MAX-PLANCK-GESELLSCHAFT

Peter Benner Grece El Khoury Miloud Sadkane

On the Squared Smith Method for Large-Scale Stein Equations

MAX–PLANCK–INSTITUT
FÜR DYNAMIK KOMPLEXER
TECHNISCHER SYSTEME
MAGDEBURG

Max Planck Institute Magdeburg
Preprints

MPIMD/12-15 September 11, 2012

Impressum:

Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg

Publisher:
Max Planck Institute for Dynamics of Complex Technical Systems

Address:
Max Planck Institute for Dynamics of
Complex Technical Systems
Sandtorstr. 1
39106 Magdeburg

www.mpi-magdeburg.mpg.de/preprints

On the Squared Smith Method for Large-Scale Stein Equations

Peter Benner* Grece El Khoury† Miloud Sadkane‡

September 11, 2012

Abstract

A squared Smith type algorithm for solving large-scale discrete-time Stein equations is developed. The algorithm uses restarted Krylov spaces to compute approximations of the squared Smith iterations in low-rank factored form. Fast convergence results when very few iterations of the alternating direction implicit method are applied to the Stein equation beforehand. The convergence of the algorithm is discussed and its performance is demonstrated by several test examples.

Key words. Stein equation; squared Smith iteration; block-Arnoldi; low-rank factor; ADI iteration.

1 Introduction

The Stein equation

$$X - AXB^T = C, \tag{1}$$

where A, B and C are given real matrices and X is the unknown plays an important role in areas such as discrete-time control, model reduction of discrete-time dynamical systems, and restoration of images, see, e.g. [12, 14, 2, 9, 5]. It has a unique solution if and only if $\lambda\mu \neq 1$ for all $\big(\lambda, \mu\big) \in \Lambda(A) \times \Lambda(B)$, where $\Lambda(S)$ denotes the set of eigenvalues of the square matrix S. When the size of B is small compared to that of A, an efficient algorithm for solving (1) is based on the alternating direction implicit method (ADI) [5]. Note that when B is small and well conditioned, equation (1) can

*Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany. benner@mpi-magdeburg.mpg.de.

†Université de Brest, CNRS – UMR 6205, Laboratoire de Mathématiques de Bretagne Atlantique. 6, Av. Le Gorgeu. 29238 Brest Cedex 3. France.

‡Université de Brest, CNRS - UMR 6205, Laboratoire de Mathématiques de Bretagne Atlantique. 6, Av. Le Gorgeu. 29238 Brest Cedex 3. France. miloud.sadkane@univ-brest.fr.

be reduced to the classical Sylvester equation $AX - X\big(B^T\big)^{-1} = -C\big(B^T\big)^{-1}$ for which efficient methods already exist, see, e.g. [11, 20, 8, 16].

In the present paper we are interested in the case when A and B are large $n \times n$ matrices and C has the low-rank form $C = EF^T$, where E and F are $n \times p$ matrices with $p = \text{rank}(E) = \text{rank}(F) \ll n$. We assume that A and B are discrete-stable (their eigenvalues lie inside the open unit disk). Under this assumption, the unique solution of (1) is given by

$$X = \sum_{j=0}^{\infty} A^j EF^T \big(B^j\big)^T. \tag{2}$$

In fact, the existence of (2) is ensured under the weaker condition that the spectral radii $\rho(A)$ and $\rho(B)$ of A and B satisfy $\rho(A)\rho(B) < 1$. Note that if $\rho(A) < 1 < \rho(B) < 1/\rho(A)$, then we can find ξ such that $\rho(B) < \xi < 1/\rho(A)$ and replace A and B by ξA and B^T/ξ in equation (1).

We assume further that the norms of A^j and B^j decrease as j increases, which ensures that the solution can be approximated as $X \approx Z_1 Z_2^T$ where Z_1 and Z_2 have ranks much smaller than n, see Section 2.

The main aim of the present paper is to describe an effective squared Smith type algorithm for computing the factors Z_1 and Z_2. The squared Smith method is an improvement of Smith's method initially devised for solving Sylevester equations [21]. In its standard from, the squared Smith method converges quadratically, but necessitates, at each iteration, $n \times n$ matrix-matrix operations, which is the reason why it is mainly confined to matrix equations of small sizes. In this paper we show that these difficulties can be bypassed. We shall use a Krylov space with a restarting scheme to generate the squared Smith iterations in low rank factors. This leads to a computational cost which is linear in n, but the quadratic convergence of the original squared Smith method may not be maintained. However, the convergence can be made fast when a simple version of the ADI method, similar to the one proposed in [4], is applied beforehand. In the special case where $B = A$ and $E = F$, an adaptation of squared Smith is developed in [19]. We show that many ideas in this paper can be adapted here but the presence of $B \neq A$ and $E \neq F$ requires a careful implementation.

The paper is organized as follows. Section 2 gives sufficient conditions under which a low-rank solution may be expected and derives low-rank squared Smith iterations from a combination of the squared Smith method and a block variant of the Arnoldi algorithm. Section 3 estimates the error between the exact solution and its low rank approximation and expresses the residual in terms of quantities readily computable. The residual is used in Section 4 to develop a cheap restarting scheme to overcome the increase in computational cost and memory requirements for the block Arnoldi bases. The acceleration of convergence due to a simple version of ADI iterations is discussed in Section 5. Section 6 is devoted to numerical tests, and concluding remarks are given in Section 7.

2 Low-rank approximation

The solution (2) can be written as

$$X = \Big(E, AE, A^2E, \dots\Big)\Big(F, BF, B^2F, \dots,\Big)^T,$$

and due to the rank properties and the Cayley-Hamilton theorem, we have

$$\operatorname{rank}(X) \leq \min\Big(\operatorname{rank}\left(E, AE, \dots, A^{n-1}E\right), \operatorname{rank}\left(F, AF, \dots, B^{n-1}F\right)\Big). \tag{3}$$

The ranks in the right-hand sides of (3) can be much smaller than n, depending on the proprieties of A, B, E, and F. For example, if the columns of E (or F) span an invariant subspace of A (or B), then $\operatorname{rank}(X) \leq p$. If the norms of powers of A (or B) decrease rapidly, then $\operatorname{rank}(X)$ is small compared to n. In the general case, X can be decomposed as

$$X = X_{k,1} + X_{k,2},$$

where $X_{k,1} = \sum_{j=0}^{k-1} A^jEF^T(B^j)^T$ and $X_{k,2} = X - X_{k,1}$ and k is chosen such that $kp \leq n-1$. Since

$$X_{k,1} = \Big(E, AE, \dots, A^{k-1}E\Big)\Big(F, AF, \dots, A^{k-1}F\Big)^T,$$

it is clear that $\operatorname{rank}(X_{1,k}) \leq pk$. Let the singular values of X be labeled so that $\sigma_1(X) \geq \sigma_2(X) \ldots \geq \sigma_n(X)$. Then the Schmidt-Mirsky theorem [22] gives $\sigma_{kp+1}(X) \leq \|X - X_{k,1}\|$, and therefore

$$\begin{aligned} \sigma_{kp+1}(X) &\leq \left\| \sum_{j=k}^{\infty} A^jEF^T(B^j)^T \right\| \\ &= \|A^kX(B^k)^T\| \leq \|A^k\|\|B^k\|\|X\|. \end{aligned}$$

This yields the upper bound

$$\frac{\sigma_{kp+1}(X)}{\sigma_1(X)} \leq \|A^k\|\|B^k\|. \tag{4}$$

None of the bounds (3), (4) is sharp but they both show that a solution with a (numerical) low-rank can be expected provided that the norms of the powers of A or B decrease rapidly.

2.1 Low-rank squared Smith approximation

The solution X can be approximated by the partial sum

$$X_k = \sum_{j=0}^{2^k-1} A^jEF^T(B^T)^j \tag{5}$$

with large k.

We see that $X_0 = EF^T$ and for $k \geq 1$

$$\begin{aligned} X_k &= \sum_{j=0}^{2^{k-1}-1} A^j EF^T (B^T)^j + A^{2^{k-1}} \Bigg(\sum_{j=0}^{2^{k-1}-1} A^j EF^T (B^T)^j \Bigg) (B^T)^{2^{k-1}}, \\ &= X_{k-1} + A^{2^{k-1}} X_{k-1} (B^T)^{2^{k-1}} \end{aligned}$$

and

$$A^{2^k} = \Big(A^{2^{k-1}} \Big)^2, \;\; (B^T)^{2^k} = \Big((B^T)^{2^{k-1}} \Big)^2.$$

Hence, the squared Smith scheme can be written as

$$X_0 = EF^T, \;\; A_0 = A, \;\; B_0 = B,$$

$$X_k = X_{k-1} + A_{k-1} X_{k-1} B_{k-1}^T, \;\; A_k = A_{k-1}^2, \;\; B_k = B_{k-1}^2, \;\; k \geq 1. \tag{6}$$

It is clear that (6) should not be used as such since the matrix sequences A_k and B_k are large and dense. However, from (5) we have

$$X_k = \big(E, AE, \ldots, A^{2^k-1}E\big)\big(F, BF, \ldots, B^{2^k-1}F\big)^T \tag{7}$$

and then

$$X_k \approx Z_k^E (Z_k^F)^T, \tag{8}$$

where Z_k^E and Z_k^F are matrices of small rank that can be constructed from the Krylov spaces

$$\begin{aligned} \mathcal{K}_k(A, E) &= \text{range}\big(E, AE, \ldots, A^{2^k-1}E\big), \\ \mathcal{K}_k(A, F) &= \text{range}\big(F, BF, \ldots, B^{2^k-1}F\big). \end{aligned}$$

A natural way to compute Z_k^E and Z_k^F is the block Arnoldi algorithm applied to A and B and starting with E and F, respectively. The following algorithm is the version applied to A and started with E.

For the implementation issues, see the discussion after Algorithm 2.

For $j = 1, \ldots, 2^m$, let

$$\mathbb{Q}_j^E = (Q_1^E, \; \ldots, \; Q_j^E), \;\; \mathbb{Q}_{j+1}^E = (\mathbb{Q}_j^E \quad Q_{j+1}^E),$$

$$\mathbb{H}_j^E = (H_{i,l}^E)_{1 \leq i,l \leq j}, \;\; \underline{\mathbb{H}}_j^E = \begin{pmatrix} \mathbb{H}_j^E \\ H_{j+1,j}^E I_j^T \end{pmatrix},$$

where $I_j^T = (0\,0 \; \ldots \; I)$ and 0 and I denote the zero and identity matrices of appropriate sizes.

Algorithm 1 Block Arnoldi method

```
INPUT: $A \in \mathbb{R}^{n,n}$, $E \in \mathbb{R}^{n,p}$, an integer $m$ such that $2^m \ll n$.
OUTPUT: Arnoldi basis $(Q_1^E, \dots, Q_{2^m+1}^E)$, and blocks $H_{i,j}^E$ of corresponding Hessenberg matrix.
$QR$ factorize $E = Q_1^E R_1^E$
for $j = 1, \dots, 2^m$ do
    $W_j^E = AQ_j^E$
    for $i = 1, \dots, j$ do
        $H_{i,j}^E = (Q_i^E)^T W_j^E$
        $W_j^E := W_j^E - Q_i^E H_{i,j}^E$
    end fori
    $QR$ factorize $W_j^E = Q_{j+1}^E H_{j+1,j}^E$
end forj
```

It is known (see, e.g. [17]) that $\mathbb{H}_j^E$ is block upper Hessenberg and the columns of $\mathbb{Q}_{2^m}^E$ form an orthonormal basis of $\mathbb{K}_m(A, Q_1^E)$ if no deflation occurs (which we assume for clarity of presentation — in the actual implementation, this is taken care of). From Algorithm 1 and the equalities above we obtain

$$A\mathbb{Q}_j^E = \mathbb{Q}_{j+1}^E \underline{\mathbb{H}}_j^E. \tag{9}$$

We will often need operations of type $A_l \mathbb{Q}_j^E$ where $A_l = A_{l-1}^2 = A^{2^l}$ is defined in (6). From (9) we have

$$A_l \mathbb{Q}_j^E = \underbrace{A(A(\dots(A\,\mathbb{Q}_j^E)))}_{2^l \ \text{times}} = \mathbb{Q}_{j+2^l}^E \prod_{k=j+2^l-1}^{j} \underline{\mathbb{H}}_k^E. \tag{10}$$

The same algorithm applied to B and started with F leads to similar formulas as above. We will refer to these by the superscript F.

The approximation (8) is then obtained as follows:

$$X_0 = EF^T \equiv Z_0^E Z_0^{F^T} \text{ with } Z_0^E = Q_1^E R_1^E \text{ and } Z_0^F = Q_1^F R_1^F, \tag{11}$$

$$X_1 = X_0 + AX_0B^T = (E, AE)(F, BF)^T \tag{12}$$

with

$$\begin{aligned} (E, AE) &= (Q_1^E R_1^E, AQ_1^E R_1^E) \\ &= (Q_1^E R_1^E, \mathbb{Q}_2^E \underline{\mathbb{H}}_1^E R_1^E) \\ &= \mathbb{Q}_2^E \left(\begin{pmatrix} R_1^E \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_1^E R_1^E \right). \end{aligned}$$

Similarly, for (F, BF) we have

$$(F, BF) = \mathbb{Q}_2^F \left(\begin{pmatrix} R_1^F \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_1^F R_1^F \right).$$

The low-rank approximation of (E, AE) and (F, AF) are then obtained through SVDs of the small matrices [1] $\left(\begin{pmatrix} R_1^E \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_1^E R_1^E \right)$ and $\left(\begin{pmatrix} R_1^F \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_1^F R_1^F \right)$ obtained by deleting the singular values which are smaller than a threshold tol_1:

$$\begin{aligned} \left(\begin{pmatrix} R_1^E \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_1^E R_1^E \right) &= U_1^E S_1^E (V_1^E)^T + \Delta_1^E, && (13) \\ \left(\begin{pmatrix} R_1^F \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_1^F R_1^F \right) &= U_1^F S_1^F (V_1^F)^T + \Delta_1^F, && (14) \end{aligned}$$

with

$$\|\Delta_1^E\| < \mathrm{tol}_1 \quad \text{and} \quad \|\Delta_1^{\mathrm{F}}\| < \mathrm{tol}_1, \tag{15}$$

where the matrices U_1^E, U_1^F, V_1^E and V_1^F have orthonormal columns, and S_1^E and S_1^F are diagonal matrices whose diagonals contain the singular values which are larger than tol_1. Here and throughout the paper, the symbol $\| \ \|$ denotes the spectral norm.

Care must be taken when deleting the smaller singular values to make the operation $(V_1^E)^T V_1^F$ possible. In our implementation, the number of singular values deleted equals the maximum of the number of singular values that are smaller than tol_1 in both $\left(\begin{pmatrix} R_1^E \\ 0 \end{pmatrix}, \mathbb{H}_1^E R_1^E \right)$ and $\left(\begin{pmatrix} R_1^F \\ 0 \end{pmatrix}, \mathbb{H}_1^F R_1^F \right)$. Then we have the first low-rank approximation:

$$X_1 \approx Z_1^E \left(Z_1^F \right)^T \text{ with } Z_1^E = \mathbb{Q}_2^E U_1^E S_1^E (V_1^E)^T V_1^F \text{ and } Z_1^F = \mathbb{Q}_2^F U_1^F S_1^F. \tag{16}$$

Note that the choice $Z_1^E = \mathbb{Q}_2^E U_1^E S_1^E (V_1^E)^T$ and $Z_1^F = \mathbb{Q}_2^F U_1^F S_1^F (V_1^F)^T$ is not recommended since the numbers of columns of the factors Z_k^E and Z_k^F are intended to increase at each iteration k, see Algorithm 2.

The same procedure leads us to the second step:

$$\begin{aligned} X_2 &= X_1 + A_1 X_1 B_1^T \\ &\approx Z_1^E {Z_1^F}^T + A_1 Z_1^E {Z_1^F}^T B_1^T \\ &= (Z_1^E, A_1 Z_1^E) \ (Z_1^F, B_1 Z_1^F)^T \end{aligned}$$

with

$$\begin{aligned} (Z_1^E, A_1 Z_1^E) &= (\mathbb{Q}_2^E U_1^E S_1^E (V_1^E)^T V_1^F, A^2 \mathbb{Q}_2^E U_1^E S_1^E (V_1^E)^T V_1^F) \\ &= (\mathbb{Q}_2^E U_1^E S_1^E (V_1^E)^T V_1^F, \mathbb{Q}_4^E \underline{\mathbb{H}}_3^E \underline{\mathbb{H}}_2^E U_1^E S_1^E (V_1^E)^T V_1^F) \\ &= \mathbb{Q}_4^E \left(\begin{pmatrix} U_1^E S_1^E (V_1^E)^T V_1^F \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_3^E \underline{\mathbb{H}}_2^E U_1^E S_1^E (V_1^E)^T V_1^F \right). \end{aligned}$$

[1] A cheaper alternative is the rank revealing QR algorihtm, see e.g. [1]

and similarly for $(Z_1^F, B_1 Z_1^F)$:

$$(Z_1^F, B_1 Z_1^F) = \mathbb{Q}_4^F \left(\begin{pmatrix} U_1^F S_1^F \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_3^F \underline{\mathbb{H}}_2^F U_1^F S_1^F \right).$$

As before the SVDs

$$\left(\begin{pmatrix} U_1^E S_1^E (V_1^E)^T V_1^F \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_3^E \underline{\mathbb{H}}_2^E U_1^E S_1^E (V_1^E)^T V_1^F \right) = U_2^E S_2^E (V_2^E)^T + \Delta_2^E, \tag{17}$$

$$\left(\begin{pmatrix} U_1^F S_1^F \\ 0 \end{pmatrix}, \underline{\mathbb{H}}_3^F \underline{\mathbb{H}}_2^F U_1^F S_1^F \right) = U_2^F S_2^F (V_2^F)^T + \Delta_2^F, \tag{18}$$

with a threshold tol_2 such that

$$\|\Delta_2^E\| < \mathrm{tol}_2 \quad \text{and} \quad \|\Delta_2^{\mathrm{F}}\| < \mathrm{tol}_2 \tag{19}$$

lead to the second low-rank approximation:

$$X_2 \approx Z_2^E {Z_2^F}^T \text{ with } Z_2^E = \mathbb{Q}_4^E U_2^E S_2^E (V_2^E)^T V_2^F \text{ and } Z_2^F = \mathbb{Q}_4^F U_2^F S_2^F. \tag{20}$$

More generally, at step k we have:

$$X_k \approx (Z_{k-1}^E, A_{k-1} Z_{k-1}^E)(Z_{k-1}^F, B_{k-1} Z_{k-1}^F)^T$$

with

$$\begin{gathered} (Z_{k-1}^E, A_{k-1} Z_{k-1}^E) = \\ (\mathbb{Q}_{2^{k-1}}^E U_{k-1}^E S_{k-1}^E (V_{k-1}^E)^T V_{k-1}^F, A^{2^{k-1}} \mathbb{Q}_{2^{k-1}}^E U_{k-1}^E S_{k-1}^E (V_{k-1}^E)^T V_{k-1}^F) = \\ \mathbb{Q}_{2^k}^E \left(\begin{pmatrix} U_{k-1}^E S_{k-1}^E (V_{k-1}^E)^T V_{k-1}^F \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}_j^E U_{k-1}^E S_{k-1}^E (V_{k-1}^E)^T V_{k-1}^F \right) \end{gathered}$$

and

$$\begin{aligned} (Z_{k-1}^F, B_{k-1} Z_{k-1}^F) &= (\mathbb{Q}_{2^{k-1}}^F U_{k-1}^F S_{k-1}^F, B^{2^{k-1}} \mathbb{Q}_{2^{k-1}}^F U_{k-1}^F S_{k-1}^F) \\ &= \mathbb{Q}_{2^k}^F \left(\begin{pmatrix} U_{k-1}^F S_{k-1}^F \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}_j^F U_{k-1}^F S_{k-1}^F \right). \end{aligned}$$

Then we compute, with a tolerance $\mathrm{tol_k}$, the reduced SVDs

$$\begin{gathered} \left(\begin{pmatrix} U_{k-1}^E S_{k-1}^E (V_{k-1}^E)^T V_{k-1}^F \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}_j^E U_{k-1}^E S_{k-1}^E (V_{k-1}^E)^T V_{k-1}^F \right) = \\ U_k^E S_k^E (V_k^E)^T + \Delta_k^E, \end{gathered} \tag{21}$$

$$\left(\begin{pmatrix} U_{k-1}^F S_{k-1}^F \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}_j^F U_{k-1}^F S_{k-1}^F \right) = U_k^F S_k^F (V_k^F)^T + \Delta_k^F \tag{22}$$

with

$$\|\Delta_k^E\| < \mathrm{tol_k} \text{ and } \|\Delta_\mathrm{k}^\mathrm{F}\| < \mathrm{tol_k} \tag{23}$$

and we obtain the low rank approximation

$$X_k \approx Z_k^E (Z_k^F)^T \text{ with } Z_k^E = \mathbb{Q}_{2^k}^E U_k^E S_k^E (V_k^E)^T V_k^F \text{and } Z_k^F = \mathbb{Q}_{2^k}^F U_k^F S_k^F. \tag{24}$$

3 Error and residual estimates

In this section we estimate the error

$$X - Z_k^E (Z_k^F)^T \tag{25}$$

between the exact and the computed solution and the residual

$$\Gamma_k = EF^T + AZ_k^E (Z_k^F)^T B^T - Z_k^E (Z_k^F)^T \tag{26}$$

associated with the computed solution.

The estimates in Propositions 1–4, though complicated to compute, will provide a reasonable indication of the error due to the Krylov space approximation and the SVD truncations. Proposition 4 shows that the norm of the residual can be computed at a lower cost. As we will see in Algorithm 2, it will be used to stop and/or restart the iterations.

3.1 Estimation of the error

From (2) and (5) we obtain

$$X - X_k = \sum_{j \geq 2^k} A^j EF^F (B^j)^T. \tag{27}$$

Since the eigenvalues of A lie in the open unit disk, they are certainly inside some circle $|z| = \rho_a$ of center 0 and radius $\rho_a < 1$. The Cauchy integral formula [10] gives

$$A^j = \frac{1}{2i\pi} \int_{|z|=\rho_a} z^j (zI - A)^{-1} dz = \frac{\rho_a^{j+1}}{2\pi} \int_0^{2\pi} e^{i(j+1)\theta} (\rho_a e^{i\theta} I - A)^{-1} d\theta$$

and a similar formula for B^j can be derived. Thus

$$\|A^j\| \leq C_a \rho_a^{j+1}, \quad \|B^j\| \leq C_b \rho_b^{j+1} \tag{28}$$

with

$$C_a = \max_{0 \leq \theta \leq 2\pi} \|(\rho_a e^{i\theta} I - A)^{-1}\|, \quad C_b = \max_{0 \leq \theta \leq 2\pi} \|(\rho_b e^{i\theta} I - B)^{-1}\|.$$

Taking the norm in (27) and using (28) we obtain the following proposition.

Proposition 1. *For $k \geq 0$ we have*

$$\|X - X_k\| \leq \frac{C_a C_b \|EF^T\|}{1 - \rho_a \rho_b} \big(\rho_a \rho_b\big)^{2^k+1}.$$

This proposition holds true if we only assume that $\rho_a \rho_b < 1$ and it clearly shows that $\|X - X_k\|$ tends to zero as k increases. However, the convergence may be slowed down if $\rho_a \rho_b$ is very close to 1 or if the constants C_a or C_b are too large. In the first case, a simple variant of the ADI iteration can be used to minimize the spectral radii, see Section 5. In the second case, the pseudospectra of A or B may significantly protrude from the unit circle [23] and a low-rank approximate solution may not exist.

Next, we estimate the error between X_k and $Z_k^E (Z_k^F)^T$.

Proposition 2. *For $k \geq 0$ we have*

$$\|X_k - Z_k^E (Z_k^F)^T\| \leq \mu_k$$

with $\mu_0 = 0$ and for $k \geq 1$

$$\mu_k \leq \delta_k + \left\| \left(\begin{pmatrix} I \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}_j^E \right) \right\| \left\| \left(\begin{pmatrix} I \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}_j^F \right) \right\| \delta_{k-1} \tag{29}$$

with

$$\delta_0 = 0 \quad \textit{and} \quad \textit{for} \quad k \geq 1, \quad \delta_k = \big(\|S_k^E\| + \|S_k^F\|\big)\mathrm{tol_k} + \mathrm{tol_k^2},$$

where $\mathrm{tol}_1, \mathrm{tol}_2, \ldots$ *are the SVD thresholds defined in (15), (19), ..., (23).*

Proof. For $k = 0$, the bound is satisfied since $X_0 = EF^T = Z_0^E (Z_0^F)^T$.

For $k = 1$, we have from (13) and (14)

$$\begin{aligned} (E, AE) &= \mathbb{Q}_2^E \big(U_1^E S_1^E (V_1^E)^T + \Delta_1^E\big), \\ (F, BF) &= \mathbb{Q}_2^F \big(U_1^F S_1^F (V_1^F)^T + \Delta_1^F\big). \end{aligned}$$

Hence

$$\begin{aligned} X_1 &= (E, AE)(F, BF)^T \\ &= \mathbb{Q}_2^E \big(U_1^E S_1^E (V_1^E)^T + \Delta_1^E\big) \big(U_1^F S_1^F (V_1^F)^T + \Delta_1^F\big)^T (\mathbb{Q}_2^F)^T \\ &= Z_1^E (Z_1^F)^T + \widetilde{X}_1 \end{aligned}$$

with

$$\widetilde{X}_1 = \mathbb{Q}_2^E \Big(U_1^E S_1^E (V_1^E)^T (\Delta_1^F)^T + \Delta_1^E V_1^F S_1^F (U_1^F)^T + \Delta_1^E (\Delta_1^F)^T \Big) (\mathbb{Q}_2^F)^T$$

and hence

$$\|\widetilde{X}_1\| \leq \big(\|S_1^E\| + \|S_1^F\|\big)\mathrm{tol}_1 + \mathrm{tol}_1^2 = \delta_1,$$

which shows the proposition for $k = 1$.

The general case is straightforward but tedious. From (21)–(23) we have

$$(E, AE, \ldots, A^{2^{k-1}}E) =$$

$$\mathbb{Q}^E_{2^k}\left(\begin{pmatrix} U^E_{k-1}S^E_{k-1}(V^E_{k-1})^T \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}^E_j U^E_{k-1}S^E_{k-1}(V^E_{k-1})^T\right) +$$

$$\mathbb{Q}^E_{2^k}\left(\begin{pmatrix} \Delta^E_{k-1} \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}^E_j \Delta^E_{k-1}\right)$$

and

$$\left(F, BF, \ldots, B^{2^{k-1}}F\right) =$$

$$\mathbb{Q}^F_{2^k}\left(\begin{pmatrix} U^F_{k-1}S^F_{k-1}(V^F_{k-1})^T \\ 0 \end{pmatrix} \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}^F_j U^F_{k-1}S^F_{k-1}(V^F_{k-1})^T\right) +$$

$$\mathbb{Q}^F_{2^k}\left(\begin{pmatrix} \Delta^F_{k-1} \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}^F_j \Delta^F_{k-1}\right).$$

Hence

$$\begin{aligned} X_k &= \left(E, AE, \ldots, A^{2^{k-1}}E\right)\left(F, BF, \ldots, B^{(2^{k-1}}F\right)^T \\ &= Z^E_k (Z^F_k)^T + \widetilde{X}_k \end{aligned}$$

with

$$\widetilde{X}_k = \mathbb{Q}^E_{2^k}\left(\Upsilon^{E,F}_0 + \Upsilon^E_1\left(\Upsilon^F_2\right)^T + \Upsilon^E_2\left(\Upsilon^F_1\right)^T + \Upsilon^E_2\left(\Upsilon^F_2\right)^T\right)(\mathbb{Q}^F_{2^k})^T,$$

where

$$\Upsilon^{E,F}_0 = U^E_k S^E_k (V^E_k)^T (\Delta^F_k)^T + \Delta^E_k V^F_k S^F_k (U^F_k)^T + \Delta^E_k (\Delta^F_k)^T$$

and for $G = E$ or F

$$\begin{aligned} \Upsilon^G_1 &= \left(\begin{pmatrix} I \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}^G_j\right)\begin{pmatrix} U^G_{k-1}S^G_{k-1}(V^G_{k-1})^T & 0 \\ 0 & U^G_{k-1}S^G_{k-1}(V^G_{k-1})^T \end{pmatrix} \\ \Upsilon^G_2 &= \left(\begin{pmatrix} I \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}^G_j\right)\begin{pmatrix} \Delta^G_{k-1} & 0 \\ 0 & \Delta^G_{k-1} \end{pmatrix}. \end{aligned}$$

We clearly have

$$\begin{aligned}
\|\Upsilon_0^{E,F}\| &\leq (\|S_k^E\| + \|S_k^F\|)\mathrm{tol_k} + \mathrm{tol_k^2} = \delta_\mathrm{k}, \\
\|\Upsilon_1^G\| &\leq \left\| \left(\begin{pmatrix} I \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \mathbb{H}_j^G \right) \right\| \|S_{k-1}^G\|, \\
\|\Upsilon_2^G\| &\leq \left\| \left(\begin{pmatrix} I \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \mathbb{H}_j^G \right) \right\| \mathrm{tol_{k-1}},
\end{aligned}$$

from which the proposition follows. □

By combining Propositions 1 and 2 we obtain

Proposition 3. *For $k \geq 0$ we have*

$$\|X - Z_k^E \left(Z_k^F\right)^T\| \leq \frac{C_a C_b \|EF^T\|}{1 - \rho_a \rho_b} \left(\rho_a \rho_b\right)^{2^k+1} + \mu_k.$$

When the constant $C_a C_b$ is not large and $\rho_a \rho_b$ is not close to 1, the proposition shows that for large k, the error and the residual norms behave respectively as μ_k and $\left\|EF^T + AX_k B^T - X_k\right\|$.

Remark 3.1. *In the special case where $B = A$ and $F = E$, the factor Z_k^E is entirely defined from $\left(E, AE, \ldots, A^{2^{k-1}} E\right)$ and this helps to improve the error estimate. To illustrate this point, consider the simple case $k = 1$. Then, from (13)–(15) we have, when $B = A$ and $F = E$*

$$\left(E, AE\right) = \mathbb{Q}_2^E \left(U_1^E S_1^E (V_1^E)^T + \Delta_1^E\right).$$

Since $\Delta_1^E V_1^E = 0$, we obtain

$$X_1 = \left(E, AE\right)\left(E, AE\right)^T = Z_1^E (Z_1^E)^T + \mathbb{Q}_2^E \Delta_1^E \left(\mathbb{Q}_2^E \Delta_1^E\right)^T$$

with $Z_1^E = \mathbb{Q}_2^E U_1^E S_1^E$.

In the general case ($B \neq A$, $F \neq E$), the expressions of Z_1^E and Z_1^F are given in (16).

Remark 3.2. *Propositions 1 and 2 produce actually worst-case bounds. In practical computations, the convergence can be much better than these bounds predict.*
The choice of $\mathrm{tol_k}$ may be difficult to tune, in our experiments we use the same tolerance threshold, that is, for all k, $\mathrm{tol_k} = \mathrm{tol_{svd}}$.

3.2 Estimation of the residual

The following proposition relates the residual to the powers of A and B and the error incurred in $X_k - Z_k^E Z_k^F$.

Proposition 4. *For $k \geq 0$, we have for the residual Γ_k defined in (26),*

$$\|\Gamma_k\| \leq \|E\|\|F\| \left\| A^{2^k}(B^{2^k})^T \right\| + (1 + \|A\|\|B\|)\mu_k,$$

where μ_k is defined in Proposition 2.

Proof.

$$\Gamma_k = EF^T + AX_kB^T - X_k + A(Z_k^E(Z_k^F)^T - X_k)B^T - (Z_k^E(Z_k^F)^T - X_k),$$

thus

$$\|\Gamma_k\| \leq \|EF^T + AX_kB^T - X_k\| + (1 + \|A\|\|B\|)\|Z_k^E(Z_k^F)^T - X_k\|.$$

From (5) we obtain

$$EF^T + AX_kB^T - X_k = A^{2^k}EF^T(B^{2^k})^T \tag{30}$$

and the proof is completed by using Proposition 2. □

The following proposition shows that the norm of the residual involves quantities readily computable. In particular, it will be used as a stopping criterion in Algorithm 2.

Proposition 5. *We have $\|\Gamma_0\| = \|\left(\underline{\mathbb{H}}_1^E R_1^E\right)\left(\underline{\mathbb{H}}_1^F R_1^F\right)^T\|$ and for $k \geq 1$,*

$$\begin{aligned}\|\Gamma_k\| = \Bigg\| & \begin{pmatrix} R_1^E \\ 0 \end{pmatrix} \begin{pmatrix} (R_1^F)^T & 0 \end{pmatrix} + \left(\underline{\mathbb{H}}_{2^k}^E U_k^E S_k^E (V_k^E)^T V_k^F\right)\left(\underline{\mathbb{H}}_{2^k}^F U_k^F S_k^F\right)^T - \\ & \begin{pmatrix} U_k^E S_k^E (V_k^E)^T V_k^F \\ 0 \end{pmatrix} \begin{pmatrix} S_k^F\,(U_k^F)^T & 0 \end{pmatrix} \Bigg\|.\end{aligned}$$

Proof. We have

$$\Gamma_0 = (AE)(BF)^T = (\mathbb{Q}_2^E \underline{\mathbb{H}}_1^E R_1^E)(\mathbb{Q}_2^F \underline{\mathbb{H}}_1^F R_1^F)^T.$$

Hence $\|\Gamma_0\| = \|\left(\underline{\mathbb{H}}_1^E R_1^E\right)\left(\underline{\mathbb{H}}_1^F R_1^F\right)\|$.

For $k \geq 1$, using (26), (24) and (9) the residual Γ_k can be written

$$\begin{aligned}\Gamma_k &= (Q_1^E R_1^E)(Q_1^F R_1^F)^T + A(\mathbb{Q}_{2^k}^E U_k^E S_k^E (V_k^E)^T V_k^F)(\mathbb{Q}_{2^k}^F U_k^F S_k^F)^T B^T - \\ &\quad (\mathbb{Q}_{2^k}^E U_k^E S_k^E (V_k^E)^T V_k^F)(\mathbb{Q}_{2^k}^F U_k^F S_k^F)^T \\ &= \mathbb{Q}_{2^k+1}^E \Bigg[\begin{pmatrix} R_1^E \\ 0 \end{pmatrix} \begin{pmatrix} (R_1^F)^T & 0 \end{pmatrix} + \left(\underline{\mathbb{H}}_{2^k}^E U_k^E S_k^E {V_k^E}^T V_k^F\right)\left(\underline{\mathbb{H}}_{2^k}^F U_k^F S_k^F\right)^T - \\ &\quad \begin{pmatrix} U_k^E S_k^E {V_k^E}^T V_k^F \\ 0 \end{pmatrix} \begin{pmatrix} S_k^F\,(U_k^F)^T & 0 \end{pmatrix} \Bigg] (\mathbb{Q}_{2^k+1}^F)^T\end{aligned}$$

and the proposition follows by taking the norm. □

In practice we cannot use Algorithm 1 with large m since the computational cost and memory storage become prohibitive. On the other hand, with a too small m, the Krylov spaces thus obtained will not contain sufficient information to allow a good approximation of the solution. To remedy this, we will use in the next two sections a restarting technique and a simple version of the ADI iteration.

4 Restarting the low-rank approximation

As in iterative methods for large linear systems [17], restarting is based on the residual. Our residual Γ_k has a special form. From (30) we see that its rank is not larger than p. Its smallest singular values decrease as the number of iterations increases. We will use this information to construct new E^{rst} and F^{rst} that will be used in place of E and F for the next restart. The construction is based on an incomplete SVD of the residual Γ_k computed with the help of Proposition 4 at a lower cost.

Consider the reduced SVD

$$\begin{pmatrix} R_1^E \\ 0 \end{pmatrix} \begin{pmatrix} (R_1^F)^T & 0 \end{pmatrix} + \left(\underline{\mathbb{H}}_{2^k}^E U_k^E S_k^E {V_k^E}^T V_k^F\right)\left(\underline{\mathbb{H}}_{2^k}^F U_k^F S_k^F\right)^T -$$
$$\begin{pmatrix} U_k^E S_k^E {V_k^E}^T V_k \\ 0 \end{pmatrix} \begin{pmatrix} U_k^F S_k^F & 0 \end{pmatrix} = U_k S_k V_k^T + \Theta_k,$$

where U_k and V_k have orthonormal columns, S_k is diagonal whose elements are larger than some convergence threshold tol_{cvg} and Θ_k contains the rest of the SVD with $\|\Theta_k\| \le \text{tol}_{\text{cvg}}$. Then from the proof of Proposition 5 we have the decomposition

$$\begin{aligned} \Gamma_k &= \mathbb{Q}_{2^k+1}^E (U_k S_k V_k^T + \Theta_k)(\mathbb{Q}_{2^k+1}^F)^T \\ &= E^{\text{rst}}(F^{\text{rst}})^T + \widetilde{\Theta}_k \end{aligned} \tag{31}$$

with $E^{\text{rst}} = \mathbb{Q}_{2^k+1}^E U_k S_k^{\frac{1}{2}}$, $F^{\text{rst}} = \mathbb{Q}_{2^k+1}^F V_k S_k^{\frac{1}{2}}$ and $\widetilde{\Theta}_k = \mathbb{Q}_{2^k+1}^E \Theta_k (\mathbb{Q}_{2^k+1}^F)^T$, $\|\widetilde{\Theta}_k\| = \|\Theta_k\| \le \text{tol}_{\text{cvg}}$.

Denote

$$Q_1^{\text{rst},E} = \mathbb{Q}_{2^k+1}^E U_k \ , \ Q_1^{\text{rst},F} = \mathbb{Q}_{2^k+1}^F V_k, \ R_1^{\text{rst},E} = R_1^{\text{rst},F} = S_k^{\frac{1}{2}},$$

then we have the QR factorizations

$$E^{\text{rst}} = Q_1^{\text{rst},\,E} R_1^{\text{rst},\,E} \text{ and } F^{\text{rst}} = Q_1^{\text{rst},\,F} R_1^{\text{rst},\,F}. \tag{32}$$

Applying Algorithm 1 with A and $Q_1^{\text{rst},E}$ and with B and $Q_1^{\text{rst},F}$ and proceeding as in Section 2, we obtain factors of the form $Z_k^{\text{rst},E}$ and $Z_k^{\text{rst},F}$. The residual associated with $Z_k^{\text{rst},E}\left(Z_k^{\text{rst},F}\right)^T$ is

$$\Gamma_k^{\text{rst}} = E^{\text{rst}}\left(F^{\text{rst}}\right)^T + AZ_k^{\text{rst},E}\left(Z_k^{\text{rst},F}\right)^T B^T - Z_k^{\text{rst},E}\left(Z_k^{\text{rst},F}\right)^T \tag{33}$$

and its norm is computed by Proposition 5.

If this norm is smaller than tol_{cvg}, then the iterations can be stopped and the new approximate solution is

$$X_k^{\text{new}} \approx \left(Z_k^E, Z_k^{\text{rst},\,E}\right)\left(Z_k^F, Z_k^{\text{rst},\,F}\right)^T. \tag{34}$$

The corresponding residual is given by

$$\begin{aligned} \Gamma_k^{\text{new}} &= EF + A\left(Z_k^E, Z_k^{\text{rst},\,E}\right)\left(Z_k^F, Z_k^{\text{rst},\,F}\right)^T B^T - \left(Z_k^E\ , Z_k^{\text{rst},\,E}\right)\left(Z_k^F, Z_k^{\text{rst},\,F}\right)^T \\ &= \Gamma_k + \Gamma_k^{\text{rst}} - E^{\text{rst}}\left(F^{\text{rst}}\right)^T \\ &= \Gamma_k^{\text{rst}} + \widetilde{\Theta}_k, \end{aligned}$$

from which we see that $\|\Gamma_k^{\text{new}}\| \leq \|\Gamma_k^{\text{rst}}\| + \|\widetilde{\Theta}_k\| \leq 2\ \text{tol}_{\text{cvg}}$, so that Γ_k^{new} may have a norm slightly larger than tol_{cvg}.

If the norm of Γ_k^{rst} is larger than tol_{cvg}, then Γ_k^{rst} is decomposed as in (31) and a new restart is used. The process is repeated until the norm of the restarted residual becomes smaller than tol_{cvg}. We summarize this discussion in the following algorithm, which will be referred to as Low-Rank Krylov Squared Smith (LRKSS).

A few comments are in order. Algorithm LRKSS computes Q_1^E and Q_1^F from QR decompositions of E and F and applies a variant Algorithm 1 to A starting with Q_1^E and to B starting with Q_1^F. In fact, our block Arnoldi implementation is based on Ruhe's version with elimination, see [17, p.197] or [16, Algorithm 6.1]. Then, at each iteration j, the matrices Q_{j+1}^E, Q_{j+1}^F, $\underline{\mathbb{H}}_j^E$ and $\underline{\mathbb{H}}_j^F$ are computed. If $j = 1$, then $Z_0^E = E$ and $Z_0^F = F$ and the residual norm $\|\Gamma_0\|$ is computed as in Proposition 5. Else, if j is a power of 2, then the reduced SVDs are computed as in (21)–(23) by eliminating the same number of singular values which are smaller than tol_{svd}. The factors Z_k^E and Z_k^F are then computed as well as the corresponding residual norm. If the residual norm is larger than tol_{cvg} and the size of the Krylov bases reaches its maximum $m_{\max}$, the factors Z_k^E and Z_k^F are updated and the algorithm is restarted with new matrices Q_1^E and Q_1^F obtained from a reduced SVD of the residual as in (32).

It seems difficult to find optimal choices of tol_{svd} and tol_{cvg}. We notice, however, that tol_{cvg} should not be chosen too small compared to tol_{svd} for the SVDs in (21) and (22) would be such that $\|S_k^E\| < \text{tol}_{\text{svd}}$ and $\|S_k^F\| < \text{tol}_{\text{svd}}$, which leads to no improvement of the approximate solution.

With this way of restarting, the proposed squared Smith version can be applied to large matrices and this was our primary objective. However, the quadratic convergence of the original squared Smith method may be lost. The purpose of the next section is to accelerate the convergence by replacing equation (1) with an equivalent one with matrices having smaller spectral radii.

5 ADI iteration

The ADI method is an important iterative process for solving Lyapunov and Sylvester equations [24, 25, 5, 6, 15, 18, 4, 3, 13, 7]. An ADI iteration suited for equation (1) is proposed in [6], see also [4]. It is given, for $i = 0, 1, \ldots,$ by

$$X_{i+\frac{1}{2}}(I - \delta_i B^T) = (A - \delta_i I)X_i B^T + EF^T,$$

$$(I - \eta_i A)X_{i+1} = AX_{i+\frac{1}{2}}(B^T - \eta_i I) + EF^T,$$

where X_0 is an initial approximate solution of (1) and μ_i and η_i are parameters chosen to accelerate the convergence. Eliminating $X_{i+\frac{1}{2}}$ from these two equations and rearranging the terms, we obtain

$$\begin{aligned} X_{i+1} \quad &= \quad (I - \eta_i A)^{-1}A(A - \delta_i I)X_i B^T(I - \delta_i B^T)^{-1}(B^T - \eta_i I) + \\ & \quad\quad (I - \eta_i A)^{-1}((1 - \delta_i\eta_i)AEF^T B^T)(I - \delta_i B^T)^{-1} + EF^T. \end{aligned}$$

Algorithm 2 LRKSS

```
INPUT: $A, B \in \mathbb{R}^{n,n}, E, F \in \mathbb{R}^{n,p}$, an integer $m_{\max}$, tolerances $\mathrm{tol}_{\mathrm{svd}}$, $\mathrm{tol}_{\mathrm{cvg}}$, and an
    initial value of $\|\Gamma\| > \mathrm{tol}_{\mathrm{cvg}}$.
OUTPUT: Approximate solution to (1) in factored form $Z_s^E \left(Z_s^F\right)^T$
$QR$ factorize $E = Q_1^E R_1^E$ and $F = Q_1^F R_1^F$.
Set $U^E = I$, $U^F = I$, $V^E = I$, $V^F = I$, $S_1^E = R_1^E$, $S_1^F = R_1^F$, $Z_s^E = [\,]$, $Z_s^F = [\,]$,
$p^E = \dim(Q_1^E)$, $p^F = \dim(Q_1^F)$,
$j = 0$, $iter = 0$, $rst = 0$.
while ($\|\Gamma\| > \mathrm{tol}_{\mathrm{cvg}}$) do
  $j := j + 1$
  Update $\underline{\mathbb{H}}_j^E$, $\underline{\mathbb{H}}_j^F$, $\mathbb{Q}_{j+1}^E$ and $\mathbb{Q}_{j+1}^F$.
  if $j = 2^k$ then
    if $k = 0$ then
      $Z^E = E,\ Z^F = F,\ \|\Gamma\| = \|\left(\underline{\mathbb{H}}_1^E R_1^E\right)\left(\underline{\mathbb{H}}_1^F R_1^F\right)^T\|$
    else
      compute the reduced SVDs
      $$U^E S^E (V^E)^T := \left(\begin{pmatrix} U^E S^E (V^E)^T V^F \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}_j^E U^E S^E (V^E)^T V^F\right)$$
      and
      $$U^F S^F (V^F)^T := \left(\begin{pmatrix} U^F S^F \\ 0 \end{pmatrix}, \prod_{j=2^k-1}^{2^{k-1}} \underline{\mathbb{H}}_j^F U^F S^F\right)$$
      by eliminating the same number of singular values of $S^E$ and $S^F$ which are
      less than $\mathrm{tol}_{\mathrm{svd}}$.
      $Z^E = \mathbb{Q}_j^E U^E S^E (V^E)^T V^F, \quad Z^F = \mathbb{Q}_j^F U^F S^F$
      $\|\Gamma\| = \left\| \begin{pmatrix} R_1^E (R_1^F)^T & 0 \\ 0 & 0 \end{pmatrix} + \left(\underline{\mathbb{H}}_j^E U^E S^E (V^E)^T V^F\right)\left(\underline{\mathbb{H}}_j^F U^F S^F\right)^T - \begin{pmatrix} U^E S^E (V^E)^T V^F (U^F)^T (S^F)^T & 0 \\ 0 & 0 \end{pmatrix} \right\|$
    end if
    $iter := iter + 1$
  end if
  if $(j + 1)p^E > m_{\max}$ or $\|\Gamma\| \leq \mathrm{tol}_{\mathrm{cvg}}$ then
    Set $Z_s^E := (Z_s^E, Z^E),\ Z_s^F := (Z_s^F, Z^F)$.
    Compute the reduced SVD
    $$U^E S^E (V^E)^T \quad := \quad \begin{pmatrix} R_1^E (R_1^F)^T & 0 \\ 0 & 0 \end{pmatrix} + \left(\underline{\mathbb{H}}_j^E U^E S^E (V^E)^T V^F\right)\left(\underline{\mathbb{H}}_j^F U^F S^F\right)^T - \begin{pmatrix} U^E S^E (V^E)^T V^F (U^F)^T (S^F)^T & 0 \\ 0 & 0 \end{pmatrix}$$
    with $\sigma_{\min}(S) > tol_{svd}$.
    Set $R_1^E = S^{\frac{1}{2}},\ R_1^F = S^{\frac{1}{2}}, Q_1^E = \mathbb{Q}_{j+1}^E U,\ Q_1^F = \mathbb{Q}_{j+1}^F V$.
    Set $U^E = I,\ U^F = I,\ V^E = I,\ V^F = I,\ S_1^E = R_1^E,\ S_1^F = R_1^F,\ j = 0$,
    $rst := rst + 1$, $p^E = \dim(Q_1^E)$, $p^F = \dim(Q_1^F)$.
  end if
end while
```

Let

$$\mathcal{A}_i = (I - \eta_i A)^{-1} A \ (A - \delta_i I), \quad \mathcal{B}_i = (I - \delta_i B)^{-1} B \ (B - \eta_i I),$$

$$\mathcal{E}_i = (E \ , \ (I - \eta_i A)^{-1} AE \ \sqrt{1 - \delta_i \eta_i}), \quad \mathcal{F}_i = (F \ , \ (I - \delta_i B)^{-1} BF \ \sqrt{1 - \delta_i \eta_i}).$$

Then the sequence $\big(X_i\big)_{i \geq 0}$ satisfies the iteration

$$X_{i+1} = \mathcal{A}_i X_i \mathcal{B}_i^T + \mathcal{E}_i \mathcal{F}_i^T. \tag{35}$$

A straightforward calculation shows that the solution X is a fixed point of this iteration and hence the error is given by

$$X_{i+1} - X = \mathcal{A}_i \big(X_i - X\big) \mathcal{B}_i^T.$$

A repetition of this iteration gives

$$X_{i+1} - X = \big(\Pi_{j=0}^{i} \mathcal{A}_i\big)\big(X_0 - X\big)\big(\Pi_{j=0}^{i} \mathcal{B}_i\big)^T.$$

The convergence $X_{i+1} - X \to 0$ is fast if the spectral radii of $\Pi_{j=0}^{i} \mathcal{A}_j$ and $\Pi_{j=0}^{i} \mathcal{B}_j$ are as small as possible. Ideally, this will be the case if the parameters μ_i and η_i, $i = 0, 1, \ldots, i$, are chosen to satisfy

$$\min_{\substack{\delta_i \in \mathbb{C} \\ \eta_i \in \mathbb{C}}} \ \max_{\substack{\lambda \in \Lambda(A) \\ \mu \in \Lambda(B)}} \left| \Pi_{j=0}^{i} \frac{\lambda(\lambda - \delta_j)\mu(\mu - \eta_j)}{(1 - \eta_j \lambda)(1 - \delta_j \mu)} \right|. \tag{36}$$

However, this problem is hard to solve and computationally expensive. Since we are only interested in parameters that help reduce the spectral radii, we will consider only two parameters μ and η which approximate the minimax problem

$$\min_{\substack{\delta \in \mathbb{C} \\ \eta \in \mathbb{C}}} \ \max_{\substack{\lambda \in \Lambda(\tilde{A}) \\ \mu \in \Lambda(\tilde{B})|}} \frac{|\lambda(\lambda - \delta)\mu(\mu - \eta)|}{|(1 - \eta\lambda)(1 - \delta\mu)|}, \tag{37}$$

where the sets $\Lambda(\tilde{A})$ and $\Lambda(\tilde{B})$ contain a few smallest and/or largest approximate eigenvalues of A and B. The numerical tests show that this simple choice gives a satisfactory acceleration.

Once δ and η are computed, the matrices A, B, E and F of equation (1) are replaced by

$$\mathcal{A} = (I - \eta A)^{-1} A (A - \delta I), \quad \mathcal{B} = (I - \delta B)^{-1} B (B - \eta I) \tag{38}$$

$$\mathcal{E} = \big(E, (I - \eta A)^{-1} AE \sqrt{1 - \delta\eta}\big), \quad \mathcal{F} = \big(F, (I - \delta B)^{-1} BF \sqrt{1 - \delta\eta}\big) \tag{39}$$

on which algorithm LRKSS is applied.

6 Numerical Tests

We present numerical tests to illustrate the key points crucial for the convergence of algorithm LRKSS and its ADI acceleration.

Test 1: This test shows the convergence behavior when the spectral radii of A and B approach 1. The matrices A and B are $n \times n$, Toeplitz tridiagonal. A has $-\alpha$, 0 and $+\alpha$ respectively on its subdiagonal, diagonal and superdiagonal and B has the same structure with $-\beta$, 0 and $+\beta$, where α and β are positive parameters to be varied. The matrix E is $n \times 2$ formed by the first two vectors of the canonical basis, that is, $E_{11} = E_{22} = 1$ and zero elsewhere, and $F = -E$. The eigenvalues of A and B are given by $\lambda_j = 2i\alpha \cos \frac{\pi j}{n+1}$ and $\mu_j = 2i\beta \cos \frac{\pi j}{n+1}$, $j = 1, ..., n$. For large n we see that $\rho(A) \approx 2\alpha$ and $\rho(B) \approx 2\beta$.

Table 1 shows the results obtained with $n = 10^3$, $\text{tol}_{\text{cvg}} = \text{tol}_{\text{svd}} = 10^{-10}$ and different values of α, β and the maximum dimension of the Krylov spaces $m_{\max}$. The table also indicates the corresponding spectral radii $\rho(A)$ and $\rho(B)$, residual norms, total number of iterations and restarts. As expected, the closer the spectral radii get to 1, the slower the convergence is. Also, note that the restart, while it remedies the problem of storage requirements and computational cost, slows down the convergence. The numbers of restarts and iterations are almost doubled when the dimension of the Krylov spaces is divided by 2.

α β	$\rho(A)$ $\rho(B)$	$m_{\max}$	res.norm	iter	rst
0.45 0.445	0.9 0.89	32	5.96×10^{-11}	20	4
		64	5.96×10^{-11}	14	2
		128	5.96×10^{-11}	10	1
0.499 0.495	0.998 0.99	32	9.94×10^{-11}	268	66
		64	8.24×10^{-11}	171	33
		128	5.35×10^{-11}	102	16
0.4999 0.499	0.9998 0.998	32	9.93×10^{-11}	1205	296
		64	9.93×10^{-11}	753	148
		128	9.93×10^{-11}	452	74

Table 1: Results of LRKSS with different values of α, β and $m_{\max}$ (Test 1)

Test 2: We use the matrices as in the previous test with fixed $\alpha = 0.499$, $\beta = 0.495$, $m_{\max} = 64$ and consider three values of n, $n = 10^3$, $n = 10^4$, and $n = 10^5$. The corresponding spectral radii are almost the same $\big(\rho(A) \approx 2\alpha$, $\rho(B) \approx 2\beta\big)$ meaning that the convergence behavior is almost the same, see Proposition 3. For the three tests, the numbers of iterations and restarts are respectively 171 and 33. The convergence behaviors are shown in Figure 1.

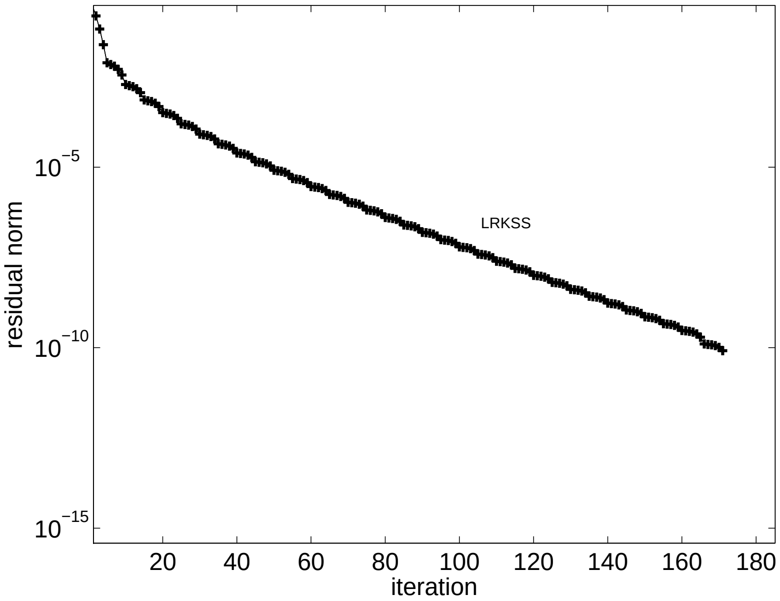

Figure 1: Residual norms vs the number of iterations (Test 2, $n = 10^p$, $p = 3, 4, 5$)

Test 3: We consider now Test 1 with ADI acceleration. The parameters δ and η of the minimax problem (37) are obtained from the 10 largest (in modulus) eigenvalues of A and B. The parameters δ and η are computed by the MATLAB function `fminsearch` as proposed in [4]. The linear systems in (38) and (39) are solved by GMRES with restart value 20 and tolerance 10^{-10}. The results are summarized in Table 2. An improvement can be noticed compared to Table 1. Note that the parameters η and μ are close to zero, which means that the matrices $\mathcal{A}$, $\mathcal{B}$, $\mathcal{E}$ and $\mathcal{F}$ in (38) and (39) are close to A^2, B^2, (E, AE) and (F, BF).

α β	δ η	$\rho(\mathcal{A})$ $\rho(\mathcal{B})$	$m_{\max}$	res.norm	iter	rst
0.45 0.445	$-7.877\ 10^{-9}$ $-7.877\ 10^{-9}$	0.8099 0.7921	32	7.15×10^{-11}	13	3
			64	7.15×10^{-11}	11	2
			128	7.15×10^{-11}	8	1
0.499 0.495	$-4.3315\ 10^{-8}$ $3.5532\ 10^{-8}$	0.9959 0.9801	32	8.30×10^{-11}	159	43
			64	9.74×10^{-11}	102	22
			128	6.70×10^{-11}	66	11
0.4999 0.499	$1.5357\ 10^{-7}$ $-1.1153\ 10^{-7}$	0.9995 0.99959	32	9.76×10^{-11}	670	173
			64	9.76×10^{-11}	424	86
			128	9.76×10^{-11}	256	42

Table 2: Results of LRKSS and ADI with different values of α, β and $m_{\max}$ (Test 3)

Test 4: We consider now the equivalent equation

$$X - A^2X(B^2)^T = \big(E, AE\big)\big(F, AF\big)^T$$

on which we apply ADI iterations and algorithm LRKSS, $n = 10^3$, $\mathrm{tol}_{\mathrm{svd}} = \mathrm{tol}_{\mathrm{cvg}} = 10^{-10}$. GMRES is used with the same parameters as in Test 3. Note that the parameters η and μ allow now a significant improvement compared to the previous results, see Table 3.

α β	δ η	$\rho(\mathcal{A})$ $\rho(\mathcal{B})$	$m_{\max}$	res.norm	iter	rst
0.45 0.445	$-8.0986\ 10^{-1}$ $-7.9195\ 10^{-1}$	0.2565 0.2453	32	9.43×10^{-11}	4	0
			64	9.43×10^{-11}	3	0
			128	9.43×10^{-11}	3	0
0.499 0.495	$-9.9582\ 10^{-1}$ $-9.7992\ 10^{-1}$	0.7427 0.7192	32	2.10×10^{-11}	16	6
			64	1.62×10^{-12}	13	3
			128	7.59×10^{-11}	9	2
0.4999 0.499	$-9.9583\ 10^{-1}$ $-9.9942\ 10^{-1}$	0.8742 0.8679	32	4.04×10^{-11}	31	12
			64	7.12×10^{-11}	24	7
			128	1.30×10^{-11}	17	3

Table 3: Results of LRKSS and ADI with different values of α, β and $m_{\max}$ (Test 4)

Figure 2 draws the convergence behaviors when $n = 10^3$, $n = 10^4$ and $n = 10^5$. For the three cases, $\alpha = 0.499$, $\beta = 0.495$ and $m_{\max} = 64$. The three indistinguishable curves in this figure show that the convergence behavior is the same.

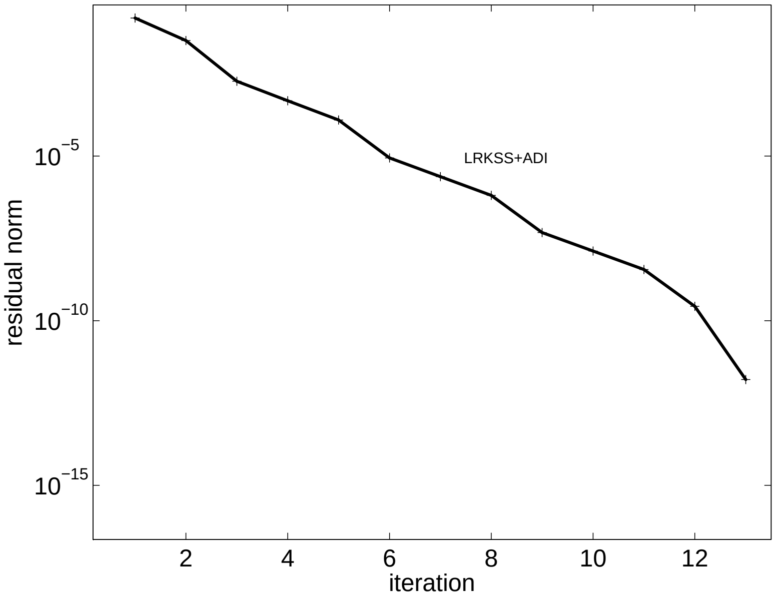

Figure 2: Residual norms vs the number of iterations, (Test 4, $n = 10^p$, $p = 3, 4, 5$)

Test 5: In this test we solve the equation

$$X - AXA = EF^T,$$

where $A = Q^T \hat{A} Q$, Q is an orthogonal matrix constructed with the MATLAB function `orth`, $\hat{A} = \mathrm{diag}\left(\left(0.999\ e^{i\frac{\pi}{n}}\right)^k,\ 1 \le k \le n \right)$, and the matrices E and F are the same as in the previous tests.

Taking $\eta = \delta$ in (37) and omitting the spectral part due to B, the minimax problem simplifies to

$$\min_{|\delta|<1} \max_{\lambda \in \Lambda(\tilde{A})} \frac{|\lambda(\lambda - \delta)|}{|(1 - \delta\lambda)|}. \tag{40}$$

The matrices in (38) and (39) become

$$\mathcal{A} = (I - \delta A)^{-1} A (A - \delta I), \tag{41}$$

$$\mathcal{E} = \left(E, (I - \delta A)^{-1} A E \sqrt{1 - \delta^2}\right), \quad \mathcal{F} = \left(F, (I - \delta A^T)^{-1} A^T F \sqrt{1 - \delta^2}\right) \tag{42}$$

Taking $n = 10^3$ we obtain $\delta = 9.8280 \times 10^{-1}$, $\rho(\mathcal{A}) = 8.9335 \times 10^{-1}$. The parameters $\mathrm{tol}_{\mathrm{svd}}$ and $\mathrm{tol}_{\mathrm{cvg}}$ are fixed at 10^{-10} and $m_{\max} = 32$. Figure 3 on the left shows the convergence of LRKSS with and without ADI preconditioning. The figure on the right shows the singular values of the exact and computed solutions. The smallest singular values differ by a factor of order $10^{-10}(= \mathrm{tol}_{\mathrm{svd}})$. Figure 4 shows the convergence with different values of $m_{\max}$ and confirms again that the larger the values of $m_{\max}$ is, the smaller the number of iterations.

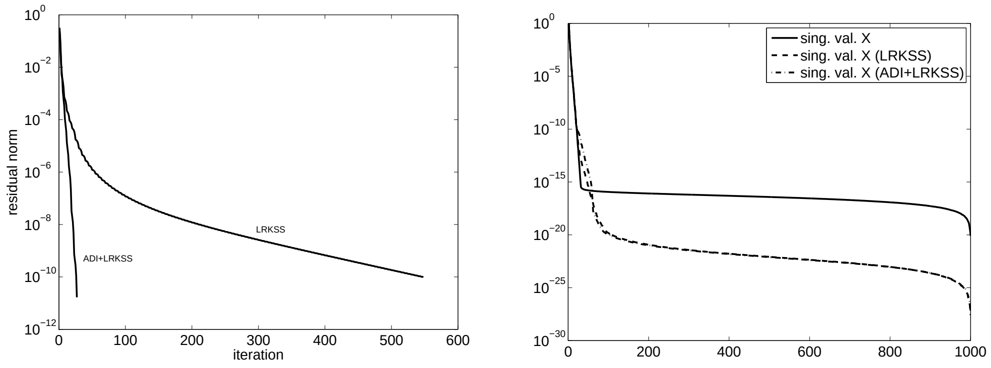

Figure 3: Residual norms and singular values of the exact and the computed solutions (Test 5)

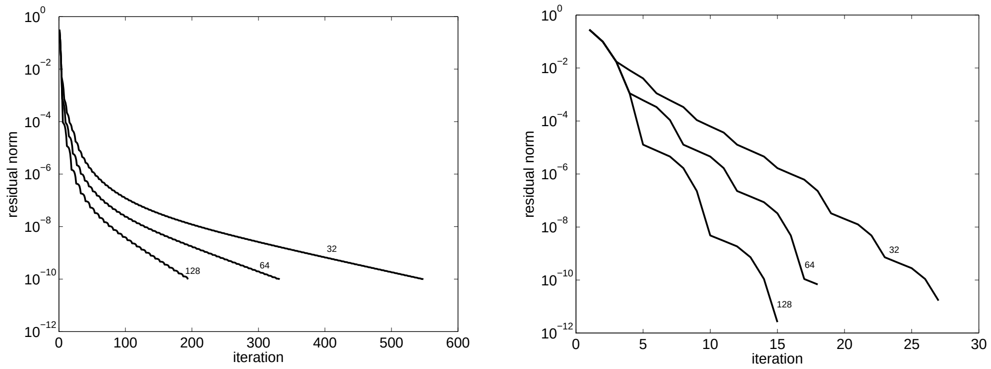

Figure 4: Convergence behaviors of LKRSS (left) and LKRSS+ADI(right) with different values of $m_{\max}$ (Test 5)

7 Conclusions

The main purpose of this work was to show one way of adapting the Squared Smith method to large-scale Stein equations. The adaptation requires the use of Krylov spaces to build approximations of the squared Smith iterates in low-rank factors. As expected, the quadratic convergence in the original squared Smith algorithm is not maintained, but the association with the proposed adaptation with a simple version of the ADI iteration as a preconditioner allows a great acceleration of the convergence.

This is consistent with the numerical results in [13, 7] where the “optimal” number of ADI iterations is less than 3 in [13] and around 4 in [7]. The acceleration depends largely on the ADI parameters and to a lesser extend on the other parameters of the algorithm. Improvements can still be made if these parameters can be chosen in a cheap and nearly optimal way.

References

[1] Björck A. Numerical Methods for Least Squares Problems. SIAM: Philadelphia, 1996.

[2] Antoulas AC. Approximation of Large-Scale Dynamical Systems, Advances in Design and Control. SIAM: Philadelphia, 2005.

[3] Benner P, Li RC, Truhar N. On the ADI method for Sylvester equations. J. Comput. Appl. Math. 2009; 233: 1035--1045.

[4] Benner P, Faßbender H. On the numerical solution of large-scale sparse discrete-time Riccati equations. Advances in Computational Mathematics. 2011; 35: 119--147.

[5] Calvetti D, Reichel L. Application of ADI iterative methods to the restoration of noisy images. SIAM J. Matrix Anal. Appl. 1996; 17: 165--186.

[6] Calvetti D, Levenberg N, Reichel L. Iterative methods for $X - AXB = C$. J. Comput. Appl. Math. 1997; 86: 73--101.

[7] Damm T. Direct methods and ADI-preconditioned Krylov subspace methods for generalized Lyapunov equations. Numer. Linear Algebra Appl. 2008; 15: 853--871.

[8] El Guennouni A, Jbilou K, Riquet J. Block Krylov subspace methods for solving large Sylvester equations. Numerical Algorithms. 2001; 29: 75--96.

[9] Gajič Z, Qureshi MTJ. Lyapunov matrix equation in system stability and control. Mineola, N. Y., Dover, 2008.

[10] Horn RA, Johnson CR. Topics in Matrix Analysis. Cambridge Universtity Press, Cambridge, 1990.

[11] Hu DY, Reichel L. Krylov subspace methods for the Sylvester equations. Linear Algebra Appl. 1992; 174: 283--314.

[12] Ionescu V, Oara C, Weiss M. Generalized Riccati theroy and robust control. A Popov function approach, John Wiley & Sons, Ltd., Chichester, 1999.

[13] Jbilou K. ADI preconditioned Krylov methods for large Lyapunov matrix equations. Linear Algebra Appl., 2010; 432: 2473--2485.

[14] Lancaster P, Rodman L. Algebraic Riccati Equations. Clarendon Press, Oxford, 1995.

[15] Penzl T. A cyclic low-rank Smith method for large sparse Lyapunov equations. SIAM J. Sci. Comput.2000; 21: 1401--1418.

[16] Robbé M, Sadkane M. A convergence analysis of GMRES and FOM methods for Sylvester equations. Numerical Algorithms. 2002; 30: 71--89.

[17] Saad Y. Iterative methods for sparse linear systems, 2nd ed. SIAM: Philadelphia, 2003.

[18] Sabino J. Solution of large-scale Lyapunov equations via the block modified Smith method. PhD thesis, Rice University, 2006.

[19] Sadkane M. A low-rank squared Smith method for large-scale discrete-time Lyapunov equations. Linear Algebra Appl. 2012; 436: 2807--2827.

[20] Simoncini V. On the numerical Solution of $AX - XB = C$. BIT. 1996; 36: 814--830.

[21] Smith R. Matrix equation $XA + BX = C$. SIAM J. Appl. Math. 1968; 16: 198--201.

[22] Stewart GW, Sun JG. Matrix Perturbation Theory. Academic Press, San Diego, CA, 1990.

[23] Trefethen LN, Embree M. Spectra and pseudospactra. The behavior of nonnormal matrices and operators. Princeton University Press, Princeton, 2005.

[24] Wachspress EL. Extended application of alternating direction implicit iteration model problem theory. Journal of SIAM. 1963; 11: 994--1016.

[25] Wachspress EL. Iterative solution of the Lyapunov matrix equations. Appl. Math. Letters. 1988; 107: 87--90.