Journal of Computational and Applied Mathematics 330 (2018) 648–659

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Computationally enhanced projection methods for symmetric Sylvester and Lyapunov matrix equations☆

Davide Palitta [a,*], Valeria Simoncini [a,b]

[a] Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, I-40127 Bologna, Italy
[b] IMATI-CNR, Pavia, Italy

ARTICLE INFO

Article history:
Received 16 February 2016
Received in revised form 1 February 2017

MSC 2010:
47J20
65F30
49M99
49N35
93B52

Keywords:
Sylvester equation
Lyapunov equation
Projection methods
Krylov subspaces

ABSTRACT

In the numerical treatment of large-scale Sylvester and Lyapunov equations, projection methods require solving a reduced problem to check convergence. As the approximation space expands, this solution takes an increasing portion of the overall computational effort. When data are symmetric, we show that the Frobenius norm of the residual matrix can be computed at significantly lower cost than with available methods, without explicitly solving the reduced problem. For certain classes of problems, the new residual norm expression combined with a memory-reducing device make classical Krylov strategies competitive with respect to more recent projection methods. Numerical experiments illustrate the effectiveness of the new implementation for standard and extended Krylov subspace methods.

1. Introduction

Consider the Sylvester matrix equation

$$AX + XB + C_1C_2^T = 0, \quad A \in \mathbb{R}^{n_1\times n_1},\ B \in \mathbb{R}^{n_2\times n_2},\ C_1 \in \mathbb{R}^{n_1\times s}, C_2 \in \mathbb{R}^{n_2\times s} \tag{1}$$

where A, B are very large and sparse, symmetric negative definite matrices, while $C_1, C_2 \neq 0$ are tall, that is $s \ll n_1, n_2$. Under these hypotheses, there exists a unique solution matrix X. This kind of matrix equation arises in many applications, from the analysis of continuous-time linear dynamical systems to eigenvalue problems and the discretization of self-adjoint elliptic PDEs; see, e.g., [1], and [2] for a recent survey. Although A and B are sparse, the solution X is in general dense so that storing it may be unfeasible for large-scale problems. On the other hand, under certain hypotheses on the spectral distribution of A and B, the singular values of X present a fast decay, see, e.g., [3], thus justifying the search for a low-rank approximation $\widetilde{X} = Z_1Z_2^T$ to X so that only these two tall matrices are actually computed and stored. To simplify the presentation of what follows, from now on we will focus on the case of the Lyapunov matrix equation, that is $B = A$ ($n \equiv n_1 = n_2$) and $C \equiv C_1 = C_2$, so that X will be square, symmetric and positive semidefinite [4]. In later sections we will describe how to naturally treat the general case with A and B distinct and not necessarily with the same dimensions, and different C_1, C_2.

☆ This research is supported in part by the FARB12SIMO grant of the Università di Bologna, and in part by INdAM-GNCS under the 2015 Project *Metodi di regolarizzazione per problemi di ottimizzazione e applicazioni.*

* Corresponding author.
E-mail addresses: davide.palitta3@unibo.it (D. Palitta), valeria.simoncini@unibo.it (V. Simoncini).

http://dx.doi.org/10.1016/j.cam.2017.08.011
0377-0427/

For the Lyapunov equation, projection methods compute the numerical solution $\widetilde{X}$ in a sequence of nested subspaces, $\mathcal{K}_m \subseteq \mathcal{K}_{m+1} \subseteq \mathbb{R}^n$, $m \geq 1$. The approximation, usually denoted by X_m, is written as the product of low-rank matrices $X_m = V_m Y_m V_m^T$ where $\mathcal{K}_m = \text{Range}(V_m)$ and the columns of V_m are far fewer than n. The quality and effectiveness of the approximation process depend on how much spectral information is captured by $\mathcal{K}_m$, without the space dimension being too large. The matrix Y_m is determined by solving a related (reduced) problem, whose dimension depends on the approximation space dimension. To check convergence, the residual matrix norm is monitored at each iteration by using Y_m but without explicitly computing the large and dense residual matrix $R_m = AX_m + X_m A + CC^T$ [2]. The solution of the reduced problem is meant to account for a low percentage of the overall computation cost. Unfortunately, this cost grows nonlinearly with the space dimension, therefore solving the reduced problem may become very expensive if a large approximation space is needed.

A classical choice for $\mathcal{K}_m$ is the (standard) block Krylov subspace $\mathbf{K}_m^{\square}(A, C) := \text{Range}\{[C, AC, \ldots, A^{m-1}C]\}$ [5], whose basis can be generated iteratively by means of the block Lanczos procedure. Numerical experiments show that $\mathbf{K}_m^{\square}(A, C)$ may need to be quite large before a satisfactory approximate solution is obtained [6,7]. This large number of iterations causes high computational and memory demands. More recent alternatives include projection onto extended or rational Krylov subspaces [7,8], or the use of explicit iterations for the approximate solution [6]; see the thorough presentation in [2]. Extended and more generally rational Krylov subspaces contain richer spectral information, that allow for a significantly lower subspace dimension, at the cost of more expensive computations per iteration, since s system solves with the coefficients matrix are required at each iteration.

We devise a strategy that significantly reduces the computational cost of evaluating the residual norm for both $\mathbf{K}_m^{\square}$ and the extended Krylov subspace $\mathbf{EK}_m^{\square}(A, C) := \text{Range}\{[C, A^{-1}C, \ldots, A^{m-1}C, A^{-m}C]\}$. In case of $\mathbf{K}_m^{\square}$ a "two-pass" strategy is implemented to avoid storing the whole basis V_m; see [9] for earlier use of this device in the same setting, and, e.g., [10] in the matrix function context.

Throughout the paper, Greek bold letters ($\boldsymbol{\alpha}$) will denote $s \times s$ matrices, while roman capital letters (A) larger ones. In particular $E_i \in \mathbb{R}^{sm \times s}$ will denote the ith block of s columns of the identity matrix $I_{sm} \in \mathbb{R}^{sm \times sm}$. Scalar quantities will be denoted by Greek letters (α).

Here is a synopsis of the paper. In Section 2 the basic tools of projection methods for solving (1) are recalled. In Section 3.1 we present a cheap residual norm computation whose implementation is discussed in Section 3.2. The two-pass strategy for $\mathbf{K}_m^{\square}(A, C)$ is examined in Section 3.3. In Section 4 we extend the residual computation to $\mathbf{EK}_m^{\square}(A, C)$. Section 5 discusses the generalization of this procedure to the case of the Sylvester equation in (1). In particular, Section 5.1 analyzes the case when both coefficient matrices are large, while Section 5.2 discusses problems where one of them has small dimensions. Numerical examples illustrating the effectiveness of our strategy are reported in Section 6, while our conclusions are given in Section 7.

2. Galerkin projection methods

Consider a subspace $\mathcal{K}_m$ spanned by the orthonormal columns of the matrix $V_m = [\mathcal{V}_1, \ldots, \mathcal{V}_m] \in \mathbb{R}^{n \times sm}$ and seek an approximate solution X_m to (1) of the form $X_m = V_m Y_m V_m^T$ with Y_m symmetric and positive semidefinite, and residual matrix $R_m = AX_m + X_m A + CC^T$. With the matrix inner product

$$\langle Q, P \rangle_F := \text{trace}(P^T Q), \qquad Q, P \in \mathbb{R}^{n_1 \times n_2},$$

the matrix Y_m can be determined by imposing an orthogonality (Galerkin) condition on the residual with respect to this inner product,

$$R_m \perp \mathcal{K}_m \quad \Leftrightarrow \quad V_m^T R_m V_m = 0. \tag{2}$$

Substituting R_m into (2), we obtain $V_m^T A X_m V_m + V_m^T X_m A V_m + V_m^T CC^T V_m = 0$, that is

$$\left(V_m^T A V_m\right) Y_m V_m^T V_m + V_m^T V_m Y_m \left(V_m^T A V_m\right) + V_m^T CC^T V_m = 0. \tag{3}$$

We assume $\text{Range}(V_1) = \text{Range}(C)$, that is $C = V_1 \boldsymbol{\gamma}$ for some nonsingular $\boldsymbol{\gamma} \in \mathbb{R}^{s \times s}$. Since V_m has orthonormal columns, $V_m^T C = E_1 \boldsymbol{\gamma}$ and Eq. (3) can be written as

$$T_m Y_m + Y_m T_m + E_1 \boldsymbol{\gamma} \boldsymbol{\gamma}^T E_1^T = 0, \tag{4}$$

where $T_m := V_m^T A V_m$ is symmetric and negative definite. The orthogonalization procedure employed in building V_m determines the sparsity pattern of T_m. In particular, for $\mathcal{K}_m = \mathbf{K}_m^{\square}(A, C)$, the block Lanczos process produces a block tridiagonal matrix T_m with blocks of size s,

$$T_m = \begin{pmatrix} \boldsymbol{\tau}_{11} & \boldsymbol{\tau}_{12} & & & \\ \boldsymbol{\tau}_{21} & \boldsymbol{\tau}_{22} & \boldsymbol{\tau}_{23} & & \\ & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \boldsymbol{\tau}_{m-1,m} \\ & & \boldsymbol{\tau}_{m,m-1} & \boldsymbol{\tau}_{m,m} \end{pmatrix}.$$

As long as m is of moderate size, methods based on the Schur decomposition of the coefficient matrix T_m can be employed to solve Eq. (4), see, e.g., [11,12].

The last s columns (or rows) of the solution matrix Y_m are employed to compute the residual norm. In particular, letting $\underline{T}_m = V_{m+1}^T A V_m$, it was shown in [13] that the norm of the residual in (2) satisfies

$$\|R_m\|_F = \sqrt{2}\|Y_m \underline{T}_m^T E_{m+1}\|_F = \sqrt{2}\|Y_m E_m \boldsymbol{\tau}_{m+1,m}^T\|_F. \tag{5}$$

The matrix Y_m is determined by solving (4), and it is again symmetric and positive semidefinite. At convergence, the backward transformation $X_m = V_m Y_m V_m^T$ is never explicitly computed or stored. Instead, we factorize Y_m as

$$Y_m = \widehat{Y}\widehat{Y}^T, \quad \widehat{Y} \in \mathbb{R}^{sm\times sm}, \tag{6}$$

from which a low-rank factor of X_m is obtained as $Z_m = V_m\widehat{Y} \in \mathbb{R}^{n\times sm}$, $X_m = Z_m Z_m^T$. The matrix Y_m may be numerically rank deficient, and this can be exploited to further decrease the rank of Z_m. We write the eigendecomposition of Y_m, $Y_m = W\Sigma W^T$ (with eigenvalues ordered non-increasingly) and discard only the eigenvalues below a certain tolerance, that is $\Sigma = \mathrm{diag}(\Sigma_1, \Sigma_2)$, $W = [W_1, W_2]$ with $\|\Sigma_2\|_F \le \epsilon$ (in all our experiments we used $\epsilon = 10^{-12}$). Therefore, we define again $Y_m \approx \widehat{Y}\widehat{Y}^T$, with $\widehat{Y} = W_1\Sigma_1^{1/2} \in \mathbb{R}^{sm\times t}$, $t \le sm$; in this way, $\|Y_m - \widehat{Y}\widehat{Y}^T\|_F \le \epsilon$. Hence, we set $Z_m = V_m\widehat{Y} \in \mathbb{R}^{n\times t}$. We notice that a significant rank reduction in Y_m is an indication that all relevant information for generating X_m is actually contained in a subspace that is much smaller than $\mathbf{K}_m^{\square}(A, C)$. In other words, the generated Krylov subspace is not efficient in capturing the solution information and a much smaller space could have been generated to obtain an approximate solution of comparable accuracy.

Algorithm 1: Galerkin projection method for the Lyapunov matrix equation
Input: $A \in \mathbb{R}^{n\times n}$, A symmetric and negative definite, $C \in \mathbb{R}^{n\times s}$
Output: $Z_m \in \mathbb{R}^{n\times t}$, $t \le sm$

1. Set $\beta = \|C\|_F$
2. Perform economy-size QR of C, $C = V_1\boldsymbol{\gamma}$. Set $\mathcal{V}_1 \equiv V_1$
3. For $m = 2, 3, \ldots,$ till convergence, **Do**
4. Compute next basis block $\mathcal{V}_m$ and set $V_m = [V_{m-1}, \mathcal{V}_m]$
5. Update $T_m = V_m^T A V_m$
6. **Convergence check:**
6.1 Solve $T_m Y_m + Y_m T_m + E_1\boldsymbol{\gamma}\boldsymbol{\gamma}^T E_1^T = 0, \quad E_1 \in \mathbb{R}^{ms\times s}$
6.2 Compute $\|R_m\|_F = \sqrt{2}\|Y_m E_m \boldsymbol{\tau}_{m+1,m}^T\|_F$
6.3 If $\|R_m\|_F/\beta^2$ is small enough **Stop**
7. EndDo
8. Compute the eigendecomposition of Y_m and retain $\widehat{Y} \in \mathbb{R}^{sm\times t}$, $t \le sm$
9. Set $Z_m = V_m\widehat{Y}$

Algorithm 1 describes the generic Galerkin procedure to determine V_m, Y_m and Z_m as m grows, see, e.g., [2]. Methods thus differ for the choice of the approximation space. If the block Krylov space $\mathbf{K}_m^{\square}(A, C)$ is chosen, the block Lanczos method can be employed in line 4 of Algorithm 1. In exact arithmetic,

$$\mathcal{V}_m\boldsymbol{\tau}_{m+1,m} = A\mathcal{V}_{m-1} - \mathcal{V}_{m-1}\boldsymbol{\tau}_{m,m} - \mathcal{V}_{m-2}\boldsymbol{\tau}_{m-1,m}. \tag{7}$$

Algorithm 2 describes this process at iteration m, with $W = A\mathcal{V}_{m-1}$, where the orthogonalization coefficients $\boldsymbol{\tau}$'s are computed by the modified block Gram–Schmidt procedure (MGS), see, e.g., [14]; to ensure local orthogonality in finite precision arithmetic, MGS is performed twice (beside each command is the leading computational cost of the operation). To simplify the presentation, we assume throughout that the generated basis is full rank. Deflation could be implemented as it is customary in block methods whenever rank deficiency is detected.

We emphasize that only the last $3s$ terms of the basis must be stored, and the computational cost of Algorithm 2 is fixed with respect to m. In particular, at each iteration m, Algorithm 2 costs $\mathcal{O}\left((19n + s)s^2\right)$ flops.

As the approximation space expands, the principal costs of Algorithm 1 are steps 4 and 6.1. In particular, the computation of the whole matrix Y_m requires full matrix–matrix operations and a Schur decomposition of the coefficient matrix T_m, whose costs are $\mathcal{O}\left((sm)^3\right)$ flops. Clearly, step 6.1 becomes comparable with step 4 in cost for $sm \gg 1$, for instance if convergence is slow, so that $m \gg 1$.

Step 9 of Algorithm 1 shows that at convergence, the whole basis must be saved to return the factor Z_m. This represents a major shortcoming when convergence is slow, since V_m may require large memory allocations.

3. Standard Krylov subspace

For the block space $\mathbf{K}_m^{\square}(A, C)$, we devise a new residual norm expression and discuss the two-pass strategy.

Algorithm 2: One step of block Lanczos with block MGS
Input: $m, W, \mathcal{V}_{m-2}, \mathcal{V}_{m-1} \in \mathbb{R}^{n\times s}$
Output: $\mathcal{V}_m \in \mathbb{R}^{n\times s}, \boldsymbol{\tau}_{m-1,m}, \boldsymbol{\tau}_{m,m}, \boldsymbol{\tau}_{m+1,m} \in \mathbb{R}^{s\times s}$

1. Set $\boldsymbol{\tau}_{m-1,m} = \boldsymbol{\tau}_{m,m} = \mathbf{0}$
2. For $l = 1, 2,$ **Do**
3. **For** $i = m-1, m,$ **Do**
3. Compute $\boldsymbol{\alpha} = \mathcal{V}_{i-1}^T W$ $\leftarrow (2n-1)s^2$ flops
5. Set $\boldsymbol{\tau}_{i,m} = \boldsymbol{\tau}_{i,m} + \boldsymbol{\alpha}$ $\leftarrow s^2$ flops
6. Compute $W = W - \mathcal{V}_{i-1}\boldsymbol{\alpha}$ $\leftarrow 2s^2 n$ flops
7. **EndDo**
8. EndDo
9. Perform economy-size QR of W, $W = \mathcal{V}_m \boldsymbol{\tau}_{m+1,m}$ $\leftarrow 3ns^2$ flops

3.1. Computing the residual norm without the whole solution $\mathbf{Y}_m$

The solution of the projected problem (4) requires the Schur decomposition of T_m. For real symmetric matrices, the Schur decomposition amounts to the eigendecomposition $T_m = Q_m \Lambda_m Q_m^T$, $\Lambda_m = \operatorname{diag}(\lambda_1, \ldots, \lambda_{sm})$, and the symmetric block tridiagonal structure of T_m can be exploited so as to use only $\mathcal{O}((sm)^2)$ flops; see Section 3.2 for further details. Eq. (4) can thus be written as

$$\Lambda_m \widetilde{Y} + \widetilde{Y}\Lambda_m + Q_m^T E_1 \boldsymbol{\gamma}\boldsymbol{\gamma}^T E_1^T Q_m = 0, \quad \text{where} \quad \widetilde{Y} := Q_m^T Y_m Q_m. \tag{8}$$

Since Λ_m is diagonal, the entries of $\widetilde{Y}$ can be computed by substitution [2, Section 4], so that

$$Y_m = Q_m \widetilde{Y} Q_m^T = -\, Q_m \left(\frac{e_i^T Q_m^T E_1 \boldsymbol{\gamma}\boldsymbol{\gamma}^T E_1^T Q_m e_j}{\lambda_i + \lambda_j} \right)_{ij} Q_m^T, \tag{9}$$

where e_k denotes the kth vector of the canonical basis of $\mathbb{R}^{sm}$. It turns out that only the quantities within parentheses in (9) are needed for the residual norm computation, thus avoiding the $\mathcal{O}\left((sm)^3\right)$ cost of recovering Y_m.

Proposition 1. *Let $T_m = Q_m \Lambda_m Q_m^T$ denote the eigendecomposition of T_m. Then*

$$\|R_m\|_F^2 = 2\left(\|e_1^T S_m D_1^{-1} W_m\|_2^2 + \cdots + \|e_{sm}^T S_m D_{sm}^{-1} W_m\|_2^2 \right), \tag{10}$$

where $S_m = Q_m^T E_1 \boldsymbol{\gamma}\boldsymbol{\gamma}^T E_1^T Q_m \in \mathbb{R}^{sm\times sm}$, $W_m = Q_m^T E_m \boldsymbol{\tau}_{m+1,m}^T \in \mathbb{R}^{sm\times s}$ and $D_j = \lambda_j I_{sm} + \Lambda_m$ for all $j = 1, \ldots, sm$.

Proof. Exploiting (5) and the representation formula (9) we have

$$\begin{aligned} \|R_m\|_F^2 &= 2\|Y_m E_m \boldsymbol{\tau}_{m+1,m}^T\|_F^2 = 2 \left\| \left(\frac{e_i^T Q_m^T E_1 \boldsymbol{\gamma}\boldsymbol{\gamma}^T E_1^T Q_m e_j}{\lambda_i + \lambda_j} \right)_{ij} Q_m^T E_m \boldsymbol{\tau}_{m+1,m}^T \right\|_F^2 \\ &= 2\sum_{k=1}^{s} \left\| \left(\frac{e_i^T S_m e_j}{\lambda_i + \lambda_j} \right)_{ij} W_m e_k \right\|_2^2 . \end{aligned} \tag{11}$$

For all $k = 1, \ldots, s$, we can write

$$\begin{aligned} \left\| \left(\frac{e_i^T S_m e_j}{\lambda_i + \lambda_j} \right)_{ij} W_m e_k \right\|_2^2 &= \left(\sum_{j=1}^{sm} \frac{e_1^T S_m e_j}{\lambda_1 + \lambda_j} e_j^T W_m e_k \right)^2 + \ldots + \left(\sum_{j=1}^{sm} \frac{e_{sm}^T S_m e_j}{\lambda_{sm} + \lambda_j} e_j^T W_m e_k \right)^2 \\ &= \left(e_1^T S_m D_1^{-1} W_m e_k \right)^2 + \cdots + \left(e_{sm}^T S_m D_{sm}^{-1} W_m e_k \right)^2 . \end{aligned} \tag{12}$$

Plugging (12) into (11) we have

$$\begin{aligned} \|R_m\|_F^2 &= 2\sum_{k=1}^{s}\sum_{i=1}^{sm} \left(e_i^T S_m D_i^{-1} W_m e_k \right)^2 = 2\sum_{i=1}^{sm}\sum_{k=1}^{s} \left(e_i^T S_m D_i^{-1} W_m e_k \right)^2 \\ &= 2\sum_{i=1}^{sm} \left\| e_i^T S_m D_i^{-1} W_m \right\|_2^2 . \quad \square \end{aligned}$$

Algorithm 3: `cTri`
Input: $T_m \in \mathbb{R}^{\ell m \times \ell m}$, $\boldsymbol{\gamma}$, $\boldsymbol{\tau}_{m+1,m} \in \mathbb{R}^{\ell \times \ell}$ (ℓ is the block size)
Output: *res* ($= \|R\|_F$)

1. Tridiagonalize $P_m^T T_m P_m = F_m$
2. Compute $F_m = G_m \Lambda_m G_m^T$
3. Compute $E_1^T Q_m = \left(E_1^T P_m\right) G_m$, $E_m^T Q_m = \left(E_m^T P_m\right) G_m$
4. Compute $S_m = \left(Q_m^T E_1 \boldsymbol{\gamma}\right)\left(\boldsymbol{\gamma}^T E_1^T Q_m\right)$ $\quad \leftarrow (2\ell-1)\ell^2 m + (2\ell-1)\ell^2 m^2$ *flops*
5. Compute $W_m = \left(Q_m^T E_m\right)\boldsymbol{\tau}_{m+1,m}^T$ $\quad \leftarrow (2\ell-1)\ell^2 m$ *flops*
6. Set $res = 0$
7. For $i = 1, \ldots, \ell m$, **Do**
8. Set $D_i = \lambda_i I_{\ell m} + \Lambda_m$
9. $res = res + \left\| \left(e_i^T S_m\right) D_i^{-1} W_m \right\|_2^2$ $\quad \leftarrow 2\ell^2 m + \ell m + \ell$ *flops*
10. EndDo
11. Set $res = \sqrt{2}\sqrt{res}$

3.2. The algorithm for the residual norm computation

Algorithm 3 summarizes the procedure that takes advantage of Proposition 1. Computing the residual norm by (11) has a leading cost of $4s^3m^2$ flops for standard Krylov (with $\ell = s$). This should be compared with the original procedure in steps 6.1 and 6.2 of Algorithm 1, whose cost is $\mathcal{O}\left(s^3m^3\right)$ flops, with a large constant. Proposition 1 also shows that only the first and last ℓ components of the eigenvectors of T_m are necessary in the residual norm evaluation and the computation of the complete eigendecomposition $T_m = Q_m \Lambda_m Q_m^T$ may be avoided. To this end, the matrix T_m can be tridiagonalized, $P_m^T T_m P_m = F_m$, explicitly computing only the first and last ℓ rows of the transformation matrix P_m, namely $E_1^T P_m$ and $E_m^T P_m$. The eigendecomposition $F_m = G_m \Lambda_m G_m^T$ is computed exploiting the tridiagonal structure of F_m. The matrices $E_1^T Q_m$ and $E_m^T Q_m$ needed in (10) are then computed as $E_1^T Q_m = \left(E_1^T P_m\right) G_m$, $E_m^T Q_m = \left(E_m^T P_m\right) G_m$.

Once the stopping criterion in step 6.3 of Algorithm 1 is satisfied, the factor Z_m can be finally computed. Once again, this can be performed without explicitly computing Y_m, which requires the expensive computation $Y_m = Q_m \widetilde{Y} Q_m^T$. Indeed, the truncation strategy discussed around (6) can be applied to $\widetilde{Y}$ by computing the matrix $\check{Y} \in \mathbb{R}^{sm \times t}$, $t \le sm$ so that $\widetilde{Y} \approx \check{Y}\check{Y}^T$. This factorization further reduces the overall computational cost, since only $(2ms-1)tms$ flops are required to compute $Q_m\check{Y}$, with no loss of information at the prescribed accuracy. The solution factor Z_m is then computed as $Z_m = V_m\left(Q_m\check{Y}\right)$.

To make fair comparisons with state-of-the-art algorithms that employ LAPACK and SLICOT subroutines (see Section 6 for more details), we used a C-compiled mex-code `cTri` to implement Algorithm 3, making use of LAPACK and BLAS subroutines. In particular, the eigendecomposition $T_m = Q_m \Lambda_m Q_m^T$ is performed as follows. The block tridiagonal matrix T_m is tridiagonalized, $P_m^T T_m P_m = F_m$, by the LAPACK subroutine `dsbtrd` that exploits its banded structure. The transformation matrix P_m is represented as a product of elementary reflectors and only its first and last ℓ rows, $E_1^T P_m$, $E_m^T P_m$, are actually computed. The LAPACK subroutine `dstevr` is employed to compute the eigendecomposition of the tridiagonal matrix F_m. This routine applies Dhillon's MRRR method [15] whose main advantage is the computation of numerically orthogonal eigenvectors without an explicit orthogonalization procedure. This feature limits to $\mathcal{O}((\ell m)^2)$ flops the computation of $F_m = G_m \Lambda_m G_m^T \in \mathbb{R}^{\ell m \times \ell m}$; see [15] for more details. Since the residual norm computation (10) requires the first and last ℓ rows of the eigenvectors matrix Q_m, we compute only those components, that is $E_1^T Q_m = \left(E_1^T P_m\right) G_m$ and $E_m^T Q_m = \left(E_m^T P_m\right) G_m$, avoiding the expensive matrix–matrix product $Q_m = P_m G_m$.

3.3. A "two-pass" strategy

While the block Lanczos method requires the storage of only $3s$ basis vectors, the whole $V_m = [\mathcal{V}_1, \ldots, \mathcal{V}_m] \in \mathbb{R}^{n \times sm}$ is needed to compute the low-rank factor Z_m at convergence (step 9 of Algorithm 1). Since

$$Z_m = V_m(Q_m\check{Y}) = \sum_{i=1}^{m} \mathcal{V}_i E_i^T (Q_m\check{Y}), \tag{13}$$

we suggest not to store V_m during the iterative process but to perform, at convergence, a second Lanczos pass computing and adding the rank-s term in (13) at the ith step, in an incremental fashion. We point out that the orthonormalization coefficients are already available in the matrix T_m, therefore $\mathcal{V}_i$ is simply computed by repeating the three-term recurrence (7), which costs $\mathcal{O}\left((4n+1)s^2\right)$ flops plus the multiplication by A, making the second Lanczos pass cheaper than the first one.

4. Extended Krylov subspace

Rational Krylov subspaces have shown to provide dramatic performance improvements over classical polynomial Krylov subspaces, because they build spectral information earlier, thus generating a much smaller space dimension to reach the

desired accuracy. The price to pay is that each iteration is more computationally involved, as it requires solves with the coefficient matrices. The overall CPU time performance thus depends on the data sparsity of the given problem; we refer the reader to [2] for a thorough discussion.

In this section we show that the enhanced procedure for the residual norm computation can be applied to a particular rational Krylov based strategy, the *Extended* Krylov subspace method, since also this algorithm relies on a block tridiagonal reduced matrix when data is symmetric. Different strategies for building the basis $V_m = [\mathcal{V}_1, \ldots, \mathcal{V}_m] \in \mathbb{R}^{n\times 2sm}$ of the extended Krylov subspace $\mathbf{EK}_m^{\square}(A, C)$ can be found in the literature, see, e.g., [7,16,17]. An intuitive key fact is that the subspace expands in the directions of A and A^{-1} at the same time. In the block case, a natural implementation thus generates two new blocks of vectors at the time, one in each of the two directions. Starting with $[V_1, A^{-1}V_1]$, the next iterations generate the blocks $\mathcal{V}_m^{(1)}, \mathcal{V}_m^{(2)} \in \mathbb{R}^{n\times s}$ by multiplication by A and solve with A, respectively, and then setting $\mathcal{V}_m = [\mathcal{V}_m^{(1)}, \mathcal{V}_m^{(2)}] \in \mathbb{R}^{n\times 2s}$. As a consequence, the block Lanczos procedure described in Algorithm 2 can be employed with $W = [A\mathcal{V}_{m-1}^{(1)}, A^{-1}\mathcal{V}_{m-1}^{(2)}]$ (with $2s$ columns). The orthogonalization process determines the coefficients of the symmetric block tridiagonal matrix H_m with blocks of size $2s$,

$$H_m = \begin{pmatrix} \boldsymbol{\vartheta}_{11} & \boldsymbol{\vartheta}_{12} & & & \\ \boldsymbol{\vartheta}_{21} & \boldsymbol{\vartheta}_{22} & \boldsymbol{\vartheta}_{23} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \boldsymbol{\vartheta}_{m-1,m} \\ & & & \boldsymbol{\vartheta}_{m,m-1} & \boldsymbol{\vartheta}_{m,m} \end{pmatrix} \in \mathbb{R}^{2sm\times 2sm},$$

such that $\mathcal{V}_m\boldsymbol{\vartheta}_{m+1,m} = [A\mathcal{V}_{m-1}^{(1)}, A^{-1}\mathcal{V}_{m-1}^{(2)}] - \mathcal{V}_{m-1}\boldsymbol{\vartheta}_{m,m} - \mathcal{V}_{m-2}\boldsymbol{\vartheta}_{m-1,m}$. The coefficients $\boldsymbol{\vartheta}$'s correspond to the $\boldsymbol{\tau}$'s in Algorithm 2, however as opposed to the standard Lanczos procedure, $H_m \neq T_m = V_m^TAV_m$. Nonetheless, a recurrence can be derived to compute the columns of T_m from those of H_m during the iterations; see [7, Proposition 3.2]. The computed T_m is block tridiagonal, with blocks of size $2s$, and this structure allows us to use the same approach followed for the block standard Krylov method as relation (5) still holds. Algorithm 3 can thus be adopted to compute the residual norm also in the extended Krylov approach with $\ell = 2s$. Moreover, it is shown in [7] that the off-diagonal blocks of T_m have a zero lower $s \times 2s$ block, that is

$$\boldsymbol{\tau}_{i,i-1} = \begin{bmatrix} \overline{\boldsymbol{\tau}}_{i,i-1} \\ 0 \end{bmatrix}, \quad \overline{\boldsymbol{\tau}}_{i,i-1} \in \mathbb{R}^{s\times 2s}, \ i = 1, \ldots, m.$$

This observation can be exploited in the computation of the residual norm as

$$\|R_m\|_F = \sqrt{2}\|Y_mE_m\boldsymbol{\tau}_{m+1,m}^T\|_F = \sqrt{2}\|Y_mE_m\overline{\boldsymbol{\tau}}_{m+1,m}^T\|_F,$$

and $\overline{\boldsymbol{\tau}}_{m+1,m}$ can be passed as an input argument to `cTri` instead of the whole $\boldsymbol{\tau}_{m+1,m}$.

The extended Krylov subspace dimension grows faster than the standard one as it is augmented by $2s$ vectors per iteration. In general, this does not create severe storage difficulties as the extended Krylov approach exhibits faster convergence than standard Krylov in terms of number of iterations. However, for hard problems the space may still become too large to be stored, especially for large s. In this case, a two-pass-like strategy may be appealing. To avoid the occurrence of sm new system solves with A, however, it may be wise to still store the second blocks, $\mathcal{V}_i^{(2)}$, $i = 1, \ldots, m$, and only save half memory allocations, those corresponding to the matrices $\mathcal{V}_i^{(1)}$, $i = 1, \ldots, m$.

Finally, we remark that if we were to use more general rational Krylov subspaces, which use rational functions other than A and A^{-1} to generate the space [2], the projected matrix T_m would lose the convenient block tridiagonal structure, so that the new strategy would not be applicable.

5. The case of the Sylvester equation

The strategy presented for the symmetric Lyapunov equation (1) can be extended to the Sylvester equation

$$AX + XB + C_1C_2^T = 0, \quad A \in \mathbb{R}^{n_1\times n_1}, B \in \mathbb{R}^{n_2\times n_2}, C_1 \in \mathbb{R}^{n_1\times s}, C_2 \in \mathbb{R}^{n_2\times s}, \tag{14}$$

where the coefficient matrices A, B are both symmetric and negative definite while C_1, C_2 are tall, that is $s \ll n_1, n_2$.

5.1. Large $\boldsymbol{A}$ and large $\boldsymbol{B}$

We consider the case when both A and B are large and sparse matrices. If their eigenvalue distributions satisfy certain hypotheses, the singular values of the nonsymmetric solution $X \in \mathbb{R}^{n_1\times n_2}$ to (14) exhibit a fast decay, and a low-rank approximation $\widetilde{X} = Z_1Z_2^T$ to X can be sought; see, e.g., [18, Th. 2.1.1], [2, Section 4.4].

Projection methods seek an approximate solution $X_m \in \mathbb{R}^{n_1\times n_2}$ to (14) of the form $X_m = V_mY_mU_m^T$ where the orthonormal columns of V_m and U_m span suitable subspaces $\mathcal{K}_m$ and $\mathcal{C}_m$ respectively[1] . The construction of two approximation spaces is

[1] The space dimensions of $\mathcal{K}_m$ and $\mathcal{C}_m$ are not necessarily equal, we limit our discussion to the same dimension for simplicity of exposition.

thus requested and, for the sake of simplicity, we limit our discussion to the standard Krylov method, that is $\mathcal{K}_m = \mathbf{K}_m^{\square}(A, C_1)$ and $\mathcal{C}_m = \mathbf{K}_m^{\square}(B, C_2)$, with obvious generalization to the extended Krylov subspace. As in the Lyapunov case, Y_m is computed by imposing a Galerkin condition on the residual matrix $R_m := AX_m + X_mB + C_1C_2^T$, that is

$$V_m^T R_m U_m = 0. \tag{15}$$

We assume $C_1 = V_1\boldsymbol{\gamma}_1$, $C_2 = U_1\boldsymbol{\gamma}_2$ for some nonsingular $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathbb{R}^{s\times s}$, and a similar discussion to the one presented in Section 2 shows that condition (15) is equivalent to solving the reduced Sylvester problem

$$T_m Y_m + Y_m J_m + E_1\boldsymbol{\gamma}_1\boldsymbol{\gamma}_2^T E_1^T = 0, \tag{16}$$

where $T_m := V_m^T A V_m$, $J_m := U_m^T B U_m = (\boldsymbol{\iota}_{ij})$. Similarly to the Lyapunov case, computing the eigendecompositions $T_m = Q_m \Lambda_m Q_m^T$, $\Lambda_m = \mathrm{diag}(\lambda_1, \ldots, \lambda_{sm})$, and $J_m = P_m \Upsilon_m P_m^T$, $\Upsilon_m = \mathrm{diag}(\upsilon_1, \ldots, \upsilon_{sm})$, the solution Y_m to (16) can be written as

$$Y_m = Q_m \widetilde{Y} P_m^T = -Q_m \left(\frac{e_i^T Q_m^T E_1 \boldsymbol{\gamma}_1 \boldsymbol{\gamma}_2^T E_1^T P_m e_j}{\upsilon_i + \lambda_j} \right)_{ij} P_m^T. \tag{17}$$

The last s rows and columns of Y_m are employed in the residual norm calculation. Indeed, letting $\underline{T}_m = V_{m+1}^T A V_m$ and $\underline{J}_m = U_{m+1}^T B U_m$, it can be shown that

$$\|R_m\|_F^2 = \|\underline{T}_m Y_m\|_F^2 + \|Y_m \underline{J}_m^T\|_F^2 = \|\boldsymbol{\tau}_{m+1,m} E_m^T Y_m\|_F^2 + \|Y_m E_m \boldsymbol{\iota}_{m+1,m}^T\|_F^2, \tag{18}$$

where $\boldsymbol{\tau}_{m+1,m} = E_{m+1}^T \underline{T}_m E_m \in \mathbb{R}^{s\times s}$ and $\boldsymbol{\iota}_{m+1,m} = E_{m+1}^T \underline{J}_m E_m \in \mathbb{R}^{s\times s}$, see, e.g., [2,19].

The same arguments of Section 3.1 can be applied to the factors in (18) leading to Algorithm 4 for the computation of the residual norm without explicitly assembling the matrix Y_m. The eigendecompositions in step 1 are not fully computed. In particular, only the spectrum and the first and last ℓ components of the eigenvectors of T_m and J_m are explicitly computed following the strategy presented in Section 3.2.

Algorithm 4: Computing the residual norm for A and B large
Input: $T_m, J_m \in \mathbb{R}^{\ell m\times \ell m}$, $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \boldsymbol{\tau}_{m+1,m}, \boldsymbol{\iota}_{m+1,m} \in \mathbb{R}^{\ell\times\ell}$
Output: res $(= \|R_m\|_F)$

1. Compute $T_m = Q_m \Lambda_m Q_m^T$ and $J_m = P_m \Upsilon_m P_m^T$
2. Compute $S_m := \left(Q_m^T E_1 \boldsymbol{\gamma}_1\right)\left(\boldsymbol{\gamma}_2^T E_1^T P_m\right)$
3. Compute $F_m := \left(Q_m^T E_m\right)\boldsymbol{\tau}_{m+1,m}^T$, $G_m := \left(P_m^T E_m\right)\boldsymbol{\iota}_{m+1,m}^T$
4. Set $res = 0$
5. For $i = 1, \ldots, \ell m$, **Do**
6. Set $D_i' := \upsilon_i I_{\ell m} + \Lambda_m$ and $D_i'' := \lambda_i I_{\ell m} + \Upsilon_m$
7. $res = res + \left\| e_i^T S_m D_i'^{-1} G_m \right\|_2^2 + \left\| e_i^T S_m^T D_i''^{-1} F_m \right\|_2^2$
8. EndDo
11. Set $res = \sqrt{res}$

At convergence, the matrix Y_m can be computed by (17). Also in the Sylvester problem the matrix Y_m may be numerically singular. In this case, factors $\widehat{Y}_1, \widehat{Y}_2 \in \mathbb{R}^{sm\times t}$, $t \le sm$, such that $\|\widetilde{Y} - \widehat{Y}_1 \widehat{Y}_2^T\|_F \le \epsilon$ can be computed via the truncated singular value decomposition of the nonsymmetric matrix $\widetilde{Y}$. The low-rank factors Z_1, Z_2 of X_m, $X_m \approx Z_1 Z_2^T$, are then computed as $Z_1 = V_m\left(Q_m \widehat{Y}_1\right)$ and $Z_2 = U_m\left(P_m \widehat{Y}_2\right)$.

If Eq. (14) is solved by the standard Krylov method, the two-pass strategy presented in Section 3.3 can be easily adapted to the Sylvester case. Indeed, denoting by $V_m = [\mathcal{V}_1, \ldots, \mathcal{V}_m]$ and $U_m = [\mathcal{U}_1, \ldots, \mathcal{U}_m]$, the low-rank factors Z_1 and Z_2 can be written as

$$Z_1 = V_m\left(Q_m \widehat{Y}_1\right) = \sum_{i=1}^{m} \mathcal{V}_i E_i^T \left(Q_m \widehat{Y}_1\right), \qquad Z_2 = U_m\left(P_m \widehat{Y}_2\right) = \sum_{i=1}^{m} \mathcal{U}_i E_i^T \left(P_m \widehat{Y}_2\right).$$

As in the Lyapunov case, the factors Z_1, Z_2 can be computed in a second Lanczos pass since the terms $\mathcal{V}_i E_i^T\left(Q_m \widehat{Y}_1\right)$ and $\mathcal{U}_i E_i^T\left(P_m \widehat{Y}_2\right)$ do not require the whole basis to be available. Therefore, for the Sylvester problem (14), the "two-pass" strategy allows us to store only $6s$ basis vectors, $3s$ vectors for each of the two bases.

5.2. Large $\boldsymbol{A}$ and small $\boldsymbol{B}$

In some applications, such as the solution of eigenvalue problems [20] or boundary value problems with separable coefficients [21], the matrices A and B in (14) could have very different dimensions. In particular, one of them, for instance, B, could be of moderate size, that is $n_2 \ll 1000$. In this case, the projection method presented in Section 5.1 can be

simplified. Indeed, a reduction of the matrix B becomes unnecessary, so that a numerical solution X_m to (14) of the form $X_m = V_m Y_m$ is sought, where the columns of V_m span $\mathcal{K}_m = \mathbf{K}_m^{\square}(A, C_1)$, as before. The Galerkin condition on the residual matrix $R_m := AX_m + X_m B + C_1 C_2^T$ thus becomes

$$V_m^T R_m = 0, \tag{19}$$

see [2, Section 4.3] for more details. The procedure continues as in the previous cases, taking into account that the original problem is only reduced "from the left". Assuming $C_1 = V_1 \boldsymbol{\gamma}_1$, we obtain

$$0 = V_m^T A X_m + V_m^T X_m B + V_m^T C_1 C_2^T = \left(V_m^T A V_m\right) Y_m + \left(V_m^T V_m\right) Y_m B + E_1 \boldsymbol{\gamma}_1 C_2^T,$$

that is

$$T_m Y_m + Y_m B + E_1 \boldsymbol{\gamma}_1 C_2^T = 0. \tag{20}$$

Computing the eigendecompositions $T_m = Q_m \Lambda_m Q_m^T$, $\Lambda_m = \operatorname{diag}(\lambda_1, \ldots, \lambda_{sm})$ and $B = P \Upsilon P^T$, $\Upsilon = \operatorname{diag}(\upsilon_1, \ldots, \upsilon_{n_2})$, the solution matrix Y_m to (20) can be written as

$$Y_m = Q_m \widetilde{Y} P^T = -Q_m \left(\frac{Q_m^T E_1 \boldsymbol{\gamma}_1 C_2^T P}{\lambda_i + \upsilon_j} \right)_{ij} P^T. \tag{21}$$

As before, the block tridiagonal structure of T_m can be exploited in the eigendecomposition computation $T_m = Q_m \Lambda_m Q_m^T$, while the eigendecomposition $B = P \Upsilon P^T$ is computed once for all at the beginning of the whole process.

The expression of the residual norm simplifies as $\|R_m\|_F = \|Y_m^T E_m^T \boldsymbol{\tau}_{m+1,m}^T\|_F$. To compute this norm without assembling the whole matrix Y_m, a slight modification of Algorithm 3 can be implemented. The resulting procedure is summarized in Algorithm 5 where only selected entries of the eigenvector matrix Q_m in step 1 are computed; see the corresponding strategy in Section 3.2.

Algorithm 5: Computing the residual norm for A large and B small
Input: $T_m \in \mathbb{R}^{\ell m \times \ell m}$, $\boldsymbol{\tau}_{m+1,m} \in \mathbb{R}^{\ell \times \ell}$, $P^T C_2 \boldsymbol{\gamma}_1^T \in \mathbb{R}^{n_2 \times \ell}$, $\{\upsilon_i\}_{i=1,\ldots,n_2}$
Output: res $(= \|R\|_F)$

1. Compute $T_m = Q_m \Lambda_m Q_m^T$
2. Compute $S_m = \left(P^T C_2 \boldsymbol{\gamma}_1^T\right) \left(E_1^T Q_m\right)$
3. Compute $W_m = \left(Q_m^T E_m\right) \boldsymbol{\tau}_{m+1,m}^T$
4. Set $res = 0$
5. For $i = 1, \ldots, n_2$, **Do**
6. Set $D_i = \upsilon_i I_{\ell m} + \Lambda_m$
7. $res = res + \left\| \left(e_i^T S_m\right) D_i^{-1} W_m \right\|_2^2$
8. EndDo
9. Set $res = \sqrt{res}$

A reduced rank approximation to the solution Y_m obtained by (21) is given as $\widetilde{Y} \approx \widehat{Y}_1 \widehat{Y}_2^T$, so that the low rank factors Z_1, Z_2 are computed as $Z_1 = V_m \left(Q_m \widehat{Y}_1\right)$ and $Z_2 = P \widehat{Y}_2$. A two-pass strategy can again be employed to avoid storing the whole matrix V_m.

6. Numerical experiments

In this section some numerical examples illustrating the enhanced algorithm are reported. All results were obtained with Matlab R2015a on a Dell machine with two 2 GHz processors and 128 GB of RAM.

The standard implementation of projection methods (Algorithm 1) and the proposed enhancement, where lines 6.1 and 6.2 of Algorithm 1 are replaced by Algorithm 3, are compared. For the standard implementation, different decomposition based solvers for line 6.1 in Algorithm 1 are considered: The Bartels–Stewart algorithm (function `lyap`), one of its variants (`lyap2`)[2] , and the Hammarling method (`lyapchol`). All these algorithms make use of SLICOT or LAPACK subroutines.

Examples with a sample of small values of the rank s of $C_1 C_2^T$ are reported. In all our experiments the convergence tolerance on the relative residual norm is $\texttt{tol} = 10^{-6}$.

Example 1. The block standard Krylov approach is tested for solving the Lyapunov equation $AX + XA + CC^T = 0$. We consider $A \in \mathbb{R}^{n \times n}$, $n = 21\,904$ stemming from the discretization by centered finite differences of the differential operator

$$\mathcal{L}(u) = (e^{-xy} u_x)_x + (e^{xy} u_y)_y,$$

[2] The function `lyap2` was slightly modified to exploit the orthogonality of the eigenvectors matrix.

Table 1
Example 1. CPU times and gain percentages. Convergence is checked every d iterations. Left: $d = 1$. Right: $d = 10$.

	Time res (s)	Gain	Time tot (s)	Gain	Time res (s)	Gain	Time tot (s)	Gain
				$s = 1$ (444 its)				
`lyap`	42.36	89.5%	45.18	83.9%	4.78	89.7%	7.87	52.9%
`lyapchol`	36.51	87.9%	38.51	81.2%	4.27	88.5%	7.59	51.25%
`lyap2`	34.27	87.1%	37.07	80.4%	3.85	87.2%	7.14	48.1%
`cTri`	**4.42**	↵	**7.25**	↵	**0.49**	↵	**3.70**	↵
				$s = 4$ (319 its)				
`lyap`	819.02	96.4%	825.44	95.6%	88.52	96.6%	95.60	91.65%
`lyapchol`	213.87	86.1%	220.51	83.6%	21.38	86.1%	26.83	70.2%
`lyap2`	212.99	86.0%	219.34	83.5%	20.28	85.3%	27.65	71.1%
`cTri`	**29.78**	↵	**36.21**	↵	**2.97**	↵	**7.98**	↵
				$s = 8$ (250 its)				
`lyap`	2823.31	97.9%	2836.29	97.6%	305.11	98.2%	313.49	95.8%
`lyapchol`	415.42	85.7%	427.21	84.1%	38.94	85.7%	46.96	71.8%
`lyap2`	424.23	86.0%	435.90	84.4%	41.39	86.5%	49.15	73.1%
`cTri`	**59.25**	↵	**67.89**	↵	**5.56**	↵	**13.22**	↵

Table 2
Example 1. Memory requirements with and without full storage, and CPU time of the second Lanczos sweep.

			Memory whole V_m	Reduced mem. alloc.	CPU time (s)
n	s	m	$s \cdot m$	$3s$	
21 904	1	444	444	3	1.44
21 904	4	319	1276	12	2.35
21 904	8	250	2000	24	3.74

on the unit square with zero Dirichlet boundary conditions, while $C = \texttt{rand}(n, s)$, $s = 1,\ 4, 8$, that is the entries of C are random numbers uniformly distributed in the interval $(0, 1)$. C is then normalized, $C = C/\|C\|_F$. Table 1 (left) reports the CPU time (in seconds) needed for evaluating the residual norm (time res) and for completing the whole procedure (time tot). Convergence is checked at each iteration. For instance, for $s = 1$, using `lyapchol` as inner solver the solution process takes 38.51 s, 36.51 of which are used for solving the inner problem of step 6.1. If we instead use `cTri`, the factors of X_m are determined in 7.25 seconds, only 4.42 of which are devoted to evaluating the residual norm. Therefore, 87.9% of the residual computation CPU time is saved, leading to a 81.2% saving for the whole procedure. An explored device to mitigate the residual norm computational cost is to check the residual only periodically. In the right-hand side of Table 1 we report the results in case the residual norm is computed every 10 iterations.

Table 2 shows that the two-pass strategy of Section 3.3 drastically reduces the memory requirements of the solution process, as already observed in [9], at a negligible percentage of the total execution time.

Example 2. The RAIL benchmark problem[3] solves the generalized Lyapunov equation

$$AXE + EXA + CC^T = 0, \tag{22}$$

where $A, E \in \mathbb{R}^{n\times n}$, $n = 79\,841$, $C \in \mathbb{R}^{n\times s}$, $s = 7$. Following the discussion in [7], Eq. (22) can be treated as a standard Lyapunov equation for E symmetric and positive definite. This is a recognized hard problem for the standard Krylov subspace, therefore the extended Krylov subspace is applied, and convergence is checked at each iteration. Table 3 collects the results. In spite of the 52 iterations needed to converge, the space dimension is large, that is $\dim\left(\mathbf{EK}_m^{\square}(A, C)\right) = 728$ and the memory-saving strategy of Section 4 may be attractive; it was not used for this specific example, but it can be easily implemented. The gain in the evaluation of the residual norm is still remarkable, but less impressive from the global point of view. Indeed, the basis construction represents the majority of the computational efforts; in particular, the linear solves $A^{-1}\mathcal{V}_i^{(2)}$, $i = 1, \ldots, 52$, required 17.60 s.

Example 3. In this example, we compare the standard and the extended Krylov approaches again for solving the standard Lyapunov equation. We consider the matrix $A \in \mathbb{R}^{n\times n}$, $n = 39\,304$, coming from the discretization by isogeometric analysis (IGA) of the 3D Laplace operator on the unit cube $[0, 1]^3$ with zero Dirichlet boundary conditions and a uniform mesh. Since high degree B-splines are employed as basis functions (here the degree is 4 but higher values are also common), this discretization method yields denser stiffness and mass matrices than those typically obtained by low degree finite element

[3] http://www.simulation.uni-freiburg.de/downloads/benchmark/Steel%20Profiles%20%2838881%29.

Table 3
Example 2. CPU times and gain percentages.

	Time res (s)	Gain	Time tot (s)	Gain
`lyap`	11.25	75.9%	75.53	7.7%
`lyapchol`	6.05	55.2%	70.76	1.5%
`lyap2`	6.68	59.4%	73.01	4.5%
`cTri`	**2.71**	↵	**69.70**	↵

Table 4
Example 3. Performance comparison of Standard and Extended Krylov methods.

	m	Whole V_m mem. alloc.	Reduced mem. alloc.	Time res (s)	Two-pass (s)	Time tot (s)
			$s = 3$			
St. Krylov	280	840	9	1.59	20.75	44.56
Ex. Krylov	30	180	180	0.09	–	85.54
			$s = 8$			
St. Krylov	260	2080	24	3.84	45.35	93.49
Ex. Krylov	27	216	216	0.57	-	347.99

Table 5
Example 4. CPU times and gain percentages.

	Time res (s)	Gain	Time tot (s)	Gain
		$s = 3$ (217 its)		
`lyap`	60.19	83.6%	65.32	76.2%
`lyap2`	74.05	86.6%	78.08	80.1%
`cTri`	**9.89**	↵	**15.51**	↵
		$s = 8$ (145 its)		
`lyap`	201.28	88.7%	208.93	81.5%
`lyap2`	140.92	83.8%	149.95	74.2%
`cTri`	**22.74**	↵	**38.65**	↵

or finite difference methods; in our experiment, 1.5% of the components of A is nonzero. See, e.g., [22] for more details on IGA.

For the right-hand side we set $C = \texttt{rand}(n, s)$, $s = 3, 8$, $C = C/\|C\|_F$. In the standard Krylov method the residual norm is computed every 20 iterations. The convergence can be checked every d iterations in the extended approach as well, with d moderate to avoid excessive wasted solves with A at convergence [7]. In our experiments the computation of the residual norm only takes a small percentage of the total execution time and we can afford taking $d = 1$. In both approaches, the residual norm is computed by Algorithm 3. Table 4 collects the results.

The standard Krylov method generates a large space to converge for both values of s. Nonetheless, the two-pass strategy allows us to store only 9 basis vectors for $s = 3$ and 24 basis vectors for $s = 8$. This feature may be convenient if storage of the whole solution process needs to be allocated in advance. By checking the residual norm every 20 iterations, the standard Krylov method becomes competitive with respect to the extended procedure, which is in turn penalized by the system solutions with dense coefficient matrices. Indeed, for $s = 3$ the operation $A^{-1}\mathcal{V}_i^{(2)}$ for $i = 1, \ldots, 30$ takes 32.75 s, that is 38.29% of the overall execution time required by the extended Krylov subspace method. Correspondingly, for $s = 8$ the same operation performed during 27 iterations takes 152.92 s, that is, 44.94% of the overall execution time. This example emphasizes the potential of the enhanced classical approach when system solves are costly, in which case rational methods pay a higher toll.

Example 4. In this example, a Sylvester equation (14) is solved. The coefficient matrices $A, B \in \mathbb{R}^{n\times n}$, $n = 16\,384$, come from the discretization by centered finite differences of the partial differential operators

$$\mathcal{L}_A(u) = (e^{-xy}u_x)_x + (e^{xy}u_y)_y \quad \text{and} \quad \mathcal{L}_B(u) = (\sin(xy)u_x)_x + \big(\cos(xy)u_y\big)_y,$$

on $[0, 1]^2$ with zero Dirichlet boundary conditions. The right-hand side is a uniformly distributed random matrix where $C_1 = \texttt{rand}(n, s)$, $C_1 = C_1/\|C_1\|_F$ and $C_2 = \texttt{rand}(n, s)$, $C_2 = C_2/\|C_2\|_F$, $s = 3, 8$. Since both A and B are large, Eq. (14) is solved by the standard Krylov method presented in Section 5.1 and 217 iterations are needed to converge for $s = 3$, and 145 iterations for $s = 8$. The residual norm is checked at each iteration and Table 5 collects the results. Two approximation spaces, $\mathbf{K}_m^{\square}(A, C_1) = \text{Range}(V_m)$, $\mathbf{K}_m^{\square}(B, C_2) = \text{Range}(U_m)$, are generated and a two-pass strategy is employed to cut down the storage demand. See Table 6.

Table 6
Example 4. Memory requirements with and without full storage, and CPU time of the second Lanczos sweep.

			Memory whole V_m, U_m	Reduced mem. alloc.	CPU time (s)
n	s	m	$2s \cdot m$	$6s$	
16 384	3	217	1032	18	2.62
16 384	8	145	2320	48	4.93

Table 7
Example 5. CPU times and gain percentages.

	Time res (s)	Gain	Time tot (s)	Gain
		$s = 3$ (190 its)		
`lyap`	15.47	75.7%	17.88	63.6%
`lyap2`	25.35	85.2%	27.50	76.3%
`cTri`	**3.76**	↵	**6.51**	↵
		$s = 8$ (150 its)		
`lyap`	36.99	68.2%	40.90	60.0%
`lyap2`	77.04	84.7%	80.91	79.8%
`cTri`	**11.77**	↵	**16.35**	↵

Table 8
Example 4. Memory requirements with and without full storage, and CPU time of the second Lanczos sweep.

			Memory whole V_m	Reduced mem. alloc.	CPU time (s)
n^2	s	m	$s \cdot m$	$3s$	
21 904	3	190	570	9	0.93
21 904	8	150	1200	24	1.31

Example 5. In this last example, we again consider the Sylvester problem (14), this time stemming from the 3D partial differential equation

$$\left(e^{-xy}u_x\right)_x + \left(e^{xy}u_y\right)_y + 10u_{zz} = f \quad \text{on } [0,1]^3, \tag{23}$$

with zero Dirichlet boundary conditions. Thanks to the regular domain, its discretization by centered finite differences can be represented by the Sylvester equation

$$AX + XB = F, \tag{24}$$

where $A \in \mathbb{R}^{n^2 \times n^2}$ accounts for the discretization in the x, y variables, while $B \in \mathbb{R}^{n \times n}$ is associated with the z variable. The right-hand side $F \in \mathbb{R}^{n^2 \times n}$ takes into account the source term f in agreement with the space discretization. See [23] for a similar construction.

In our experiment, $n = 148$ (so that $n^2 = 21\,904$) and Eq. (24) falls into the case addressed in Section 5.2. The right-hand side is $F = -C_1 C_2^T$ where C_1, C_2 are two different normalized random matrices, $C_j = \texttt{rand}(n, s)$, $C_j = C_j / \|C_j\|_F$, $j = 1, 2$, and $s = 3, 8$. Convergence is checked at each iteration and Table 7 collects the results. The method requires 190 iterations to converge below 10^{-6} for $s = 3$ and 150 for $s = 8$, and a two-pass strategy allows us to avoid the storage of the whole basis $V_m \in \mathbb{R}^{n^2 \times sm}$. See Table 8.

7. Conclusions

We have presented an expression for the residual norm that significantly reduces the cost of monitoring convergence in projection methods based on $\mathbf{K}_m^{\square}$ and $\mathbf{EK}_m^{\square}$ for Sylvester and Lyapunov equations and symmetric data. For the standard Krylov approach, the combination with a two-pass strategy makes this classical algorithm appealing compared with recently developed methods, both in terms of computational costs and memory requirements, whenever data do not allow for cheap system solves. The proposed enhancements rely on the symmetric block tridiagonal structure of the projected matrices. In case this pattern does not arise, as is the case for instance in the nonsymmetric setting, different approaches must be considered.

Acknowledgment

We thank Mattia Tani for providing us with the data of Example 3.

References

[1] A.C. Antoulas, Approximation of Large-Scale Dynamical Systems, in: Advances in Design and Control, vol. 6, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.
[2] V. Simoncini, Computational methods for linear matrix equations, SIAM Rev. 58 (2016) 377–441.
[3] T. Penzl, Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case, Systems Control Lett. 40 (2000) 139–144.
[4] J. Snyders, M. Zakai, On nonnegative solutions of the equation $AD + DA' = -C$, SIAM J. Appl. Math. 18 (1970) 704–714.
[5] M.H. Gutknecht, Krylov subspace algorithms for systems with multiple right hand sides: an introduction, 2006. Available at http://www.sam.math.ethz.ch/~mhg/pub/delhipap.pdf.
[6] T. Penzl, A cyclic low-rank Smith method for large sparse Lyapunov equations, SIAM J. Sci. Comput. 21 (2000) 1401–1418.
[7] V. Simoncini, A new iterative method for solving large-scale Lyapunov matrix equations, SIAM J. Sci. Comput. 29 (2007) 1268–1288.
[8] V. Druskin, V. Simoncini, Adaptive rational Krylov subspaces for large-scale dynamical systems, Systems Control Lett. 60 (2011) 546–560.
[9] D. Kressner, Memory-efficient Krylov subspace techniques for solving large-scale Lyapunov equations, in: IEEE International Symposium on Computer-Aided Control Systems, San Antonio, 2008, pp. 613–618.
[10] A. Frommer, V. Simoncini, Stopping criteria for rational matrix functions of Hermitian and symmetric matrices, SIAM J. Sci. Comput. 30 (2008) 1387–1412.
[11] R.H. Bartels, G.W. Stewart, Algorithm 432: Solution of the matrix equation $AX + XB = C$, Comm. ACM. 15 (1972) 820–826.
[12] S.J. Hammarling, Numerical solution of the stable, nonnegative definite Lyapunov equation, IMA J. Numer. Anal. 2 (1982) 303–323.
[13] I.M. Jaimoukha, E.M. Kasenally, Krylov subspace methods for solving large Lyapunov equations, SIAM J. Numer. Anal. 31 (1994) 227–251.
[14] Y. Saad, Iterative Methods for Sparse Linear Systems, second ed., SIAM, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003.
[15] I.S. Dhillon, A new O(n^2) algorithm for the symmetric tridiagonal eigenvalue/eigenvector problem (Ph.D. thesis), University of California, Berkeley, 1997.
[16] C. Jagels, L. Reichel, The extended Krylov subspace method and orthogonal Laurent polynomials, Linear Algebra Appl. 431 (2009) 441–458.
[17] C. Mertens, R. Vandebril, Short recurrences for computing extended Krylov bases for Hermitian and unitary matrices, Numer. Math. 131 (2015) 303–328.
[18] J. Sabino, Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Method (Ph.D. thesis), Rice University, 2006.
[19] T. Breiten, V. Simoncini, M. Stoll, Low-rank solvers for fractional differential equations, Electron. Trans. Numer. Anal. 45 (2016) 107–132.
[20] D.S. Watkins, The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods, first ed., SIAM, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2007.
[21] E.L. Wachspress, Extended application of alternating direction implicit iteration model problem theory, J. Soc. Ind. Appl. Math. 11 (1963) 994–1016.
[22] J.A. Cottrell, T.J.R. Hughes, Y. Bazilevs, Isogeometric Analysis: Toward Integration of CAD and FEA, first ed., Wiley Publishing, 2009.
[23] D. Palitta, V. Simoncini, Matrix-equation-based strategies for convection–diffusion equations, BIT 56 (2016) 751–776.