Numerical Algorithms **30:** 71–89, 2002.

A convergence analysis of GMRES and FOM methods for Sylvester equations

Mickaël Robbé [a] and Miloud Sadkane [b]

[a] *Free Field Technologies, 16 place de l'Université, B-1348 Louvain-la-Neuve, Belgium*
E-mail: mickael.robbe@fft.be

[b] *Université de Bretagne Occidentale, Département de Mathématiques, 6, Av. Le Gorgeu, BP 809, 29285 Brest Cedex, France*
E-mail: sadkane@univ-brest.fr

Received 24 April 2001; revised 18 December 2001
Communicated by C. Brezinski

We discuss convergence properties of the GMRES and FOM methods for solving large Sylvester equations of the form $AX - XB = C$. In particular we show the importance of the separation between the fields of values of A and B on the convergence behavior of GMRES. We also discuss the stagnation phenomenon in GMRES and its consequence on FOM. We generalize the issue of breakdown in the block-Arnoldi algorithm and explain its consequence on FOM and GMRES methods. Several numerical tests illustrate the theoretical results.

Keywords: Sylvester equation, GMRES, FOM, block Krylov subspace, stagnation, breakdown

AMS subject classification: 65F10

1. Introduction

Consider the Sylvester equation

$$\mathcal{S}(X) = C, \quad \text{where } \mathcal{S}(X) = AX - XB \tag{1.1}$$

and $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{p\times p}$, $C \in \mathbb{R}^{N\times p}$ with $N \gg p$.

Linear control theory provides an important application for such Sylvester equations (see, e.g., the papers by Datta [4], by Datta and Saad [5] and by Calvetti et al. [3]).

In the literature, several methods have been proposed to solve (1.1). Let us mention some of them. When the size N of A is small, direct methods based on Schur factorizations of A and B or on the Schur factorization of A and the Hessenberg factorization of B have been proposed respectively by Bartels and Stewart [1] and by Golub et al. [8]. When N is large, Saad [13] proposed a Krylov subspace method of Galerkin type for computing low rank solutions of the Lyapunov equation

$$AX - XA^* = uu^* \quad \text{with } u \in \mathbb{R}^N, \tag{1.2}$$

where here and throughout this note, the notation u^* is used for complex and real cases to denote the transpose conjugate of u. Jaimoukha and Kasenally [11] extended Saad's method to the case where the right-hand side of (1.2) has the form UU^*, where $U \in \mathbb{R}^{N\times p}$. They obtained two block methods of Galerkin and Petrov–Galerkin type that compute low rank solutions and gave the residual expressions of these methods. In [10], Hu and Reichel presented two methods of Galerkin and Petrov–Galerkin types for the Sylvester equation (1.1). Their method starts by finding a rank one approximation of C in the form fg^*, where $f \in \mathbb{R}^N$ and $g \in \mathbb{R}^p$. It then constructs two Krylov subspaces from A and f and from B^* and g on which approximate solutions of (1.1) are found. In [17], Simoncini extended the work of Hu and Reichel to block form using the idea developed in [11]. In [7], El Guennouni et al. introduced new Krylov methods to solve (1.1) based on block-Arnoldi and nonsymmetric block-Lanczos algorithms. They gave perturbation results for their methods and simple expressions for the Frobenius norm of the residuals.

In this note we discuss convergence analysis of block GMRES and block FOM methods to solve the Sylvester equation (1.1). A particular attention is given to the issues of breakdown and stagnation in these methods. The note is organized as follows: After a brief discussion in section 2 on the relationship between the block Krylov subspaces constructed with the operator $\mathcal{S}$ and with the matrix A, we explain in section 3 how GMRES and FOM methods reduce the computation of the approximate solution of (1.1) to that of a small least squares Sylvester problem and a small Sylvester equation respectively. In section 4 we discuss some convergence properties of these methods. We show, in particular, the importance of the separation between the fields of values of A and B on the convergence behavior of GMRES. We also discuss the stagnation phenomenon in GMRES and its consequence on FOM. A rather surprising result discussed in section 5 concerns the issue of breakdown in the block-Arnoldi algorithm and its positive consequence on FOM. This leads to an efficient implementation of FOM. Numerical illustrations and comparisons with GMRES are given in section 6.

Throughout this note, the operator $\mathcal{S}$ is assumed to be nonsingular. This is equivalent to assuming that the spectrums of A and B are disjoint (see, e.g., [18]). Its norm will be defined by

$$\|\mathcal{S}\| = \max_{\|Y\|_{\mathrm{F}}=1} \big\|\mathcal{S}(Y)\big\|_{\mathrm{F}}, \tag{1.3}$$

where $\|Y\|_{\mathrm{F}} = \sqrt{\mathrm{trace}(Y^*Y)}$ denotes the Frobenius norm of the matrix Y associated with the inner product $(Y, Z)_{\mathrm{F}} = \mathrm{trace}(Z^*Y)$, $Y, Z \in \mathbb{R}^{N\times p}$. The adjoint of $\mathcal{S}$ with respect to the inner product $(.,.)_{\mathrm{F}}$ is the operator $\mathcal{S}^{\mathrm{adj}}$ that satisfies for all $Y, Z \in \mathbb{R}^{N\times p}$, $(\mathcal{S}(Y), Z)_{\mathrm{F}} = (Y, \mathcal{S}^{\mathrm{adj}}(Z))_{\mathrm{F}}$. It is easy to see that $\mathcal{S}^{\mathrm{adj}}$ is defined by $\mathcal{S}^{\mathrm{adj}}(Y) = A^*Y - YB^*$.

If x is a vector, its euclidean norm is $\|x\|_2 = \sqrt{x^*x}$. If E is a matrix, $E(:, i)$ denotes its ith column, $\mathrm{Span}\{E\}$ is the space spanned by the columns of E, $\mathrm{Null}(E)$, $\mathrm{Range}(E)$ and $\mathrm{rank}(E)$ denote respectively its null space, its range and its rank. When E is square, its field of values (or numerical range) is the set $\mathcal{F}(E) = \{(u^*Eu)/\|u\|_2^2\colon\ u \in \mathbb{C}^N,\ u \neq 0\}$. Moreover, when E is real symmetric (or complex hermitian) of order n, its

eigenvalues are denoted by $\lambda_{\max}(E) = \lambda_1(E) \geqslant \lambda_2(E) \geqslant \cdots \geqslant \lambda_n(E) = \lambda_{\min}(E)$. The identity (null) matrix of order n is denoted by $I_n(0_n)$ or just $I(0)$ if the order is clear from the context.

Finally, notice that the matrix A in the Sylvester equation (1.1) may be assumed to be nonsingular since (1.1) can always be written $(A-\lambda I)X - X(B-\lambda I) = C$, where λ is an arbitrary scalar. If A is singular, it suffices to choose λ outside the spectrum of A. The assumption of nonsingularity of A will be used in lemmas 5.2 and 5.3.

2. Relationship between the Krylov subspaces applied to $\mathcal{S}$ and to A

2.1. Krylov subspaces with $\mathcal{S}$ and with A

The mth block Krylov subspace $\mathcal{K}_m(\mathcal{S}, C)$ associated with C and the Sylvester operator $\mathcal{S}$ is defined by

$$\mathcal{K}_m(\mathcal{S}, C) = \operatorname{Span}\big\{C, \mathcal{S}(C), \ldots, \mathcal{S}^{m-1}(C)\big\},$$

where $\mathcal{S}^i(C) = \mathcal{S}(\mathcal{S}^{i-1}(C))$ and $\mathcal{S}^0(C) = C$.

The subspace $\mathcal{K}_m(A, C)$ is defined in an analogous way. The following lemma shows that the Krylov subspaces constructed with the matrix A and with the operator $\mathcal{S}$ are the same.

Lemma 2.1. For all $m \geqslant 1$, $\mathcal{K}_m(\mathcal{S}, C) = \mathcal{K}_m(A, C)$.

Proof. A direct calculation shows that

$$\forall k \geqslant 0 \quad \mathcal{S}^k(C) = \sum_{i=0}^{k} \binom{k}{i} A^{k-i} C (-B)^i \quad \text{with } \binom{k}{i} = \frac{k!}{i!(k-i)!}.$$

Therefore $\mathcal{K}_m(\mathcal{S}, C) \subset \mathcal{K}_m(A, C)$.

The converse inclusion can be proved by induction. The inclusion trivially holds when $m = 1$. Now, assume that $\mathcal{K}_k(A, C) \subset \mathcal{K}_k(\mathcal{S}, C)$, $k = 1, \ldots, m$. Then,

$$\begin{aligned} A^m C &= \sum_{i=0}^{m} \binom{m}{i} A^{m-i} C (-B)^i - \sum_{i=1}^{m} \binom{m}{i} A^{m-i} C (-B)^i \\ &= \mathcal{S}^m(C) - \sum_{i=1}^{m} \binom{m}{i} A^{m-i} C (-B)^i \in \mathcal{K}_{m+1}(\mathcal{S}, C). \end{aligned}$$

Thus $\mathcal{K}_{m+1}(A, C) \subset \mathcal{K}_{m+1}(\mathcal{S}, C)$. □

Lemma 2.1 is mentioned in [17] in the particular case where C is a full rank matrix.

The following lemma characterizes the space $\mathcal{K}_{m+1}(A, C)$. It will be used later (see proposition 5.4).

Lemma 2.2. For all $m \geqslant 1$, $\mathcal{K}_{m+1}(A, C) = \mathrm{Span}\{C, \mathcal{S}(\mathcal{K}_m(A, C))\}$, where

$$\mathcal{S}\big(\mathcal{K}_m(A, C)\big) = \big\{\mathcal{S}\big(CZ_0 + ACZ_1 + \cdots + A^{m-1}CZ_{m-1}\big),\ Z_0, \ldots, Z_{m-1} \in \mathbb{R}^{p\times p}\big\}.$$

Proof. By induction

$$\begin{aligned}\mathrm{Span}\big\{C, \mathcal{S}\big(\mathcal{K}_1(A, C)\big)\big\} &= \mathrm{Span}\big\{C, \mathcal{S}(CZ_0)\big\} = \mathrm{Span}\{C, ACZ_0 - CZ_0B\}\\ &= \mathrm{Span}\{C, AC\} = \mathcal{K}_2(A, C).\end{aligned}$$

Assume that $\mathrm{Span}\{C, \mathcal{S}(\mathcal{K}_m(A, C))\} = \mathcal{K}_{m+1}(A, C)$. Then

$$\begin{aligned}\mathrm{Span}\big\{C, \mathcal{S}\big(\mathcal{K}_{m+1}(A, C)\big)\big\} &= \mathrm{Span}\big\{C, \mathcal{S}\big(CZ_0 + \cdots + A^mCZ_m\big)\big\}\\ &= \mathrm{Span}\big\{C, \mathcal{S}\big(CZ_0 + \cdots + A^{m-1}CZ_{m-1}\big) + \mathcal{S}\big(A^mCZ_m\big)\big\}\\ &= \mathrm{Span}\big\{\mathcal{K}_{m+1}(A, C), \mathcal{S}\big(A^mCZ_m\big)\big\} = \mathcal{K}_{m+2}(A, C). \qquad \square\end{aligned}$$

2.2. *Block-Arnoldi applied to the Sylvester operator $\mathcal{S}$ and to the matrix A*

The block-Arnoldi method (see, e.g., [14,16]) can be applied either to the operator $\mathcal{S}$ or to the matrix A. In the former case, it constructs, after m steps, an orthonormal basis of the block-Krylov space $\mathcal{K}_m(\mathcal{S}, V_1^{(\mathcal{S})})$ where $V_1^{(\mathcal{S})} \in \mathbb{R}^{N\times p}$ with $(V_1^{(\mathcal{S})})^*V_1^{(\mathcal{S})} = I_p$ is the starting block. In the latter case, it constructs the standard block-Krylov space $\mathcal{K}_m(A, V_1^{(A)})$, where $V_1^{(A)} \in \mathbb{R}^{N\times p}$ with $(V_1^{(A)})^*V_1^{(A)} = I_p$. The corresponding algorithm is given in algorithm 2.3 where d denotes either $\mathcal{S}$ (the $\mathcal{S}$ version) or A (the A version):

Algorithm 2.3 (Block-Arnoldi on $\mathcal{S}$ or on A).
Choose $V_1^{(d)} \in \mathbb{R}^{N\times p}$ with $(V_1^{(d)})^*V_1^{(d)} = I_p$.
for $j = 1, 2 \ldots, m$
  $W_j = AV_j^{(d)} - V_j^{(d)}B$ if $d = \mathcal{S}$ and $W_j = AV_j^{(d)}$ if $d = A$
  for $i = 1, 2, \ldots, j$
    $H_{i,j}^{(d)} = (V_i^{(d)})^*W_j$
    $W_j := W_j - V_i^{(d)}H_{i,j}^{(d)}$
  end for i
  $W_j = V_{j+1}^{(d)}H_{j+1,j}^{(d)}$ (QR factorization)
end for j

It is clear that the only apparent difference between the $\mathcal{S}$ and A versions of algorithm 2.3 is the first step in loop j. If $d = \mathcal{S}$ then $W_j = \mathcal{S}(V_j^{(\mathcal{S})})$ whereas $W_j = AV_j^{(A)}$ if $d = A$. An important difference however is that the block sizes in the $\mathcal{S}$ version must be equal to p since otherwise the operation $\mathcal{S}(V_j^{(\mathcal{S})})$ cannot be done. This needs not necessarily be the case for the A version. We will go back to this point later.

Let us collect the matrices $V_j^{(d)}$ and $H_{i,j}^{(d)}$ constructed by algorithm 2.3 in the $N \times mp$ and $N \times (m+1)p$ orthonormal matrices

$$\mathcal{V}_m^{(d)} = \left[V_1^{(d)} \ \vdots \ V_2^{(d)} \ \vdots \ \dots \ \vdots \ V_m^{(d)} \right] \quad \text{and} \quad \mathcal{V}_{m+1}^{(d)} = \left[\mathcal{V}_m^{(d)} \ \vdots \ V_{m+1}^{(d)} \right]$$

and in the $mp \times mp$ and $(m+1)p \times mp$ block-upper Hessenberg matrices

$$\mathcal{H}_m^{(d)} = \begin{pmatrix} H_{1,1}^{(d)} & H_{1,2}^{(d)} & \dots & \dots & H_{1,m}^{(d)} \\ H_{2,1}^{(d)} & H_{2,2}^{(d)} & \dots & \dots & H_{2,m}^{(d)} \\ 0 & H_{3,2}^{(d)} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & H_{m-1,m}^{(d)} & H_{m,m}^{(d)} \end{pmatrix} \quad \text{and}$$

$$\overline{\mathcal{H}}_m^{(d)} = \begin{pmatrix} & & \mathcal{H}_m^{(d)} & \\ 0 & \dots & 0 & H_{m+1,m}^{(d)} \end{pmatrix}.$$

Then we have the following properties whose verification is straightforward.

Proposition 2.4. Assume that $V_1^{(\mathcal{S})} = V_1^{(A)}$. Then after m iterations of algorithm 2.3, the following properties hold:

- The columns of $\mathcal{V}_{m+1}^{(d)}$ are orthonormal and $V_i^{(\mathcal{S})} = V_i^{(A)}$ for all $i = 1, \dots, m+1$.
- $\forall i \ H_{i,i}^{(\mathcal{S})} = H_{i,i}^{(A)} - B$ and $\forall i \neq j \ H_{i,j}^{(\mathcal{S})} = H_{i,j}^{(A)}$.

The second property of proposition 2.4 shows that the matrices $\mathcal{H}_m^{(\mathcal{S})}$ and $\mathcal{H}_m^{(A)}$ differ only in their diagonal blocks.

The analysis done in this note assumes that exact arithmetic is used. We denote by

$$V_i = V_i^{(d)}, \qquad \mathcal{V}_m = \mathcal{V}_m^{(d)} \quad \text{and} \quad H_{i,j} = H_{i,j}^{(d)}, \quad i \neq j, \ \text{with } d = \mathcal{S} \text{ or } A.$$

Proposition 2.5. After m iterations of algorithm 2.3, the following properties hold:

- The columns of $\mathcal{V}_m$ form an orthonormal basis of $\mathcal{K}_m(\mathcal{S}, V_1) = \mathcal{K}_m(A, V_1)$.
- $\mathcal{V}_m^* V_{m+1} = 0$.
- $\mathcal{S}(\mathcal{V}_m) = \mathcal{V}_{m+1} \overline{\mathcal{H}}_m^{(\mathcal{S})}$ where $\mathcal{S}(\mathcal{V}_m) = [\mathcal{S}(V_1) \vdots \dots \vdots \mathcal{S}(V_m)]$.
- $A\mathcal{V}_m = \mathcal{V}_{m+1} \overline{\mathcal{H}}_m^{(A)}$.
- $\mathcal{V}_m^* \mathcal{S}(\mathcal{V}_m) = \mathcal{H}_m^{(\mathcal{S})}$, $\mathcal{V}_m^* A \mathcal{V}_m = \mathcal{H}_m^{(A)}$.

Let us denote by q with $q \leqslant p$ the rank of C. The full rank QR factorization of C gives:

$$C = V_1 \Lambda_1, \quad \text{with } V_1 \in \mathbb{R}^{N \times q}, \ V_1^* V_1 = I_q \text{ and } \Lambda_1 \in \mathbb{R}^{q \times p}, \ \operatorname{rank}(\Lambda_1) = q. \quad (2.1)$$

From V_1, the A version of algorithm 2.3 can be applied. We obtain the matrices $V_i \in \mathbb{R}^{N\times q}$, $H_{i,j} \in \mathbb{R}^{q\times q}$, $\mathcal{V}_m \in \mathbb{R}^{N\times mq}$, $\mathcal{H}_m^{(A)} \in \mathbb{R}^{mq\times mq}$. Note that, in this situation, the $\mathcal{S}$ version cannot be used.

The following proposition is a consequence of proposition 2.5. It will be used in the GMRES and FOM methods.

Proposition 2.6. Let $\mathcal{Y} \in \mathbb{R}^{mq\times p}$, then

$$\mathcal{S}(\mathcal{V}_m\mathcal{Y}) = \mathcal{V}_{m+1}\overline{\mathcal{S}}_m(\mathcal{Y}) \tag{2.2}$$

with

$$\overline{\mathcal{S}}_m(\mathcal{Y}) = \overline{\mathcal{H}}_m^{(A)}\mathcal{Y} - \begin{pmatrix}\mathcal{Y}\\ 0\end{pmatrix}B. \tag{2.3}$$

3. GMRES and FOM applied to Sylvester equations

Let $X_0 \in \mathbb{R}^{N\times p}$ be an initial guess to the solution of (1.1) and $R_0 = C - \mathcal{S}(X_0)$. Assume that $\operatorname{rank}(R_0) = q$, then the full rank QR factorization of R_0 gives $R_0 = V_1\Lambda_1$ with $V_1 \in \mathbb{R}^{N\times q}$, $q \leqslant p$, $V_1^*V_1 = I_q$ and $\Lambda_1 \in \mathbb{R}^{q\times p}$ upper triangular.

The GMRES method consists in finding the solution X_m^g to (1.1) that solves the problem

$$\text{Find}\quad X_m^g \in X_0 + \mathcal{K}_m(\mathcal{S}, V_1)\quad \text{such that } C - \mathcal{S}\big(X_m^f\big) \perp_{\mathrm{F}} \mathcal{S}\big(\mathcal{K}_m(\mathcal{S}, V_1)\big), \tag{3.1}$$

where the symbol $\perp_{\mathrm{F}}$ denotes the orthogonality with respect to the Frobenius inner product. Relation (3.1) is equivalent to the minimization problem $\|C - \mathcal{S}(X_m^g)\|_{\mathrm{F}} = \min_{Z\in X_0+\mathcal{K}_m(\mathcal{S},V_1)} \|C - \mathcal{S}(Z)\|_{\mathrm{F}}$.

The FOM method consists in finding the solution X_m^f to (1.1) that solves the problem

$$\text{Find}\quad X_m^f \in X_0 + \mathcal{K}_m(\mathcal{S}, V_1)\quad \text{such that } C - \mathcal{S}\big(X_m^f\big) \perp_2 \mathcal{K}_m(\mathcal{S}, V_1), \tag{3.2}$$

where the symbol $\perp_2$ denotes the orthogonality with respect to the Euclidean inner product. Relation (3.2) is equivalent to $Z^*(C - \mathcal{S}(X_m^f)) = 0$ for all $Z \in \mathcal{K}(\mathcal{S}, V_1)$.

Proposition 3.1. The solution X_m^g of problem (3.1) can be written

$$X_m^g = X_0 + \mathcal{V}_m\mathcal{Y}_m^g, \tag{3.3}$$

where $\mathcal{Y}_m^g \in \mathbb{R}^{mq\times p}$ solves the least squares Sylvester equation

$$\big\|\overline{\Lambda} - \overline{\mathcal{S}}_m\big(\mathcal{Y}_m^g\big)\big\|_{\mathrm{F}} = \min_{\mathcal{Y}\in\mathbb{R}^{mq\times p}} \big\|\overline{\Lambda} - \overline{\mathcal{S}}_m(\mathcal{Y})\big\|_{\mathrm{F}} \tag{3.4}$$

with

$$\overline{\Lambda} = \begin{pmatrix}\Lambda_1\\ 0\end{pmatrix} \in \mathbb{R}^{(m+1)q\times p} \quad \text{and} \quad \overline{\mathcal{S}}_m \text{ is defined in (2.3).}$$

Proposition 3.2. The solution X_m^f of problem (3.2) can be written

$$X_m^f = X_0 + \mathcal{V}_m \mathcal{Y}_m^f, \tag{3.5}$$

where $\mathcal{Y}_m^f \in \mathbb{R}^{mq\times p}$ solves the Sylvester equation

$$\mathcal{S}_m\big(\mathcal{Y}_m^f\big) = \Lambda \tag{3.6}$$

with

$$\Lambda = \begin{pmatrix} \Lambda_1 \\ 0 \end{pmatrix} \in \mathbb{R}^{mq\times p} \quad \text{and} \quad \mathcal{S}_m(\mathcal{Y}) = \mathcal{H}_m^{(A)}\mathcal{Y} - \mathcal{Y}B.$$

We see that the difference between GMRES and FOM is the solution $\mathcal{Y}_m^g$ and $\mathcal{Y}_m^f$ of (3.4) and (3.6), respectively. The least squares problem (3.4) has always a solution whereas the solution $\mathcal{Y}_m^f$ of (3.6) may not exist if the operator $\mathcal{S}_m$ is singular. The above discussion can be summarized in the following algorithm.

Algorithm 3.3 (GMRES/FOM for $\mathcal{S}(X) = C$).

1. Choose X_0 and compute $R_0 = C - \mathcal{S}(X_0)$. Compute $R_0 = V_1\Lambda_1$ using the full rank QR factorization.
2. From V_1 compute via the A version of algorithm 2.3 the matrices $\mathcal{V}_m$ and $\overline{\mathcal{H}}_m^{(A)}$.
3. **GMRES:** compute the approximate solution $X_m^g = X_0 + \mathcal{V}_m\mathcal{Y}_m^g$, where $\mathcal{Y}_m^g$ solves the minimization problem (3.4).
4. **FOM:** compute the approximate solution $X_m^f = X_0 + \mathcal{V}_m\mathcal{Y}_m^f$ where $\mathcal{Y}_m^f$ solves equation (3.6).

Using propositions 3.1 and 3.2, we see that the residuals $R_m^g = C - \mathcal{S}(X_m^g)$ and $R_m^f = C - \mathcal{S}(X_m^f)$ of GMRES and FOM can be expressed as follows:

$$R_m^g = \mathcal{V}_{m+1}\big(\overline{\Lambda} - \overline{\mathcal{S}}_m\big(\mathcal{Y}_m^g\big)\big), \tag{3.7}$$

$$\big\|R_m^g\big\|_{\mathrm{F}} = \big\|\overline{\Lambda} - \overline{\mathcal{S}}_m\big(\mathcal{Y}_m^g\big)\big\|_{\mathrm{F}}, \tag{3.8}$$

$$R_m^f = -V_{m+1}H_{m+1,m}Y_m^f, \tag{3.9}$$

$$\big\|R_m^f\big\|_{\mathrm{F}} = \big\|H_{m+1,m}Y_m^f\big\|_{\mathrm{F}}, \tag{3.10}$$

where Y_m^f is the last $q \times p$ block of $\mathcal{Y}_m^f$.

Moreover, it is clear that

$$\big\|R_m^g\big\|_{\mathrm{F}} \leqslant \big\|\overline{\Lambda} - \overline{\mathcal{S}}_m\big(\mathcal{Y}_m^f\big)\big\|_{\mathrm{F}} = \big\|H_{m+1,m}Y_m^f\big\|_{\mathrm{F}} = \big\|R_m^f\big\|_{\mathrm{F}}. \tag{3.11}$$

Remark 3.4.

1. To solve the small least squares Sylvester problem (3.4), we can, for instance, proceed as follows:

- First, we reduce the $p \times p$ matrix B to Schur form: $B = QTQ^*$, where Q is unitary and $T = (t_{ij})$ is upper triangular. Then we have

$$\forall \mathcal{Y} \in \mathbb{R}^{mq \times p} \quad \left\| \overline{\Lambda} - \overline{\mathcal{S}}_m(\mathcal{Y}) \right\|_{\mathrm{F}} = \left\| \overline{\Gamma} - \overline{\mathcal{T}}_m(\mathcal{Z}) \right\|_{\mathrm{F}},$$

  where $\overline{\Gamma} = \overline{\Lambda} Q \in \mathbb{C}^{(m+1)q \times p}$, $\mathcal{Z} = \mathcal{Y} Q \in \mathbb{C}^{mq \times p}$ and $\overline{\mathcal{T}}_m(\mathcal{Z}) = \overline{\mathcal{H}}_m^{(A)} \mathcal{Z} - \binom{\mathcal{Z}}{0} T$.
- Using the tensor product (see, e.g., [18]), it is easy to see that (3.4) is equivalent to the linear least squares system

$$\min_{z \in \mathbb{C}^{mqp}} \|\gamma - \mathcal{L} z\|_2 \tag{3.12}$$

  where

$$\begin{aligned} \gamma &= \left(\Gamma(:,1)^*, \ldots, \Gamma(:,p)^*\right)^*, \\ z &= \left(\mathcal{Z}(:,1)^*, \ldots, \mathcal{Z}(:,p)^*\right)^*, \\ \mathcal{L} &= I_{mq} \otimes \overline{\mathcal{H}}_m^{(A)} - B^* \otimes \begin{pmatrix} I_{mq} \\ 0 \end{pmatrix}, \end{aligned}$$

$$\mathcal{L} = \begin{pmatrix} L_{11} & & & \\ L_{21} & \ddots & & \\ \vdots & & \ddots & \\ L_{p1} & \cdots & & L_{pp} \end{pmatrix} \quad \text{with} \quad \begin{cases} L_{ii} = \overline{\mathcal{H}}_m^{(A)} - \bar{t}_{ii} \begin{pmatrix} I_{mq} \\ 0 \end{pmatrix}, \\ L_{ij} = -\bar{t}_{ji} \begin{pmatrix} I_{mq} \\ 0 \end{pmatrix}, \ i < j, \end{cases}$$

  where $\bar{t}_{kl}$ denotes the conjugate of t_{kl}.

  The problem (3.12) can be solved with standard methods based on the QR or the SVD factorizations (see [2;9, chapter V]).

2. The small Sylvester equation (3.6) can be solved using the methods developed in [1,8].

3. Because of the expense of storage requirements, algorithm 3.3 must be restarted periodically with $X_0 := X_m^g$ or $X_0 := X_m^f$, where X_m^g / X_m^f is the last computed approximate solution obtained with GMRES/FOM. This will be used in our numerical tests.

4. Convergence analysis

As we have just mentioned, algorithm 3.3 must be restarted periodically. The restart alleviates the problems due to storage requirements but does not ensure the convergence. In this section we analyze the behavior of algorithm 3.3 during one restart.

The following proposition shows that GMRES algorithm applied to (1.1) converges when the sets $\mathcal{F}(A)$ and $\mathcal{F}(B)$ are disjoint.

Proposition 4.1. Assume that $\mathcal{F}(A) \cap \mathcal{F}(B) = \emptyset$.
If $d = \operatorname{dist}(\mathcal{F}(A), \mathcal{F}(B)) := \min_{u\in\mathcal{F}(A), v\in\mathcal{F}(B)} |u - v| > 0$, then

$$\left\| R_m^g \right\|_{\mathrm{F}} \leqslant \left(1 - \frac{d^2}{\|\mathcal{S}\|^2}\right)^{m/2} \|R_0\|_{\mathrm{F}}.$$

Proof. We first note that since the matrices A and B are real, the condition $d > 0$ can be written

$$\gamma_1 := \lambda_{\min}\left(\frac{A + A^*}{2}\right) - \lambda_{\max}\left(\frac{B + B^*}{2}\right) > 0 \quad \text{or}$$
$$\gamma_2 := \lambda_{\min}\left(\frac{B + B^*}{2}\right) - \lambda_{\max}\left(\frac{A + A^*}{2}\right) > 0.$$

We have

$$\begin{aligned}\left\| R_m^g \right\|_{\mathrm{F}} = \min_{Z\in\mathcal{K}_m(\mathcal{S},R_0)} \left\| R_0 - \mathcal{S}(Z) \right\|_{\mathrm{F}} &\leqslant \min_{k\geqslant 0} \left\| (I - k\mathcal{S})^m (R_0) \right\|_{\mathrm{F}} \\ &\leqslant \min_{k\geqslant 0} \|I - k\mathcal{S}\|^m \|R_0\|_{\mathrm{F}}.\end{aligned}$$

But

$$\begin{aligned}\|I - k\mathcal{S}\|^2 &= \max_{\|X\|_{\mathrm{F}}=1} \left\| (I - k\mathcal{S})(X) \right\|_{\mathrm{F}}^2 \\ &= \max_{\|X\|_{\mathrm{F}}=1} \operatorname{trace}\left(\left(X - k\mathcal{S}(X)\right)^* \left(X - k\mathcal{S}(X)\right)\right) \\ &\leqslant \max_{\|X\|_{\mathrm{F}}=1} \left(1 - k \operatorname{trace}\left(X^*\mathcal{S}(X) + \left(\mathcal{S}(X)\right)^* X\right) + k^2 \left\| \mathcal{S}(X) \right\|_{\mathrm{F}}^2\right) \\ &\leqslant 1 - k \min_{\|X\|_{\mathrm{F}}=1} \operatorname{trace}\left(X^*\mathcal{S}(X) + \left(\mathcal{S}(X)\right)^* X\right) + k^2 \|\mathcal{S}\|^2.\end{aligned}$$

We now give a lower bound for $\min_{\|X\|_{\mathrm{F}}=1} \operatorname{trace}(X^*\mathcal{S}(X) + (\mathcal{S}(X))^* X)$. If $\|X\|_{\mathrm{F}} = 1$, then

$$\begin{aligned}\operatorname{trace}\left(\frac{X^*\mathcal{S}(X) + (\mathcal{S}(X))^* X}{2}\right) &= \operatorname{trace}\left(\frac{X^*(AX - XB) + (AX - XB)^* X}{2}\right) \\ &= \operatorname{trace}\left(X^* \frac{A + A^*}{2} X - X^* X \frac{B + B^*}{2}\right)\end{aligned}$$

and

$$\begin{aligned}\operatorname{trace}\left(X^* \frac{A + A^*}{2} X\right) &= \sum_i \lambda_i \left(X^* \frac{A + A^*}{2} X\right) \\ &\geqslant \lambda_{\min}\left(\frac{A + A^*}{2}\right) \sum_i \lambda_i\left(X^*X\right) = \lambda_{\min}\left(\frac{A + A^*}{2}\right),\end{aligned}$$

$$\operatorname{trace}\left(X^*X\frac{B+B^*}{2}\right)=\sum_i \lambda_i\left(X\frac{B+B^*}{2}X^*\right)$$
$$\leqslant \lambda_{\max}\left(\frac{B+B^*}{2}\right)\sum_i \lambda_i\left(X^*X\right)=\lambda_{\max}\left(\frac{B+B^*}{2}\right).$$

Hence $\|I-k\mathcal{S}\|^2 \leqslant 1-2k\gamma_1+k^2\|\mathcal{S}\|^2$. This expression is minimized by taking $k=\gamma_1/\|\mathcal{S}\|^2$ and then $\|I-k\mathcal{S}\|^2\leqslant 1-\gamma_1^2/\|\mathcal{S}\|^2$. It is clear that

$$\operatorname{trace}\left(\frac{X^*\mathcal{S}(X)+(\mathcal{S}(X))^*X}{2}\right)\leqslant\|\mathcal{S}\|,$$

hence $\gamma_1/\|\mathcal{S}\|\leqslant 1$.

We also have

$$\left\|R_m^g\right\|_{\mathrm{F}}=\min_{Z\in\mathcal{K}_m(R_0)}\left\|R_0-\mathcal{S}(Z)\right\|_{\mathrm{F}}\leqslant\min_{k>0}\left\|(I+k\mathcal{S})^m(R_0)\right\|_{\mathrm{F}}\leqslant\min_{k>0}\|I+k\mathcal{S}\|^m\|R_0\|_{\mathrm{F}}.$$

By repeating the argument above, we obtain $\|I+k\mathcal{S}\|^2\leqslant 1-\gamma_2^2/\|\mathcal{S}\|^2$. □

Remark 4.2.

1. When $B=0$ and $p=1$, then proposition 4 in [15] is recovered.

2. Proposition 4.1 shows that the ideal situation arises when the distance d is large (see table 1). We mention that the role played by the field of values in the convergence analysis of iterative methods have been pointed out by Eiermann [6].

3. To simplify the notation, we have considered one restart of algorithm 3.3. The general case may be done as follows:
   If we denote by $X_{0,s}^g$, $X_{m,s}^g$, $R_{m,s}^g=C-\mathcal{S}(X_{m,s}^g)$ respectively the initial guess, the approximate solution and the residual obtained at the sth restart of GMRES, then proposition 4.1 gives

$$\left\|R_{m,s}^g\right\|_{\mathrm{F}}\leqslant\rho^{m/2}\left\|R_{0,s}^g\right\|_{\mathrm{F}},\quad\text{with }\rho=1-\frac{d^2}{\|\mathcal{S}\|^2}.$$

   But $R_{0,s}^g=C-\mathcal{S}(X_{0,s}^g)=C-\mathcal{S}(X_{m,s-1}^g)=R_{m,s-1}^g$.
   Hence

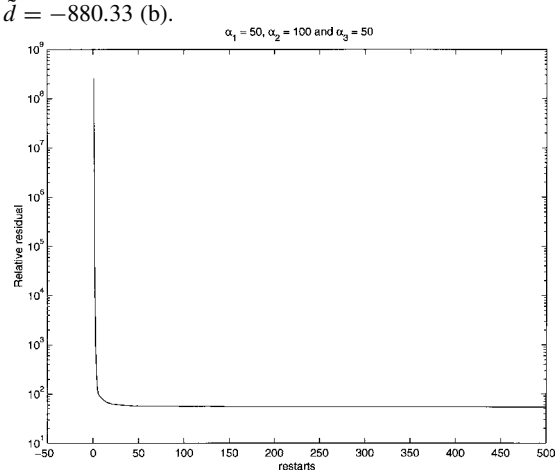
$$\left\|R_{m,s}^g\right\|_{\mathrm{F}}\leqslant\rho^{m/2}\left\|R_{m,s-1}^g\right\|_{\mathrm{F}}.$$

   Therefore the sequence $(\|R_{m,s}^g\|_{\mathrm{F}})_s$ should generally converge but not necessarily towards 0 (when $d=0$). What happens when it converges to a nonzero constant? This situation is referred to as the stagnation of GMRES. It means that for (eventually) large s, the norm of $R_{m,s}^g$ becomes constant. The consequences of stagnation are analyzed in the next subsection.

Table 1
GMRES applied to (4.7) with different separations between $\mathcal{F}(A)$ and $\mathcal{F}(B_\alpha)$ (a); (*) stagnation in the case $\tilde{d} = -880.33$ (b).

α	$\tilde{d} = \alpha e$	# of restarts	$\|R_m^g\|_2$
-1	-880.33	500 (*)	53.98 (*)
0	0.0	15	$2.33 \cdot 10^{-4}$
1	880.33	7	$1.68 \cdot 10^{-4}$
3	2640.99	4	$2.18 \cdot 10^{-4}$
6	5281.97	3	$1.89 \cdot 10^{-4}$

(a)



(b)

4.1. Stagnation in GMRES

In order to simplify the notation, we drop the subscript s and consider again one restart of algorithm 3.3 where the stagnation occurs. In other words, we consider the situation where $\|R_m^g\|_{\mathrm{F}} = \|R_0\|_{\mathrm{F}}$.

Proposition 4.3. If GMRES stagnates (i.e., $\|R_m^g\|_{\mathrm{F}} = \|R_0\|_{\mathrm{F}}$), then

$$R_0^* A R_0 = B R_0^* R_0, \tag{4.1}$$
$$R_0^* A V_i = 0, \quad i = 2, \ldots, m. \tag{4.2}$$

Moreover, in this case, $\mathcal{F}(A) \cap \mathrm{sp}(B) \neq \emptyset$ and $\Lambda \in \mathrm{Null}(\mathcal{S}_m^{\mathrm{adj}})$.

Proof. The stagnation means that (see proposition 3.1):

$$\|\Lambda_1\|_{\mathrm{F}} = \|R_0\|_{\mathrm{F}} = \left\|R_m^g\right\|_{\mathrm{F}} = \min_{\mathcal{Y} \in \mathbb{R}^{mq \times p}} \left\|V_1 \Lambda_1 - \mathcal{S}(\mathcal{V}_m \mathcal{Y})\right\|_{\mathrm{F}}$$

which is equivalent to the following

$$\begin{aligned}
&\mathrm{trace}\big((V_1\Lambda_1)^* \mathcal{S}(\mathcal{V}_m Z)\big) = 0 && \forall Z \in \mathbb{R}^{mq \times p},\\
&\mathrm{trace}\big((V_1\Lambda_1)^* (A\mathcal{V}_m Z - \mathcal{V}_m Z B)\big) = 0 && \forall Z \in \mathbb{R}^{mq \times p},\\
&\mathrm{trace}\big((V_1\Lambda_1)^* (A\mathcal{V}_m Z)\big) = \mathrm{trace}\big(B(V_1\Lambda_1)^* \mathcal{V}_m Z\big) && \forall Z \in \mathbb{R}^{mq \times p}.
\end{aligned}$$

Thus

$$(V_1\Lambda_1)^* A \mathcal{V}_m = B(V_1\Lambda_1)^* \mathcal{V}_m$$

or

$$(V_1\Lambda_1)^* A V_1 = B\Lambda_1^* \tag{4.3}$$

and

$$(V_1\Lambda_1)^* A V_i = 0, \quad i = 2, \ldots, m. \tag{4.4}$$

Hence relations (4.1) and (4.2).

As the $q \times p$ matrix Λ_1 has full rank, relation (4.4) implies that $V_1^* A V_i = 0$, $i = 2, \ldots, m$, or

$$V_1 \mathcal{S}(V_i) = H_{1,i} = 0, \quad i = 2, \ldots, m. \tag{4.5}$$

Relation (4.3) can be written $\Lambda_1^* H_{1,1}^{(A)} = B\Lambda_1^*$ which means that the Sylvester equation $U H_{11}^{(A)} = BU$ possesses the non-null solution $U = \Lambda_1^*$. Therefore the spectrum of $H_{11}^{(A)}$ intersects that of B and thus $\mathcal{F}(A) \cap \mathrm{sp}(B) \neq \emptyset$.

Moreover, from (4.3) and (4.5), we also obtain

$$\left(\mathcal{H}_m^{(A)}\right)^* \Lambda - \Lambda B^* = 0 \tag{4.6}$$

which means that $\Lambda \in \mathrm{Null}(\mathcal{S}_m^{\mathrm{adj}})$. □

An illustration of propositions 4.1 and 4.3 is given in table 1 where the solved Sylvester equation originates from the convection–diffusion operator (6.1) with the following parameters $\alpha_1 = \alpha_3 = 50$, $\alpha_2 = 100$, $m = 3$, $p = 14$, $l = mp = 42$ and $N = 200$. We will go back to this operator and the implementation issues in section 6. For this example, the matrices A and B satisfy

$$\lambda_{\min}\left(\frac{A + A^*}{2}\right) = -49.90, \qquad \lambda_{\max}\left(\frac{A + A^*}{2}\right) = 1.56\ 10^3,$$
$$\lambda_{\min}\left(\frac{B + B^*}{2}\right) = -8.40\ 10^2, \qquad \lambda_{\max}\left(\frac{B + B^*}{2}\right) = 40.16.$$

Let

$$e = \lambda_{\max}\left(\frac{B + B^*}{2}\right) - \lambda_{\min}\left(\frac{B + B^*}{2}\right) = 8.8033 \times 10^2$$

and

$$B_0 = B + \left(\lambda_{\max}\left(\frac{A + A^*}{2}\right) - \lambda_{\min}\left(\frac{B + B^*}{2}\right)\right) I_p.$$

It is clear that $dist(\mathcal{F}(A), \mathcal{F}(B_0) = 0$. Table 1 shows the results obtained when GMRES is applied to the Sylvester equation

$$AX - XB_\alpha = C \quad \text{with } B_\alpha = B_0 + \alpha e I_p, \tag{4.7}$$

where the parameter $\tilde{d} = \alpha e$ varies. We see from this table that the larger the distance between the field of values of A and B_α, the faster the convergence.

The following proposition shows that the convergence of GMRES and FOM methods are interrelated.

Corollary 4.4. If FOM converges then the same is true for GMRES. On the other hand, if GMRES stagnates, then FOM fails.

Proof. The proof of the first part follows from the fact that $\|R_m^g\|_{\mathrm{F}} \leqslant \|R_m^f\|_{\mathrm{F}}$ (see 3.11). The second part is a consequence of proposition 4.3: $\Lambda \in \mathrm{Null}(\mathcal{S}_m^{\mathrm{adj}}) = \mathrm{Range}(\mathcal{S}_m)^{\perp_{\mathrm{F}}}$, where $\perp_{\mathrm{F}}$ denotes the orthogonality with respect to the Frobenius inner product. □

Remark 4.5. It is important to notice that in the proposition above the initial residual R_0 and the parameter m are the same in FOM and GMRES methods. We will go back to this point in the numerical test section.

5. Breakdown in block-Arnoldi

In this section, we study some peculiarities of algorithm 3.3 when the block-Arnoldi algorithm breaks down. This arises at iteration m of algorithm 2.3 when $H_{m+1,m} = 0$. The case where $H_{m+1,m}$ is a nonzero singular matrix will also be studied.

5.1. *Case where* $H_{m+1,m} = 0$

From proposition 2.6 we have

$$\forall \mathcal{Y} \in \mathbb{R}^{mq\times p} \quad \mathcal{S}(\mathcal{V}_m\mathcal{Y}) = \mathcal{V}_{m+1}\overline{\mathcal{S}}_m(\mathcal{Y}) = \mathcal{V}_m\mathcal{S}_m(\mathcal{Y}).$$

Therefore the spectrum of the Sylvester operator $\mathcal{S}_m$ (i.e., the set of scalars λ such that $\mathcal{S}_m(\mathcal{Y}) = \lambda\mathcal{Y}$, $\mathcal{Y} \neq 0$) is included in that of $\mathcal{S}$. We conclude that $\mathcal{S}_m$ is nonsingular and therefore FOM is well defined. From (3.10) we have $\|R_m^f\|_{\mathrm{F}} = 0$ and from (3.8) we have

$$\|R_m^g\|_{\mathrm{F}} = \|\overline{\Lambda} - \overline{\mathcal{S}}_m(\mathcal{Y}_m^g)\|_{\mathrm{F}} = \|\Lambda - \mathcal{S}_m(\mathcal{Y}_m^g)\|_{\mathrm{F}} = 0 \quad \text{(since } \mathcal{S}_m \text{ is nonsingular)}.$$

We thus have the following proposition.

Proposition 5.1. If at iteration m of algorithm 2.3, $H_{m+1,m} = 0$, then the approximate solutions computed by GMRES or FOM are exact: $X_m^g = X_m^f = X$.

5.2. *Case where* $\dim(\mathrm{Null}(H_{m+1,m})) = k$ *with* $0 < k < q$

We have the following lemmas.

Lemma 5.2. Assume that A and $H_{j+1,j}$ are nonsingular for $j = 1, \ldots, m-1$ then

$$\dim\big(\mathrm{Span}\{R_0\} \cap A\mathcal{K}_m(A, R_0)\big) = \dim\big(\mathrm{Null}(H_{m+1,m})\big) = k,$$

where

$$A\mathcal{K}_m(A, R_0) = \mathrm{Span}\big\{AR_0, A^2R_0, \ldots, A^mR_0\big\}.$$

Proof. The proof is given in [12]. □

Lemma 5.3. Assume that A and $H_{j+1,j}$ are nonsingular for $j = 1, \ldots, m-1$ then

$$\dim\big(\mathcal{K}_{m+1}(A, R_0)\big) = (m+1)q - k.$$

Proof. As in lemma 2.2, we have

$$\mathcal{K}_{m+1}(A, R_0) = \mathrm{Span}\big\{R_0, A\mathcal{K}_m(A, R_0)\big\}.$$

Therefore

$$\begin{aligned}\dim\big(\mathcal{K}_{m+1}(A, R_0)\big) = {} & \dim\big(\mathrm{Span}\{R_0\}\big) + \dim\big(A\big(\mathcal{K}_m(A, R_0)\big)\big) \\ & - \dim\big(\mathrm{Span}\{R_0\}\big) \cap A\big(\mathcal{K}_m(A, R_0)\big).\end{aligned}$$

The proof follows by noticing that:

- $\dim(\mathrm{Span}\{R_0\}) = q$.
- $\dim(A(\mathcal{K}_m(A, R_0))) = mq$ (since A is nonsingular and $H_{j+1,j}$ nonsingular for $j < m$).
- $\dim(\mathrm{Span}\{R_0\}) \cap A(\mathcal{K}_m(A, R_0)) = \dim(\mathrm{Null}(H_{m+1,m})) = k$ (see lemma 5.2). □

For the FOM method, we have the following proposition.

Proposition 5.4. Assume that the FOM method is defined (e.g., $\mathcal{S}_m$ is nonsingular) and $H_{j+1,j}$ is nonsingular for $j < m$. Then

$$\mathrm{rank}\big(R_m^f\big) = \mathrm{rank}(H_{m+1,m}) = q - k.$$

Proof. From (3.9), we have $\mathrm{rank}(R_m^f) \leqslant \mathrm{rank}(H_{m+1,m})$.

If $\mathrm{rank}(R_m^f) = q - \tilde{k} < q - k$, then from lemmas 2.2 and 5.3, we obtain

$$\begin{aligned}(m+1)q - k &= \dim\big(\mathcal{K}_{m+1}(A, R_0)\big) = \dim\big(\mathrm{Span}\big\{R_0, \mathcal{S}\big(\mathcal{K}_m(A, R_0)\big)\big\}\big) \\ &= \dim\big(\mathrm{Span}\{R_0\}\big) + \dim\big(\mathcal{S}\big(\mathcal{K}_m(A, R_0)\big)\big) \\ &\quad - \dim\big(\mathrm{Span}\{R_0\}\big) \cap \mathcal{S}\big(\mathcal{K}_m(A, R_0)\big).\end{aligned}$$

As in the proof of lemma 5.3, we have

- $\dim(\mathrm{Span}\{R_0\}) = q$ and $\dim(\mathcal{S}(\mathcal{K}_m(A, R_0))) = mq$ (since $\mathcal{S}$ is nonsingular and $H_{j+1,j}$ nonsingular for $j < m$).
- $\dim(\mathrm{Span}\{R_0\}) \cap \mathcal{S}(\mathcal{K}_m(A, R_0)) \geqslant \tilde{k}$ (since $\mathrm{rank}(R_m^f) = q - \tilde{k}$ means that $\mathrm{Span}\{R_0\} \cap \mathcal{S}(\mathcal{K}_m(A, R_0))$ contains a vector space of dimension $\tilde{k}$).

Hence, the contradiction $(m+1)q - k \leqslant q + mq - \tilde{k}$. □

Unfortunately, the following example shows that the residual R_m^g computed by the GMRES method does not satisfy $\mathrm{rank}(R_m^g) = \mathrm{rank}(H_{m+1,m})$.

Example 5.5. Consider

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{pmatrix} \quad \text{and} \quad C = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

With $V_1 = 1/\sqrt{2}C$ and $\Lambda_1 = \sqrt{2}I_3$ and $m = 1$, we have

$$H_{11}^{(A)} = \frac{1}{2}\begin{pmatrix} 0 & 0 & 1 \\ 2 & 0 & 0 \\ 0 & 2 & 0 \end{pmatrix} \quad \text{and} \quad H_{21} = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\sqrt{3} \end{pmatrix}.$$

The residuals R_m^f and R_m^g satisfy

$$\sigma_i\big(R_m^f\big) = \big\{2.9047, 2.9005\text{e}{-}15, 2.1858\text{e}{-}15\big\} \quad \text{and}$$
$$\sigma_i\big(R_m^g\big) = \big\{0.7822, 0.1759, 0.0209\big\}$$

where σ_i denotes the singular values. We see that $\operatorname{rank}(H_{2,1}) = \operatorname{rank}(R_m^f) = 1$ but $\operatorname{rank}(R_m^g) = 3$.

6. Numerical illustrations

The aim of this section is to illustrate the benefits that the rank reduction (see proposition 5.4) can bring to the FOM method. The block-Arnoldi used in our numerical tests is the version developed by Ruhe (see [13, p. 117]) with elimination. The main steps of this version are written below using MATLAB style notations:

Algorithm 6.1 (Block-Arnoldi on A-Ruhe's version with elimination).
Choose $V \in \mathbb{R}^{N\times q}$ with $V_1^* V_1 = I_q$, an integer $l_{\max} > 0$ and $\varepsilon > 0$
Set $l = q$
for $j = q : l_{\max} - 1$
  $k = j - q + 1$
  if $k > l$ **exit**
  $w = AV(:, k)$
    for $i = 1, 2, \ldots, l$
      $H(i,k) = V(:,i)^* w$
      $w := w - H(i,k)V(:,i)$
    end for i
  if $\|w\|_2 > \varepsilon$
    $l = l + 1$

$H(l,k) = \|w\|_2$
$V(:,l) = w/\|w\|_2$
end if
end for j

The test "if $\|w\|_2 > \varepsilon$" allows to drop vectors that are not "numerically" linearly independent. We have used $\varepsilon = \sqrt{\varepsilon_{\text{mach}}}q \approx 1.5 \times 10^{-8}q$. At the end of algorithm 6.1, the matrix V contains l vectors "numerically" linearly independent.

We replace in step 2 of algorithm 3.3, algorithm 2.3 by algorithm 6.1. Both GMRES and FOM methods need initial guesses X_0, a maximum size l for the Krylov basis and an initial block size q. We take $X_0 = 0$. That is $R_0 = C$. The SVD factorization of R_0 gives $R_0 = U_l \Sigma U_r^*$ with U_l, U_r orthonormal and $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots)$ with $\sigma_1 \geqslant \sigma_2 \geqslant \cdots$ the singular values of R_0. We fix $\text{tol} = 10^{-12}\|R_0\|_2$ and choose the numerical rank q of R_0 such that $\sigma_q \geqslant \text{tol} > \sigma_{q+1}$. Then we start algorithm 6.1 with $V = U_l(:, 1:q)$ and use two options:

1. construction of a block Krylov subspace of *variable size* $l = mq$, where q is the numerical rank of R_0,
2. construction of a block Krylov subspace of *fixed size* $l = mp$, where p is the size of B.

The term "variable size" in the first construction follows from the fact that the parameter $q = \operatorname{rank}(R_0)$ may vary at each restart. In the second construction termed as "fixed size", the parameter p is fixed during the restarts. As in the first option, we start algorithm 6.1 with the $N \times q$ matrix V but continues the construction until the Krylov basis reaches the size $l = mp$. Since $q \leqslant p$ the second option generally allows the construction of larger Krylov basis.

Let $\Omega = [0:a] \times [0:b]$ and the family of grids

$$\Omega_h = \left\{(x_i, y_j) \in \Omega:\ x_i = ih_x,\ y_j = 1 - jh_y\right\}$$

where $h_x = 1/(n+1)$ and $h_y = 1/(p+1)$. We consider the convection–diffusion equation:

$$\begin{cases} -\Delta u + 2\alpha_1 u_x + 2\alpha_2 u_y - 2\alpha_3 u = f & \text{in } \Omega, \\ u = 0 & \text{on the boundary of } \Omega, \end{cases} \tag{6.1}$$

with non-negative parameters α_1, α_2 and α_3 and the function f is chosen such that $u(x,y) = xe^{xy}\sin(\pi x)\sin(\pi y)$ is the exact solution.

With a five-point stencil and centered finite differences discretization, we obtain the Sylvester equation (1.1) with

$$A = \frac{1}{h_x^2}\operatorname{tridiag}\left(-1 - \alpha_1 h_x, 2 - \alpha_3 h_x^2, -1 + \alpha_1 h_x\right),$$
$$B = \frac{-1}{h_y^2}\operatorname{tridiag}\left(-1 - \alpha_2 h_y, 2 - \alpha_3 h_y^2, -1 + \alpha_2 h_y\right)$$

and

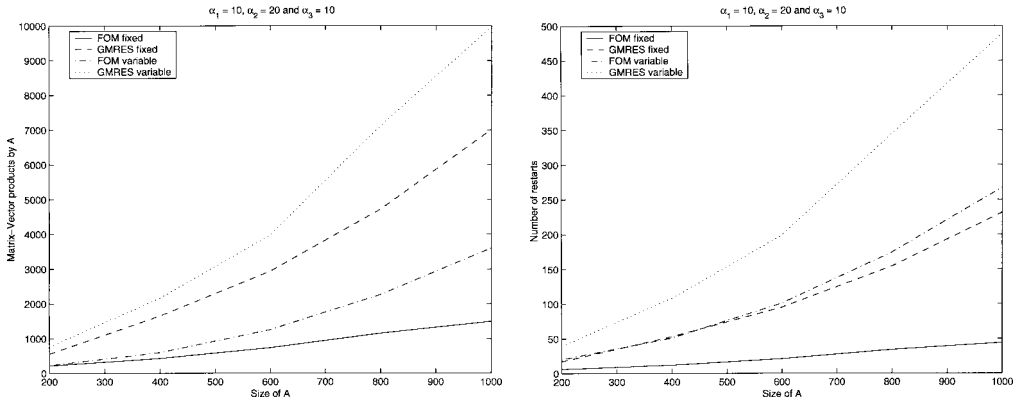
$$C(i, j) = f(x_i, y_j), \quad 1 \leqslant i \leqslant n \text{ and } 1 \leqslant j \leqslant p.$$

This convection diffusion problem is taken from [10]. It was also used in [7,17] but without the denominators h_x^2, h_y^2 in A, B and C was chosen randomly.

All the numerical tests have been done with the following parameters: $a = 10$, $b = 1$, $m = 3$, $p = 14$ and the stopping criteria

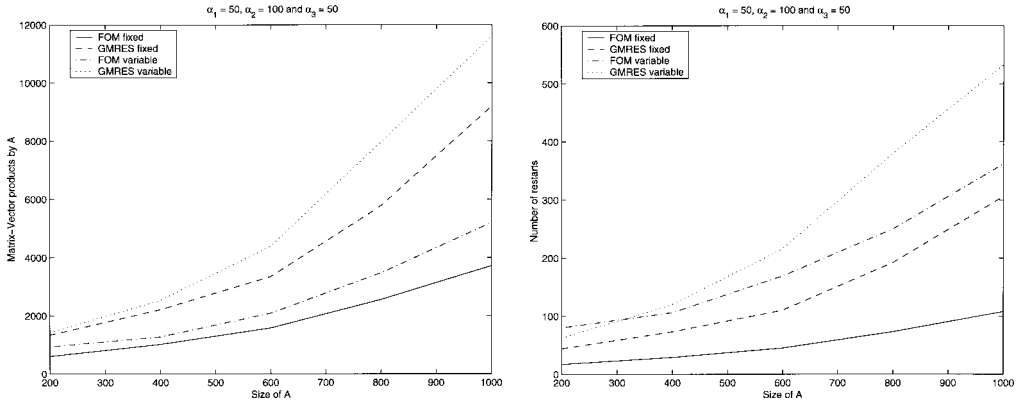
$$\left\|C - \mathcal{S}(X_m)\right\|_2 \leqslant \text{tol} \quad \text{with tol} = 10^{-12}\|C\|_2.$$

The convergence behaviors of GMRES and FOM (using the two options described above) are compared in figures 1 and 2 in terms of matrix–vector products (a) and number of restarts (b). Notice that the tests given with the number of restarts may be misleading since the size of the Krylov basis changes at each restart. For this reason, the



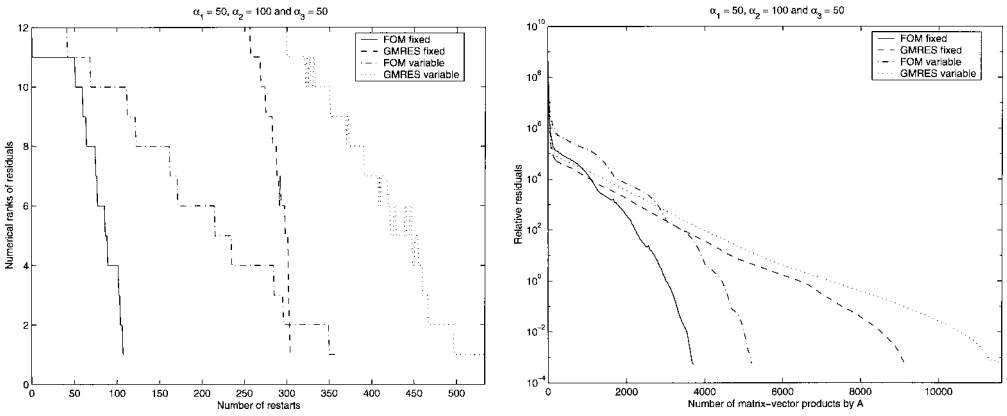
(a) (b)

Figure 1. Comparisons between GMRES and FOM with fixed and variable sizes.



(a) (b)

Figure 2. Comparisons between GMRES and FOM with fixed and variable sizes.



(a) (b)

Figure 3. Illustration of rank reduction (a) and behaviors of $\|R_m^f\|_2$ and $\|R_m^g\|_2$ (b). Examples with $N = 1000$.

tests with matrix–vector products are more reliable. Finally, note that the experiments shown in these figures do not contradict corollary 4.4 (see remark 4.5).

Figure 3 shows the numerical rank of the residuals R_m^f and R_m^g and the behaviors of their 2-norm during the iterations of FOM and GMRES.

7. Conclusion

The purpose of this note was the adaptation of block GMRES and block FOM methods to solve large Sylvester equations. These methods reduce the approximate solution of the large Sylvester equation to that of a small least squares Sylvester problem and a small Sylvester equation, respectively. The main contribution of this note can be summarized as follows: the rank of the residual matrix of FOM is equal to that of the matrix $H_{m+1,m}$ computed in the block-Arnoldi algorithm (provided that the matrices $H_{j+1,j}$, $j = 1, \ldots, m-1$, are nonsingular). In practice, this translates into the efficient use of a block-Arnoldi version (see algorithm 6.1) where the blocks do not necessarily have the same size. This rank property which is not satisfied by the GMRES method allows a numerical superiority of FOM over GMRES. On the other hand, the GMRES method enjoys nice mathematical properties. In particular, its convergence depends upon the separation between the fields of values of the matrices A and B that define the Sylvester equation.

References

[1] R. Bartels and G.W. Stewart, Algorithm 432: Solution of the matrix equation $AX + XB = C$, Comm. ACM 15 (1972) 820–826.

[2] A. Björck, *Numerical Methods for Least Squares Problems* (SIAM, Philadelphia, PA, 1996).

[3] D. Calvetti, B. Lewis and L. Reichel, On the solution of large Sylvester-obesrver equations, Numer. Linear Algebra Appl. 8 (2001) 435–451.
[4] B. Datta, Numerical linear algebra in control theory, Linear Algebra Appl. 198 (1994) 755–790.
[5] B. Datta and Y. Saad, Arnoldi methods for large Sylvester-like observer matrix equations, and an associated algorithm for partial spectrum assignment, Linear Algebra Appl. 154–156 (1991) 225–244.
[6] M. Eiermann, Fields of values and iterative methods, Linear Algebra Appl. 180 (1993) 167–197.
[7] A. El Guennouni, K. Jbilou and J. Riquet, Block Krylov subspace methods for solving large Sylvester equations, Preprint LMPA, No. 132 (2000), Université du Littoral, to appear in Numer. Algorithms.
[8] G.H. Golub, S. Nash and C. Van Loan, A Hessenberg–Schur method for the problem $AX + XB = C$, IEEE Trans. Automat. Control 24 (1979) 909–913.
[9] G.H. Golub and C.F. Van Loan, *Matrix Computation*, 2nd edn (Johns Hopkins Univ. Press, Baltimore, MD, 1989).
[10] D.Y. Hu and L. Reichel, Krylov subspace methods for the Sylvester equations, Linear Algebra Appl. 174 (1992) 283–314.
[11] I.M. Jaimoukha and E.M. Kasenally, Krylov subspace method for solving large Lyapunov equation, SIAM J. Numer. Anal. 31 (1994) 227–251.
[12] M. Robbé and M. Sadkane, Breakdown and stagnation in block FOM and block GMRES methods, Manuscript (2001).
[13] Y. Saad, Numerical solution of large Lyapunov equation, in: *Signal Processing, Scattering, Operator Theory and Numerical Methods* (Birkhäuser, Boston, 1990) pp. 503–511.
[14] Y. Saad, *Iterative Methods for Sparse Linear Systems* (PWS, Boston, 1996).
[15] Y. Saad and M.H. Schultz, GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, SIAM J. Sci. Statist. Comput. 7 (1986) 856–869.
[16] M. Sadkane, A block Arnoldi–Chebyshev method for computing the leading eigenpairs of large sparse unsymmetric matrices, Numer. Math. 64 (1993) 181–193.
[17] V. Simoncini, On the numerical solution of $AX - XB = C$, BIT 36 (1996) 814–830.
[18] G. Stewart and J. Sun, *Matrix Perturbation Theory* (Academic Press, San Diego, CA, 1990).