Optimality Properties of Galerkin and Petrov–Galerkin Methods for Linear Matrix Equations

OPTIMALITY PROPERTIES OF GALERKIN AND PETROV-GALERKIN METHODS FOR LINEAR MATRIX EQUATIONS*

DAVIDE PALITTA† AND VALERIA SIMONCINI‡

Dedicated to Volker Mehrmann on the occasion of his 65th birthday

Abstract. Galerkin and Petrov-Galerkin methods are some of the most successful solution procedures in numerical analysis. Their popularity is mainly due to the optimality properties of their approximate solution. We show that these features carry over to the (Petrov-)Galerkin methods applied for the solution of linear matrix equations.

Some novel considerations about the use of Galerkin and Petrov-Galerkin schemes in the numerical treatment of general linear matrix equations are expounded and the use of constrained minimization techniques in the Petrov-Galerkin framework is proposed.

Key words. Linear matrix equations. Large scale equations. Sylvester equation.

AMS subject classifications. 65F10, 65F30, 15A06

1. Introduction. Many state-of-the-art solution procedures for algebraic linear systems of the form

$$\mathcal{M}x = f, \tag{1.1}$$

where $\mathcal{M} \in \mathbb{R}^{N\times N}$ and $f \in \mathbb{R}^N$, are based on projection. Given a subspace $\mathcal{K}_m$ of dimension m, and a matrix $\mathcal{V}_m$ whose orthonormal columns span $\mathcal{K}_m$, these methods seek an approximate solution $x_m = \mathcal{V}_m y_m$ for some $y_m \in \mathbb{R}^m$ by imposing certain conditions. The most successful projection procedures impose either a *Galerkin* or a *Petrov-Galerkin* condition on the residual $r_m = f - \mathcal{M}x_m$. See, e.g., [40]. These conditions are very general, and they are at the basis of many approximation methods, beyond the algebraic context of interest here; any approximation strategy associated with an inner product can determine the projected solution by one of such a condition. Finite element methods, both at the continuous and discrete levels, strongly rely on this methodology; see, e.g., [49], but also eigenvalue problems [39].

It is very important to realize that this is a methodology, not a single method: the approximation space can be generated independently of the condition, and in a way to make the computation of y_m more effective, while obtaining a sufficiently accurate approximation with the smallest possible space dimension.

A fundamental property of the Galerkin methodology is obtained whenever the coefficient matrix $\mathcal{M}$ is symmetric and positive definite (spd): the Galerkin condition on the residual corresponds to minimizing the error vector in the norm associated with $\mathcal{M}$ over the approximation space. This property is at the basis of the convergence analysis of methods such as the Conjugate Gradient (CG) [21], and it ensures monotonic convergence, in addition to finite termination, in exact precision arithmetic.

When $\mathcal{M}$ is not spd, the application of the Galerkin method does not automatically imply a minimization of the error norm. Nevertheless, a certain family of Petrov-Galerkin procedures still

*Version of November 14, 2019

†Research Group Computational Methods in Systems and Control Theory (CSC), Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, 39106 Magdeburg, Germany. palitta@mpi-magdeburg.mpg.de

‡Dipartimento di Matematica, Alma Mater Studiorum Università di Bologna, Piazza di Porta San Donato 5, I-40127 Bologna, Italy, and IMATI-CNR, Pavia, Italy. valeria.simoncini@unibo.it

fulfills an optimality property. Indeed, these methods minimize the residual norm over the space $\mathcal{MK}_m$. See, e.g., [40]. Some of the most popular solvers for linear systems such as MINRES [34] and GMRES [41] belong to this collection of methods.

In the past decades, projection techniques have been successfully used to solve linear matrix equations of the form

$$A_1XB_1 + A_2XB_2 + \ldots + A_\ell XB_\ell = F, \tag{1.2}$$

that have arisen as a natural algebraic model for discretized partial differential equations (PDEs), possibly including stochastic terms or parameter dependent coefficient matrices [5, 12, 35, 38], for PDE-constrained optimization problems [48], data assimilation [19], and many other applied contexts, including building blocks of other numerical procedures [30]; see also [14, 44] for further references.

The general matrix equation (1.2) covers two well known cases, the (generalized) Sylvester equation (for $\ell = 2$), and the Lyapunov equation

$$AX + XA^T = F, \tag{1.3}$$

which plays a crucial role in many applications such as control and system theory [2, 13], and in the solution of Riccati equations by the Newton method, in which a Lyapunov equation needs to be solved at each Newton step. See, e.g., [33].

The aim of this paper is to generalize the optimality properties of the Galerkin and Petrov-Galerkin methods to matrix equations, and to extend other convergence properties of CG and some related schemes to the matrix setting. Some of the proposed results are new, some others can be found in articles scattered in the literature in different contexts. We thus provide a more uniform presentation of these results.

To introduce a matrix version of the error and residual minimization, we first recall the relation between matrix-matrix operations and Kronecker products. Indeed, if $\otimes$ denotes the Kronecker product and $\mathcal{T} := B^T \otimes A$, then

$$Y = AXB \quad \Leftrightarrow \quad y = \mathcal{T}x, \quad x = \mathrm{vec}(X),\ y = \mathrm{vec}(Y),$$

where the usual "vec$(\cdot)$" operator stacks the columns of the argument matrix one after the other into a long vector.

2. The Galerkin condition. In this section we first recall the result connecting the Galerkin condition on the residual with the minimization of the error norm when this is applied to the solution of linear systems, and then we show that similar results can be obtained also in the matrix equation setting. For the rest of the section we assume that $\mathcal{M}$ in (1.1) is symmetric and positive definite.

2.1. The linear system setting. Let $x_m = V_m y_m$ be an approximation to the true solution of (1.1), and let $e_m = x - x_m$, $r_m = f - \mathcal{M}x_m$ be the associated error and residual, respectively. We recall that imposing the Galerkin condition yields

$$V_m^T r_m = 0 \quad \Leftrightarrow \quad V_m^T \mathcal{M} V_m y_m = V_m^T f. \tag{2.1}$$

Note that the coefficient matrix $V_m^T \mathcal{M} V_m$ is symmetric and positive definite. Solving this system yields the "projected" vector y_m, so as to completely define x_m.

Let $\|e_m\|^2_{\mathcal{M}} := e_m^T \mathcal{M} e_m$ be the $\mathcal{M}$-norm associated with the spd matrix $\mathcal{M}$. For the error we thus have

$$\|e_m\|^2_{\mathcal{M}} = \|\mathcal{M}^{1/2}(x - x_m)\|^2 = \|\mathcal{M}^{1/2}x - \mathcal{M}^{1/2}V_m y_m\|^2. \tag{2.2}$$

The minimization of the error $\mathcal{M}$-norm thus corresponds to solving the least squares problem on the right, which gives

$$(\mathcal{M}^{1/2}V_m)^T \mathcal{M}^{1/2} V_m y_m = (\mathcal{M}^{1/2}V_m)^T \mathcal{M}^{1/2} x,$$

which, upon simplifications of the transpositions yields $V_m^T \mathcal{M} V_m y_m = V_m^T f$, that is, using (2.1), $V_m^T r_m = 0$.

2.2. Galerkin method and error minimization for matrix equations. To simplify the presentation, we first discuss Galerkin projection with the Lyapunov equation. Given the equation (1.3) with A spd and $F = F^T$, then it can be shown that X is symmetric. Letting range(V_k) be an approximation space, we can determine an approximation to X as $X_k = V_k Y_k V_k^T$, which in vector notation is written as $\mathrm{vec}(X_k) = (V_k \otimes V_k)\mathrm{vec}(Y_k)$. The matrix Y_k is obtained by imposing the Galerkin condition in a matrix sense to the residual matrix $R_k = F - (AX_k + X_k A)$, that is

$$V_k^T R_k V_k = 0 \quad \Leftrightarrow \quad (V_k \otimes V_k)^T r_k = 0,$$

where $r_k = \mathrm{vec}(R_k)$. Therefore, if one writes the Lyapunov equation by means of the Kronecker formulation, the obtained approximation space is $\mathcal{K}_m = \mathrm{range}(V_k \otimes V_k)$.

We explicitly notice that X_k belongs to range(V_k), which is much smaller than range$(V_k \otimes V_k)$. Therefore, by sticking to the matrix equation formulation, we expect to build a much smaller approximation space than if a blind use of the Kronecker form were used. In other words, by exploiting the original matrix structure, no redundant information is sought after. In section section 4 we provide a rigorous analysis of this argument. See also [27]. To be able to exploit the derivation in (2.2) we will define an error matrix and the associated inner product.

The generalization to the multiterm linear equation (1.2) requires the definition of two approximation spaces, since the right and left coefficient matrices are not necessarily the same. Therefore, let range(V_k) and range(W_k) be two approximation spaces of dimension k each[1], and let us write the approximation to X as $X_k = V_k Y_k W_k^T$. With the residual matrix $R_k = F - \sum_{j=1}^{\ell} A_j X_k B_j$, the Galerkin condition now takes the form

$$V_k^T R_k W_k = 0 \quad \Leftrightarrow \quad (W_k \otimes V_k)^T r_k = 0,$$

where $r_k = \mathrm{vec}(R_k)$, so that $\mathcal{K}_m = \mathrm{range}(W_k \otimes V_k)$ with $m = k^2$ in the Kronecker formulation.

To adapt the error minimization procedure to the matrix equation setting we first introduce a matrix norm, that allows us to make a connection with the $\mathcal{M}$-norm of the error vector. A corresponding derivation for $\ell = 2$ can be found, for instance, in [52, p. 2557] and [11, p. 149].

Definition 2.1. *Let*

$$\begin{aligned} \mathcal{S}: \quad \mathbb{R}^{n\times p} &\to \mathbb{R}^{n\times p} \\ X &\mapsto \sum_{j=1}^{\ell} A_j X B_j, \end{aligned} \tag{2.3}$$

[1] In principle, we can have $\dim(\mathrm{range}(V_k)) \neq \dim(\mathrm{range}(W_k))$. Here we consider $\dim(\mathrm{range}(V_k)) = \dim(\mathrm{range}(W_k)) = k$ for the sake of simplicity in the presentation.

and $\mathcal{S}_\ell = \sum_{j=1}^{\ell} B_j^T \otimes A_j$. *We say that the operator* $\mathcal{S}$ *is symmetric and positive definite if for any* $0 \neq x \in \mathbb{R}^{np}$, $x = \mathrm{vec}(X)$, *with* $X \in \mathbb{R}^{n\times p}$, *it holds that* $\mathcal{S}_\ell = \mathcal{S}_\ell^T$ *and* $x^T \mathcal{S}_\ell x > 0$, *where*

$$x^T \mathcal{S}_\ell x = \mathrm{trace}\left(\sum_{j=1}^{\ell} X^T A_j X B_j \right).$$

The norm induced by this operator will be denoted by $\|X\|_{\mathcal{S}}$.

Note that any linear operator $\mathcal{L} : \mathbb{R}^{n\times p} \to \mathbb{R}^{n\times p}$ can be written in the form (2.3) with a uniquely defined minimum number of terms ℓ called the *Sylvester index*. See [25].

Assuming $\mathcal{S}$ to be spd, in the following proposition we show that the error matrix is minimized in the $\mathcal{S}$-norm.

PROPOSITION 2.2. *Let* $\mathcal{S}(X) = F$ *with* $\mathcal{S} : X \mapsto \sum_j A_j X B_j$ *spd, and let range*(V_k), *range*(W_k) *be the constructed approximation spaces, so that* $X_k = V_k Y_k W_k^T$ *is the Galerkin approximate solution. Then*

$$\|X - X_k\|_{\mathcal{S}} = \min_{\substack{Z = V_k Y W_k^T \\ Y \in \mathbb{R}^{k\times k}}} \|X - Z\|_{\mathcal{S}}.$$

Proof. Let $e_k = \mathrm{vec}(X - X_k)$ be the error vector, $r_k = \mathrm{vec}(F - \sum_j A_j X_k B_j)$ the residual vector, $\mathcal{K}_m = \mathrm{range}(W_k \otimes V_k)$ the approximation space and $\mathcal{S}_\ell = \sum_j B_j^T \otimes A_j$ the coefficient matrix. Then, since $\mathcal{S}$ is spd by assumption, also $\mathcal{S}_\ell$ is spd, and the Galerkin condition $\mathcal{V}_m^T r_k = 0$, $\mathcal{V}_m = W_k \otimes V_k$, corresponds to the minimization of the error. More precisely, it holds

$$\|e_k\|_{\mathcal{S}}^2 = e_k^T \mathcal{S}_\ell e_k = \mathrm{trace}((X - X_k)^T \mathcal{S}(X - X_k)) = \|X - X_k\|_{\mathcal{S}}^2,$$

and the proof is completed. ∎

Proposition 2.2 states that as long as the approximation spaces are expanded, the error will decrease monotonically in the considered norm. A Galerkin approach for a multiterm linear matrix equation was for instance employed in [38]; the proposition above thus ensures that under the stated hypotheses on the data the method will minimize the error as the approximation spaces grow. See also Example 2.4.

A result similar to the one stated in Proposition 2.2 can be found in [27] where the authors consider specific approximation spaces and assume $\mathcal{S}$ to be a so-called *Laplace-like* operator. Proposition 2.2 shows the strength of the Galerkin method, also in the general matrix equation setting. Indeed, the optimality condition of the Galerkin method does neither depend on the adopted approximation spaces nor on the definition of $\mathcal{S}$, as long as this is spd.

Given a general linear matrix equation (1.2) written in the form $\mathcal{S}(X) = F$, one would like to characterize the symmetry and positive definiteness of $\mathcal{S}$ by looking only at the properties of the matrices A_j and B_j and avoid the construction of the large matrix $\mathcal{S}_\ell$.

Assuming ℓ to be the Sylvester index of $\mathcal{S}$, it is easy to show that $\mathcal{S}$ is a symmetric operator if and only if the matrices A_j and B_j are symmetric for all $j = 1, \ldots, \ell$, whereas, in general, it is not possible to identify the positive definiteness of $\mathcal{S}$ by examining the spectral distributions of A_j and B_j, even when these are completely known. See, e.g., [28]. Note that for $\mathcal{S}$ to be spd it is not necessary for all the A_j's and B_j's to be positive definite. Nevertheless, if A_j, B_j are positive definite for all $j = 1, \ldots, \ell$, then $\mathcal{S}$ is positive definite; see, e.g., [52, Proposition 3.1] for $\ell = 2$.

Therefore, in the case of the Lyapunov equation with A spd, also the operator $\mathcal{S}$ is spd and it holds that

$$\|X\|_{\mathcal{S}}^2 = 2\,\mathrm{trace}(X^T A X).$$

Another case where the properties of $\mathcal{S}$ can be determined in terms of the (symmetric) coefficient matrices A_j and B_j is the Sylvester operator $\mathcal{S} : X \mapsto AX + XB$. By exploiting the property of the Kronecker product, it holds that $\mathcal{S}$ is positive definite if and only if $\lambda_i(A) + \lambda_j(B) > 0$ for all is and js. Moreover, the norm $\|\cdot\|_{\mathcal{S}}$ can be written as $\|X\|_{\mathcal{S}}^2 = \mathrm{trace}(X^T A X) + \mathrm{trace}(XBX^T)$.

REMARK 2.3. *Consider the Lyapunov equation (1.3) with the spd coefficient matrix A, and let $E_k := X - X_k$ be the corresponding error matrix. Then, the previous discussion shows that*

$$\|E_k\|_{\mathcal{S}}^2 = \min_{\substack{Z = V_k Y W_k^T \\ Y \in \mathbb{R}^{k\times k}}} \|X - Z\|_{\mathcal{S}}^2 = 2\,\mathrm{trace}(E_k^T A E_k).$$

In the remark above we have not specified whether the known term F in (1.3) needs to be symmetric. If F is symmetric, then indeed the two spaces can coincide, and E_k is also symmetric. On the other hand, if F has the form $F = F_1 F_2^T$, possibly low rank, natural choices as approximation spaces are such that $\mathrm{range}(F_1) \subseteq \mathrm{range}(V_k)$ and $\mathrm{range}(F_2) \subseteq \mathrm{range}(W_k)$, so that the (vector) residual is orthogonal to $F_2 \otimes F_1$. A possible alternative could use $V_k = W_k$ such that $\mathrm{range}(F_1), \mathrm{range}(F_2) \subseteq \mathrm{range}(V_k)$, where however in general we expect $\mathrm{range}(V_k)$ to have larger dimension than in the previous case.

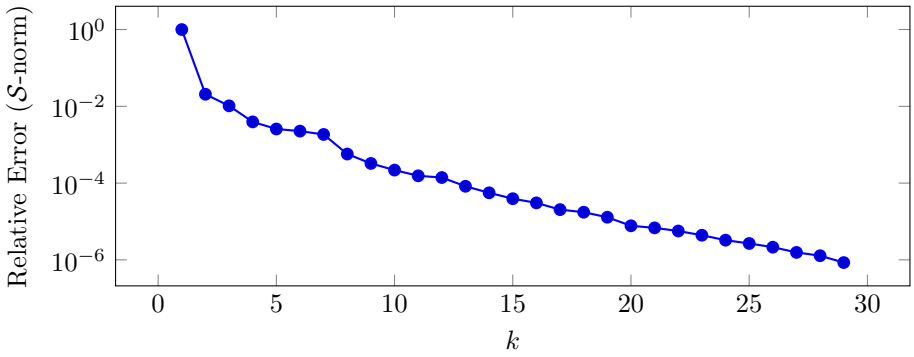
EXAMPLE 2.4. By applying the stochastic Galerkin methodology for the discretization of elliptic stochastic PDEs [3], the resulting algebraic formulation can be written as the linear matrix equation (1.2) with typically $\ell > 2$. When dealing with the stochastic steady-state diffusion problem with homogeneous Dirichlet boundary conditions, the symmetric matrices A_j and B_j may not all be positive definite; nonetheless, the associated operator $\mathcal{S}$ is symmetric and indeed *positive definite* (see, e.g., [37]), so that the previous theory applies. In the following we consider the Galerkin approach developed in [38] – based on the rational Krylov subspace – to illustrate the monotonic decrease of the error $\mathcal{S}$-norm as the approximation space increases[2].We generate A_j and B_j as the second test case in the S-IFISS package [43] with the default setting for all the requested parameters. This yields *a linear matrix equation of the form (1.2) with $\ell = 6$, $A_j \in \mathbb{R}^{n\times n}$, $n = 225$, and $B_j \in \mathbb{R}^{p\times p}$, $p = 56$. The right-hand side F has rank 1. Thanks to the small problem dimension, we could compute the vectorized solution $x \in \mathbb{R}^{np}$ as $x = \mathrm{vec}(X) = \mathcal{S}_\ell^{-1} f$ (Matlab function "\"), to be used as a reference "exact" solution. In particular, if X_k denotes the approximate solution obtained after k iterations of the Galerkin method, we compute $\|X - X_k\|_{\mathcal{S}}/\|X\|_{\mathcal{S}}$ until this falls below 10^{-6}. Figure 2.1 displays the history of this relative error $\mathcal{S}$-norm, illustrating the expected monotonically non-increasing curve.*

3. Convergence properties. In the previous section we have shown that the Galerkin condition leads to a minimization of the error $\mathcal{S}$-norm and this property does not depend on the selected space $\mathcal{K}_k = \mathrm{range}(V_k)$. In actual computations, a measurable estimate of the error is needed and in [45] an upper bound on the Euclidean norm of the error is provided in the case of the Lyapunov equation (1.3) with rank-one right-hand side $F = bb^T$ with $\|b\| = 1$, and a positive definite but not necessarily symmetric A. By exploiting the closed-form of the solution X, the authors showed that

$$\|X - X_k\|_2 \leqslant 2\int_0^\infty e^{-t\alpha_{\min}(A)} \|x - x_m\|_2 dt, \quad \alpha_{\min}(A) = \lambda_{\min}((A + A^T)/2),$$

[2]The Matlab code is available at `http://www.dm.unibo.it/~simoncin/software.html`.

FIG. 2.1. *Example 2.4. Relative error energy norm.*



where $x = e^{-tA}b$, $x_k = V_k e^{-tA_k} e_1$, $A_k := V_k^T A V_k$, and $\|\cdot\|_2$ denotes the Euclidean norm.

This led to the following proposition when the selected approximation space is the Krylov subspace $\text{range}(V_k) = K_k(A,b) = \text{span}\{b, Ab, \dots, A^{k-1}b\}$ and A is symmetric.

PROPOSITION 3.1 ([45]). *Let A be spd, and let $\lambda_{\max}$ and $\lambda_{\min}$ be the largest and the smallest eigenvalue of A, respectively. Denoting by $\hat{\kappa} = (\lambda_{\max} + \lambda_{\min})/2\lambda_{\min}$ the condition number of the spd matrix $A + \lambda_{\min} I$, then the Galerkin approximate solution $X_k = V_k Y_k V_k^T$ satisfies*

$$\|X - X_k\|_2 \leqslant 2\frac{\sqrt{\hat{\kappa}}+1}{\lambda_{\min}\sqrt{\hat{\kappa}}}\left(\frac{\sqrt{\hat{\kappa}}-1}{\sqrt{\hat{\kappa}}+1}\right)^k. \tag{3.1}$$

This bound, in terms of slope as k increases, was shown to be sharp in [45]. Notice that the bound (3.1) holds also for the Frobenius norm of the error, namely $\|X - X_k\|_F$. Indeed, we can still write $\|X - X_k\|_F \leqslant 2\int_0^\infty e^{-t\alpha_{\min}(A)}\|x - x_m\|_F dt$ and the rest of the proof of Proposition 3.1 makes use of bounds for norms of vectors only, for which the Euclidean and the Frobenius norms coincide. See [45, Proposition 3.1] for more details. The bound can be generalized to the use of other spaces, such as rational Krylov subspaces, see, e.g., [6,7,18,24].

We generalize the bound presented in Proposition 3.1 to the case of the Sylvester equation,

$$AX + XB = b_1 b_2^T, \tag{3.2}$$

with A and B symmetric and positive definite; without loss of generality, we can assume that $\|b_1\|_\star = \|b_2\|_\star = 1$ where $\|\cdot\|_\star$ denotes either the Euclidean or the Frobenius norms.

We first recall the Cauchy representation of the solution matrix X to (3.2). Let us for now only assume that A and B are positive definite, and not necessarily symmetric. We can write (see, e.g., [28])

$$X = \int_0^\infty e^{-tA} b_1 b_2^T e^{-tB} dt.$$

Consider the approximation $X_k = V_k Y_k W_k^T$ where V_k and W_k span suitable subspaces and both have orthonormal columns. The matrix Y_k is obtained by imposing the Galerkin condition on $R_k = AX_k + X_k B - b_1 b_2^T$, that is

$$V_k^T R_k W_k = 0 \quad \Leftrightarrow \quad (V_k^T A V_k)Y_k + Y_k(W_k^T B W_k) - (V_k^T b_1)(b_2^T W_k) = 0.$$

Let $A_k := V_k^T A V_k$, $B_k := W_k^T B W_k$. Thus Y_k is obtained by solving a reduced Sylvester equation, whose size depends on the approximation space dimensions. Since the spectrum of A_k (B_k) is contained in the spectral region of A (B) , we have that $\Lambda(A_k) + \Lambda(B_k) \subset \mathbb{C}_+$ and the matrix Y_k can be written in integral form as $Y_k = \int_0^\infty e^{-tA_k}(V_k^T b_1)(b_2^T W_k)e^{-tB_k}dt$ so that

$$X_k = V_k \int_0^\infty e^{-tA_k}(V_k^T b_1)(b_2^T W_k)e^{-tB_k}dt\, W_k^T.$$

Let $x := e^{-tA}b_1, x_k := V_k e^{-tA_k}(V_k^T b_1)$, $y := e^{-tB}b_2, y_k := W_k e^{-tB_k}(W_k^T b_2)$. Then, using $\|x\|_\star \leqslant e^{-t\alpha_{\min}(A)}$ (see, e.g., [16, Lemma 3.2.1]), and since $\alpha_{\min}(A_k) \geqslant \alpha_{\min}(A)$, it holds that $\|x_k\|_\star \leqslant e^{-t\alpha_{\min}(A)}$. Similarly, $\|y\|_\star, \|y_k\|_\star \leqslant e^{-t\alpha_{\min}(B)}$. Therefore, (see also [27, Lemma 4.7])

$$\begin{aligned}
\|X - X_k\|_\star &= \left\| \int_0^\infty (xy^T - x_k y_k^T)dt \right\|_\star \\
&= \frac{1}{2}\left\| \int_0^\infty (x + x_k)(y - y_k)^T + (x - x_k)(y + y_k)^T)dt \right\|_\star \\
&\leqslant \frac{1}{2}\int_0^\infty \Big((\|x\|_\star + \|x_k\|_\star)\|y - y_k\|_\star + \|x - x_k\|_\star(\|y\|_\star + \|y_k\|_\star) \Big)dt \\
&\leqslant \int_0^\infty \Big(e^{-t\alpha_{\min}(A)}\|y - y_k\|_\star + e^{-t\alpha_{\min}(B)}\|x - x_k\|_\star \Big)dt \\
&= \int_0^\infty (\|\hat{y} - \hat{y}_k\|_\star + \|\hat{x} - \hat{x}_k\|_\star)dt, \qquad (3.3)
\end{aligned}$$

where $\hat{y} = e^{-t(B+\lambda_{\min}(A)I)}b_2$, $\hat{x} = e^{-t(A+\lambda_{\min}(B)I)}b_1$, and analogously for $\hat{y}_k, \hat{x}_k$. The inequality in (3.3) states that the $\star$-norm of the error associated with the Galerkin solution can be bounded by integrating over $[0, \infty)$ the errors obtained in the approximation of the exponential of the shifted matrices $B + \lambda_{\min}(A)I$ and $A + \lambda_{\min}(B)I$.

In the next proposition we specialize the bound above when the Krylov subspaces range$(V_k) = K_k(A, b_1)$ and range$(W_k) = K_k(B, b_2)$ are adopted as approximation spaces and A, B are both symmetric and positive definite. To this end, let us define $\lambda_{\min}(A)$, $\lambda_{\max}(A)$, $\lambda_{\min}(B)$, and $\lambda_{\max}(B)$ to be the extreme eigenvalues of A and B, respectively, and

$$\hat{\kappa}_A = \frac{\lambda_{\max}(A) + \lambda_{\min}(B)}{\lambda_{\min}(A) + \lambda_{\min}(B)}, \qquad \hat{\kappa}_B = \frac{\lambda_{\max}(B) + \lambda_{\min}(A)}{\lambda_{\min}(B) + \lambda_{\min}(A)},$$

the condition numbers of $A + \lambda_{\min}(B)I$ and $B + \lambda_{\min}(A)I$, respectively.

PROPOSITION 3.2. *Let A and B be spd and range$(V_k) = K_k(A, b_1)$, range$(W_k) = K_k(B, b_2)$. Then the Galerkin approximate solution $X_k = V_k Y_k W_k^T$ to (3.2) is such that*

$$\|X - X_k\|_\star \leqslant \frac{2}{\lambda_{\min}(A) + \lambda_{\min}(B)} \left(\frac{\sqrt{\hat{\kappa}_A} + 1}{\sqrt{\hat{\kappa}_A}} \left(\frac{\sqrt{\hat{\kappa}_A} - 1}{\sqrt{\hat{\kappa}_A} + 1} \right)^k + \frac{\sqrt{\hat{\kappa}_B} + 1}{\sqrt{\hat{\kappa}_B}} \left(\frac{\sqrt{\hat{\kappa}_B} - 1}{\sqrt{\hat{\kappa}_B} + 1} \right)^k \right),$$

where $\|\cdot\|_\star$ denotes either the Euclidean or the Frobenius norm.

Proof. The proof can be obtained by applying the same arguments of the proof of [45, Proposition 3.1] to the single integrals $\int_0^\infty \|\hat{y} - \hat{y}_k\|_\star dt$, $\int_0^\infty \|\hat{x} - \hat{x}_k\|_\star dt$ in (3.3). ∎

Convergence results for generic matrix equations of the form (1.2) are difficult to derive as no easy-to-handle closed-form solution is known in general. The main difficulty is given by the fact that the exponential of a Kronecker sum $\sum_{j=1}^{\ell} B_j^T \otimes A_j$ cannot be separated in the product of the exponentials of the single terms if no further assumptions on A_j and B_j are considered.

By adapting the reasonings proposed in this section, one may be able to deduct error estimates for some special equations of the form

$$\sum_{j,k=0}^{\ell} \alpha_{j,k} A^j X B^k = F,$$

where the coefficient matrices are given as powers of two *seed* matrices A and B, and $\alpha_{j,k} \in \mathbb{R}$ for all j, k. Indeed, in this case, the exact solution X can be written in integral form as illustrated in [28, Theorem 4]. However, such derivations deserve a separate analysis.

4. Comparison with the Kronecker formulation. Given a linear matrix equation of the form (1.2), the simplest-minded numerical procedure for its solution consists in applying well-established iterative schemes to the *vector* linear system obtained from (1.2) by Kronecker transformations, namely

$$\left(\sum_{j=1}^{\ell} B_j^T \otimes A_j \right) \mathrm{vec}(X) = \mathrm{vec}(F). \tag{4.1}$$

Sometimes this is the only option as effective algorithms to solve (1.2) in its *natural* matrix equation form are still lacking in the literature in the most general case. The methods developed so far require some additional assumptions on the coefficient matrices A_j, B_j; see, e.g., [10, 23, 26, 38, 42].

In this section we show that exploiting the matrix structure of equation (1.2) not only leads to numerical algorithms with lower computational costs per iteration and modest storage demands, but they also avoid some spectral redundancy encoded in the problem formulation (4.1). Such a redundancy often leads to a delay in the convergence of the adopted solution scheme when iterative procedures are applied to (4.1). A similar discussion can be found in [27, Remark 4.5] for more general tensor structured problems.

To illustrate this phenomenon we consider a Lyapunov equation of the form (1.3) with $A \in \mathbb{R}^{n \times n}$ spd and $F = bb^T$, $b \in \mathbb{R}^n$, $\|b\| = 1$. We compare the Galerkin method applied to the matrix equation (1.3) with the CG method applied to the linear system

$$\mathcal{A}\mathrm{vec}(X) = \mathrm{vec}(bb^T), \quad \mathcal{A} = A \otimes I + I \otimes A \in \mathbb{R}^{n^2 \times n^2}. \tag{4.2}$$

Notice that since A is spd, $\mathcal{A}$ is also spd. Let $x = \mathrm{vec}(X)$ be the exact solution to (4.2). Let the CG initial guess be equal to the zero vector, and let x_k^{cg} be the approximate solution to x obtained after k CG iterations. Then the following classical bound for the energy-norm of the error $x - x_k^{cg}$ holds

$$\frac{\|x - x_k^{cg}\|_{\mathcal{A}}}{\|x\|_{\mathcal{A}}} \leqslant 2 \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^k, \tag{4.3}$$

where $\kappa = \lambda_{\max}(\mathcal{A})/\lambda_{\min}(\mathcal{A}) = \lambda_{\max}(A)/\lambda_{\min}(A)$. See, e.g., [20, Theorem 10.2.6]. This bound may be rather pessimistic since it takes into account neither the role of the right-hand side nor the actual spectral distribution of $\mathcal{A}$. See, e.g., [8, 9, 29, 51].

We want to compare the bound in (4.3) with the estimate proposed in Proposition 3.1, using the same norms and relative quantities. To this end, we recall that for any vector v it holds that

$$\sqrt{2\lambda_{\min}(A)}\|v\|_2 \leqslant \|v\|_{\mathcal{A}} \leqslant \sqrt{2\lambda_{\max}(A)}\|v\|_2.$$

In particular, letting $X_k^{cg} \in \mathbb{R}^{n\times n}$ be such that $\mathrm{vec}(X_k^{cg}) = x_k^{cg}$, we have

$$\frac{\|X - X_k^{cg}\|_F}{\|X\|_F} = \frac{\|x - x_k^{cg}\|_2}{\|x\|_2} \leqslant \sqrt{\kappa}\frac{\|x - x_k^{cg}\|_{\mathcal{A}}}{\|x\|_{\mathcal{A}}} \leqslant 2\sqrt{\kappa}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k. \tag{4.4}$$

Therefore, to obtain a relative error (in Frobenius norm) of less than ε, a sufficient number $k_*^{(cg)}$ of CG iterations is given by

$$k_*^{(cg)} := \frac{\log\left(\varepsilon/(2\sqrt{\kappa})\right)}{\log\left((\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)\right)}.$$

If X_k denotes the approximate solution computed after k iterations of the Galerkin-based method with $K_k(A,b)$ as approximation space, the error norm bound in (3.1) can be written in relative terms as

$$\frac{\|X - X_k\|_F}{\|X\|_F} \leqslant 4(\sqrt{\hat\kappa}+1)\sqrt{\hat\kappa}\left(\frac{\sqrt{\hat\kappa}-1}{\sqrt{\hat\kappa}+1}\right)^k, \tag{4.5}$$

where we used $\|x\|_2 \geqslant \lambda_{\min}(\mathcal{A}^{-1})\,\|\mathrm{vec}(bb^T)\|_F = 1/(\lambda_{\max}+\lambda_{\min})$. Once again, to obtain a relative error (in Frobenius norm) of less than ε, a sufficient number $k_*^{(G)}$ of iterations is given by

$$k_*^{(G)} := \frac{\log\left(\varepsilon/(4\sqrt{\hat\kappa}(\sqrt{\hat\kappa}+1))\right)}{\log\left((\sqrt{\hat\kappa}-1)/(\sqrt{\hat\kappa}+1)\right)}.$$

The bounds (4.4)–(4.5) show that the asymptotic behavior of the relative error norms of CG and the Galerkin method are guided by κ and $\hat\kappa$, respectively, where $\hat\kappa$ is always smaller than κ, for $\kappa > 1$. Indeed,

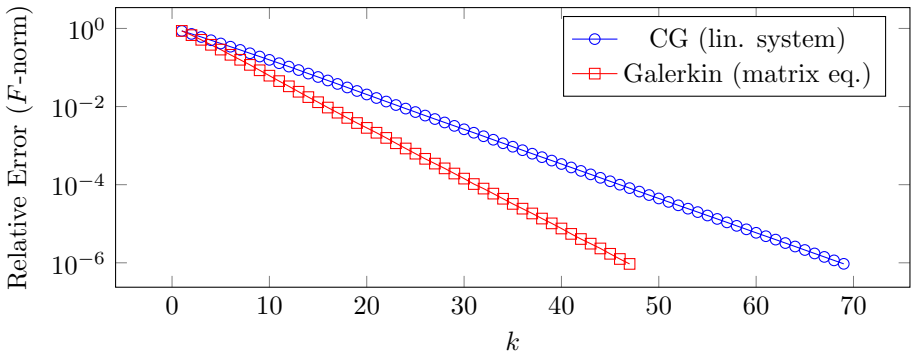
$$\hat\kappa = \frac{\lambda_{\max}(A)+\lambda_{\min}(A)}{2\lambda_{\min}(A)} = \frac{1}{2}\kappa + \frac{1}{2}.$$

The worse conditioning of the linear system formulation (4.2) may lead to a delay in the convergence of CG so that, for a fixed threshold, CG may require more iterations to converge than the Galerkin method applied to the matrix equation (1.3). This is numerically illustrated in the examples below.

We once again stress that the similarities of the two formulations (matrix equation and Kronecker form) highlight the fact that what makes the matrix equation context more efficient than CG on $\mathcal{A}x = b$ is the special choice of the approximation space, that is $\mathcal{K}_m = \mathrm{range}(V_k \otimes V_k)$, which heavily takes into account the Kronecker sum structure of $\mathcal{A}$. On the other hand, CG applied blindly on $\mathcal{A}$ generates a redundant approximation space.

Example 4.1. We consider the spd matrix $A = QDQ^T \in \mathbb{R}^{n\times n}$, where D is a diagonal matrix whose diagonal entries are uniformly distributed (in logarithmic scale) values between 1 and 100, and Q is orthogonal. This means that $\kappa = 100$ and $\hat\kappa = 50.5$ for any n. The vector $b \in \mathbb{R}^n$ is a random vector with unit norm.

FIG. 4.1. *Example 4.1. Relative error norms produced by the Galerkin and CG methods.*



For $\varepsilon = 10^{-6}$, a direct computation shows that $k_*^{(G)} = 68$ iterations of the Galerkin method are sufficient to get $\|X - X_{k_*^{(G)}}\|_F / \|X\|_F \leqslant \varepsilon$, whereas according to the bound (4.4), $k_*^{(cg)} = 84$ iterations are required for CG to reach the same accuracy when solving $\mathcal{A}x = b$. In practice, the number of actual iterations can be lower, since this estimate is obtained from a bound.

Figure 4.1 reports the error convergence history of the two iterations, using logarithmic scale for $n = 1000$. The two methods are stopped as soon as the relative error norm becomes smaller than ε. The "exact" solution X was computed with the Bartels-Stewart method [4], which was feasible due to the small problem size.

Both methods require slightly fewer iterations than predicted by the bounds. Nonetheless, we can still appreciate that CG applied to the linear system (4.2) requires more iterations than the Galerkin method applied to the matrix equation (1.3) to achieve the same prescribed accuracy.

EXAMPLE 4.2. We modify the data of Example 4.1 by replacing $\lambda_{\min}(A) = 1$ with $\lambda_{\min}(A) = \lambda_1 = 0.001$, while the other eigenvalues are such that $\lambda_2 = 2, \dots, \lambda_n = n$. Here $b \in \mathbb{R}^n$ is the vector of all ones normalized. The relative error energy norm obtained by CG and the Galerkin method is reported in the left plot of Figure 4.2 for $n = 100$. Notice that with such a n we have $\kappa = 10^5, \hat{\kappa} \approx 5 \cdot 10^4$.

Both methods stagnate in the initial phase of the solution process, followed by a rapid convergence afterwards. The stagnation phase is significantly longer for CG, contributing to the overall CG delay. A closer look at the convergence history of the Ritz values towards the eigenvalues of the corresponding coefficient matrices provides a better understanding. In this setting, the Ritz values for the Galerkin and CG methods are the eigenvalues of $V_k^T A V_k$ and of $Q_k^T \mathcal{A} Q_k$, respectively, where the columns of Q_k are the orthonormal basis of the space generated by CG (here we used the Arnoldi procedure to compute Q_k). Recalling that $\lambda_{\min}(A) = 0.001$, $\lambda_{\max}(A) = 100$ so that $\lambda_{\min}(\mathcal{A}) = 0.002$, $\lambda_{\max}(\mathcal{A}) = 200$, the right plot of Figure 4.2 reports the convergence history of the extreme Ritz values computed at the k-th iteration of both CG and the Galerkin method for $k = 1, \dots, 82$ (the dashed lines indicate the target eigenvalues). The Ritz value tending to the largest eigenvalue converges in very few iterations. For each approach, the Ritz value approximating the smallest eigenvalue takes many more iterations to converge, and these iterations seem to match the stagnation phase observed in the left plot. It appears that the matrix Galerkin approximation space is able to implicitly capture the Kronecker structure of the eigenvector associated with $\lambda_{\min}(\mathcal{A})$ much earlier than what CG can do by using the unstructured basis Q_k. Once again,

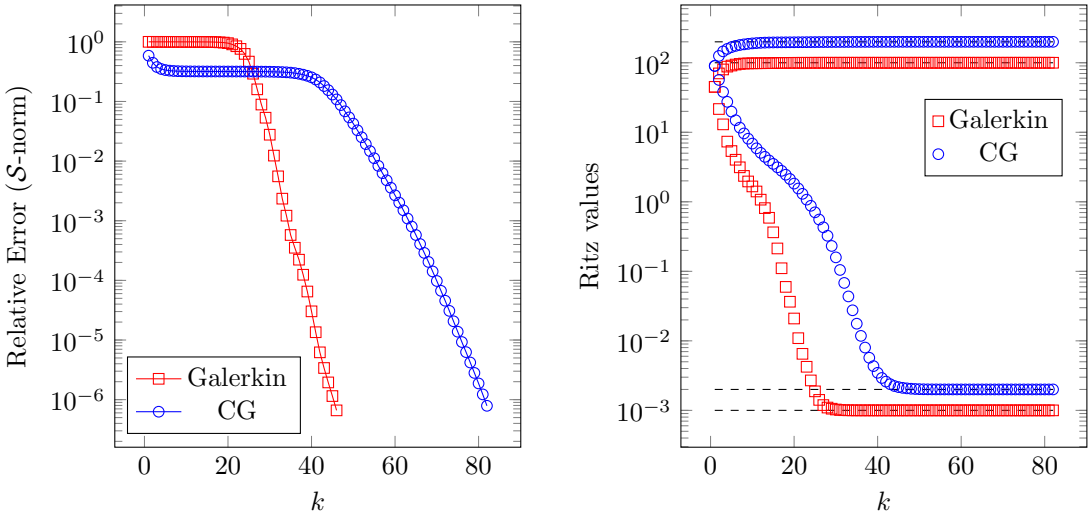


FIG. 4.2. *Example 4.2. Left: Relative error energy norms produced by the Galerkin and CG methods. Right: Maximum and minimum Ritz values computed at the k-th iteration of CG and the Galerkin method. The dashed black line represents the quantity to be approximated, namely* $\lambda_{\min}(A)$, $\lambda_{\max}(A)$, $\lambda_{\min}(\mathcal{A})$, *and* $\lambda_{\max}(\mathcal{A})$.

this emphasizes the importance of the Kronecker basis determined by the matrix Galerkin method.

5. Petrov-Galerkin method and residual minimization. Whenever the linear operator $\mathcal{S}$ is not spd, the Galerkin method does not necessarily lead to a minimization of the error norm. As in the linear system setting, a numerical procedure fulfilling an optimality condition can be obtained by imposing a Petrov-Galerkin condition on the residual also when solving linear matrix equations. For the case of Lyapunov and Sylvester equations, this strategy has been already explored in, e.g., [22,30], and in this section we are going to present some considerations about the application of Petrov-Galerkin methods to the solution of generic linear matrix equations of the form (1.2).

We first recall the Petrov-Galerkin framework applied to the solution of the linear system (1.1). If the columns of $\mathcal{V}_m \in \mathbb{R}^{N\times m}$ constitute an orthonormal basis for the selected trial space $\mathcal{K}_m$, we want to compute a solution $x_m = \mathcal{V}_m y_m$, where $y_m \in \mathbb{R}^m$ is calculated by imposing a Petrov-Galerkin condition on the residual vector $r_m = f - \mathcal{M}x_m$. In its full generality, such a condition reads

$$r_m \perp \mathcal{L}_m, \text{ i.e., } \mathcal{W}_m^T r_m = 0, \text{ range}(\mathcal{W}_m) = \mathcal{L}_m, \tag{5.1}$$

where $\mathcal{L}_m$ is the chosen test space. See, e.g., [40, Chapter 5].

For the particular choice $\mathcal{L}_m = \mathcal{M}\mathcal{K}_m$, the condition in (5.1) is equivalent to computing x_m as the minimizer of the residual norm over $\mathcal{K}_m$, namely

$$x_m = \underset{x\in\mathcal{K}_m}{\text{argmin}} \|f - \mathcal{M}x\|.$$

See, e.g., [40, Proposition 5.3]. With the selection $\mathcal{K}_m = K_m(\mathcal{M}, f)$, the minimization problem above can be significantly simplified by exploiting the Arnoldi relation; this is the foundation of some of the most popular minimal residual methods for linear systems such as, e.g., MINRES [34] and GMRES [41].

A similar approach can be pursued for the solution of linear matrix equations. Indeed, let $N = np$ and consider $V_k \in \mathbb{R}^{n\times k}$, $W_k \in \mathbb{R}^{p\times k}$ with full column rank [3], and let range(V_k), range(W_k), be the corresponding left and right approximation spaces. With $\mathcal{S}_\ell$ as in Definition 2.1, we can formally set $\mathcal{K}_m = \text{range}(W_k \otimes V_k)$ and $\mathcal{L}_m = \mathcal{S}_\ell \mathcal{K}_m$. An approximate solution in the form $X_k = V_k Y_k W_k^T$, with $Y_k \in \mathbb{R}^{k\times k}$, can be determined by imposing the condition (5.1) to the vector form of the residual matrix $R_k = \mathcal{S}(V_k Y_k W_k^T) - F$.

Petrov-Galerkin methods for (1.2) thus seek a solution $X_k = V_k Y_k W_k^T$ by solving

$$\min_{x\in\text{range}(W_k\otimes V_k)} \|\text{vec}(F) - \mathcal{S}_\ell x\|_2 = \min_{y\in\mathbb{R}^{k^2}} \|\text{vec}(F) - \mathcal{S}_\ell(W_k \otimes V_k)y\|_2,$$

that is

$$\min_{X=V_k Y W_k^T} \|F - \mathcal{S}(X)\|_F = \min_{Y\in\mathbb{R}^{k\times k}} \|F - \mathcal{S}(V_k Y W_k^T)\|_F. \tag{5.2}$$

In spite of their appealing minimization property, minimal residual methods are not very popular in the matrix equation literature. This is mainly due to the difficulty in dealing with the numerical solution of the minimization problem (5.2). In general, one can apply an operator-oriented (preconditioned) CG method to the normal equations as

$$Y_k = \operatorname*{argmin}_{Y\in\mathbb{R}^{k\times k}} \|F - \mathcal{S}(V_k Y W_k^T)\|_F \qquad \Leftrightarrow \qquad \mathcal{S}^*(F - \mathcal{S}(V_k Y_k W_k^T)) = 0, \tag{5.3}$$

where $\mathcal{S}^*$ is the adjoint of $\mathcal{S}$, namely

$$\begin{aligned} \mathcal{S}^* : \quad \mathbb{R}^{n\times p} &\to \mathbb{R}^{n\times p} \\ X &\mapsto \sum_{j=1}^{\ell} A_j^T X B_j^T. \end{aligned}$$

If range(V_k) and range(W_k) are general spaces, the solution of (5.3) can be very expensive in terms of both computational time and memory requirements.

In [30], the authors consider (5.3) in the case of the Lyapunov equation (1.3) with F low-rank and negative semidefinite. In particular, if $F = -bb^T$, $b \in \mathbb{R}^{n\times q}$, $q \ll n$, they employ the approximation spaces range(V_k) = range(W_k) such that $b = V_1 L_b$ for some $L_b \in \mathbb{R}^{q\times q}$, $q = \text{rank}(C)$, and satisfying an Arnoldi-like relation of the form

$$AV_k = [V_k, \breve{V}_{k+1}]\underline{H}_k,$$

for $[V_k, \breve{V}_{k+1}] \in \mathbb{R}^{n\times(k+1)q}$ having orthonormal columns and $\underline{H}_k \in \mathbb{R}^{(k+1)q\times kq}$. In this case, the minimization problem (5.2) can be written as

$$Y_k = \operatorname*{argmin}_{Y\in\mathbb{R}^{kq\times kq}} \left\| \underline{H}_k Y [I_{kq}, 0] + \begin{bmatrix} I_{kq} \\ 0 \end{bmatrix} Y \underline{H}_k^T + \begin{bmatrix} L_b L_b^T & 0 \\ 0 & 0 \end{bmatrix} \right\|_F, \tag{5.4}$$

and three different methods for its solution are illustrated.

[3] Once again, the two matrices may have different column dimensions, that is $V_{k_1} \in \mathbb{R}^{n\times k_1}$, $W_{k_2} \in \mathbb{R}^{p\times k_2}$. For the sake of clarity in the exposition, we limit our presentation to the case $k_1 = k = k_2$.

If the coefficient matrix A in (1.3) is stable (antistable) and F is symmetric negative semidefinite, the exact solution X is symmetric positive (negative) semidefinite. See, e.g., [47]. However, as reported in [30], the numerical solution $X_k = V_k Y_k V_k^T$ is not guaranteed to be semidefinite if Y_k is computed as in (5.4).

In [30, Section 3.4] it is shown that (5.4) is equivalent to computing Y_k as the solution of the *generalized* Sylvester equation

$$\underline{H}_k^T \underline{H}_k Y + Y \underline{H}_k^T \underline{H}_k + H_k Y H_k + H_k^T Y H_K^T + D = 0, \tag{5.5}$$

where

$$D := \underline{H}_k^T \begin{bmatrix} L_b L_b^T & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} I_{kq} \\ 0 \end{bmatrix} + [I_{kq}, 0] \begin{bmatrix} L_b L_b^T & 0 \\ 0 & 0 \end{bmatrix} \underline{H}_k = H_k^T \begin{bmatrix} L_b L_b^T & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} L_b L_b^T & 0 \\ 0 & 0 \end{bmatrix} H_k,$$

so that D is symmetric but indefinite. This is one of the main obstacles in proving the semidefiniteness of Y_k through the matrix formulation (5.5). Without further hypotheses, the symmetric matrix Y_k solving (5.5) is indefinite in general, thus preventing Y_k from preserving the semidefiniteness property of the solution to be approximated.

From a computational viewpoint, if resorting to a Kronecker form is excluded, the generalized Sylvester equation (5.5) can be solved by means of the methods described in [30] and its references. In addition, setting $\mathfrak{L}(Z) = \underline{H}_k^T \underline{H}_k Z + Z \underline{H}_k^T \underline{H}_k$ and $\mathfrak{N}(Z) = H_k Z H_k + H_k^T Z H_k^T$, fixed point iterations can be used whenever the spectral radius of the operator $\mathfrak{L}^{-1}(\mathfrak{N}(\cdot))$ is less than one; see, e.g., [17, 23, 42] for various implementations.

5.1. A constrained residual minimization approach for Lyapunov equations. To cope with the lack of semidefiniteness in the least squares problem approach, we propose to explicitly impose the semidefiniteness as a constraint. For instance, if a negative semidefinite solution is sought, the problem becomes

$$Y_k = \underset{\substack{Y \in \mathbb{R}^{kq \times kq} \\ Y \leqslant 0}}{\operatorname{argmin}} \left\| \underline{H}_k Y [I_{kq}, 0] + \begin{bmatrix} I_{kq} \\ 0 \end{bmatrix} Y \underline{H}_k^T + \begin{bmatrix} L_b L_b^T & 0 \\ 0 & 0 \end{bmatrix} \right\|_F. \tag{5.6}$$

To numerically solve this inequality constrained least squares problem, we consider a linear matrix inequalities (LMI) approach, which suits very well the matrix equation framework [15, 46]; other general purpose methods could also be considered [1, 32].

In the LMI context, (5.6) can be stated as the following semidefiniteness matrix inequalities

$$Y \leqslant 0, \qquad \begin{bmatrix} I & \operatorname{vec}(\underline{H}_k Y J^T + J Y \underline{H}_k^T + M) \\ \operatorname{vec}(\underline{H}_k Y J^T + J Y \underline{H}_k^T + M)^T & \gamma \end{bmatrix} \geq 0,$$

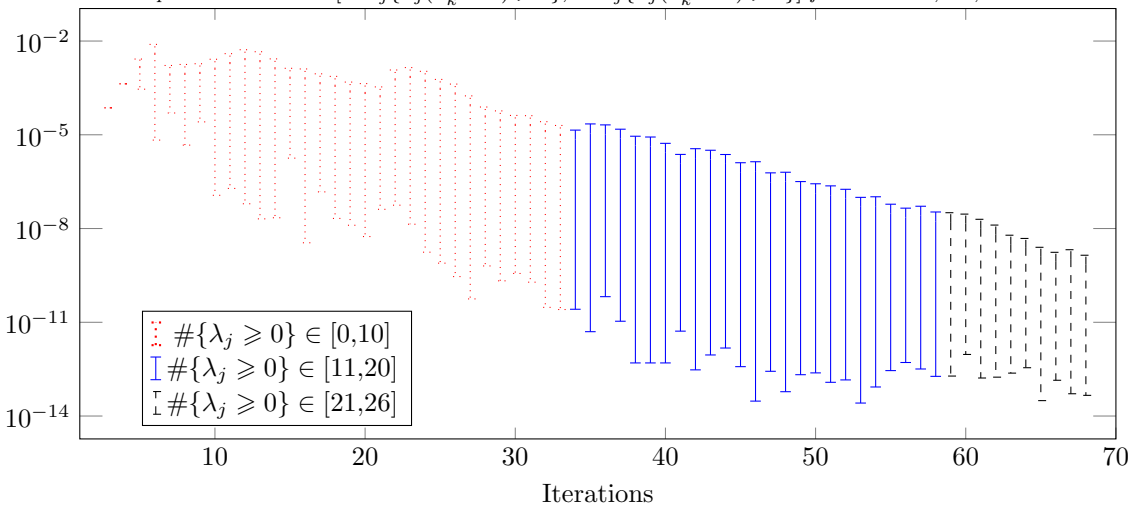
for the unknown matrix Y and scalar $\gamma > 0$; here $J = [I_{kq}; 0]$ and $M = [L_b L_b^T, 0; 0, 0]$.

Example 5.1. We consider the Lyapunov equation (1.3) with $A = QDQ^{-1} \in \mathbb{R}^{n \times n}$, D as in Example 4.1, Q a random matrix, and $F = -bb^T$, where $b \in \mathbb{R}^n$ is a random vector with unit norm.

Since A is antistable and the right-hand side is symmetric negative semidefinite, the solution X is symmetric negative semidefinite and we thus expect the approximate solution $X_k = V_k Y_k V_k^T$ to be so as well.

We apply the Petrov-Galerkin method discussed in this section in the solution process and we adopt the Krylov subspace as approximation space, i.e., $\operatorname{range}(V_k) = K_k(A, b)$. The matrix Y_k

FIG. 5.1. *Example 5.1. Intervals* $[\min_j\{\lambda_j(Y_k^{uncon}) \geqslant 0\}, \max_j\{\lambda_j(Y_k^{uncon}) \geqslant 0\}]$ *for all* $k = 1, \dots, 68$. $n = 1000$.



is computed in two different ways. We first solve the unconstrained minimization problem (5.4) getting the matrix Y_k^{uncon}. In particular, Y_k^{uncon} is computed by applying a (preconditioned) CG method to the matrix equation (5.5). See, e.g., [30]. Then, we compute Y_k^{const} by solving the constrained minimization problem (5.6). The Petrov-Galerkin method is stopped as soon as the relative residual norm becomes smaller than 10^{-6}.

In Figure 5.1 we plot the intervals $[\min_j\{\lambda_j(Y_k^{\text{uncon}}) \geqslant 0\}, \max_j\{\lambda_j(Y_k^{\text{uncon}}) \geqslant 0\}]$ of the undesired positive eigenvalues of Y_k^{uncon} for all k for the case $n = 1000$. For $k = 1, 2$, Y_k^{uncon} has all negative eigenvalues, while it starts being indefinite for $k \geqslant 3$ so that $X_k^{\text{uncon}} = V_k Y_k^{\text{uncon}} V_k^T$ is an indefinite approximation to the negative semidefinite X. Nonetheless, for $k = 68$, the (undesired) positive eigenvalues of X_k^{uncon} are small enough so as to still allow a sufficiently accurate approximation, in terms of relative residual norm. On the other hand, this problem is not encountered with Y_k^{constr}, thanks to the explicit negative semidefiniteness constraint in the formulation (5.6).

From the legend of Figure 5.1, we can see that the number of positive eigenvalues of Y_k^{uncon} increases as the iteration proceed, even though they diminish in magnitude. The latter trend is not surprising. Indeed, even if Y_k^{uncon} is computed by (5.4), the Petrov-Galerkin method is converging towards the negative semidefinite solution X and, for an approximation space spanning the whole $\mathbb{R}^n$, the method would retrieve the exact solution, regardless of the minimization problem (5.4).

We would like to point out that both tested variants of the Petrov-Galerkin method needed 68 iterations to converge and the actual values of the residual norm provided by (5.4) and (5.6) were always very similar to each other, during the whole convergence history. This phenomenon surely deserves further studies as, in principle, (5.6) leads to a residual norm that is greater or equal than the one provided by (5.4), while the two solutions (constrained and unconstrained) do not necessarily have to be close to each other.

In our computational experiments, we have used the Yalmip software [31] running the algorithm Sedumi in Matlab [50]. This algorithm is rather expensive and computing the solution Y_k to (5.6) at each Krylov iteration k often leads to a very time consuming solution procedure. We think this issue can be fixed in different ways. For instance, one may compute Y_k, and thus check the residual

norm, only periodically, say every $d \geqslant 1$ iterations. Moreover, the explicit solution Y_k is required only at convergence while we just need the value of the residual norm during the Krylov routine. It may be possible to compute such a residual norm without calculating the whole Y_k as it is done in [36] for the Galerkin method and in [30] for the Petrov-Galerkin technique equipped with the unconstrained minimization problem (5.4).

The study of the aforementioned enhancements and, more in general, the employment of constrained minimization procedures in the solution of linear matrix equations will be the topic of future research.

6. Conclusions. We have shown that the optimality properties of Galerkin and Petrov-Galerkin methods naturally extend to the general linear matrix equation setting. Such features do not depend on the adopted approximation spaces even though, in actual computations, fast convergence depends on the suitable subspace selection. Identifying effective subspaces for general (multiterm) linear matrix equations depends on the problem at hand, and it may seem easier to recast the solution in terms of a large vector linear system. On the other hand, the vector form can be extremely memory consuming, while the vector linear system encodes some spectral redundancy which may cause a delay in the converge of the adopted iterative solution scheme.

Petrov-Galerkin schemes require to solve a matrix minimization problem at each iteration and we have suggested to explicitly incorporate a semidefiniteness constraint in its formulation. To the best of our knowledge, such approach has never been proposed in the literature and the employment of constrained optimization techniques in the context of Petrov-Galerkin methods for linear matrix equations opens many new research directions.

Acknowledgements. Both authors are members of the Italian INdAM Research group GNCS. We thank the two anonymous reviewers for their insightful remarks.

REFERENCES

[1] M. F. Anjos and J. B. Lasserre, eds., *Handbook on semidefinite, conic and polynomial optimization*, vol. 166 of International Series in Operations Research & Management Science, Springer, New York, 2012.

[2] A. C. Antoulas, *Approximation of large-scale dynamical systems*, vol. 6 of Advances in Design and Control, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.

[3] I. Babuška, R. Tempone, and G. E. Zouraris, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825.

[4] R. H. Bartels and G. W. Stewart, *Algorithm 432: Solution of the Matrix Equation $AX + XB = C$*, Comm. ACM, 15 (1972), pp. 820–826.

[5] M. Baumann, R. Astudillo, Y. Qiu, E. Y. M. Ang, M. B. van Gijzen, and R.-É. Plessix, *An MSSS-preconditioned matrix equation approach for the time-harmonic elastic wave equation at multiple frequencies*, Computational Geosciences, 22 (2018), pp. 43–61.

[6] B. Beckermann, *An Error Analysis for Rational Galerkin Projection applied to the Sylvester Equation*, SIAM J. Numer. Anal., 49 (2011), pp. 2430–2450.

[7] B. Beckermann, D. Kressner, and Ch. Tobler, *An error analysis of Galerkin projection methods for linear systems with tensor product structure*, SIAM J. Numer. Anal., 51 (2013), pp. 3307–3326.

[8] B. Beckermann and A. B. J. Kuijlaars, *Superlinear convergence of conjugate gradients*, SIAM J. Numer. Anal., 39 (2001), pp. 300–329.

[9] ———, *Superlinear CG convergence for special right-hand sides*, vol. 14, 2002, pp. 1–19. Orthogonal polynomials, approximation theory, and harmonic analysis (Inzel, 2000).

[10] P. Benner and T. Breiten, *Low rank methods for a class of generalized Lyapunov equations and related issues*, Numer. Math., 124 (2013), pp. 441–470.

[11] ———, *Rational interpolation methods for symmetric Sylvester equations*, Electron. Trans. Numer. Anal., 42 (2014), pp. 147–164.

[12] P. Benner and T. Damm, *Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems*, SIAM J. Control Optim., 49 (2011), pp. 686–711.
[13] P. Benner, V. Mehrmann, and D. C. Sorensen, eds., *Dimension reduction of large-scale systems*, vol. 45 of Lecture Notes in Computational Science and Engineering, Springer, Berlin, 2005.
[14] P. Benner and J. Saak, *Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey*, GAMM-Mitt., 36 (2013), pp. 32–52.
[15] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*, vol. 15 of SIAM Studies in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994.
[16] M. J. Corless and A. E. Frazho, *Linear Systems and Control – An Operator Perspective*, Pure Appl. Math., Marcel Dekker, New York, Basel, 2003.
[17] T. Damm, *Direct methods and ADI-preconditioned Krylov subspace methods for generalized Lyapunov equations*, Num. Lin. Alg. with Appl., 15 (2008), pp. 853–871. Special issue on Matrix equations.
[18] V. Druskin, L. 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.
[19] M. A. Freitag and D. L. H. Green, *A low-rank approach to the solution of weak constraint variational data assimilation problems*, J. Comput. Phys., 357 (2018), pp. 263–281.
[20] G. H. Golub and C. F. Van Loan, *Matrix computations*, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, fourth ed., 2013.
[21] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436 (1953).
[22] D. Y. Hu and L. Reichel, *Krylov-subspace methods for the Sylvester equation*, Linear Algebra Appl., 172 (1992), pp. 283–313. Second NIU Conference on Linear Algebra, Numerical Linear Algebra and Applications (DeKalb, IL, 1991).
[23] E. Jarlebring, G. Mele, D. Palitta, and E. Ringh, *Krylov methods for low-rank commuting generalized sylvester equations*, Numerical Linear Algebra with Applications, 25 (2018). e2176.
[24] L. Knizhnerman and V. Simoncini, *Convergence analysis of the Extended Krylov Subspace Method for the Lyapunov equation*, Numerische Mathematik, 118 (2011), pp. 567–586.
[25] M. Konstantinov, V. Mehrmann, and P. Petkov, *On properties of Sylvester and Lyapunov operators*, Linear Algebra Appl., 312 (2000), pp. 35–71.
[26] D. Kressner and P. Sirković, *Truncated low-rank methods for solving general linear matrix equations*, Numer. Linear Algebra Appl., 22 (2015), pp. 564–583.
[27] D. Kressner and C. Tobler, *Krylov subspace methods for linear systems with tensor product structure*, SIAM J. Matrix Analysis and Appl., 31 (2010), pp. 1688–1714.
[28] P. Lancaster, *Explicit solutions of linear matrix equations*, SIAM Rev., 12 (1970), pp. 544–566.
[29] Joerg Liesen and Zdenek Strakos, *Krylov Subspace Methods. Principles and Analysis*, Oxford University Press, 2013.
[30] Y. Lin and V. Simoncini, *Minimal residual methods for large scale Lyapunov equations*, Appl. Numer. Math., 72 (2013), pp. 52–71.
[31] J. Löfberg, *YALMIP : A toolbox for modeling and optimization in MATLAB*, in Proceedings of the CACSD Conference, Taipei, Taiwan, 2004.
[32] J. Malick, *A dual approach to semidefinite least-squares problems*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 272–284.
[33] V. L. Mehrmann, *The autonomous linear quadratic control problem*, vol. 163 of Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, 1991. Theory and numerical solution.
[34] C. C. Paige and M. A. Saunders, *Solutions of sparse indefinite systems of linear equations*, SIAM J. Numer. Anal., 12 (1975), pp. 617–629.
[35] D. Palitta and V. Simoncini, *Matrix-equation-based strategies for convection-diffusion equations*, BIT, 56 (2016), pp. 751–776.
[36] D. Palitta and V. Simoncini, *Computationally enhanced projection methods for symmetric Sylvester and Lyapunov equations*, J. Comput. Appl. Math., 330 (2018), pp. 648–659.
[37] C. E. Powell and H. C. Elman, *Block-diagonal preconditioning for spectral stochastic finite-element systems*, IMA J. Numer. Anal., 29 (2009), pp. 350–375.
[38] C. E. Powell, D. Silvester, and V. Simoncini, *An efficient reduced basis solver for stochastic Galerkin matrix equations*, SIAM J. Scient.Comput., 39 (2017), pp. A141–A163.
[39] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halstead Press, New York, 1992.
[40] Y. Saad, *Iterative methods for sparse linear systems*, SIAM, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2nd ed., 2003.

[41] Y. Saad and M. H. Schultz, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM Journal on Scientific and Statistical Computing, No. 7 (1986), pp. pp. 856–869.
[42] S. D. Shank, V. Simoncini, and D. B. Szyld, *Efficient low-rank solution of generalized Lyapunov equations*, Numer. Math., 134 (2016), pp. 327–342.
[43] D. J. Silvester, A. Bespalov, and C. E. Powell, *S-IFISS version 1.04*, 2017.
[44] V. Simoncini, *Computational methods for linear matrix equations*, SIAM Rev., 58 (2016), pp. 377–441.
[45] V. Simoncini and V. Druskin, *Convergence analysis of projection methods for the numerical solution of large Lyapunov equations*, SIAM J. Numer. Anal., Vol. 47, No. 2 (2009), pp. pp. 828–843.
[46] R. E. Skelton, T. Iwasaki, and K. M. Grigoriadis, *A unified algebraic approach to linear control design*, The Taylor & Francis Systems and Control Book Series, Taylor & Francis, Ltd., London, 1998.
[47] J. Snyders and M. Zakai, *On nonnegative solutions of the equation $AD + DA' = -C$*, SIAM J. Appl. Math., 18 (1970), pp. 704–714.
[48] M. Stoll and T. Breiten, *A low-rank in time approach to PDE-constrained optimization*, SIAM J. Sci. Comput., 37 (2015), pp. B1–B29.
[49] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, New York, 1973.
[50] J. F. Sturm, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, vol. 11/12, 1999, pp. 625–653. Interior point methods.
[51] A. van der Sluis and H. A. van der Vorst, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.
[52] B. Vandereycken and S. Vandewalle, *A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations*, SIAM J. Matrix Analysis and Appl., 31 (2010), pp. 2553–2579.