norm used is the 2-norm unless stated otherwise. The full orthogonalization method (FOM) seeks an approximate solution $x^{(m)}$ of the form $x^{(m)} = x^{(0)} + z^{(m)}$ such that $z^{(m)}$ satisfies the following orthogonal projection:

$$\left\{ \begin{array}{l} z^{(m)} \in \mathcal{K}_m(r^{(0)}, A), \\ r^{(m)} \perp \mathcal{K}_m(r^{(0)}, A), \end{array} \right. \tag{2}$$

where $r^{(m)} = b - Ax^{(m)} = r^{(0)} - Az^{(m)}$.

Now let π_m be the orthogonal projector on $\mathcal{K}_m(r^{(0)}, A)$. Then it is easy to see that the following relations are equivalent:

$$\hat{H}_m y^{(m)} = B_m^{-1} V_m^H r^{(0)}, \tag{3}$$

$$\pi_m(r^{(0)} - Az^{(m)}) = 0, \tag{4}$$

where $z^{(m)} = V_m y^{(m)}$, the superscript "H" denotes the conjugate transpose, and $\hat{H}_m = B_m^{-1} C_m$, where $B_m = V_m^H V_m$ and $C_m = V_m^H A V_m$. Here $\hat{H}_m$ is called the matrix representation of the restriction of A to $\mathcal{K}_m(r^{(0)}, A)$ in the basis $\{v_i\}_1^m$.

It has been shown in [31, 21] that the sequence $\{x^{(m)}\}$ usually converges to x^* rapidly as m increases, provided that the eigenvalue distribution of A is favorable, A not highly defective and the eigenvector or Jordan basis of A not ill-conditioned.

In implementations, FOM usually exploits Arnoldi's process [1, 30] to generate an orthonormal basis $\{v_i\}_1^m$ of $\mathcal{K}_m(r^{(0)}, A)$ and restarts per m steps. In this case, $B_m = I_m, V_m^H r^{(0)} = \beta e_1$, and $\hat{H}_m = H_m$ is an $m \times m$ upper Hessenberg matrix, where I_m is the $m \times m$ identity matrix, $\beta = \|r^{(0)}\|$ and e_1 is the first standard unit vector of dimension m. Refer to [31] for details.

Algorithm 1. The restarted FOM

1. Start: Choose the number of steps m, a tolerance tol and an initial guess $x^{(0)}$. Then compute $r^{(0)} = b - Ax^{(0)}$, and take $v_1 := r^{(0)}/\|r^{(0)}\|$.

2. For $i = 1, 2, \ldots, m$ do Arnoldi's process

$$w := Av_i - \sum_{j=1}^{i} h_{ji} v_j \qquad \text{with } h_{ji} := v_j^H A v_i, \tag{5}$$

$$h_{i+1i} := \|w\|, \qquad v_{i+1} := w/h_{i+1i}. \tag{6}$$

3. Solve the $m \times m$ linear system

$$H_m y^{(m)} = \beta e_1, \tag{7}$$

   where H_m is an $m \times m$ upper Hessenberg matrix generated by Step 2.

4. Form $z^{(m)} = V_m y^{(m)}$ and $x^{(m)} = x^{(0)} + z^{(m)}$.

5. Compute the residual norm $\|r^{(m)}\|$ based on

$$\|r^{(m)}\| = \|r^{(0)} - Az^{(m)}\| = h_{m+1m} \mid e_m^H y^{(m)} \mid, \tag{8}$$

   where e_m is the mth standard unit vector of dimension m. If $\|r^{(m)}\| \leq tol$, then stop; otherwise, set $x^{(0)} := x^{(m)}$, and return to Step 1.

For practical considerations of Algorithm 1, refer to [31, 33, 40].

IOMs(q) proposed by Saad [31] are the algorithms which are but Step 2 the same as Algorithm 1. In Step 2 of IOMs(q), Av_i is only orthogonalized against the previous q vectors $v_{i-q+1}, \ldots, v_i$. IOMs(q) have two versions, one of which is IOMs(q) without correction and another is IOMs(q) with correction. The latter essentially belongs to FOM in the sense of (2) because it computes the solution $z^{(m)}$ of (2) exactly, so its convergence theory has been established in [31, 21]. In the context of this paper, we are only concerned with the version without correction, which realizes an oblique projection process over $\mathcal{K}_m(r^{(0)}, A)$ [32] and is more attractive, valuable and interesting both in theory and practice.

Algorithm 2. The restarted IOMs(q) without correction

1. Start: Choose the number of steps m, the parameter q satisfying $2 \leq q \leq m$, a tolerance tol and an initial guess $x^{(0)}$. Then compute $r^{(0)} = b - Ax^{(0)}$, and set $v_1 := r^{(0)}/\|r^{(0)}\|$.

2. For $i = 1, 2, \ldots, m$ do the incomplete orthogonalization process

$$\begin{aligned} i_0 &:= \max\{1, i-q+1\}, \\ w &:= Av_i - \sum_{j=i_0}^{i} h_{ji} v_j \text{ with } h_{ji} := v_j^H A v_i, \quad (9) \\ h_{i+1i} &:= \|w\|, \ v_{i+1} := w/h_{i+1i}. \quad (10) \end{aligned}$$

3. Solve the $m \times m$ linear system

$$\tilde{H}_m \tilde{y}^{(m)} = \beta e_1, \tag{11}$$

where $\tilde{H}_m$ is an $m \times m$ banded upper Hessenberg matrix whose nonzero entries are h_{ji} generated by Step 2, and $\beta = \|r^{(0)}\|$.

4. Form $\tilde{z}^{(m)} = V_m \tilde{y}^{(m)}$ and $\tilde{x}^{(m)} = x^{(0)} + \tilde{z}^{(m)}$.

5. Compute the residual norm $\|\tilde{r}^{(m)}\|$ based on

$$\|\tilde{r}^{(m)}\| = \|r^{(0)} - A\tilde{z}^{(m)}\| = h_{m+1m} \mid e_m^H \tilde{y}^{(m)} \mid . \tag{12}$$

If $\|\tilde{r}^{(m)}\| \leq tol$, then stop; otherwise, set $x^{(0)} := \tilde{x}^{(m)}$, and return to Step 1.

Two notes are made in order: First, we only need to save $v_{i-q+1}, \ldots, v_i$ in primary memory at the ith step of Step 2, and recall all the $\{v_i\}_1^m$ to compute $\tilde{x}^{(m)}$ until the convergence of Algorithm 2 occurs; second, when $q = m$, Algorithm 2 is identical with Algorithm 1.

Step 2 of Algorithm 2 can be written in matrix form

$$AV_m = V_m \tilde{H}_m + h_{m+1m} v_{m+1} e_m^H, \tag{13}$$

which is very useful later.

3 Basic properties and convergence analysis

Saad [31, 33] has stated the following results.

Theorem 1 *1. Assume that Step 2 of Algorithm 2 does not break down. Then the sequence* $\{v_i\}_1^m$ *obtained by Algorithm 2 is a basis of* $\mathcal{K}_m(r^{(0)}, A)$ *and satisfies the incomplete orthogonality*

$$v_i^H v_j = \delta_{ij}, \quad \mid i-j \mid \leq q, \tag{14}$$

where δ_{ij} is the Kronecker delta.

2. Let $B_m = V_m^H V_m$ and $s_m = h_{m+1m} B_m^{-1} V_m^H v_{m+1} = h_{m+1m} V_m^+ v_{m+1}$. Then

$$\hat{H}_m = \tilde{H}_m + s_m e_m^H, \tag{15}$$

where V_m^+ is the pseudoinverse of V_m.

Three notes are made on the properties of V_m, B_m and s_m: First, B_m is an identity matrix except for the right upper and left lower corners containing nonzero entries [30]. Second, V_m is orthonormal, B_m is an identity matrix and s_m is a zero vector if the matrix A is symmetric and $q \geq 2$. In this case, Algorithm 2 reduces to the Lanczos algorithm [28] and $\tilde{H}_m$ is symmetric tridiagonal. Third, obviously, V_m is orthonormal, B_m is an identity matrix, and s_m is a zero vector for an unsymmetric A if $q = m$. Therefore, based on a continuity argument, there are some claims in the literature: V_m is nearly orthonormal, B_m nearly an identity matrix and s_m nearly a zero vector if A is nearly symmetric and $q \geq 2$. As will be shown, these claims, though correct, are too rough and thus provide no useful information both in theory and practice. We are required to investigate them in detail in theory and give much more insights into them in order to expose how IOMs(q) converge.

Saad [33] conjectured in 1984 that IOMs(q) should converge in general provided that q is sufficiently large, namely, q is close to $m-1$. So far, however, no convergence results have been established for IOMs(q). Of course, mathematically speaking, in the finite N dimensional space and exact arithmetic, IOMs(q) must converge at some step $m \leq N$, which is trivial. Thus, our aim consists in deriving some properties, which show that in what restrictions the residual norm for IOMs(q) is very near to that for FOM, or in other words in what restrictions IOMs(q) behave much like FOM.

To this end, we have to carry out two tasks: The first is to expose the properties of V_m and B_m; the second is to analyze how V_m and B_m affect the convergence of IOMs(q). First of all, we need a lemma.

Lemma 1 *If $j+2 \leq i \leq m+1$, then*

$$v_i^H A v_j = \sum_{k=j-q+1}^{\min\{i-q-1,j+1\}} h_{kj} v_i^H v_k. \tag{16}$$

Proof. According to Part 1 of Theorem 1 and the structure of $\tilde{H}_m$, we can prove through careful manipulations that

$$\begin{aligned} v_i^H A v_j &= \sum_{k=1}^{m+1} h_{kj} v_i^H v_k \\ &= \sum_{k=1}^{i-q-1} h_{kj} v_i^H v_k + h_{ij} + \sum_{k=i+q+1}^{m+1} h_{kj} v_i^H v_k \end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=1}^{\min\{i-q-1,j+1\}} h_{kj} v_i^H v_k \text{ due to } h_{kj} = 0 \text{ for } k > j+1 \\
&= \sum_{k=j-q+1}^{\min\{i-q-1,j+1\}} h_{kj} v_i^H v_k \text{ due to } h_{kj} \neq 0 \text{ for } j-k \leq q-1.
\end{aligned}
$$

Here if $k \leq 0$, define $v_i^H v_k = 0$ and $h_{kj} = 0$. $\square$

Having this lemma, we can now prove the first main result in this paper.

Theorem 2 *Assume $q \geq 2$. Then for all $i - j \geq q + 1$, if Step 2 of Algorithm 2 does not break down, the following relation holds:*

$$
| v_j^H v_i | \leq c_i \|A - A^H\| h_{ii-1}^{-1}, \tag{17}
$$

where c_i is defined by

$$
c_i = 1 + \sum_{k=1}^{i-q-2} f_{k,i}(\tilde{H}_m) (\prod_{l=i-k}^{i-1} h_{ll-1})^{-1}, \tag{18}
$$

and $f_{k,i}(\tilde{H}_m), k = 1, 2, \ldots, i-q-2$ are combinations of the products of k entries of $\tilde{H}_m$. If $k \leq 0$, define $f_{k,i}(\tilde{H}_m) = 0$.

Proof. Based on Algorithm 2 and Lemma 1, we have for all $i + 1 - j \geq q + 1$

$$
\begin{aligned}
v_j^H v_{i+1} h_{i+1i} &= v_j^H (A v_i - \sum_{k=i_0}^{i} h_{ki} v_k) \\
&= v_j^H (A - A^H) v_i + \sum_{k=j-q+1}^{\min\{i-q-1,j+1\}} \bar{h}_{kj} v_k^H v_i - \sum_{k=i_0}^{i} h_{ki} v_j^H v_k,
\end{aligned} \tag{19}
$$

where $i_0 = \max\{1, i - q + 1\}$, and the bar "–" denotes the complex conjugate of a scalar.

It thus follows from (19) and Lemma 1 that the $(i+1)$th column in the right upper corner of B_{m+1} depends on only its first i columns. Hence we can use induction on columns in the right upper corner of B_{m+1} to prove (17).

First let us look at the first nonzero column in this corner, which is the $(q+2)$th column whose only nonzero entry is $v_1^H v_{q+2}$.

According to Algorithm 2 and Theorem 1, for $i = q + 2 \leq m + 1$ and $j = 1$, we have from (19) and Lemma 1

$$
v_1^H v_{q+2} h_{q+2q+1} = v_1^H (A - A^H) v_{q+1}.
$$

Since Step 2 of Algorithm 2 does not break down, we get

$$
\begin{aligned}
| v_1^H v_{q+2} | &= | v_1^H (A - A^H) v_{q+1} | \, h_{q+2q+1}^{-1} \\
&\leq \|v_1\| \|A - A^H\| \|v_{q+1}\| h_{q+2q+1}^{-1} \\
&= \|A - A^H\| h_{q+2q+1}^{-1},
\end{aligned}
$$

which shows that (17) holds, where $c_{q+2} = 1$.

Suppose that (17) holds for the first i columns in the right upper corner of B_{m+1}. We then estimate its $(i+1)$th column for all $1 \leq j \leq i-q$ satisfying $i - j \geq q$.

For the second term of (19), we obtain by induction

$$\begin{aligned} \left| \sum_{k=j-q+1}^{\min\{i-q-1,j+1\}} \bar{h}_{kj} v_k^H v_i \right| &\leq \sum_{k=j-q+1}^{\min\{i-q-1,j+1\}} | \bar{h}_{kj} v_k^H v_i | \\ &\leq c_i \sum_{k=j-q+1}^{\min\{i-q-1,j+1\}} | h_{kj} | h_{ii-1}^{-1} \|A - A^H\|. \end{aligned}$$

For the third term of (19), clearly $1 \leq k - j \leq i - j$. We thus have by Lemma 1 and induction

$$\begin{aligned} \left| \sum_{k=i_0}^{i} h_{ki} v_j^H v_k \right| &\leq \sum_{k=i_0}^{i} | h_{ki} v_j^H v_k | = \sum_{k=j+q+1}^{i} | h_{ki} v_j^H v_k | \\ &\leq \sum_{k=j+q+1}^{i} c_k | h_{ki} | h_{kk-1}^{-1} \|A - A^H\|. \end{aligned}$$

Combining the above relations yields

$$\begin{aligned} | v_j^H v_{i+1} | &\leq \Big(\|v_j\| \|A - A^H\| \|v_i\| + \Big| \sum_{k=j-q+1}^{\min\{i-q-1,j+1\}} \bar{h}_{kj} v_k^H v_i \Big| \\ &\quad + \Big| \sum_{k=i_0}^{i} h_{ki} v_j^H v_k \Big| \Big) h_{i+1i}^{-1} \\ &\leq \Big[1 + \Big(c_i \sum_{k=j-q+1}^{\min\{i-q-1,j+1\}} | h_{kj} | \Big) h_{ii-1}^{-1} \\ &\quad + \Big(\sum_{k=j+q+1}^{i} c_k | h_{ki} | h_{kk-1}^{-1} \Big) \Big] \|A - A^H\| h_{i+1i}^{-1} \\ &= \gamma_{i+1,j} \|A - A^H\| h_{i+1i}^{-1}. \end{aligned}$$

Define $c_{i+1} = \max_{1 \leq j \leq i-q} \gamma_{i+1,j}$. Then by induction and simple combinations, it is easy to verify that c_{i+1} satisfies (18). $\Box$

Theorem 3 *All singular values $\sigma(B_m)$ of B_m satisfy the relation*

$$\max\{0, 1 - (m-q-1)c\|A - A^H\|\} \leq \sigma(B_m) \leq 1 + (m-q-1)c\|A - A^H\|, \tag{20}$$

where c is a function depending on $\tilde{H}_m$.

Proof. Noticing that B_m is Hermitian positive definite, its eigenvalues are identical with its singular values. According to the Gerschgorin theorem [41, p. 71–72] and Part 1 of Theorem 1, for any $\sigma(B_m)$, there exists some i such that

$$|\sigma(B_m)-1| \le \sum_{j\neq i} |v_i^H v_j| = \sum_{|i-j|\ge q+1} |v_i^H v_j|\,.$$

Hence

$$1-\sum_{|i-j|\ge q+1} |v_i^H v_j| \le \sigma(B_m) \le 1+\sum_{|i-j|\ge q+1} |v_i^H v_j|\,.$$

Let us now analyze the last relation. Clearly, we have

$$\sum_{|i-j|\ge q+1} |v_i^H v_j| = \sum_{j=1}^{i-q-1} |v_i^H v_j| + \sum_{j=i+q+1}^{m} |v_i^H v_j|\,. \tag{21}$$

If $i \le q+1$, the first term of (21) is zero. In terms of (17), we thus have

$$\begin{aligned}\sum_{|i-j|\ge q+1} |v_i^H v_j| &\le (m-i-q)\max_{i+q+1\le j\le m} |v_i^H v_j| \\ &\le (m-q-1)\max_{q+2\le j\le m} |v_i^H v_j| \\ &\le (m-q-1)c\|A-A^H\|,\end{aligned}$$

where $c=\max_{q+2\le j\le m} c_j h_{jj-1}^{-1}$, which shows together the positive definiteness of B_m that (20) holds.

If $i \ge m-q$, the second term of (21) is zero. Also, note that $|v_i^H v_j| = |v_j^H v_i|$. As a result, we get by (17)

$$\begin{aligned}\sum_{|i-j|\ge q+1} |v_i^H v_j| &\le (i-q-1)\max_{1\le j\le i-q-1} |v_i^H v_j| \\ &\le (m-q-1)\max_{1\le j\le m-q-1} |v_j^H v_i| \\ &\le (m-q-1)c\|A-A^H\|,\end{aligned}$$

where $c=\max_{m-q\le i\le m} c_i h_{ii-1}^{-1}$. Noticing that B_m is positive definite, (20) holds.

If $q+2 \le i \le m-q-1$, then by (17), we obtain

$$\begin{aligned}\sum_{|i-j|\ge q+1} |v_i^H v_j| &\le (i-q-1)\max_{1\le j\le i-q-1} |v_i^H v_j| + (m-i-q)\max_{i+q+1\le j\le m} |v_i^H v_j| \\ &\le (i-q-1)\|A-A^H\| c_i h_{ii-1}^{-1} \\ &\quad +(m-i-q)\|A-A^H\|\max_{i+q+1\le j\le m} c_j h_{jj-1}^{-1} \\ &\le (m-2q-1)c\|A-A^H\|,\end{aligned}$$

where

$$c = \max_{q+2 \le i \le m-q-1} \{c_i h_{ii-1}^{-1}, \max_{i+q+1 \le j \le m} c_j h_{jj-1}^{-1}\},$$

which, together the positive definiteness of B_m, proves (20). In addition, it can be seen from the above relation that B_m has some unit eigenvalues if $q \geq [\frac{m-1}{2}]$, where $[\cdot]$ denotes integer part. □

We make a few comments on Theorems 2–3.

Remarks.

1. It is obvious from Step 2 of Algorithm 2 that the left hand side of (17) is invariant under a scaling of A. In view of (18), we can verify that c_i and the right hand side of (17) are also invariant under a scaling of A. So (17) indeed has nothing to do with scaling.

2. When $A = A^H$, it can also follow from Theorems 2–3 that V_m is orthonormal, $B_m = I_m$ and $V_m^H v_{m+1} = 0$ provided that $q \geq 2$.

3. Theorems 2–3 reveal that V_m might deviate from orthonormality completely and B_m might be far from an identity matrix even though we take $q = m-2$ if A is very unsymmetric, which means that $\|A - A^H\|$ is not a number near to zero, e.g. $\|A - A^H\| > 1$. Therefore, when A is very unsymmetric, it is possible that V_m will never be close to orthonormality except for $q = m-1$, in which case V_m and B_m are naturally orthonormal and an identity matrix, respectively. Furthermore, for very unsymmetric matrices, the columns of V_m even become almost linearly dependent such that $\|B_m^{-1}\|$ will be very large, say 10^8, except for $q = m-1$. On the other hand, keep in mind that for any $| i - j | \geq q+1$ we all have $| v_i^H v_j | < \|v_i\| \|v_j\| = 1$ if Step 2 of Algorithm 2 does not break down. Therefore, (17) will possibly lose practical value if A is very unsymmetric.

4. From the process of proof, we can see that the right hand side of (17) is of order $\|A - A^H\|$ usually if A is nearly symmetric and all $h_{ii-1}, i = q+2, \ldots, m+1$ are not small. However, in view of (17), if some h_{ii-1} for $q+2 \leq i \leq m$ are small, then the columns of V_m may become nearly linearly dependent and B_m thus nearly singular.

Example. Construct an 7×7 upper Hessenberg matrix A as follows:

$$A = \begin{pmatrix} 1 & 10^{-5} & 0 & 0 & 10^{-3} & 10^{-3} & 0 \\ 10^{-5} & 2 & 10^{-5} & 0 & 0 & 0 & 0 \\ 0 & 10^{-5} & 3 & 10^{-5} & 0 & 0 & 0 \\ 0 & 0 & 10^{-5} & 4 & 10^{-5} & 0 & 0 \\ 0 & 0 & 0 & 10^{-5} & 5 & 10^{-5} & 0 \\ 0 & 0 & 0 & 0 & 10^{-5} & 6 & 10^{-5} \\ 0 & 0 & 0 & 0 & 0 & 10^{-5} & 7 \end{pmatrix}.$$

It is easy to find $\|A - A^H\| = \sqrt{2} \times 10^{-3}$. So A is nearly symmetric. Let us take $x^{(0)} = 0, m = 6, q = m - 2 = 4$ and $b = e_1$. Then in Step 2 of Algorithm 2, $v_1 = e_1$, and the computed results were $v_2 = e_2, \ldots, v_{m-1} = e_{m-1}$,

$$v_m = (\frac{10^{-3}}{\sqrt{(10^{-3})^2 + (10^{-5})^2}}, 0, 0, 0, 0, \frac{10^{-5}}{\sqrt{(10^{-3})^2 + (10^{-5})^2}}, 0)^H,$$

$$v_{m+1} = (-9.9994998 \times 10^{-3}, 2.0000104 \times 10^{-4}, 0, 0, 0, 0.99994998, 2.0000104 \times 10^{-6})^H.$$

Therefore,

$$v_1^H v_m = \frac{10^{-3}}{\sqrt{(10^{-3})^2 + (10^{-5})^2}} \approx 0.99995000047,$$

$$\|B_m^{-1}\| \approx 20001.49999, \|s_m\| \approx 7.070523392, \|V_m^H v_{m+1}\| \approx 0.0100014999,$$

$$\|V_m^+\| \approx 141.426659395, \|(I - \pi_m)v_{m+1}\| \approx 2. \times 10^{-6}.$$

For $q = 2, 3$, it can be verified that the corresponding results computed are the same as above.

We can see from the above results that V_m lost orthonormality completely and B_m was far from an identity matrix even when A is nearly symmetric and $q = m - 2$. Therefore, Theorems 2–3 with together the above comments, as well as this example, show that IOMs(q) may generate a very ill-conditioned basis of $\mathcal{K}_m(v_1, A)$ for any unsymmetric A and $q \leq m - 2$.

Having established the above results, we are now in a position to analyze the convergence of IOMs(q).

Theorem 4 *1. The residual of IOMs(q) is*

$$\tilde{r}^{(m)} = r^{(0)} - A\tilde{z}^{(m)} = -h_{m+1m} v_{m+1} e_m^H \tilde{y}^{(m)}. \tag{22}$$

2. The residual of FOM equals

$$r^{(m)} = r^{(0)} - Az^{(m)} = -h_{m+1m}(I - \pi_m)v_{m+1} e_m^H y^{(m)}. \tag{23}$$

3. If Step 2 of Algorithm 2 does not break down, and $\hat{H}_m^{-1}, \tilde{H}_m^{-1}$ exist, then

$$\|\tilde{r}^{(m)}\| = \frac{|\ e_m^H \tilde{y}^{(m)}\ |}{\|(I - \pi_m)v_{m+1}\|\ |\ e_m^H y^{(m)}\ |} \|r^{(m)}\|, \tag{24}$$

$$\|\tilde{r}^{(m)}\| = \frac{1}{\|(I - \pi_m)v_{m+1}\|\ |\ 1 - e_m^H \hat{H}_m^{-1} s_m\ |} \|r^{(m)}\|, \tag{25}$$

$$\|\tilde{r}^{(m)}\| = \frac{|\ 1 + e_m^H \tilde{H}_m^{-1} s_m\ |}{\|(I - \pi_m)v_{m+1}\|} \|r^{(m)}\|. \tag{26}$$

4. If Algorithm 2 breaks down, then $\tilde{x}^{(m)} = x^*$.

Proof. Part 1. Noticing that $V_m(\beta e_1) = r^{(0)}$ and in terms of (11) and (13), we obtain

$$\begin{aligned}
\tilde{r}^{(m)} &= r^{(0)} - A\tilde{z}^{(m)} = r^{(0)} - AV_m\tilde{y}^{(m)} \\
&= r^{(0)} - (V_m\tilde{H}_m + h_{m+1m}v_{m+1}e_m^H)\tilde{y}^{(m)} \\
&= r^{(0)} - V_m(\beta e_1) - h_{m+1m}v_{m+1}e_m^H\tilde{y}^{(m)} \\
&= -h_{m+1m}v_{m+1}e_m^H\tilde{y}^{(m)},
\end{aligned}$$

which is just (22). By the way, taking norms on two sides proves (12).

Part 2. Due to $B_m(\beta e_1) = V_m^H V_m(\beta e_1) = V_m^H r^{(0)}$, we have $B_m^{-1}V_m^H r^{(0)} = \beta e_1$. In terms of (3) and (15), we get

$$\begin{aligned}
r^{(m)} &= r^{(0)} - Az^{(m)} = r^{(0)} - AV_m y^{(m)} \\
&= r^{(0)} - (V_m\tilde{H}_m + h_{m+1m}v_{m+1}e_m^H)y^{(m)} \\
&= r^{(0)} - [V_m(\hat{H}_m - s_m e_m^H) + h_{m+1m}v_{m+1}e_m^H]y^{(m)} \\
&= r^{(0)} - V_m(\beta e_1) + V_m s_m e_m^H y^{(m)} - h_{m+1m}v_{m+1}e_m^H y^{(m)} \\
&= r^{(0)} - r^{(0)} + h_{m+1m}\pi_m v_{m+1}e_m^H y^{(m)} - h_{m+1m}v_{m+1}e_m^H y^{(m)} \\
&= -h_{m+1m}(I - \pi_m)v_{m+1}e_m^H y^{(m)},
\end{aligned}$$

which proves (23). Taking norms on two hand sides yields

$$\|r^{(m)}\| = h_{m+1m}\|(I - \pi_m)v_{m+1}\| \mid e_m^H y^{(m)} \mid, \tag{27}$$

which corrects a result of [31], where the factor $\|(I - \pi_m)v_{m+1}\|$ was lost.

Part 3. By assumption, both (22) and (23) cannot be zero vectors. So dividing the norm of (22) by (27) immediately yields (24).

Based on the Shermann–Morrison formula [16, p. 51] and (3)(15), we obtain

$$\begin{aligned}
\tilde{y}^{(m)} &= \beta\tilde{H}_m^{-1}e_1 = \beta(\hat{H}_m - s_m e_m^H)^{-1}e_1 \\
&= \beta(\hat{H}_m^{-1} + \frac{\hat{H}_m^{-1}s_m e_m^H\hat{H}_m^{-1}}{1 - e_m^H\hat{H}_m^{-1}s_m})e_1 \\
&= \beta\hat{H}_m^{-1}e_1 + \frac{\hat{H}_m^{-1}s_m e_m^H(\beta\hat{H}_m^{-1}e_1)}{1 - e_m^H\hat{H}_m^{-1}s_m} \\
&= y^{(m)} + \frac{e_m^H y^{(m)}}{1 - e_m^H\hat{H}_m^{-1}s_m}\hat{H}_m^{-1}s_m.
\end{aligned}$$

Therefore, we get

$$e_m^H\tilde{y}^{(m)} = e_m^H y^{(m)} + \frac{e_m^H y^{(m)}}{1 - e_m^H\hat{H}_m^{-1}s_m}e_m^H\hat{H}_m^{-1}s_m$$

$$
\begin{aligned}
&= e_m^H y^{(m)}\left(1+\frac{e_m^H \hat{H}_m^{-1} s_m}{1-e_m^H \hat{H}_m^{-1} s_m}\right) \\
&= \frac{e_m^H y^{(m)}}{1-e_m^H \hat{H}_m^{-1} s_m},
\end{aligned}
$$

which proves with together (24) that (25) holds.

Similarly, we have

$$
\begin{aligned}
e_m^H \hat{H}_m^{-1} s_m &= e_m^H (\tilde{H}_m + s_m e_m^H)^{-1} s_m \\
&= e_m^H \left(\tilde{H}_m^{-1} - \frac{\tilde{H}_m^{-1} s_m e_m^H \tilde{H}_m^{-1}}{1+e_m^H \tilde{H}_m^{-1} s_m}\right) s_m \\
&= e_m^H \tilde{H}_m^{-1} s_m - \frac{(e_m^H \tilde{H}_m^{-1} s_m)(e_m^H \tilde{H}_m^{-1} s_m)}{1+e_m^H \tilde{H}_m^{-1} s_m} \\
&= \frac{e_m^H \tilde{H}_m^{-1} s_m}{1+e_m^H \tilde{H}_m^{-1} s_m},
\end{aligned}
$$

from which it follows that

$$
\frac{1}{1-e_m^H \hat{H}_m^{-1} s_m} = 1 + e_m^H \tilde{H}_m^{-1} s_m.
$$

Thus, combining with (25) shows that (26) holds.

Part 4. It is direct from (22). □

According to the a-priori bounds in [31, 21] for the diagonalizable and defective A, respectively, $\|r^{(m)}\|$ converge to zero as m increases, and it can be very small even for $m \ll N$ provided that the distribution of eigenvalues of A is favorable, A not highly defective and the eigenvector or Jordan basis of A not ill-conditioned. Therefore, $\|\tilde{r}^{(m)}\|$ will also converge to zero and behave much like FOM provided that the factor

$$
\frac{|\ 1+e_m^H \tilde{H}_m^{-1} s_m\ |}{\|(I-\pi_m) v_{m+1}\|} \tag{28}
$$

is moderate, say 10.

Now let us analyze Theorem 4.

Theorem 5 *1. Let $Q_m R_m = \tilde{H}_m$ be the QR decomposition of $\tilde{H}_m$. Then if Step 2 of Algorithm 2 does not break down and $\tilde{H}_m^{-1}$ exists, we have*

$$
\|\tilde{r}^{(m)}\| \leq \frac{1 + h_{m+1m}\|V_m^{+}\|/r_{mm}}{\|(I-\pi_m)v_{m+1}\|}\|r^{(m)}\|. \tag{29}
$$

2. Let $Q_m R_m = \tilde{H}_m$ as above, and assume A to be nearly symmetric. Then if Step 2 of Algorithm 2 does not break down, $\tilde{H}_m^{-1}$ exists and

$$
1 - \|V_m\|\|B_m^{-1}\|\|V_m^H v_{m+1}\| > 0, \tag{30}
$$

then the following relation holds:

$$\|\tilde{r}^{(m)}\| \leq \frac{1 + h_{m+1m}\|B_m^{-1}\|\|V_m^H v_{m+1}\|/r_{mm}}{1 - \|V_m\|\|B_m^{-1}\|\|V_m^H v_{m+1}\|}\|r^{(m)}\|. \tag{31}$$

Proof. Part 1. Let $Q_m = (q_1, q_2, \ldots, q_m)$. Then we get via $Q_m R_m = \tilde{H}_m$

$$\begin{aligned}
\mid 1 + e_m^H \tilde{H}_m^{-1} s_m \mid &\leq 1 + \mid e_m^H \tilde{H}_m^{-1} s_m \mid = 1 + h_{m+1m} \mid e_m^H R_m^{-1} Q_m^H V_m^+ v_{m+1} \mid \\
&= 1 + h_{m+1m} \mid (q_m^H / r_{mm}) V_m^+ v_{m+1} \mid \\
&\leq 1 + h_{m+1m}\|q_m\|\|V_m^+\|\|v_{m+1}\|/r_{mm} \\
&= 1 + h_{m+1m}\|V_m^+\|/r_{mm},
\end{aligned}$$

from which and (26) it follows that (29) holds.

Part 2. Under the assumption, we obtain

$$\begin{aligned}
\|(I - \pi_m)v_{m+1}\| &\geq \|v_{m+1}\| - \|\pi_m v_{m+1}\| = 1 - \|V_m B_m^{-1} V_m^H v_{m+1}\| \\
&\geq 1 - \|V_m\|\|B_m^{-1} V_m^H v_{m+1}\| \\
&\geq 1 - \|V_m\|\|B_m^{-1}\|\|V_m^H v_{m+1}\|.
\end{aligned}$$

On the other hand, in a similar way to the proof of Part 1, we have

$$\begin{aligned}
\mid 1 + e_m^H \tilde{H}_m^{-1} s_m \mid &\leq 1 + h_{m+1m} \mid q_m^H B_m^{-1} V_m^H v_{m+1} \mid / r_{mm} \\
&\leq 1 + h_{m+1m}\|q_m\|\|B_m^{-1}\|\|V_m^H v_{m+1}\|/r_{mm} \\
&= 1 + h_{m+1m}\|B_m^{-1}\|\|V_m^H v_{m+1}\|/r_{mm},
\end{aligned}$$

which and (26) give (31). $\square$

Now let us analyze Theorems 4–5. According to Theorems 2–3 as well as the comments there, we know that if the matrix A is generally unsymmetric, the columns of V_m may have weak linear independence except for $q = m-1$, so that $\|V_m^+\|$ will be large, say 10^8. On the other hand, note that $\|(I - \pi_m)v_{m+1}\|$ measures the distance between v_{m+1} and the subspace $\mathcal{K}_m(r^{(0)}, A)$. Therefore, if v_{m+1} nearly lies in $\mathcal{K}_m(r^{(0)}, A)$, then $\|(I - \pi_m)v_{m+1}\|$ is near to zero, say 10^{-6}; conversely, the more strongly v_{m+1} is linearly independent of $\mathcal{K}_m(r^{(0)}, A)$, the closer to one $\|(I - \pi_m)v_{m+1}\|$ is. Trivially, we have $\|(I - \pi_m)v_{m+1}\| = 1$ and $\mid 1 + e_m^H \tilde{H}_m^{-1} s_m \mid = 1$ for $q = m$. In a word, we can conclude that if $\|V_m^+\|$ is large and $\|(I - \pi_m)v_{m+1}\|$ is small, then IOMs(q) may not behave like FOM since the multiplicative factor of the right hand side of (26) may be very large, say 10^{12}. In this case, $\|\tilde{r}^{(m)}\|$ may not be near to a zero vector though $\|r^{(m)}\|$ is very small, say 10^{-9}, unless $r^{(m)} = 0$. Otherwise, IOMs(q) behave much like FOM once the sequence $\{v_i\}_1^{m+1}$ is well conditioned.

Also, we should point out that if a very small r_{mm} occurs, then it will be numerically difficult to solve (11). This case corresponds to a nearly singular $\tilde{H}_m$ in the way H_m of Algorithm 1 does, e.g. [33, 7]. In implementations, we will avoid using such a

$\tilde{H}_m$ by replacing it by its some submatrix $\tilde{H}_k$, $k \leq m-1$ or enlarging the steps m until a numerically nonsingular $\tilde{H}_k$, $k \leq m-1$ or $k > m$, is found. Note that it is very convenient and cheap to do so due to the update formulas of the LU or the QR decomposition of $\tilde{H}_m$ at each step; for details, refer to [33, 40].

Now we shed light on (31). Before doing this, we must explain the validity of (30). We see from Theorems 2–3 as well as the comments there that (30) is easily met usually except for some extremal cases such as the example described earlier, provided that A is nearly symmetric, q is chosen suitably and no small h_{ii-1}, $q+2 \leq i \leq m$ occurs. In such a case, if q is chosen suitably, we then have from (20)

$$\|B_m^{-1}\| \leq \frac{1}{1-(m-q-1)c\|A-A^H\|}$$

will be moderate and from (17)

$$\begin{aligned} \|V_m^H v_{m+1}\| &= \sqrt{\sum_{j=1}^{m-q} \mid v_j^H v_{m+1} \mid^2} \leq \sum_{j=1}^{m-q} \mid v_j^H v_{m+1} \mid \\ &\leq (m-q) \max_{1 \leq j \leq m-q} \mid v_j^H v_{m+1} \mid \\ &\leq (m-q) c_{m+1} \|A - A^H\| h_{m+1m}^{-1} \end{aligned}$$

will be small if h_{m+1m} is not small. At the same time note that $1 \leq \|V_m\| = \sqrt{\|B_m\|} \leq \sqrt{1+m-q-1} = \sqrt{m-q}$ is a slowly increasing factor with m if $q \leq m-1$. Consequently, (30) cannot be small usually once q is chosen suitably and no small h_{ii-1}, $q+2 \leq i \leq m+1$ occurs, such that the multiplicative factor of the right hand side of (31) will be usually moderate unless r_{mm} is very small, namely, $\tilde{H}_m$ is nearly singular. These arguments expose that IOMs(q) can behave much like FOM under these conditions.

In summary, we can say that IOMs(q) can succeed for solving (1) under some restrictions but are possibly subject to failure, especially for general unsymmtric matrices no matter how q is chosen except for $r^{(m)} = 0$ in theory. Meanwhile, IOMs(q) may not behave like FOM without some assumptions even when A is nearly symmetric.

Finally, we point out that due to the mathematical equivalence between IOMs(q) and ORTHORES(q) established by [33] we essentially give some insights into the convergence of the latter in the mathematical sense.

4 Choice of q

First of all, we have to point out that q is problem dependent and has something to do with the process of building $\tilde{H}_m$. In particular, Theorems 2–3 will possibly lose practical values if A is generally unsymmetric. Therefore, it is impossible for us to give a global strategy which covers all real problems and treat all potential cases. Based on the theoretical analysis in Section 3, to some extent we attempt to suggest a reliable strategy. To this end, we will make the following assumptions:

- A is nearly symmetric, namely, $\|A - A^H\|$ is about 10^{-1} or less.
- No occurrence of small h_{ii-1}, $i = q+2, \ldots, m+1$ and r_{mm}, and $h_{m+1m}/r_{mm} = O(1)$, which means that (17) will be small and of order $\|A - A^H\|$. Note that from (12) a very small h_{m+1m} implies the convergence of Algorithm 2.

Under these assumptions, we can roughly simplify the results of Theorems 2–3 as follows:

$$\begin{aligned} \mid v_i^H v_j \mid &\approx \|A - A^H\|, \text{ for } \mid i-j \mid \geq q+1, \\ \|B_m^{-1}\| &\approx \frac{1}{1-(m-q-1)\|A-A^H\|}, \\ \|V_m^H v_{m+1}\| &\approx (m-q)\|A - A^H\|. \end{aligned}$$

Note that $1 \leq \|V_m\| = \sqrt{\|B_m\|} \approx \sqrt{1+(m-q-1)\|A-A^H\|} < 1.3$ if we require $(m-q-1)\|A-A^H\| \leq 0.5$. Therefore, we have

$$\frac{1 + h_{m+1m}\|B_m^{-1}\|\|V_m^H v_{m+1}\|/r_{mm}}{1-\|V_m\|\|B_m^{-1}\|\|V_m^H v_{m+1}\|} \approx \frac{1+\|B_m^{-1}\|\|V_m^H v_{m+1}\|}{1-\|B_m^{-1}\|\|V_m^H v_{m+1}\|} \approx \frac{1+\mu_{mq}}{1-\mu_{mq}}, \tag{32}$$

where

$$\mu_{mq} = \frac{(m-q)\|A-A^H\|}{1-(m-q-1)\|A-A^H\|}.$$

In order to make Algorithm 2 comparable to Algorithm 1, we have to consider two aspects: First, choice of q must keep $\|B_m^{-1}\|$ below some moderate tolerance tol_1, e.g. $tol_1 \leq 10$; second, we must restrict (32) to some moderate tolerance tol_2, say $tol_2 \leq 100$. At the same time, keep in mind that $q \geq 2$. Remembering these considerations, we now solve the following problem:

$$\left\{ \begin{array}{ccc} \frac{1}{1-(m-q-1)\|A-A^H\|} & \leq & tol_1 \\ \frac{1+\mu_{mq}}{1-\mu_{mq}} & \leq & tol_2 \\ q & \geq & 2. \end{array} \right. \tag{33}$$

The minimal integer q_{min} satisfying (33) is

$$q_{min} = [\max\{m-1-\frac{tol_1-1}{tol_1\|A-A^H\|}, m - \frac{(tol_2-1)(1+\|A-A^H\|)}{2tol_2\|A-A^H\|}, 2\}]. \tag{34}$$

(34) is not so visual. We test it for $m = 30$, $\|A - A^H\| = 0.1$, $tol_1 = 10$ and $tol_2 = 50$. Then it is easy to justify $q_{min} = 24$. If $\|A - A^H\| = 0.01$ and other

parameters remain unchanged, then $q_{min} = 2$. Numerical experiments show that (34) is reliable and usually conservative.

There remains an important problem to be solved. (34) involves the computation of $\|A - A^H\|$, which is impractical to obtain. To get rid of this, we simply replace it by using

$$\|A - A^H\| \leq \sqrt{\|A - A^H\|_1 \|A - A^H\|_\infty} = \|A - A^H\|_1, \tag{35}$$

where $\|\cdot\|_1$ and $\|\cdot\|_\infty$ denote, respectively, the 1-norm and the ∞-norm. Since A is usually very sparse, the cost of computing $\|A - A^H\|_1$ is negligible, compared with the total cost of solving (1).

5 Numerical experiments

We report some numerical results to show convergence behaviors of IOMs(q) and their restarted versions. All the experiments were run using *Matlab*4.1 on DEC RISC workstation with the machine precision $eps \approx 2.22 \times 10^{-16}$. For simplicity we do not use any preconditioner in experiments.

Example. Consider the following two dimensional convection-diffusion partial equations:

$$-\triangle u(x,y) + \alpha \frac{\partial}{\partial x} u(x,y) = f(x,y) \text{ on } \Omega = [0,1] \times [0,1] \tag{36}$$

with the boundary condition $u(x,y) = 0$.

Discretizing (36) by centered differences with mesh size $h = 1/(n+1)$ gives rise to an unsymmetric matrix $A(\alpha)$ of dimension $N = n^2$. The parameter α is used to control the unsymmetry of $A(\alpha)$, ranging from nearly symmetric to very unsymmetric. $A(\alpha)$ is of the block tridiagonal form $A(\alpha) = \text{tri}(I, B(\alpha), I)$, where $B(\alpha) = \text{tri}(-1-\delta, 4, -1+\delta)$ is a tridiagonal matrix with subdiagonal entries $-1-\delta = -1-\alpha/2(n+1)$, main diagonal entries 4 and superdiagonal entries $-1+\delta = -1+\alpha/2(n+1)$. The right hand side b of (1) was set to $A(\alpha)x^*$, where $x^* = (1,1,\ldots,1)^H$, and the initial vector $x^{(0)}$ was all taken as zero vector. First, we examined Algorithm 2 for various m, q, α and fixed $n = 22$, respectively. In the following, it denotes the number of iterations, $CPU(s)$ and CPU_a, respectively, denote the total CPU consuming with the unit "second" and the average CPU cost per restart, and *ratio* is the ratio of CPU time of Algorithm 2 to that of Algorithm 1. We required that the stopping criterion be $\|\tilde{r}^{(m)}\| \leq 10^{-5}$. Also, we remind readers that the secondary storage was not exploited for the simplicity of comparison and Algorithm 2 reduces to Algorithm 1 for $q = m$. Tables 1–3 display the results obtained by Algorithm 2.

m	q	it	$CPU(s)$	CPU_a	$ratio$	Residual norm
10	2	15	108.45	7.23	0.8871	$7.D-6$
10	5	14	111.67	7.97	0.9134	$8.5D-6$
10	8	15	128.73	8.58	1.0530	$2.8D-6$
10	10	14	122.25	8.73	1	$9.8D-6$
20	5	5	93.85	18.77	1.0858	$2.D-6$
20	10	4	78.61	19.65	0.9095	$8.7D-6$
20	15	4	82.16	20.54	0.9506	$1.7D-6$
20	20	4	86.43	21.61	1	$2.1D-6$
30	5	3	84.26	28.08	0.7638	$4.2D-6$
30	10	3	91.44	30.48	0.8288	$6.4D-6$
30	15	3	102.2	34.06	0.9264	$2.8D-6$
30	20	3	105.47	35.15	0.9560	$3.3D-6$
30	25	3	107.82	35.94	0.9773	$1.8D-6$
30	30	3	110.32	36.77	1	$6.7D-6$

Table 1: *The results obtained by Algorithm 2 for* $N = n^2 = 484$ *and* $\alpha = 1$, $\|A(1) - A(1)^H\|_1 = 0.08696$

m	q	it	$CPU(s)$	CPU_a	$ratio$	Residual norm
10	5	6	25.75	4.29	0.9690	$2.5D-6$
10	8	6	25.27	4.21	0.9507	$2.6D-6$
10	10	6	26.58	4.43	1	$9.2D-6$
20	5	4	33.22	8.31	0.7302	$3.1D-6$
20	10	4	34.79	8.69	0.7647	$1.2D-6$
20	15	4	36.03	9.00	0.7918	$7.6D-7$
20	20	5	45.50	9.10	1	$9.6D-8$
30	5	3	39.04	13.01	0.5822	$1.D-8$
30	10	3	41.92	13.97	0.6252	$3.8D-8$
30	15	3	42.13	14.04	0.6283	$5.7D-6$
30	20	3	47.94	15.97	0.7149	$3.6D-6$
30	25	4	57.13	14.28	0.8520	$4.5D-7$
30	30	4	67.06	16.76	1	$6.9D-7$

Table 2: *The results obtained by Algorithm 2 for* $N = n^2 = 484$ *and* $\alpha = 25$, $\|A(25) - A(25)^H\|_1 = 2.1739$

m	q	it	$CPU(s)$	CPU_a	$ratio$	Residual norm
10	2	45	182.72	4.06	0.9591	$9.2D-6$
10	5	45	185.35	4.12	0.9729	$8.2D-6$
10	8	45	188.35	4.19	0.9886	$7.D-6$
10	10	45	190.52	4.23	1	$5.9D-6$
20	2	19	149.65	7.88	0.9408	$5.8D-6$
20	5	17	143.27	8.42	0.9007	$9.4D-6$
20	10	17	142.27	8.37	0.8944	$5.D-6$
20	15	17	151.82	8.93	0.9544	$7.5D-6$
20	20	17	159.07	9.36	1	$8.9D-6$
30	5	12	158.86	13.22	0.9542	$5.D-6$
30	10	12	161.06	13.42	0.9684	$3.6D-6$
30	15	12	163.07	13.59	0.9805	$3.D-6$
30	20	11	154.07	14.00	0.9263	$2.1D-6$
30	25	11	158.27	14.39	0.9516	$4.D-6$
30	30	11	166.32	15.12	1	$2.6D-6$

Table 3: *The results obtained by Algorithm 2 for* $N = n^2 = 484$ *and* $\alpha = 1000$, $\|A(1000) - A(1000)^H\|_1 = 86.9565$

For $\alpha = 1$, the matrix $A(\alpha)$ is nearly symmetric. Hence, according to the theoretical analysis, IOMs(q) behave like FOM usually, and Algorithm 2 can thus be competitive with and even more efficient than Algorithm 1 usually. These expectations were confirmed by Table 1, and (34) worked well if tol_1 and tol_2 were all set to 100.

For $\alpha = 25$ and $\alpha = 1000$, $A(\alpha)$ are mild unsymmetric and very unsymmetric. In terms of the a-priori convergence theory presented by [31, 21], it can be expected that Algorithm 1 may use many restarts to achieve the desired accuracy. So if successful, Algorithm 2 may also converge slowly and are possibly less efficient than Algorithm 1 for some q. Compared with Table 1, Table 2–3 indeed coincided with the theoretical analysis. Further, we found that in most cases Algorithm 2 was competitive with Algorithm 1, and even more efficient, this is due to the moderate $\|V_m^+\|$ observed in experiments. In addition, there occurred a similar phenomenon to that found in [33]: Algorithm 2 with smaller q may result in faster convergence than those with larger q, e.g. $m = 30, q = 5$ and $\alpha = 25$ versus $m = 30, q = 25$ and $\alpha = 25$. Moreover, Algorithm 2 may also use fewer restarts than Algorithm 1, e.g. $m = 20, q = 5$ and $\alpha = 25$ versus the corresponding Algorithm 1.

Below we take $n = 9$, namely $N = 81$, and then show how the non-restarted IOMs(q) converge with the steps k, $k = 1, 2, \ldots, m$ and various q, where m was set to 35. Table 4 displays the final computed results with $k = m$.

Several notes are made in order: First, it can be seen from Table 4 that V_m lost

m	q	α	$\sigma_{min}(V_m)$	$\|A(\alpha) - A(\alpha)^H\|_1$	Residual norm
35	2	0.01	$2.6D-9$	0.002	$5.6D-11$
35	10	0.01	$2.1D-8$	0.002	$9.1D-12$
35	18	0.01	$2.1D-5$	0.002	$2.8D-13$
35	25	0.01	0.0956	0.002	$3.7D-13$
35	35	0.01	1	0.002	$3.1D-16$
35	2	1	$4.2D-10$	0.2	$2.3D-4$
35	10	1	$1.9D-8$	0.2	$6.8D-8$
35	18	1	$4.3D-6$	0.2	$1.7D-8$
35	25	1	$3.D-3$	0.2	$4.5D-10$
35	35	1	1	0.2	$1.3D-13$
35	2	10	$3.3D-12$	2	$6.5D-4$
35	10	10	$1.6D-10$	2	$1.3D-4$
35	18	10	$9.6D-7$	2	$7.2D-8$
35	25	10	$1.3D-4$	2	$1.8D-8$
35	35	10	1	2	$2.D-13$
35	2	1000	$5.3D-3$	200	7.223779742
35	10	1000	0.2911	200	1.095512851
35	18	1000	0.8898	200	0.415744345
35	25	1000	0.9955	200	0.403716911
35	35	1000	1	200	0.403735555

Table 4: *Final results obtained by IOMs(q) with* $m = 35$ *for* 81×81 $A(\alpha)$

orthonormality completely and became almost column linearly dependent rapidly with small q for nearly symmetric matrices; second, $\|\tilde{r}^{(k)}\|$ often oscillated with k though they showed the decreasing tendency on the average; third, for the very unsymmetric $A(1000)$, both FOM and IOMs(q) converged very slowly, as can be expected; fourth, IOMs(q) cannot behave like FOM for many q even when $A(\alpha)$ is nearly symmetric.

All the above experiments were run on $A(\alpha)$ with positive definite real parts. Now we report some results on some shifted $A(\alpha) := A(\alpha) - 2I$ with indefinite real parts using the non-restarted IOM(q) for various q. As is well known [33], for this class of matrices, compared with the problems earlier, solving (1) may become much more difficult. Table 5 shows the final computed results with $m = 35$.

Obviously, Table 5 shows that solving such (1) became much more difficult using either FOM or IOMs(q) and IOMs(q) cannot behave like FOM for many q even when $A(\alpha)$ is nearly symmetric. It was observed that both FOM and IOMs(q) exhibited rather irregular convergence behaviors with wild oscillations in the residual norms $\|\tilde{r}^{(k)}\|$, $k = 1, 2, \ldots, m$ though they tended to decrease in a very slow manner, this is due to the lack of minimization property of FOM and IOMs(q) over $\mathcal{K}_m(r^{(0)}, A)$.

m	q	α	$\sigma_{min}(V_m)$	$\|A(\alpha) - A(\alpha)^H\|_1$	Residual norm
35	2	0.01	$8.D-11$	0.002	$5.1D-4$
35	10	0.01	$2.1D-10$	0.002	$3.D-4$
35	18	0.01	$1.4D-6$	0.002	$6.1D-6$
35	25	0.01	0.0300	0.002	$9.D-7$
35	30	0.01	0.9184	0.002	$1.1D-7$
35	35	0.01	1	0.002	$8.5D-8$
35	2	1	$2.2D-11$	0.2	0.523776717
35	10	1	$1.1D-9$	0.2	0.296784521
35	18	1	$7.1D-7$	0.2	0.036461062
35	25	1	$1.D-4$	0.2	0.041961062
35	35	1	1	0.2	$9.D-4$
35	2	10	$1.5D-13$	2	$3.9D+3$
35	10	10	$1.1D-10$	2	83.82767282
35	20	10	$7.2D-7$	2	$1.16D+2$
35	33	10	0.9611	2	0.011178199
35	35	10	1	2	$7.2D-3$

Table 5: *Final results obtained by IOMs(q) with* $m = 35$ *for some* 81×81 $A(\alpha)$ *with indefinite real parts*

6 Conclusions and remarks

In this paper, we have analyzed in detail the convergence of IOMs(q) for solving (1) and have exposed some important features of IOMs(q). The results have exposed that IOMs(q) can behave like FOM under some restrictions and why they may not behave like FOM even when A is nearly symmetric. Due to the mathematical equivalence between IOMs(q) and ORTHORES(q), we essentially give insights into the convergence of the latter. Also, we have suggested one strategy of choosing q, which has shown its reliability. Numerical experiments have confirmed our theoretical analysis, and have shown that the restarted IOMs(q) are usually competitive with and even more efficient than the restarted FOM. Here, we want to point out that it seems likely that IOMs(q) for solving (1) are less sensitive to the value of q than IOMs(q) for solving large unsymmetric eigenproblems [22]. Also, I think that this paper may provide some ingredients for a convergence analysis of some other truncated versions mentioned in Section 1.

There remain many problems to be studied. For example, how to choose adaptively q during the process of iterations, especially when A is generally unsymmetric? How to improve Algorithm 2 in order to avoid potential failures? Final note, probably the most important one: From the previous theoretical analysis it seems that the strategy of truncating used is not reliable to overcome the drawbacks that FOM has. It will be probably best to take some completely different strategies to generate a well conditioned basis but keep arithmetic operations and primary storage as low as possible, such that the modified IOM(q) will behave much like FOM and thus their restarted versions can be expected to be more efficient than the restarted FOM in practice. All these problems will constitute future topics of research.

Acknowledgement. I thank Professor L. Elsner for reading this paper and useful discussions.

References

[1] A. E. Arnoldi, *The principle of minimized iteration in the solution of the matrix eigenvalue problem*, Quart. Appl. Math., 9 (1951), pp. 11–29.

[2] O. Axelsson, *On preconditioned conjugate gradient methods*, CNA–143, Center for Numerical Analysis, Univ. of Texas, Austin, 1978.

[3] O. Axelsson, *Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations*, Linear Algebra Appl., 29 (1980), pp. 1–16.

[4] O. Axelsson, *A generalized conjugate gradient direction method and its application on a singular perturbation problem*, Proceedings of 8th Biennial Numerical Analysis Conference, G. A. Watson, ed., Lecture Notes in Mathematics, Springer–Verlag, No. 773, 1980.

[5] P. N. Brown and A. C. Hindmarsh, *Reduced storage methods in stiff ODE systems*, Appl. Math. Comput., 31 (1989), pp. 40–91.

[6] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Stat. Comput., 11 (1990), pp. 450–481.

[7] P. N. Brown, *A theoretical comparison of the Arnoldi and GMRES algorithms*, SIAM J. Sci. Stat. Comput., 12 (1991), pp. 58–78.

[8] S. C. Eisenstat, H. C. Elman and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.

[9] H. C. Elman, *Iterative methods for large sparse nonsymmetric systems of linear equations*, Ph.D. Thesis, Computer Science Dept., Yale University, 1982.

[10] V. Faber and T. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.

[11] V. Faber and T. Manteuffel, *Orthogonal error methods*, SIAM J. Numer. Anal., 24 (1987), pp. 170–187.

[12] R. Fletcher, *Conjugate gradient methods for indefinite systems*, Proceedings, Numerical Analysis Conference, Dundee, G. A. Watson, ed., Lecture Notes in Mathematics, Springer–Verlag, No. 506, 1976, pp.73–89.

[13] R. W. Freund and N. M. Nachtigal, *QMR: a quasi-minimal residual method for non-Hermitian linear systems*, Numer. Math., 60 (1991), pp. 291–313.

[14] R. W. Freund, G. H. Golub and N. M. Nachtigal, *Iterative solution of linear systems*, Acta. Numerica, (1992), pp.1–44.

[15] G. W. Gear and Y. Saad, *Iterative solution of linear equations in ODE codes*, SIAM J. Sci. Stat. Comput., 4 (1983), pp. 583–601.

[16] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 2nd Edition, the Hopkins University Press, Baltimore, 1989.

[17] G. H. Golub and D. P. O'Leary, *Some history of the conjugate gradient and Lanczos algorithms: 1946–1976*, SIAM Review, 31 (1989), pp. 50–102.

[18] J. F. Grcar, *Operator coefficient methods for linear equations*, Technical Report Sand89-8691, Sandia National Laboratories, New Mexico 87185 and Livermore, California 94551, 1989.

[19] M. H. Gutknecht, *A completed theory of the unsymmetric Lanczos process and related algorithms, Part I*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 594–639.

[20] M. H. Gutknecht, *A completed theory of the unsymmetric Lanczos process and related algorithms, Part II*, SIAM J. Matrix Anal. Appl., 15 (1994), pp. 15–58.

[21] Z. Jia, *The convergence of Krylov subspace methods for large unsymmetric systems*, Preprint, University of Bielefeld, 1994.

[22] Z. Jia, *Some numerical methods for large unsymmetric eigenproblems*, Ph.D. Thesis, University of Bielefeld, 1994.

[23] W. D. Joubert and D. M. Young, *Necessary and sufficient conditions for simplifications of generalized conjugate-gradient algorithms*, Linear Algebra Appl., 88/89 (1987), pp. 449–485.

[24] W. D. Joubert, *On the convergence behavior of the restarted GMRES algorithm for solving nonsymmetric linear systems*, Submitted to J. Numer. linear Algebra Appl., 1992.

[25] W. D. Joubert and G. F. Carey, *Parallelizable restarted iterative methods for nonsmmetric linear systems, Part I: Theory and Part II: Parallel implementation*, Submitted to Inter. J. Computer Math., 1992.

[26] C. Lanczos, *Solution of systems of linear equations by minimized iterations*, J. Res. Natur. Bur. Stand., 49 (1952), pp. 33–53.

[27] R. B. Morgan, *Restarting GMRES for large nonsymmetric systems of linear equations*, Preprint, 1992.

[28] B. N. Parlett, *A new look at the Lanczos algorithm for solving symmetric systems of linear equations*, Linear Algebra Appl., 29 (1980), pp. 323–346.

[29] B. N. Parlett, *Reduction to tridiagnoal form and minimal realizations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 567–593.

[30] Y. Saad, *Variations on Arnoldi's method for computing eigenelements of large unsymmetric matrices*, Linear Algebra Appl., 34 (1980), pp. 269–295.

[31] Y. Saad, *Krylov subspace methods for solving large unsymmetric linear systems*, Math. Comput., 37 (1981), pp. 105–126.

[32] Y. Saad, *The Lanczos biorthogonalization algorithm and other oblique projection methods for solving large unsymmetric systems*, SIAM J. Numer. Anal., 19 (1982), pp. 485–506.

[33] Y. Saad, *Practical use of some Krylov subspace methods for solving indefinite and nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 856–869.

[34] Y. Saad and M. H. Schultz, *Conjugate gradient-like algorithms for solving nonsymmetric linear systems*, Math. Comput., 44 (1985), pp. 417–424.

[35] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving non-symmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[36] Y. Saad, *Krylov subspace methods on supercomputers*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 1200–1232.

[37] J. N. Shadid and R. S. Tuminaro, *A comparison of preconditioned nonsymmetric Krylov methods on a large-scale MIND machine*, SIAM J. Sci. Comput., 15 (1994), pp. 440–459.

[38] C. H. Tong, *A comparative study of preconditioned Lanczos methods for nonsymmetric linear systems*, Technical Report Sand91-8240, Sandia National Laboratories, New Mexico 87185 and Livermore, California 94551, 1991.

[39] P. K. W. Vinsome, *ORTHOMIN, an iterative method for solving sparse sets of simultaneous linear systems*, 4th Symposium of Numerical Simulation of Reservior Performance, Society of Petroleum Engineers of the AIME, 1976, pp. 149–160.

[40] H. F. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 152–163.

[41] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.

[42] D. M. Young and K. C. Jea, *Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods*, Linear Algebra Appl., 34 (1980), pp. 159–194.