Electronic Transactions on Numerical Analysis.
Volume 62, pp. 119–137, 2024.

ISSN 1068–9613.
DOI: 10.1553/etna_vol62s119

ETNA
Kent State University and
Johann Radon Institute (RICAM)

INEXACT LINEAR SOLVES IN THE LOW-RANK ALTERNATING DIRECTION IMPLICIT ITERATION FOR LARGE SYLVESTER EQUATIONS*

PATRICK KÜRSCHNER†

Abstract. We consider iteration for approximately solving large-scale algebraic Sylvester equations. Inside every iteration step of this iterative process, a pair of linear systems of equations has to be solved. We investigate the situation when those inner linear systems are solved inexactly by an iterative method such as, for example, preconditioned Krylov subspace methods. The main contribution of this work are thresholds for the required accuracies regarding the inner linear systems, which dictate when the employed inner Krylov subspace methods can be safely terminated. The goal is to save computational effort by solving the inner linear system as inaccurately as possible without endangering the functionality of the low-rank Sylvester–ADI method. Ideally, the inexact ADI method mimics the convergence behavior of the more expensive exact ADI method, where the linear systems are solved directly. Alongside the theoretical results, strategies for an actual practical implementation of the stopping criteria are also developed. Numerical experiments confirm the effectiveness of the proposed strategies.

Key words. Sylvester equation, alternating direction implicit, low-rank approximation, inner–outer methods

AMS subject classifications. 15A06, 15A24, 65F45, 65F55

1. Introduction.

We consider the numerical solution of large-scale algebraic Sylvester equations of the form

$$AX + XB = -fg^* \tag{1.1}$$

with large coefficient matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times m}$, the sought-after matrix $X \in \mathbb{R}^{m\times n}$, and a factorized right-hand side with $f \in \mathbb{R}^{n\times r}$ and $g \in \mathbb{R}^{m\times r}$ which have full column rank $r \ll n, m$. Sylvester equations play a vital role in many application areas, for instance, control theory, model order reduction [1, 38], and image processing [11], and they often arise as a crucial ingredient in algorithms in more complicated matrix equations [10, 20, 29, 39]. We refer to the survey article [35] for further details and examples.

For equations defined by small to medium-sized coefficient matrices, methods based on matrix factorizations can be used, such as the Bartels–Stewart and related methods [2]. When only one coefficient matrix (A or B) is large and sparse, while the other one is small and dense, special methods are applicable [5, 38]; and see [17, Section 7.6.3].

In this work we treat the case when both coefficients, A and B, are large and sparse matrices and the right-hand side is of low rank. For this scenario, one can show that the singular values of the solution X typically decay rapidly towards zero [13, 18, 31]. This motivates the computation of a solution approximation of low rank in factored form $X \approx Z\Gamma Y^*$, where Z and Y are thin rectangular matrices and the middle matrix Γ has conforming size. In recent years, different (rational) Krylov subspace projection methods have been proposed for this purpose; see, e.g., [10, 12, 16, 19, 21, 30]. Another method for this situation, and the focus of this study, is the low-rank (LR) version of the alternating direction implicit (ADI) iteration [6, 8, 11, 40, 41].

Inside every step of the Sylvester ADI iteration, linear systems of equations must be solved. Solving these inner linear systems is computationally the most expensive part of every iteration step, and, if direct solvers are not applicable, a further, inner iteration can be

*Received November 30, 2023. Accepted August 2, 2024. Published online on September 19, 2024. Recommended by Valeria Simoncini.

†Leipzig University of Applied Sciences (HTWK Leipzig), Center for Mathematics and Natural Sciences, Leipzig, Germany (patrick.kuerschner@htwk-leipzig.de).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

used in order to obtain approximate solutions. Typically, one uses preconditioned Krylov subspace methods for this purpose. We develop estimates for the required accuracies regarding those linear systems which will dictate when the inner Krylov subspace methods can be safely terminated, thus potentially saving some computational effort without endangering the functionality of the low-rank Sylvester ADI method. A peculiarity of the Sylvester ADI approach is that a pair of linear systems must be solved in every step. These linear systems are defined by shifted versions of the coefficient matrices A and B. Hence, one must determine two connected stopping criteria, for which we present strategies.

The paper is structured as follows. In Section 2 we review the Sylvester ADI iteration and its properties, which we then modify to incorporate inexact solutions of the inner linear systems. Afterward in Section 3, we develop dynamic inner stopping criteria for the iterative solution of the arising linear systems and analyze their effect on the Sylvester ADI iteration. We also discuss the actual implementation of these stopping criteria inside the ADI iteration. Section 4 demonstrates the findings by some numerical experiments, and Section 5 concludes and gives some potential future research directions.

1.1. Notation. Throughout the paper, if not stated otherwise, we use $\|\cdot\|$ to denote the Euclidean vector and associated induced matrix norm; and $(\cdot)^*$ stands for the transpose (respectively, complex conjugate transpose) for real (respectively, complex) matrices and vectors. The identity matrix of dimension k is denoted by I_k with the subscript omitted if the dimension is clear from the context. The kth column of the identity matrix is denoted e_k and $\mathbf{1}_k := [1,\dots,1]^* \in \mathbb{R}^k$. The spectrum of a matrix $A \in \mathbb{C}^{n\times n}$ and a matrix pair (A,M) is denoted by $\Lambda(A)$ and $\Lambda(A,M)$, respectively. The smallest (respectively, largest) eigenvalue in magnitude of a matrix A is denoted by $\lambda_{\min}(A)$ (respectively, $\lambda_{\max}(A)$), $\rho(A) = \max\{|\lambda|,\ \lambda \in \Lambda(A)\}$ is the spectral radius, and $\mathcal{W}(A) = \{z = x^*Ax : 0 \neq x \in \mathbb{C}^n,\ \|x\| = 1\}$ is the field of values. The symbol $\otimes$ denotes the Kronecker product. For $A \in \mathbb{C}^{n\times n}$, $\beta \in \mathbb{C}$, and $\alpha \notin -\Lambda(A)$, a two-parameter Cayley transformation is given by

$$\mathcal{C}(A,\alpha,\beta) = (A-\beta I_n)(A+\alpha I_n)^{-1} = I_n - (\beta+\alpha)(A+\alpha I_n)^{-1}. \tag{1.2}$$

2. Inexact low-rank ADI iteration for Sylvester equations. The Sylvester equations (1.1) have a unique solution, if and only if $\Lambda(A)\cap\Lambda(-B) = \emptyset$. This is, in particular, fulfilled if $\Lambda(A),\Lambda(B) \subset \mathbb{C}_-$ (or $\Lambda(A),\Lambda(B) \subset \mathbb{C}_+$), which is assumed in the remainder of the paper.

2.1. Derivation and review of basic properties. For every $\beta \notin \Lambda(A)$, $\alpha \notin \Lambda(B)$, $\alpha \neq \beta$, the continuous-time Sylvester equation (1.1) is equivalent to the discrete-time Sylvester equation

$$X = \mathcal{C}(A,\beta,\alpha)X\mathcal{C}(B,\alpha,\beta) + T(\alpha,\beta),$$

where

$$T(\alpha,\beta) := -(\beta+\alpha)(A+\beta I_n)^{-1} f g^* (B+\alpha I_m)^{-1}$$

and $\mathcal{C}(\cdot,\cdot,\cdot)$ are two-parameter Cayley transformations (1.2). This motivates the non-stationary iteration for $k = 1,2,\dots,$

$$\begin{aligned} X_k &= \mathcal{C}(A,\ \beta_k,\ \alpha_k)X_{k-1}\mathcal{C}(B,\ \alpha_k,\ \beta_k) + T(\alpha_k,\beta_k) \\ &= \mathcal{C}_k(A)X_{k-1}\mathcal{C}_k(B) + T_k, \end{aligned} \tag{2.1}$$

where we have introduced varying parameters α_k and β_k throughout the iteration. This is the alternating direction implicit iteration for algebraic Sylvester equations [40].

ETNA
Kent State University and
Johann Radon Institute (RICAM)

The error and the Sylvester residual matrix after k steps of the Sylvester ADI scheme (2.1) are given by

$$X_k - X = \mathcal{A}_k(X_0 - X)\mathcal{B}_k, \quad \mathcal{R}_k = AX_k + X_kB + fg^* = \mathcal{A}_k\mathcal{R}_0\mathcal{B}_k, \tag{2.2}$$
$$\mathcal{A}_k := \prod_{i=1}^{k} \mathcal{C}(A, \beta_i, \alpha_i), \quad \mathcal{B}_k := \prod_{i=1}^{k} \mathcal{C}(B, \alpha_i, \beta_i).$$

The iteration is convergent, for example, when $\rho(\mathcal{C}_k(A))\rho(\mathcal{C}_k(B)) < 1$ for all $k \geq 1$. A rapid decrease of error and residual can thus be achieved by minimizing this product of the spectral radii of $\mathcal{A}_k$ and $\mathcal{B}_k$, which leads to the ADI shift parameter problem,

$$\min_{\alpha_i, \beta_i \in \mathbb{C}} \left(\max_{\substack{1 \leq \ell \leq n \\ 1 \leq j \leq m}} \prod_{i=1}^{k} \left| \frac{(\lambda_\ell - \alpha_i)(\mu_j - \beta_i)}{(\lambda_\ell + \beta_i)(\mu_j + \alpha_i)} \right| \right), \quad \lambda_\ell \in \Lambda(A),\ \mu_j \in \Lambda(B). \tag{2.3}$$

Various approaches have been proposed for (2.3), e.g., pre-computing shift parameters [8, 31, 41] in an offline phase before the actual ADI iteration, based on either elliptic function regions or heuristic strategies. In contrast, more recent developments were focused on computing one pair (α_k, β_k) of parameters at a time online during the running ADI iteration [7, 22]. In this study, we assume that the pairs (α_k, β_k) that guarantee convergence of (2.1) are given in advance. This is purely for reasons of simplification, as the upcoming results on the stopping criteria do not depend on the way the shifts are generated.

A low-rank version[1] of the ADI iteration [8] is obtained by setting $X_0 = 0$ in (2.1) and exploiting that the matrices $(A + \beta_k I)^{-1}$ and $(A - \alpha_k I)$ as well as $(B + \alpha_k I)^{-1}$ and $(B - \beta_k I)$ commute. This allows us to formulate (2.1) for $k \geq 1$ as

$$z_1 = (A + \beta_1 I_n)^{-1} f, \quad s_1 = (B + \alpha_1 I_m)^{-H} g, \tag{2.4a}$$
$$z_k = z_{k-1} + (\beta_k - \alpha_{k-1})(A + \beta_k I_n)^{-1} z_{k-1}, \tag{2.4b}$$
$$y_k = y_{k-1} + \overline{(\alpha_k - \beta_{k-1})}(B + \alpha_k I_m)^{-H} y_{k-1}, \tag{2.4c}$$

which produces solution approximations in low-rank format:

$$X \approx X_k = Z_k \Gamma_k Y_k^* \quad \text{with} \quad Z_k = [z_1, \ldots, z_k] \in \mathbb{C}^{n \times kr},$$
$$\Gamma_k = \operatorname{diag}(\Gamma_{k-1}, (\beta_k - \alpha_k) I_r) \in \mathbb{C}^{kr \times kr}, \quad Y_k = [y_1, \ldots, y_k] \in \mathbb{C}^{m \times kr}.$$

The column dimensions of Z_k and Y_k grow by r columns after each iteration step, so that also their ranks grow. To reduce the memory demand, column compression techniques can be used, for instance after the iteration terminates.

It can be shown [6, 22] that the residual matrix (2.2) at step k has at most rank r and is given by the low-rank factorization

$$\mathcal{R}_k = w_k t_k^*, \tag{2.5a}$$

where

$$w_k := \mathcal{A}_k f = w_0 + Z_k \boldsymbol{\gamma}_k = w_{k-1} + \gamma_k (A + \beta_k I)^{-1} w_{k-1} \in \mathbb{C}^{n \times r}, \tag{2.5b}$$
$$t_k := \mathcal{B}_k^* g = t_0 + Y_k \overline{\boldsymbol{\gamma}_k} = t_{k-1} + \overline{\gamma_k} (B + \alpha_k I)^{-*} t_{k-1} \in \mathbb{C}^{m \times r}, \tag{2.5c}$$

with

$$w_0 := f, \quad t_0 := g, \quad \boldsymbol{\gamma}_k := [\gamma_1, \ldots, \gamma_k]^* \otimes I_r, \quad \gamma_k := -(\beta_k + \alpha_k). \tag{2.5d}$$

[1] This version is also called factored ADI iteration (fADI).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

This allows one to compute the residual norm $\|\mathcal{R}_k\|_2 = \|w_k t_k^*\|_2$ more efficiently, and the residual factors w_k and t_k can be directly integrated into the low-rank Sylvester ADI iteration (2.4), which then becomes

$$z_k = (A+\beta_k I)^{-1} w_{k-1}, \quad w_k = w_{k-1} + \gamma_k z_k, \quad w_0 := f, \tag{2.6a}$$

$$y_k = (B+\alpha_k I)^{-*} t_{k-1}, \quad t_k = t_{k-1} + \overline{\gamma_k} y_k, \quad t_0 := g. \tag{2.6b}$$

This is the form of the iteration used nowadays.

Generalized Sylvester equations. For two given additional non-singular matrices $M \in \mathbb{C}^{n\times n}$ and $C \in \mathbb{C}^{m\times m}$, generalized Sylvester equations are of the form

$$AXC + MXB = -fg^*. \tag{2.7}$$

The low-rank ADI iteration can be generalized in a straightforward manner to

$$X_k = \mathcal{C}(AM^{-1}, \beta_k, \alpha_k) X \mathcal{C}(C^{-1}B, \beta_k, \alpha_k) + \hat{T}_k,$$

with

$$\hat{T}_k := -\gamma_k (A+\beta_k M)^{-1} f g^* (B+\alpha_k C)^{-1}.$$

The corresponding low-rank iteration (2.6) transforms in that case [6, 22] to

$$z_k = (A+\beta_k M)^{-1} w_{k-1}, \quad w_k = w_{k-1} + \gamma_k M z_k, \quad w_0 := f, \tag{2.8a}$$

$$y_k = (B+\alpha_k C)^{-*} t_{k-1}, \quad t_k = t_{k-1} + \overline{\gamma_k} C^* y_k, \quad t_0 := g. \tag{2.8b}$$

The spectra $\Lambda(A)$ and $\Lambda(B)$ in (2.3) have to be replaced by the spectra $\Lambda(A, M)$ and $\Lambda(B, C)$, respectively.

REMARK 2.1. Note that

$$\begin{aligned} w_k &= w_{k-1} + \gamma_k M (A+\beta_k M)^{-1} w_{k-1} = (A-\alpha_k M)(A+\beta_k M)^{-1} w_{k-1} \\ &= (AM^{-1} - \alpha_k I)(AM^{-1} + \beta_k I)^{-1} w_{k-1} = \mathcal{C}_k(AM^{-1}) w_{k-1}, \\ t_k &= t_{k-1} + \overline{\gamma_k} C^* (B+\alpha_k C)^{-*} t_{k-1} = (B-\beta_k C)^* (B+\alpha_k C)^{-*} t_{k-1} \\ &= (C^{-1}B - \beta_k I)^* (C^{-1}B + \alpha_k I)^{-*} t_{k-1} = \mathcal{C}_k^*(C^{-1}B) t_{k-1}, \end{aligned}$$

indicating that the matrices of AM^{-1} and $C^{-1}B$ appear only for notational purposes and will not be formed explicitly in an actual implementation.

We will mostly consider the more general version (2.8) in the remainder of the paper.

The most expensive parts inside each step of the outer ADI iteration (2.6) and (2.8) are the solutions of the inner shifted linear system with $(A+\beta_k M)$, $(B+\alpha_k C)^*$, and r right-hand sides for z_k and y_k.

In this study, we discuss the situation when z_k and y_k are only approximate solutions of the linear systems but the remaining steps in (2.6) and (2.8) are kept unchanged. This gives the *inexact low-rank ADI iteration* illustrated in Algorithm 1. Let

$$\begin{aligned} r_k^A &:= w_{k-1} - (A+\alpha_k M) z_k \quad \text{with} \quad \|r_k^A\| \le \delta_k^A, \\ r_k^B &:= t_{k-1} - (B+\beta_k C)^* y_k \quad \text{with} \quad \|r_k^B\| \le \delta_k^B \end{aligned} \tag{2.9}$$

be the residual vectors with respect to the linear systems. We generally refer to r_k^A and r_k^B as inner residuals. The quantities $\delta_k^A, \delta_k^B > 0$ indicate the residual tolerances with respect to the inner linear systems at outer step k of the inexact LR-ADI iteration.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Algorithm 1: Inexact low-rank ADI (LR-ADI) method for Sylvester equations

Input : A, B, M, C, f, g as in (2.7), shift parameters $\{\alpha_1, \dots, \alpha_k\}$, $\{\beta_1, \dots, \beta_k\}$, and tolerance $0 < \tau \ll 1$.

Output : $Z_k \in \mathbb{C}^{n\times rk}$, $Y_k \in \mathbb{C}^{m\times rk}$, $\Gamma_k \in \mathbb{C}^{rk\times rk}$ such that $Z_k\Gamma_k Y_k^H \approx X$.

$w_0 = f,\ t_0 = g,\ Z_0 = \Gamma_0 = Y_0 = [\,],\ j = 1.$

while $\|w_{k-1}T_{k-1}^H\| \geq \tau\|FG^*\|$ **do**

Approximately solve the linear systems for $z_k,\ y_k$:

$$(A + \beta_k M)z_k = w_{k-1}, \quad \|r_k^A\| = \|w_{k-1} - (A + \beta_k M)z_k\| \leq \delta_k^A,$$
$$(B + \alpha_k C)^* y_k = t_{k-1}, \quad \|r_k^B\| = \|t_{k-1} - (B + \alpha_k C)^* y_k\| \leq \delta_k^B.$$

$\gamma_k := -(\beta_k + \alpha_k).$

$w_k = w_{k-1} + \gamma_k M z_k, \quad t_k = t_{k-1} + \overline{\gamma_k} C^* y_k.$

$Z_k = [Z_{k-1},\ z_k],\ Y_k = [Y_{k-1},\ y_k],\ \Gamma_k = \operatorname{diag}(\Gamma_{k-1},\ \gamma_k I_r).$

$k = k + 1.$

REMARK 2.2. In this work, "inexact" means that the solution process of the linear systems is the only source of inexactness in the LR-ADI iteration. We do not consider other errors introduced, e.g., by the finite-precision arithmetic.

REMARK 2.3. An inexact version of the dense iteration (2.1) for positive definite $A,\ B$, $M = I_n,\ C = I_m$, and fixed shifts $\alpha_k = \alpha,\ \beta_k = \beta,\ \forall\ k \geq 1$, is discussed in [27]. The motivation in [27] is to ensure asymptotic convergence of (2.1) under inexact inner solves. In contrast, our analysis is focused on the low-rank iteration with variable shifts and, moreover, pursues the goal to make the behavior of the inexact ADI iteration as close as possible to that of the exact ADI iteration. Note that, by our assumptions on the shift parameters, the exact ADI iteration converges.

2.2. Properties of inexact low-rank Sylvester ADI iteration. Before we can investigate stopping criteria for the inexact low-rank Sylvester ADI iteration, we need to generalize some results for the exact iteration [22, Corollary 3.16] as well as for the inexact iteration for Lyapunov equations [24, Theorem 3.2].

THEOREM 2.4. *The low-rank solution factors Z_k and Y_k constructed after j steps of the inexact LR-ADI iteration (Algorithm 1) satisfy the identities*

$$AZ_k = MZ_k\boldsymbol{\sigma}_k^{\alpha} + w_k E_k^* - S_k^A, \qquad B^*Y_k = C^*Y_k\boldsymbol{\sigma}_k^{\beta} + t_k E_k^* - S_k^B, \tag{2.10}$$

with

$$\boldsymbol{\sigma}_k^{\alpha} := \begin{bmatrix} \alpha_1 & & \\ -\gamma_2 & \alpha_2 & \\ \vdots & \ddots & \\ -\gamma_k & \cdots & -\gamma_k\ \alpha_k \end{bmatrix} \otimes I_r, \qquad \boldsymbol{\sigma}_k^{\beta} := \begin{bmatrix} \overline{\beta_1} & & \\ -\overline{\gamma_2} & \overline{\beta_2} & \\ \vdots & \ddots & \\ -\overline{\gamma_k} & \cdots & -\overline{\gamma_k}\ \overline{\beta_k} \end{bmatrix} \otimes I_r \in \mathbb{C}^{jr\times jr},$$

$$E_k := \mathbf{1}_k \otimes I_r \in \mathbb{R}^{jr\times r},$$

$$S_k^A := [r_1^A, \dots, r_k^A] \in \mathbb{C}^{n\times rk}, \qquad S_k^B := [r_1^B, \dots, r_k^B] \in \mathbb{C}^{m\times rk}$$

containing the residuals of the linear systems (2.9).

Proof. The result for $S_k^A = 0$ and $S_k^B = 0$ has been established in [22]. By construction, it holds that $w_{i-1} = w_i - \gamma_i M z_i$ and $Az_i = w_{i-1} - \beta_i M z_i - r_i^A$ for $i = 1, \dots, k$. Combining the two relations yields $Az_i = w_i + \alpha_i M z_i - r_i^A$. Moreover, successively inserting the defining relation (2.8a) for the previous residual factors w_j, with $j = k-1, \dots, i \geq 1$, into

ETNA
Kent State University and
Johann Radon Institute (RICAM)

the formula for w_k gives, after rearranging: $w_i = w_k - M\sum_{j=1}^{k-i}\gamma_{k-j+1}z_{k-j+1}$. Inserting this into the expression for Az_i gives

$$Az_i = w_k + \alpha_i M z_i - M\sum_{j=1}^{k-i}\gamma_{k-j+1}z_{k-j+1} - r_i^A, \quad i = 1,\dots,k,$$

such that

$$A[z_1,\dots,z_k] = [w_k,\dots,w_k] + M[z_1,\dots,z_k]\left(\begin{bmatrix}\alpha_1 & & \\ -\gamma_2 & \alpha_2 & \\ \vdots & & \ddots \\ -\gamma_k & \cdots & -\gamma_k \ \alpha_k\end{bmatrix}\otimes I_r\right) - S_k^A,$$

which is the result in (2.10) for AZ_k. The result for B^*Y_k is developed in the same way using the corresponding relations for t_k, y_k, and r_k^B. ∎

THEOREM 2.5. *After k iteration steps of inexact LR-ADI applied to (2.7), the Sylvester residual matrix is given by*

$$\begin{aligned}\mathcal{R}_k &= AZ_k\Gamma_k Y_k^* C + MZ_k\Gamma_k Y_k^* B + fg^* = w_k t_k^* - \eta_k^A - \eta_k^B,\\ \eta_k^A &:= S_k^A\Gamma_k Y_k^* C, \quad \eta_k^B := MZ_k\Gamma_k(S_k^B)^*.\end{aligned} \tag{2.11}$$

Proof. By construction, it follows from (2.5) that $f = w_k - MZ_k\boldsymbol{\gamma}_k$ and $g = t_k - C^*Y_k\overline{\boldsymbol{\gamma}_k}$. Plugging this and the identities of Theorem 2.4 into the Sylvester residual matrix yields

$$\mathcal{R}_k = MZ_k\big(\boldsymbol{\sigma}_k^\alpha\Gamma_k + \Gamma_k(\boldsymbol{\sigma}_k^\beta)^* + \boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^*\big)Y_k^*C + w_k t_k^* - MZ_k\Gamma_k(S_k^B)^* - S_k^A\Gamma_k Y_k^* C.$$

The claim follows upon realizing that Γ_k is the solution of the following Sylvester equation: $\boldsymbol{\sigma}_k^\alpha\Gamma + \Gamma(\boldsymbol{\sigma}_k^\beta)^* + \boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^* = 0$. ∎

In the following we call $\mathcal{R}_k^{\text{comp}} = w_k t_k^*$ the *computed residual*, which refers to the computation of its norm by means of the outer product of the residual factors w_k and t_k, just as done in the exact LR-ADI. As stated earlier, the norm of this outer product can be computed efficiently. Plugging $X_k = Z_k\Gamma_k Y_k^*$ into the Sylvester equation gives the *true residual* $\mathcal{R}_k^{\text{true}}$ from (2.11). The above properties show that, in the presence of inexact linear solves, there is a discrepancy between the computed Sylvester residuals $\mathcal{R}_k^{\text{comp}}$ and the true residuals $\mathcal{R}_k^{\text{true}}$ given by (2.11).

DEFINITION 2.6 (Residual gap). *The residual gap after k iteration steps of the low-rank Sylvester ADI is defined as*

$$\Delta\mathcal{R}_k := \mathcal{R}_k^{comp} - \mathcal{R}_k^{true} = S_k^A\Gamma_k Y_k^* C + MZ_k\Gamma_k(S_k^B)^* = \eta_k^A + \eta_k^B.$$

Hence, the efficiently computable quantity $\|\mathcal{R}_k^{\text{comp}}\| = \|w_k t_k^*\|$ is not the correct value of the norm of the Sylvester residual for the current solution approximation $Z_k\Gamma_k Y_k^*$. Using it might give a wrong impression of the iteration's progress. One would need to estimate the true residual norm, e.g., by means of a Lanczos process applied to $\mathcal{R}^*\mathcal{R}$, which is more costly (see comparison in [6]).

3. Dynamic residual thresholds for inner linear systems. In this section, we investigate strategies to make the inner residual norms $\|r_k^A\|$ and $\|r_k^B\|$ as large as possible without endangering the functionality of the outer ADI iteration. The goal is to achieve that the inexact low-rank ADI iteration mimics the exact counterpart up to a very small deviation.

Similar to the low-rank ADI iteration for Lyapunov equations, it can be shown that the low-rank factors Z_k and Y_k span (block) rational Krylov subspaces for AM^{-1} and, respectively,

ETNA
Kent State University and
Johann Radon Institute (RICAM)

$C^{-*}B^*$ [3, 15, 26]. Hence, the low-rank ADI iteration can be seen as a (two-sided) rational Krylov subspace method. A commonly used approach in inexact (rational) Krylov and related methods is to enforce a small norm of the residual gap: $\|\Delta\mathcal{R}_k\| \leq \varepsilon$. If $\|\mathcal{R}_k^{\text{comp}}\| \leq \varepsilon$, then a small true residual norm $\|\mathcal{R}_k^{\text{true}}\| = \|\mathcal{R}_k^{\text{comp}} + \Delta\mathcal{R}_k\| \leq 2\varepsilon$ can be expected. This technique is often coined *relaxation* because it usually leads to increasing (relaxed) inner residual norms if the outer residual norm decreases [9, 34]. The next theorem generalizes [24, Theorem 3.3] and provides a theoretical strategy for reducing $\|\Delta\mathcal{R}_k\|$ in the inexact low-rank Sylvester ADI below a desired threshold after a pre-specified number of outer steps.

THEOREM 3.1 (Theoretical inner stopping criterion for the inexact Sylvester-ADI). *Let the residual gap be given by Definition 2.6 with w_k, t_k, and γ_k as in (2.8). Let $k_{\max}$ be the maximum number of steps of Algorithm 1 and $0 < \varepsilon < 1$ a small threshold. Furthermore, let $c_k^A := \|\mathcal{C}_k(AM^{-1})\|$, $c_k^B := \|\mathcal{C}_k(C^{-1}B)\|$, and $\check{c}_k := \max(c_k^A,\ c_k^B) + 1$. If, for $1 \leq k \leq k_{\max}$, the inner residuals satisfy*

$$\check{c}_k(\|r_k^A\|\|t_{k-1}\| + \|r_k^B\|\|w_{k-1}\| + 2\|r_k^B\|\|r_k^A\|) \leq \frac{\varepsilon}{k_{\max}}, \tag{3.1}$$

then $\|\Delta\mathcal{R}_{k_{\max}}\| \leq \varepsilon$.

Proof. Consider the following estimate:

$$\begin{aligned}\|\Delta\mathcal{R}_{k_{\max}}\| &\leq \|MZ_k\Gamma_k(S_k^B)^*\| + \|S_k^A\Gamma_k Y_k^* C\| \\ &\leq \sum_{k=1}^{k_{\max}} \|r_k^B\|\|\gamma_k M z_k\| + \|r_k^A\|\|\overline{\gamma_k} C^* y_k\|.\end{aligned} \tag{3.2}$$

Moreover, we can bound

$$\begin{aligned}\|\gamma_k M z_k\| &= \|\gamma_k M(A+\beta_k M)^{-1}(w_{k-1} - r_k^A)\| = \|(\mathcal{C}_k(AM^{-1}) - I_n)(w_{k-1} - r_k^A)\| \\ &\leq (c_k^A + 1)(\|w_{k-1}\| + \|r_k^A\|), \\ \|\overline{\gamma_k} C^* y_k\| &= \|\overline{\gamma_k} C^*(B+\alpha_k C)^{-*}(t_{k-1} - r_k^B)\| = \|(\mathcal{C}_k^*(C^{-1}B) - I_m)(t_{k-1} - r_k^B)\| \\ &\leq (c_k^B + 1)(\|t_{k-1}\| + \|r_k^B\|),\end{aligned}$$

and get

$$\begin{aligned}\|\Delta\mathcal{R}_{k_{\max}}\| &\leq \sum_{k=1}^{k_{\max}} (c_k^A + 1)\|r_k^B\|(\|w_{k-1}\| + \|r_k^A\|) + (c_k^B + 1)\|r_k^A\|(\|t_{k-1}\| + \|r_k^B\|) \\ &\leq \sum_{k=1}^{k_{\max}} \check{c}_k(\|r_k^B\|\|w_{k-1}\| + \|r_k^A\|\|t_{k-1}\| + 2\|r_k^A\|\|r_k^B\|) \leq \sum_{k=1}^{k_{\max}} \frac{\varepsilon}{k_{\max}} = \varepsilon\end{aligned}$$

if (3.1) holds. ∎

Discussion and consequences. We have already observed one striking difference compared to the Lyapunov situation: We do not get a single combination for the largest possible inner residual norms but instead infinitely many admissible combinations of $\|r_k^A\|$ and $\|r_k^B\|$ that satisfy (3.1). We introduce the following notation:

$$\check{\varepsilon} := \frac{\varepsilon}{k_{\max}} \quad \text{from (3.1).}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

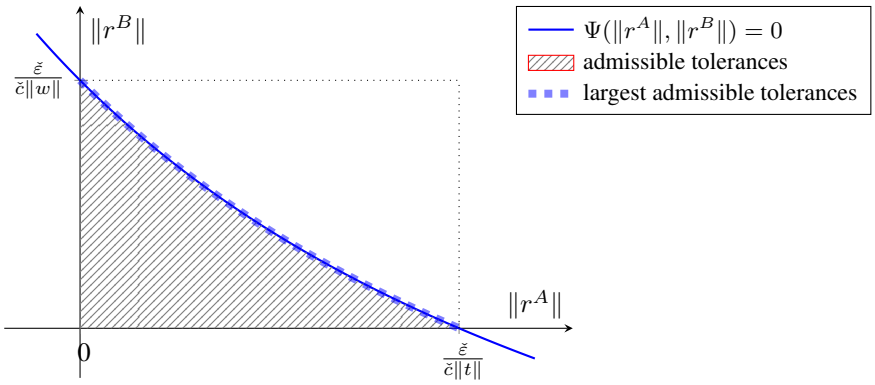

FIG. 3.1. *Illustration of the region of admissible inner residual norms with the boundary curve $\Psi = 0$, where the thick dashed portion is the set of admissible combinations of largest tolerances. The ADI iteration indices were left out for readability. The data were generated here with $\check{c} = 3$, $\check{\varepsilon} = 10^{-8}$, $\|t\| = 10^{-5}$, and $\|w\| = 2 \times 10^{-3}$.*

Due to the non-negativity of the residual norms, the bounds (3.1) dictate that admissible values are from the region bounded by

$$\begin{aligned} &0 \le \|r_k^A\| \le \frac{\check{\varepsilon}}{\check{c}_k\|t_{k-1}\|}, \quad 0 \le \|r_k^B\| \le \frac{\check{\varepsilon}}{\check{c}_k\|w_{k-1}\|}, \quad \text{and} \quad \psi(\|r_k^A\|, \|r_k^B\|) \le \check{\varepsilon}, \\ &\psi(x, y) := \check{c}_k \left(x(\|t_{k-1}\| + y) + y(\|w_{k-1}\| + x)\right). \end{aligned} \tag{3.3}$$

The largest admissible inner residual norms are located on the planar curve

$$0 = \Psi(\|r_k^A\|, \|r_k^B\|) := \psi(\|r_k^A\|, \|r_k^B\|) - \check{\varepsilon},$$

which is illustrated in Figure 3.1. How to select one particular combination will be discussed later.

By solving $\Psi = 0$, e.g., for $\|r_k^B\|$,

$$\|r_k^B\| = \frac{\check{\varepsilon} - \check{c}_k\|r_k^A\|\|t_{k-1}\|}{\check{c}_k(2\|r_k^A\| + \|w_{k-1}\|)}, \tag{3.4}$$

we can infer an upper bound for one inner residual from the other. While obeying the constraints (3.3), the smaller that $\|r_k^A\|$ is chosen, the larger $\|r_k^B\|$ can be, and vice versa.

If the computed Sylvester residual decreases in norm, the product of the norms of the residual factors $\|w_{k-1}\|\|t_{k-1}\|$ will decrease as well but not necessarily the norms of the individual residual factors $\|w_{k-1}\|$ and $\|t_{k-1}\|$. Hence, the absolute inner residual norms do not need to increase as the outer residual norm $\|\mathcal{R}_k^{\text{comp}}\|$ decreases, which forms another difference to the Lyapunov situation. However, we observe increasing relative inner residual norms

$$\frac{\|r_k^A\|}{\|w_{k-1}\|} \le \frac{\check{\varepsilon}}{\check{c}_k\|w_{k-1}\|\|t_{k-1}\|} \quad \text{and} \quad \frac{\|r_k^B\|}{\|t_{k-1}\|} \le \frac{\check{\varepsilon}}{\check{c}\|w_{k-1}\|\|t_{k-1}\|},$$

since $\check{c}_k$ can be bounded by a moderate constant, as we will discuss later. Note that, for $r = 1$, we even have $\|\mathcal{R}_{k-1}^{\text{comp}}\| = \|w_{k-1}\|\|t_{k-1}\|$, so that

$$\frac{\|r_k^A\|}{\|w_{k-1}\|}, \frac{\|r_k^B\|}{\|t_{k-1}\|} \le \frac{\check{\varepsilon}}{\check{c}_k\|\mathcal{R}_{k-1}^{\text{comp}}\|}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

3.1. Distance between exact and inexact Sylvester ADI with dynamic inner solve tolerances. We now establish a relation between the (norms of the) computed Sylvester residual matrices of both exact and inexact Sylvester-ADI, where the dynamic inner stopping criteria from Theorem 3.1 are used in the latter.

THEOREM 3.2. *Assume $k_{\max}$ steps of both exact and inexact Sylvester-ADI are applied to a Sylvester equation using the same set of shift parameters (α_k, β_k), $j = 1, \dots, k_{\max}$. We denote the quantities of the exact LR-ADI iteration by superscript $^{\text{exact}}$ and $0 < \varepsilon \ll 1$ is a small threshold. If the condition (3.1) is used in the inexact ADI iteration, then*

$$\|\mathcal{R}_{k_{\max}}^{\text{comp}}\| \leq \|\mathcal{R}_{k_{\max}}^{\text{exact}}\| + \mathcal{O}(\varepsilon).$$

Proof. First, following the construction of the residual factors w_k and t_k in Algorithm 1, we find that

$$\begin{aligned}
w_k &= \mathcal{C}_k(AM^{-1})w_{k-1} + (I - \mathcal{C}_k(AM^{-1}))r_k^A \\
&= \cdots = w_k^{\text{exact}} + \sum_{j=1}^{k} \left[\prod_{i=j+1}^{k} \mathcal{C}_i(AM^{-1}) \right] (I - \mathcal{C}_j(AM^{-1}))r_j^A, \\
t_k &= \mathcal{C}_k(C^{-1}B)^* t_{k-1} + (I - \mathcal{C}_k(C^{-1}B)^*)r_k^B \\
&= \cdots = t_k^{\text{exact}} + \sum_{j=1}^{k} \left[\prod_{i=j+1}^{k} \mathcal{C}_i(C^{-1}B)^* \right] (I - \mathcal{C}_j(C^{-1}B)^*)r_j^B,
\end{aligned}$$

which generalizes [24, Lemma 3.1]. Then,

$$\begin{aligned}
\|\mathcal{R}_{k_{\max}}^{\text{comp}}\| &= \|w_k t_k^*\| \\
&\leq \|(\mathcal{C}_k(AM^{-1}))w_{k-1}t_{k-1}^*\mathcal{C}_k(C^{-1}B)\| \\
&\quad + (c_k^A + 1)c_k^B\|r_k^A\|\|t_{k-1}\| + (c_k^B + 1)c_k^A\|r_k^B\|\|w_{k-1}\| \\
&\quad + (c_k^B + 1)(c_k^A + 1)\|r_k^A\|\|r_k^B\|,
\end{aligned}$$

where we have used the constants c_k^A and c_k^B introduced in Theorem 3.1. Consequently,

$$\begin{aligned}
&(c_k^A + 1)c_k^B\|r_k^A\|\|t_{k-1}\| + (c_k^B + 1)c_k^A\|r_k^B\|\|w_{k-1}\| + (c_k^B + 1)(c_k^A + 1)\|r_k^A\|\|r_k^B\| \\
&\leq \check{c}_k^2(\|r_k^A\|\|t_{k-1}\| + \|r_k^B\|\|w_{k-1}\| + 2\|r_k^A\|\|r_k^B\|) \leq \check{c}_k \frac{\varepsilon}{k_{\max}}.
\end{aligned}$$

Repeating this process again for w_{k-1} and t_{k-1} for $1 \leq k \leq k_{\max} - 1$ eventually yields

$$\begin{aligned}
\|\mathcal{R}_{k_{\max}}^{\text{comp}}\| &= \left\| \left(\prod_{k=1}^{k_{\max}} \mathcal{C}_k(AM^{-1}) \right) w_0 t_0^* \left(\prod_{k=1}^{k_{\max}} \mathcal{C}_k(C^{-1}B) \right) \right\|^2 + \sum_{j=1}^{k_{\max}} \check{c}_k \frac{\varepsilon}{k_{\max}} \\
&= \|\mathcal{R}_{k_{\max}}^{\text{exact}}\| + \frac{\varepsilon}{k_{\max}} \sum_{k=1}^{k_{\max}} \check{c}_k. \qquad \square
\end{aligned}$$

Combining Theorems 3.1 and 3.2 yields the following conclusion.

COROLLARY 3.3. *Under the same conditions as in Theorem 3.2 we have*

$$\|\mathcal{R}_{k_{\max}}^{\text{true}}\| \leq \|\mathcal{R}_{k_{\max}}^{\text{comp}}\| + \|\Delta\mathcal{R}_{k_{\max}}\| \leq \|\mathcal{R}_{k_{\max}}^{\text{exact}}\| + \left(1 + \sum_{k=1}^{k_{\max}} \frac{\check{c}_k}{k_{\max}} \right) \varepsilon.$$

Hence, if (3.1) is used, then the true Sylvester residual norms in the inexact LR-ADI are a small perturbation of the residuals of the exact method, provided that the $\check{c}_k$ are bounded by moderate constants, which we will discuss next.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

3.2. Practical and implementational considerations. Here, we discuss some ways for the practical usage of the stopping criterion (3.1) in an actual implementation of the low-rank Sylvester ADI.

3.2.1. Estimating the spectral norms. First, we are going to bound the constants c_k^A, c_k^B, and $\check{c}_k$. For this, we require some additional assumptions:

- The rational functions defining both Cayley transforms,

$$\phi_k^A(x) := \frac{x-\alpha_k}{x+\beta_k}, \quad \phi_k^B(x) := \frac{x-\beta_k}{x+\alpha_k}, \quad k \geq 1,$$

  are, for all $k \geq 1$, analytic on $\mathcal{W}(AM^{-1})$ and, respectively, $\mathcal{W}(C^{-1}B)$.
- It holds that

$$q^A := \max_{z \in \mathcal{W}(AM^{-1})} |\phi_k^A(z)| < 1, \quad q^B := \max_{z \in \mathcal{W}(C^{-1}B)} |\phi_k^B(z)| < 1.$$

Then, by [14],

$$c_k^A = \|\mathcal{C}_k(AM^{-1})\| \leq \psi^A q^A < \psi^A, \quad c_k^B = \|\mathcal{C}_k(C^{-1}B)\| \leq \psi^B q^B < \psi^B,$$

where $\psi^A = 1+\sqrt{2}$ (but $\psi^A = 1$ if AM^{-1} is normal) and similarly for ψ^B and $C^{-1}B$. As a consequence, we can bound the constants as

$$\check{c}_k \leq c := 2+\sqrt{2}.$$

Practical stopping criteria. Additionally, if we use $\varepsilon/(2\check{c}_k k_{\max})$ instead of $\varepsilon/k_{\max}$ for the condition in Theorem 3.1, then we achieve

$$\|\mathcal{R}_{k_{\max}}^{\text{comp}}\| \leq \|\mathcal{R}_{k_{\max}}^{\text{exact}}\| + \frac{\varepsilon}{2} \quad \text{and} \quad \|\mathcal{R}_{k_{\max}}^{\text{true}}\| \leq \|\mathcal{R}_{k_{\max}}^{\text{exact}}\| + \varepsilon$$

in Corollary 3.3. Replacing $\check{c}_k$ by the bound c leads to

$$\|r_k^A\|\|t_{k-1}\| + \|r_k^B\|\|w_{k-1}\| + 2\|r_k^B\|\|r_k^A\| \leq \xi \frac{\varepsilon}{2c^2 k_{\max}} =: \hat{\varepsilon} \tag{3.5}$$

as the practical realization of (3.1). Here, $0 < \xi \leq 1$ is a safeguard constant for situations when the above assumptions are mildly violated. Note that similar small safeguard constants are common in inexact (rational) Krylov methods [24, 34, 36]. In most of our experiments, $\xi = 1$ was sufficient.

3.2.2. Incorporating inner residual norms and residual gaps. The proposed stopping criterion requires the norms of the inner residuals, which can often be directly obtained from the employed Krylov subspace methods.

Back-looking . The previous inner residual norms can be used to further refine the dynamic stopping strategy. Assume the inner solvers achieved

$$\|r_j^B\|\|\gamma_j M z_j\| + \|r_j^A\|\|\overline{\gamma_j} C^* y_j\| \leq \frac{\varepsilon}{k_{\max}}$$

for $1 \leq j \leq k-1$. In the bound (3.2) we can at step k try to achieve

$$\|\Delta\mathcal{R}_k\| \leq \|\Delta\mathcal{R}_{k-1}\| + \|r_k^B\|\|\gamma_k M z_k\| + \|r_k^A\|\|\overline{\gamma_k} C^* y_k\| \leq \frac{k\varepsilon}{k_{\max}}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

which leads to

$$\check{c}_k(\|r_k^A\|\|t_{k-1}\| + \|r_k^B\|\|w_{k-1}\| + 2\|r_k^B\|\|r_k^A\|) + \|\Delta\mathcal{R}_{k-1}\| \le \frac{k\varepsilon}{k_{\max}}. \tag{3.6}$$

As in [24], the reasoning behind this approach is to look back at all previous inner residuals and also to incorporate the previous residual gap $\|\Delta\mathcal{R}_{k-1}\|$. Hence, we will refer to this strategy as "back-looking". This might allow the incorporation of instances when, at some of the earlier steps $j \le k-1$, smaller inner residuals than requested were achieved by the inner solvers, allowing us to use slightly larger inner tolerances at step k.

The back-looking strategy (3.6) requires the previous residual gap to be $\|\Delta\mathcal{R}_{k-1}\| = \|\eta_{k-1}^A + \eta_{k-1}^B\|$, which might be expensive to compute and, moreover, would require storing all previous inner residuals. Here, we simply use, for $k \ge 2$, the approximations

$$\begin{aligned} &\|\Delta\mathcal{R}_{k-1}\| \le \|\eta_{k-1}^A\| + \|\eta_{k-1}^B\| \le u_{k-1} + v_{k-1}, \\ &\qquad u_{k-1} := u_{k-2} + |\gamma_{k-1}|\|Mz_{k-1}\|\|r_{k-1}^B\|, \quad u_0 := 0, \\ &\qquad v_{k-1} := v_{k-2} + |\gamma_{k-1}|\|C^*y_{k-1}\|\|r_{k-1}^A\|, \quad v_0 := 0. \end{aligned} \tag{3.7}$$

The matrix–vector products Mz_{k-1} and C^*y_{k-1} can be reused from step 4 of Algorithm 1.

As a practical realization of (3.6) we propose

$$\|r_k^A\|\|t_{k-1}\| + \|r_k^B\|\|w_{k-1}\| + 2\|r_k^B\|\|r_k^A\| \le \frac{1}{c}\left|\xi\frac{k\varepsilon}{2ck_{\max}} - u_{k-1} - v_{k-1}\right| =: \hat{\varepsilon}_k. \tag{3.8}$$

Here, we have again replaced $\check{c}_k$ by c, divided the right-hand side of (3.6) by $2c$, and introduced the safeguard constant ξ. The subscript k in $\hat{\varepsilon}_k$ now indicates the dependence on k on the right-hand side of inequality (3.8).

3.2.3. Selecting one particular combination of solve tolerances. From the infinitely many possibilities for the inner tolerances, we have to select one combination in an actual implementation. First, for reasons of feasibility, we may restrict the inner accuracies to some minimal and maximal levels via

$$0 < \delta_{\min}^A \le \delta_k^A \le \delta_{\max}^A, \quad 0 < \delta_{\min}^B \le \delta_k^B \le \delta_{\max}^B.$$

Then, one simple selection for δ_k^A could be to pick it somewhere from the middle of the boundary curve of (3.3) and compute δ_k^B via (3.4), as illustrated in Figure 3.2. A strategy that worked well in our experiments is to set

$$\begin{aligned} \delta_k^A &= \max\left(\frac{1}{2}\left(\min\left(\delta_{\max}^A, \frac{\hat{\varepsilon}}{\|t_{k-1}\|}\right) - \delta_{\min}^A\right), \delta_{\min}^A\right), \\ \delta_k^B &= \max\left(\min\left(\frac{\hat{\varepsilon} - \delta_k^A\|t_{k-1}\|}{2\delta_k^A + \|w_{k-1}\|}, \delta_{\max}^B\right), \delta_{\min}^B\right), \end{aligned} \tag{3.9}$$

where $\hat{\varepsilon}$ is from (3.5). If the back-looking strategy is used, then $\hat{\varepsilon}_k$ from (3.8) is inserted in (3.9) instead.

Since any point of the curve $\Psi = 0$ is admissible, we may move this point so that smaller inner residuals are favored for the linear system, which is easier and/or less costly to solve, and thus allow larger residuals for the other linear system. For example, if the linear system with $A + \beta_k M$ can be solved more easily and faster than the one with $B + \alpha_k C$, we set

$$\delta_k^A = \delta_{\min}^A \quad \text{and} \quad \delta_k^B \quad \text{via (3.9)}. \tag{3.10}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

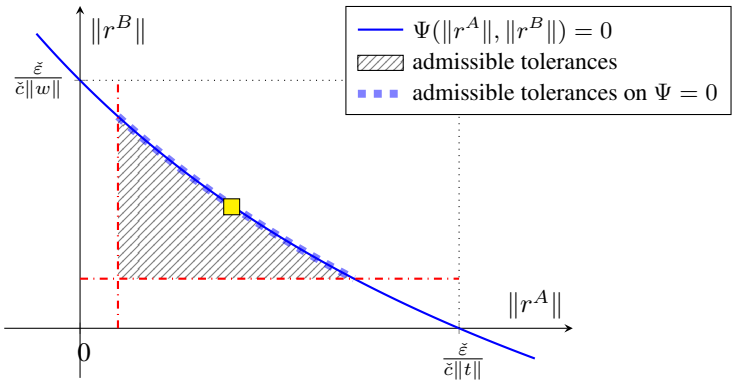

FIG. 3.2. *Extension of Figure 3.1 illustrating the region of admissible inner residual norms with the set of admissible combinations of largest tolerances (thick dashed line). The square marks one possible combination from this set. The straight dashed-dotted lines indicate minimal bounds* $\delta^A_{\min}$ *and* $\delta^B_{\min}$ *for the inner residual norms.*

In the reverse situation, we analogously select $\delta_k^B = \delta_{\min}^B$ and the largest admissible δ_k^A. As indicators as to how difficult or costly the iterative solution of a linear system is, one could look at, for example, the size of the matrix, the cost of the matrix–vector products (sparsity density), the condition number of the matrix, or a mixture thereof. A more general idea for selecting a combination might be to minimize some cost function (encoding the cost of solving both systems) constrained to the set of admissible largest tolerances. Experiments with these strategies led to working tolerances, but, unfortunately, there was hardly any performance gain at all compared to the simpler selections mentioned before. Additionally, extra costs plus several further tuning parameters were introduced for solving the constrained minimization. Hence, these strategies are not pursued further here.

Combination of direct and iterative linear solves. A similar situation arises when the linear systems of one sequence, e.g., those with $A + \beta_k M$, can be solved by sparse direct solvers, and the other linear systems, i.e., those with $B + \alpha_k C$, require iterative methods, or vice versa. Since direct solvers do not need stopping criteria, selecting a residual threshold is only required in the other sequence. For example, if the systems with $A + \beta_k M$ are solved directly and, ideally, we have $\|r^A\| \approx 0$, then the proposed practical stopping criteria (3.5) and (3.8) simplify to

$$\|r_k^B\| \lesssim \frac{\hat{\varepsilon}}{\|w_{k-1}\|}.$$

Note that this bears some similarity with the simplified stopping criterion for the inexact LR-ADI iteration for Lyapunov equations [24]. There, the bounds are of the form $\|r_k\| \lesssim \hat{\varepsilon}/\|\mathcal{R}_{k-1}^{\text{comp}}\|^{1/2}$,where $\|w_{k-1}\| = \|\mathcal{R}_{k-1}^{\text{comp}}\|^{1/2}$. For the Sylvester ADI and $r = 1$, the computed Sylvester residual norms are $\|\mathcal{R}_{k-1}^{\text{comp}}\| = \|w_{k-1}\|\|t_{k-1}\|$, so that $\|w_{k-1}\| = \|\mathcal{R}_{k-1}^{\text{comp}}\|^p$ for some $0 < p < 1$.

3.2.4. Further implementation aspects.

Choice of inner iterative solver and preconditioning. The motivation behind low-rank solvers for large matrix equations was to compute the approximate solution in a memory-efficient way. This should be maintained also in the inexact low-rank method, and we therefore strongly advocate the use of short-recurrence Krylov methods for solving (non-symmetric) inner linear systems. Working with a (non-restarted) long-recurrence method such as GMRES for solving non-symmetric inner linear systems requires storing the full Krylov basis and

ETNA
Kent State University and
Johann Radon Institute (RICAM)

hence might hinder working in a memory-efficient way (especially if the memory requirements for GMRES exceed those for the low-rank factors). Of course, having a very effective preconditioner would limit the memory requirements.

If $\operatorname{rank} fg^* = r > 1$, every inner linear system has r right-hand sides. One could then use special block Krylov methods (see, e.g., [28, 37]). In our experiments, this had no advantage over simply employing the single vector methods to every column $w_{k-1}(:,\ell)$, $t_{k-1}(:,\ell)$, $\ell = 1,\dots,r$. As stopping criterion, we simply used $\|r_k(:,\ell)\| \le \delta_k/r$.

The proposed stopping criteria require the norms of the residuals r_k of the underlying inner linear systems. If left or two-sided preconditioning is used, the preconditioned Krylov method will internally work with the preconditioned residuals, which might be different from the true inner residuals. Hence, we therefore used mostly right preconditioning, which does not suffer from this issue. In the other cases, we computed the true inner residual norms after termination of the Krylov method and ensured that $\|r_k\| \le \delta_k$ was met.

Complex shift parameters. For Sylvester equations defined by real but non-symmetric coefficient matrices, the sets of shifts $\{\alpha_j\}_{j=1}^k$ and $\{\beta_j\}_{j=1}^k$ can include pairs of complex conjugate shifts. In order to minimize the amount of complex arithmetic operations and to generate real low-rank solution factors Z_k, Γ_k, and Y_k, the results in [6, 22] can be used. The main idea is to perform a double iteration step when a complex pair of α-shifts meets a complex pair of β-shifts or two real β-shifts or vice versa. In this way, only one complex linear system needs to be solved per complex conjugate pair of shifts, but the arising variant of the low-rank Sylvester ADI involves rather cumbersome formulas. Therefore, here we stick to the possible complex formulation of the method. Note that in most of the upcoming experiments the used shifts were entirely real. As in the inexact low-rank Lyapunov ADI iteration, using the real formulation of the Sylvester ADI only involves some minor adjustments for the estimate (3.7) when the back-looking strategy (3.8) is used and a double step occurs.

Related matrix equations. In addition to the generalized Lyapunov equation ($B = A^*$, $C = M^*$, $g = f$), also cross-Gramian Sylvester equations ($B = A$, $C = M$) and symmetric Stein matrix equations ($B = M^*$, $C = -A^*$, $g = f$) are special cases of (2.7). Hence, with the appropriate adaptations, Algorithm 1 can be employed as well (see [6, 22]) including the proposed dynamic stopping criteria of this work.

4. Numerical experiments. The following experiments were carried out in MATLAB 2023a on an Intel Core 2 i7-7500U CPU @ 2.7 GHz with 16 GB RAM using the implementation of Algorithm 1 from [23]. We wish to obtain an approximate solution such that the scaled Sylvester residual norm satisfies

$$\mathfrak{R} := \|\mathcal{R}^{\text{true}}\|/\|fg^*\| \le \tilde{\varepsilon}, \quad 0 < \tilde{\varepsilon} \ll 1.$$

In all the upcoming experiments $\tilde{\varepsilon} = 10^{-8}$ is used.

As inner Krylov subspace solvers, we use BiCGstab for non-symmetric coefficients $A + \beta_k M$, $B + \alpha_k C$, and we use MINRES in the symmetric case. Note that, if $\beta_k \in \mathbb{C}$, then $A + \beta_k M$ is non-symmetric even if A and M are symmetric, and thus this requires a Krylov method for non-symmetric linear systems. Sparse-direct solves are carried out by the MATLAB `backslash` routine.

Shift parameters for the Sylvester-ADI are generated either by the heuristic approach from [8] (using 10 Ritz values for each of A and B and 20 inverse Ritz values of A^{-1} and B^{-1}) or for real spectra also by the analytic approach of Sabino [31, Algorithm 2.1] (using approximations of the extremal eigenvalues from both spectra obtained with the MATLAB routine `eigs`).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

TABLE 4.1
Examples used in the experiments with matrix properties, settings for preconditioners, and shift selection. Here, $iLU(X,\nu)$ and $iC(X,\nu)$ refer, respectively, to incomplete LU and Cholesky factorization of the matrix X with drop tolerance ν. The last column indicates which ADI shift selection strategy (analytic approach by Sabino or heuristic method) is used.

Ex.	n, m	Coefficients	r	Sym.	Prec.	Shifts
1	125000	A: $\omega = 0$, $M = I_n$	5	yes	iC$(-A, 0.1)$	Sab.
	27000	B: $\omega = 0$, $C = I_m$	5	yes	iC$(-B, 0.1)$	Sab.
2	512000	A: $\omega = [x\sin(x),\ y\cos(y),\ \mathrm{e}^{z^2-1}]^T$, $M = I_n$	3	no	iLU$(A, 0.1)$	heur.
	27000	B: $\omega = [zy(x^2-1), 1/(y^2+1), \mathrm{e}^z]^T$, $C = I_m$	3	no	iLU$(B^*, 0.1)$	heur.
3	125000	A from Example 1, $M = I_n$	5	yes	iC$(-A, 0.1)$	heur.
	22500	B: two-dim. version of (4.1) with $\omega = 0$, $C = I_m$	5	yes	—	heur.
4	106641	A, M from simplifiedMachineToolFineSA1 [32]	2	yes	iC$(-A-\beta_k M, 0.1)$	heur.
	35408	B, C from simplifiedMachineToolFineSA2 [32]	2	yes	iC$(-B-\alpha_k C, 0.1)$	heur.

The coefficients in some of our experiments come from finite-difference discretizations of convection–diffusion operators

$$\mathcal{L}(u) = -\Delta u + \omega \cdot \nabla u \quad \text{on } (0,1)^3, \tag{4.1}$$

with different $\omega \in \mathbb{R}^3$. A uniform grid with n_0 points in each spatial dimension was used, leading to matrices of size $n_0^3 \times n_0^3$. The right-hand side factors f and g are drawn from a normal distribution and are rescaled so that $\|f\| = \|g\|$. Table 4.1 gives an overview of the examples and the settings for preconditioners and ADI shift selection methods.

Example 2 is set up similarly to an example in [21]. The matrix B in Example 3 comes from a two-dimensional finite-difference discretization, and the arising linear systems can be efficiently solved by direct methods. Hence, only the inner tolerances δ_k^A must be chosen. The matrices of the generalized Sylvester equation in Example 4 come from a finite-element discretization of the heat transfer across a machine tool [33]. Here, updating the preconditioners in every ADI iteration step is required to achieve a reasonable performance of the inner MINRES solver. In all other experiments, fixed preconditioners were sufficient.

We will monitor the performance of Sylvester-ADI with fixed tolerances, using $\delta_k^{A,B} = \tilde{\varepsilon}/20$, and with the proposed dynamically chosen inner solve tolerances. Unless stated otherwise, $\delta_{\min} = \tilde{\varepsilon}/20$ and $\delta_{\max} = 0.1$ are set as minimal and maximal linear solve tolerances, and $j_{\max} = 50$ and $\xi = 1$ are used. Only Example 4 required slightly stricter settings: $\delta_{\min} = \tilde{\varepsilon}/100$, $j_{\max} = 100$, and $\xi = 0.1$.

The following settings for the dynamic stopping criteria are tested:

- dynamic, no BL, mid: strategy (3.5) without back-looking, combination (3.9) of δ_k^A and δ_k^B from the middle of the admissible set;
- dynamic, BL, mid: like above, but with back-looking (3.8);
- dynamic, no BL, B: strategy (3.5) without back-looking, inner iterations on the smaller system (defined by $B^* + \overline{\alpha}C^*$) are preferred via (3.10); and
- dynamic, BL, B: like above, but with back-looking (3.8).

Inside Algorithm 1, the scaled computed Sylvester residual norm $\|\mathcal{R}^{\text{comp}}\|/\|fg^*\|$ is used for the outer stopping criterion. After termination, we also estimate the (scaled) norm of the

ETNA
Kent State University and
Johann Radon Institute (RICAM)

TABLE 4.2
Experimental results. The columns denote the used inner stopping criterion (fixed or dynamic versions), the number of required outer iterations it^{out}, the column dimension of the low-rank solution factors (dim), the final obtained scaled residual norm $\mathfrak{R}_k$, the total number of inner iteration steps $\sum it^{in,A}$ and $\sum it^{in,B}$ for the two linear systems, the computing times in seconds, and the obtained savings in computing time compared to inexact LR-ADI with fixed inner tolerances. The best results in the categories regarding the inner solves are shown in bold.

Ex.	Settings inner tol.	it^{out}	dim	$\mathfrak{R}_k$	$\sum \text{it}^{\text{in},A}$	$\sum \text{it}^{\text{in},B}$	Time	Save (%)
1	direct solves	29	145	1.9e−09	—	—	114.6	
	fixed	29	145	1.9e−09	581	738	34.3	
	dynamic, no BL, mid	29	145	1.9e−09	471	462	26.3	23.32
	dynamic, BL, mid	29	145	2.0e−09	436	**437**	**25.8**	**24.78**
	dynamic, no BL, B	29	145	1.9e−09	450	597	27.9	18.66
	dynamic, BL, B	29	145	1.9e−09	**424**	597	28.2	17.78
2	fixed	25	75	4.1e−10	836	428	234.6	
	dynamic, BL, mid	25	75	4.1e−10	602	**315**	165.4	29.92
	dynamic, BL, B	25	75	4.2e−10	**596**	350	**162.8**	**30.61**
3	direct solves	20	100	5.3e−09	—	—	77.2	
	fixed	20	100	5.3e−09	638	—	24.2	
	dynamic, BL	20	100	5.3e−09	**430**	—	**18.0**	**25.62**
4	direct solves	85	170	1.6e−09	—	—	261.5	
	fixed	85	170	1.6e−09	4204	3099	242.9	
	dynamic, BL, mid	85	170	8.9e−09	1934	**1871**	**144.8**	**40.14**
	dynamic, BL, B	85	170	7.7e−09	**1818**	2301	176.4	27.4

true Sylvester residual matrix $\|\mathcal{R}^{\text{true}}\|$ by using the `eigs` routine to get an approximation of $\lambda_{\max}((\mathcal{R}^{\text{true}})^*\mathcal{R}^{\text{true}})$. The results are summarized in Table 4.2.

For Example 1 we tested the inexact LR-Sylvester-ADI with all of the above dynamic stopping criteria as well as with fixed inner tolerances and also used the exact ADI (with direct inner solvers). The data collected in Table 4.2 indicate that, with an appropriate choice for the inner tolerances, the inexact Sylvester-ADI needs the same number of outer steps and achieves similar final Sylvester residuals as the exact counterpart, but requires significantly less computing time. This is also the case for most of the other examples. Secondly, using the dynamic stopping criteria leads overall to smaller numbers of required inner iteration steps (columns $\sum \text{it}^{\text{in},A}$ and $\sum \text{it}^{\text{in},B}$ in the table) compared to fixed inner tolerances. This leads to reduced computing times with savings between approximately 17% and 25% for Example 1.

Now comparing the various different versions of the dynamic stopping criteria, we observe that the plain version (3.5) requires slightly more inner iterations than the version with back-looking (3.8). Since using (3.8) comes with almost no additional costs, we therefore always use back-looking for the remaining examples. The strategies that allow more inner iteration steps with the smaller linear system indeed achieve this goal: the numbers $\sum \text{it}^{\text{in},A}$ are slightly decreased while $\sum \text{it}^{\text{in},B}$ are slightly increased. However, for Example 1 this does not lead to a reduction in the computing time. Some more fine-tuning regarding the selection of a combination δ_k^A and δ_k^B might be needed here.

Figure 4.1 shows that, when the scaled computed Sylvester residual norm $\mathfrak{R}_k^{\text{comp}}$ decreases in the course of the Sylvester-ADI iteration, the dynamic stopping criteria lead to increasing inner residual norms $\|r_k^A\|$ and $\|r_k^B\|$. Note that the curves for $\mathfrak{R}_k^{\text{comp}}$ were visually indistinguishable for all the tested variants, and therefore only one curve is shown in Figure 4.1. In Figure 4.2 the cumulative sum of the inner iteration steps is illustrated for different inner stopping criteria. We clearly see that the dynamic criteria lead to a significantly reduced slope

ETNA
Kent State University and
Johann Radon Institute (RICAM)

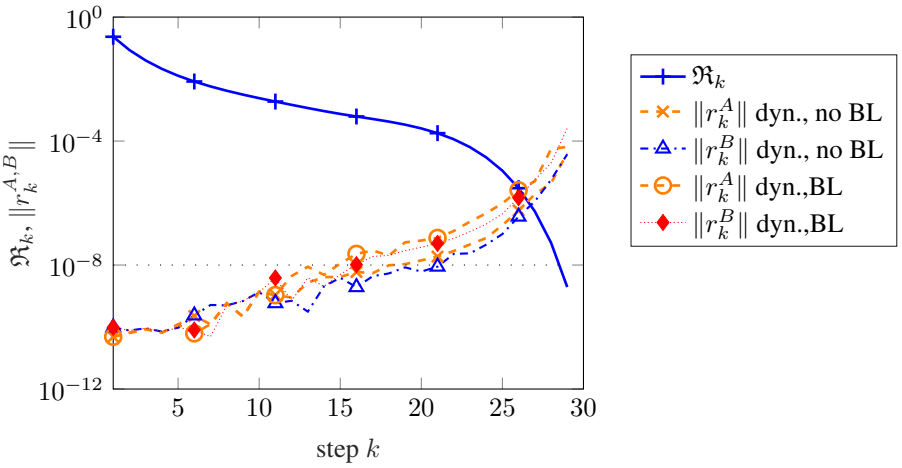

FIG. 4.1. *Residual norms for Example 1: Scaled computed residual norms* $\mathfrak{R}_k^{\text{comp}}$ *and inner residual norms* $\|r_k^A\|$ *and* $\|r_k^B\|$ *against the outer iteration number for different dynamic stopping criteria. Only one curve for* $\mathfrak{R}_k^{\text{comp}}$ *is shown.*

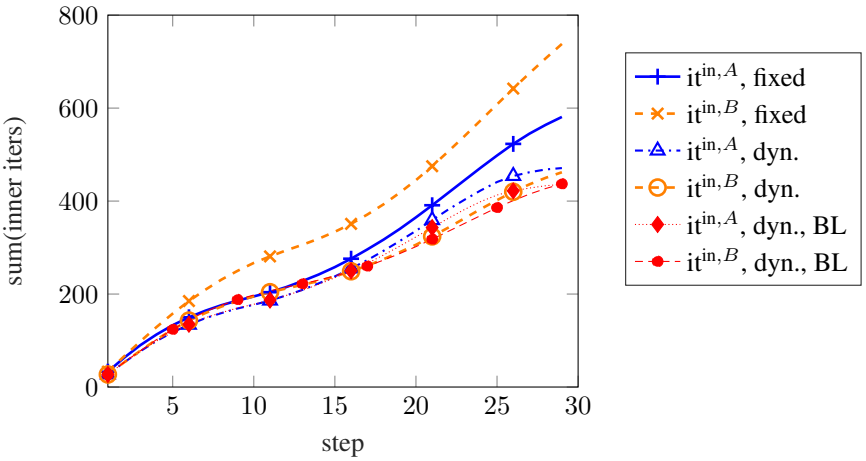

FIG. 4.2. *Inner iteration numbers for Example 1: Cumulative sum of the inner iteration steps against the outer iteration number for fixed inner tolerances and different dynamic stopping criteria.*

compared to fixed tolerances, which is reduced slightly further when back-looking (3.8) is equipped.

For Example 2 we can make a similar observation from the data in Table 4.2. Using the dynamic stopping criteria leads to savings in the computing times of roughly 30%. Figure 4.3 illustrates again the history of the scaled computed Sylvester residual norms $\mathfrak{R}_k^{\text{comp}}$ and the inner residual norms $\|r_k^A\|$ and $\|r_k^A\|$. For Example 2, the strategy [dynamic, BL, B] actually leads to a very small reduction in the computing time due to a small change in the inner iteration numbers $\sum \text{it}^{\text{in},A}$ and $\sum \text{it}^{\text{in},B}$. This is also visible in Figure 4.3, where the inner residual norms $\|r^A\|$ in this variant are slightly larger but the $\|r_k^B\|$ (corresponding to the much smaller linear systems) are kept at a much lower level.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

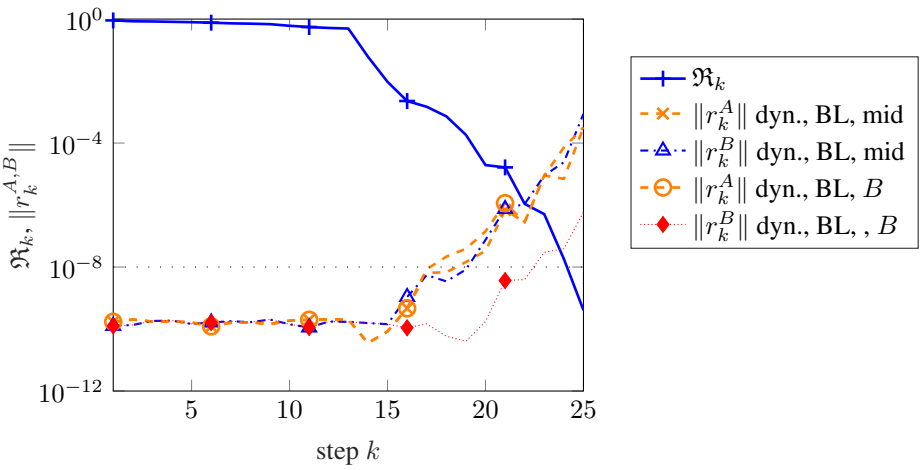

FIG. 4.3. *Residual norms for Example 2: Scaled computed residual norms* $\mathfrak{R}_k^{\text{comp}}$ *and inner residual norms* $\|r_k^A\|$ *and* $\|r_k^A\|$ *against the outer iteration number for different dynamic stopping criteria.*

In Example 3 only the sequence of linear systems defined by A is solved iteratively, while direct solvers are used for the other one defined by B. Hence, only the threshold for $\|r_k^A\|$ must be adjusted by simply setting $\|r_k^B\| = 0$ in (3.8). The results in Table 4.2 show again reduced numbers of inner iterations gained with the dynamic stopping strategies and savings of approximately 25% in the computing time.

The results for the generalized Sylvester equation of Example 4 are in line with those of the previous examples. Here, the dynamic criteria lead to runtime savings between roughly 27% and 40%. The strategy [dynamic, BL, B] does lead to higher computation times than [dynamic, BL, mid], but overall the total number of inner iteration steps as well as the computing time are lower compared to using fixed inner tolerances. The arising linear systems in Example 4 seem to be harder for the inner solver (MINRES) compared to the other examples. That is the reason why we opted to update the preconditioners in every step. Using smaller drop tolerances for the incomplete Cholesky factorization did not lead to significant improvements.

A further noteworthy observation in these and other experiments (not reported here) is that sometimes, in the inexact ADI iteration with fixed tolerances, the required residual thresholds could not or could hardly be achieved by the inner solver. This was much less frequently an issue with the proposed dynamic criteria because there the smallest inner residual norms are typically only required in the first outer iteration steps. Moreover, for some examples, looser fixed inner tolerance might work, too, potentially at an increase of the number of outer iteration steps. With this setting, however, the residual gap might be larger so that $\|\mathcal{R}^{\text{comp}}\|$ might not be a correct indicator for the achieved accuracy. One would have to estimate the true Sylvester residual norm (e.g., via a Lanczos process), which is more costly than using the norm of the computed Sylvester residual. Choosing the fixed tolerances too loose can also lead to a stagnation of $\|\mathcal{R}^{\text{true}}\|$, although $\|\mathcal{R}^{\text{comp}}\|$ indicates a decrease, which is a common issue in inexact (rational) Krylov methods.

5. Conclusions and future research perspectives. We considered the inexact low-rank ADI iteration for large-scale Sylvester equations and proposed dynamic stopping criteria for the inner solvers which are used to iteratively solve the arising linear systems. We provided theoretical results showing that, with an appropriate choice for the inner accuracies, the residuals in inexact ADI iteration are only a small perturbation of the exact ADI iteration.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Moreover, the practical implementation of the dynamic stopping criteria was discussed, and numerical experiments confirmed the effectiveness of these strategies, leading to fewer inner iteration steps and, hence, shorter computing times compared to the case when constant inner tolerances were used.

A potential next research direction might be subspace recycling techniques for the sequences of shifted linear systems by, for example, storing the Krylov basis obtained from solving one linear system. This was discussed, for example, for the Lyapunov LR-ADI in [25], and a similar idea was developed in [4], leading to an efficient hybrid method. Corresponding strategies for the Sylvester-ADI iteration are promising future research directions. Conducting similar studies for the linear systems inside rational Krylov projection methods for Sylvester equations [30] is, of course, also a further worthwhile research direction.

REFERENCES

[1] A. C. ANTOULAS, *Approximation of Large-Scale Dynamical Systems*, SIAM, Philadelphia, 2005.
[2] R. H. BARTELS AND G. W. STEWART, *Solution of the matrix equation $AX + XB = C$: Algorithm 432*, Comm. ACM, 15 (1972), pp. 820–826.
[3] B. BECKERMANN, *An error analysis for rational Galerkin projection applied to the Sylvester equation*, SIAM J. Numer. Anal., 49 (2011), pp. 2430–2450.
[4] D. BENNER, P. PALITTA, AND J. SAAK, *On an integrated Krylov-ADI solver for large-scale Lyapunov equations*, Numer. Algorithms, 92 (2023).
[5] P. BENNER, M. KÖHLER, AND J. SAAK, *Sparse-dense Sylvester equations in H_2-model order reduction*, Preprint MPIMD/11-11, Max Planck Institute Magdeburg, Magdeburg, 2011.
[6] P. BENNER AND P. KÜRSCHNER, *Computing real low-rank solutions of Sylvester equations by the factored ADI method*, Comput. Math. Appl., 67 (2014), pp. 1656–1672.
[7] P. BENNER, P. KÜRSCHNER, AND J. SAAK, *Self-generating and efficient shift parameters in ADI methods for large Lyapunov and Sylvester equations*, Electron. Trans. Numer. Anal., 43 (2014), pp. 142–162. https://etna.ricam.oeaw.ac.at/vol.43.2014/pp142-162.dir/pp142-162.pdf
[8] P. BENNER, R.-C. LI, AND N. TRUHAR, *On the ADI method for Sylvester equations*, J. Comput. Appl. Math., 233 (2009), pp. 1035–1045.
[9] A. BOURAS AND V. FRAYSSÉ, *Inexact matrix-vector products in Krylov methods for solving linear systems: a relaxation strategy*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 660–678.
[10] T. BREITEN, V. SIMONCINI, AND M. STOLL, *Low-rank solvers for fractional differential equations*, Electron. Trans. Numer. Anal., 45 (2016), pp. 107–132. https://etna.ricam.oeaw.ac.at/vol.45.2016/pp107-132.dir/pp107-132.pdf
[11] D. CALVETTI AND L. REICHEL, *Application of ADI iterative methods to the restoration of noisy images*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 165–186.
[12] A. CASULLI AND L. ROBOL, *An efficient block rational Krylov solver for Sylvester equations with adaptive pole selection*, SIAM J. Sci. Comput., 46 (2024), pp. A798–A824.
[13] R. CLOUÂTRE, B. KLIPPENSTEIN, AND R. M. SLEVINSKY, *Lifting Sylvester equations: singular value decay for non-normal coefficients*, Preprint on arXiv, 2023. https://arxiv.org/abs/2308.11533
[14] M. CROUZEIX AND C. PALENCIA, *The numerical range is a $(1+\sqrt{2})$-spectral set*, SIAM J. Matrix Anal. Appl., 38 (2017), pp. 649–655.
[15] V. DRUSKIN, L. A. KNIZHNERMAN, AND V. SIMONCINI, *Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation, SIAM* J. Numer. Anal., 49 (2011), pp. 1875–1898.
[16] A. EL GUENNOUNI, K. JBILOU, AND A. RIQUET, *Block Krylov subspace methods for solving large Sylvester equations*, Numer. Algorithms, 29 (2002), pp. 75–96.
[17] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, 4th ed., Johns Hopkins University Press, Baltimore, 2013.
[18] L. GRASEDYCK, *Existence of a low rank or $\mathcal{H}$-matrix approximant to the solution of a Sylvester equation*, Numer. Linear Algebra Appl., 11 (2004), pp. 371–389.
[19] D. Y. HU AND L. REICHEL, *Krylov-subspace methods for the Sylvester equation*, Linear Algebra Appl., 172 (1992), pp. 283–313.
[20] E. JARLEBRING, G. MELE, D. PALITTA, AND E. RINGH, *Krylov methods for low-rank commuting generalized Sylvester equations*, Numer. Linear Algebra Appl, 25 (2018), Paper No. e2176 (17 pages).
[21] D. KRESSNER, K. LUND, S. MASSEI, AND D. PALITTA, *Compress-and-restart block Krylov subspace methods for Sylvester matrix equations*, Numer. Linear Algebra Appl., 28 (2021), Paper No. e2339 (17 pages).

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[22] P. KÜRSCHNER, *Efficient Low-Rank Solution of Large-Scale Matrix Equations*, PhD. Thesis, Fakultät für Mathematik, Otto von Guericke Universität, Magdeburg, April 2016.
[23] ———, *Low-rank Sylvester ADI implementations (v1.0)*, Software, 2024. https://zenodo.org/records/10454177
[24] P. KÜRSCHNER AND M. FREITAG, *Inexact methods for the low rank solution to large scale Lyapunov equations*, BIT, 60 (2020), pp. 1221–1259.
[25] J.-R. LI, *Model Reduction of Large Linear Systems via Low Rank System Gramians*, PhD. Thesis, Dept. of Math., Massachusettes Institute of Technology, Cambridge, September 2000.
[26] J.-R. LI AND J. WHITE, *Low rank solution of Lyapunov equations*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 260–280.
[27] Z. LIU, Y. ZHOU, AND Y. ZHANG, *On inexact alternating direction implicit iteration for continuous Sylvester equations*, Numer. Linear Algebra Appl., 27 (2020), Paper No. e2320 (13 pages).
[28] D. P. O'LEARY, *The block conjugate gradient algorithm and related methods*, Linear Algebra Appl., 29 (1980), pp. 293–322.
[29] D. PALITTA AND P. KÜRSCHNER, *On the convergence of low-rank Krylov methods*, Numer. Algorithms, 88 (2021), pp. 1383–1417.
[30] D. PALITTA AND V. SIMONCINI, *Computationally enhanced projection methods for symmetric Sylvester and Lyapunov matrix equations*, J. Comput. Appl. Math., 330 (2018), pp. 648–659.
[31] J. SABINO, *Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Method*, PhD. Thesis, Rice University, Houston, June 2007.
Available from: https://hdl.handle.net/1911/102054.
[32] S. SAUERZAPF, A. NAUMANN, J. VETTERMANN, AND J. SAAK, *Matrices for a simplified machine tool model*, website hosted at MORwiki—Model Order Reduction Wiki, 2023.
[33] S. SAUERZAPF, J. VETTERMANN, A. NAUMANN, J. SAAK, M. BEITELSCHMIDT, AND P. BENNER, *Simulation of the thermal behavior of machine tools for efficient machine development and online correction of the Tool Center Point (TCP)-displacement*, in Conference Proceedings on Thermal Issues, Aachen, 26-27 February, euspen, 2020, pp. 135–138.
[34] V. SIMONCINI, *Variable accuracy of matrix-vector products in projection methods for eigencomputation*, SIAM J. Numer. Anal., 43 (2005), pp. 1155–1174.
[35] ———, *Computational methods for linear matrix equations*, SIAM Rev., 58 (2016), pp. 377–441.
[36] V. SIMONCINI AND L. ELDÉN, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.
[37] K. SOODHALTER, *A block MINRES algorithm based on the banded Lanczos method*, Numer. Algorithms, 69 (2015), pp. 473–494.
[38] D. C. SORENSEN AND A. C. ANTOULAS, *The Sylvester equation and approximate balanced reduction*, Linear Algebra Appl., 351/352 (2002), pp. 671–700.
[39] M. STOLL AND T. BREITEN, *A low-rank in time approach to PDE-constrained optimization*, SIAM J. Sci. Comput., 37 (2015), pp. B1–B29.
[40] E. L. WACHSPRESS, *Iterative solution of the Lyapunov matrix equation*, Appl. Math. Lett., 1 (1988), pp. 87–90.
[41] ———, *The ADI Model Problem*, Springer, New York, 2013.