View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Infoscience - École polytechnique fédérale de Lausanne

An error analysis of Galerkin projection methods for linear systems with tensor product structure

Bernhard Beckermann* Daniel Kressner† Christine Tobler‡

July 2, 2013

Abstract

Recent results on the convergence of a Galerkin projection method for the Sylvester equation are extended to more general linear systems with tensor product structure. In the Hermitian positive definite case, explicit convergence bounds are derived for Galerkin projection based on tensor products of rational Krylov subspaces. The results can be used to optimize the choice of shifts for these methods. Numerical experiments demonstrate that the convergence rates predicted by our bounds appear to be tight.

Linear system, Kronecker product structure, Sylvester equation, tensor projection, Galerkin projection, rational Krylov subspaces.

15A24, 65F10.

1 Introduction

This paper is concerned with Galerkin methods on tensor product subspaces for particularly structured large-scale linear systems. These structures are motivated by the Sylvester equation

$$A_1 X + X A_2^T = C, \tag{1}$$

with coefficient matrices $A_1 \in \mathbb{C}^{n_1 \times n_1}, A_2 \in \mathbb{C}^{n_2 \times n_2}$, a right-hand side matrix $C \in \mathbb{C}^{n_1 \times n_2}$ and the solution matrix $X \in \mathbb{C}^{n_1 \times n_2}$. Using Kronecker products, the matrix equation (1) can be reformulated as a linear system

$$(I_{n_2} \otimes A_1 + A_2 \otimes I_{n_1})x = c, \tag{2}$$

where $c = \mathrm{vec}(C)$ and $x = \mathrm{vec}(X)$. Sylvester equations arise in the context of numerical methods for eigenvalue problems [9, 13], algebraic Riccati equations [5], and model reduction [1]. In the last two cases, the right-hand side C often has low rank.

Given orthonormal bases $V_1 \in \mathbb{C}^{n_1 \times k_1}$ and $V_2 \in \mathbb{C}^{n_2 \times k_2}$, the Galerkin method on the tensor product subspace $\mathrm{span}(V_1) \otimes \mathrm{span}(V_2)$ constructs an approximation to (2) as $\widetilde{x} = (V_2 \otimes V_1)y$, where y is chosen such that

$$(I_{k_2} \otimes \widetilde{A}_1 + \widetilde{A}_2 \otimes I_{k_1})y = (V_2^* \otimes V_1^*)c, \tag{3}$$

*Laboratoire Painlevé UMR 8524 (ANO-EDP), UFR Mathématiques – M3, UST Lille, F-59655 Villeneuve d'Ascq CEDEX, France. E-mail: bbecker@math.univ-lille1.fr.

†ANCHP, MATHICSE, EPF Lausanne, Switzerland. daniel.kressner@epfl.ch

‡ANCHP, MATHICSE, EPF Lausanne, Switzerland. christine.tobler@epfl.ch

with $\widetilde{A}_1 = V_1^* A_1 V_1$ and $\widetilde{A}_2 = V_2^* A_2 V_2$. A number of numerical methods for solving Sylvester and Lyapunov equations can be seen as special cases of such a Galerkin approach. This includes methods based on Krylov subspaces [17, 24], extended Krylov subspaces [26], and rational Krylov subspaces [6, 18].

The convergence of the Galerkin method on tensor products of (rational) Krylov subspaces for Lyapunov and Sylvester equations has been analysed in [2, 10, 21, 20, 27, 28]. For the extensions considered in this paper the framework developed in [2] appears to be most suitable. It is based on a decomposition of the residual into a sum of three orthogonal vectors, as follows:

$$r := c - (I \otimes A_1 + A_2 \otimes I)(V_2 \otimes V_1)y = (V_2 \otimes I)r_1 + (I \otimes V_2)r_2 + \widehat{c}, \tag{4}$$

where

$$\begin{aligned} r_1 &= (V_2^* \otimes I)c - (I \otimes A_1 + \widetilde{A}_2 \otimes I)(I \otimes V_1)y, \\ r_2 &= (I \otimes V_1^*)c - (I \otimes \widetilde{A}_1 + A_2 \otimes I)(V_2 \otimes I)y, \\ \widehat{c} &= ((I - V_2 V_2^*) \otimes (I - V_1 V_1^*))\, c. \end{aligned}$$

The usual choices for V_1 and V_2 yield $\widehat{c} = 0$. The partial residuals r_1 and r_2 can be analysed separately and more easily compared to r as a whole. In particular, it becomes relatively straightforward to derive convergence bounds based on the fields of values of A_1 and A_2.

A natural extension of (2) is given by the linear system

$$\left(\sum_{\mu=1}^{d} I_{n_d} \otimes \cdots \otimes I_{n_{\mu+1}} \otimes A_\mu \otimes I_{n_{\mu-1}} \otimes \cdots \otimes I_{n_1} \right) x = c, \tag{5}$$

with coefficient matrices $A_\mu \in \mathbb{C}^{n_\mu \times n_\mu}$. Such linear systems arise, for example, from the discretization of a separable d-dimensional PDE with tensorized finite elements [14, 21]. Moreover, methods for (5) can be used as preconditioners in iterative methods for more general linear systems [19, 4] and eigenvalue problems [22]. Note that the solution vector $x \in \mathbb{R}^{n_1 n_2 \cdots n_d}$ quickly grows in size as d increases. For large d, this growth will exclude the application of any standard linear solver to (5). In fact, for a general right-hand side c, it is questionable whether the solution of (5) can be approached at all for large d. However, in the special case when c can be written as a Kronecker product (or as a short sum of Kronecker products), a number of algorithms have recently been developed that are capable of dealing even with $d = 50$ and larger. A method that approximates x by a sum of Kronecker products of vectors was proposed in [14], based on the approximation of the scalar inverse function by a sum of exponentials. A Galerkin method on the tensor subspace spanned by $V_d \otimes \cdots \otimes V_1$ for orthonormal bases $V_\mu \in \mathbb{C}^{n_\mu \times k_\mu}$ was proposed in [21]. This approach leads to a smaller linear system of size $k_1 \cdots k_d$, which is solved by the method from [14]. More recently, an ADI-like method, applying low-rank tensor approximations in each ADI iteration was proposed [23].

In this paper, we provide an error analysis of such a Galerkin method based on the ideas of Beckermann [2]. Compared to the results from [21], our analysis allows for more elegant and improved convergence results when using standard or extended Krylov subspaces. It also gives some insight into a good choice of shifts when using rational Krylov subspaces.

The rest of this paper is organized as follows. In Section 2, a decomposition of the residual into $d + 1$ orthogonal vectors is given for the case of projections into arbitrary subspaces. In Section 3, this result is made more concrete for the case of rational Krylov subspaces, and a

bound is given for such subspaces, based only on the fields of values of A_μ. Section 4 gives bounds on the residuals for three specific cases of rational Krylov subspaces. In Section 5, numerical experiments are given to confirm these bounds.

2 Galerkin projection on tensor product subspaces

To study Galerkin projection methods on tensor product subspaces for the solution of (5), we define the (huge) system matrix $\mathcal{A} \in \mathbb{C}^{n_1 \cdots n_d \times n_1 \cdots n_d}$ as

$$\mathcal{A} := \left(\sum_{\mu=1}^{d} I_{n_d} \otimes \cdots \otimes I_{n_{\mu+1}} \otimes A_\mu \otimes I_{n_{\mu-1}} \otimes \cdots \otimes I_{n_1} \right). \tag{6}$$

Throughout the rest of this paper, we assume that $\mathcal{A}$ is invertible. The matrix $\mathcal{A}$ will never be constructed explicitly. Given matrices $V_\mu \in \mathbb{R}^{n_\mu \times k_\mu}$, $\mu = 1, \ldots, d$, with orthonormal columns, the Galerkin projection method computes an approximation

$$\widetilde{x} = \mathcal{V} y,$$

where $\mathcal{V} = V_d \otimes V_{d-1} \otimes \cdots \otimes V_1$ and $y \in \mathbb{R}^{k_1 k_2 \cdots k_d}$ is the solution of

$$\mathcal{V}^* \mathcal{A} \mathcal{V} y = \mathcal{V}^* c, \tag{7}$$

provided that $\mathcal{V}^* \mathcal{A} \mathcal{V}$ is invertible. Note that the projected matrix $\mathcal{V}^* \mathcal{A} \mathcal{V}$ has the same Kronecker product structure as $\mathcal{A}$:

$$\mathcal{V}^* \mathcal{A} \mathcal{V} = \sum_{\mu=1}^{d} I_{k_d} \otimes \cdots \otimes I_{k_{\mu+1}} \otimes \widetilde{A}_\mu \otimes I_{k_{\mu-1}} \otimes \cdots \otimes I_{k_1},$$

where $\widetilde{A}_\mu = V_\mu^* A_\mu V_\mu$. It is assumed that (7) is uniquely solvable, which is always the case if each $A_\mu + A_\mu^*$ is Hermitian positive definite.

An equivalent characterization of $\widetilde{x} \in \operatorname{span}(\mathcal{V})$ is given by the Galerkin orthogonality condition

$$r \equiv r(\mathcal{V}, A_1, \ldots, A_d, c) := c - \mathcal{A}\widetilde{x} \perp \operatorname{span}(\mathcal{V}). \tag{8}$$

In this section, we will study general properties of the approximate solution $\widetilde{x}$ without making any further assumptions on the choice of V_μ. For this purpose, we will require the following notation:

$$\begin{aligned} \mathcal{V}_\mu &:= I_{k_d} \otimes \cdots \otimes I_{k_{\mu+1}} \otimes V_\mu \otimes I_{k_{\mu-1}} \otimes \cdots \otimes I_{k_1} \\ \overline{\mathcal{V}}_\mu &:= V_d \otimes \cdots \otimes V_{\mu+1} \otimes I_{n_\mu} \otimes V_{\mu-1} \otimes \cdots \otimes V_1. \end{aligned}$$

In particular, this implies $\mathcal{V} = \overline{\mathcal{V}}_\mu \mathcal{V}_\mu$ for every $\mu = 1, \ldots, d$. The following proposition reveals a useful relation between the orthogonal projections $\overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*$ and $\mathcal{V}\mathcal{V}^*$, using the fact these projectors commute.

Proposition 2.1. *Consider $d \geq 2$ matrices $V_1 \in \mathbb{R}^{n_1 \times k_1}, \ldots, V_d \in \mathbb{R}^{n_d \times k_d}$ having orthonormal columns. Then the following equality holds:*

$$\prod_{\mu=1}^{d} (I - \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*) = I - \sum_{\mu=1}^{d} \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^* + (d-1)\mathcal{V}\mathcal{V}^*. \tag{9}$$

Proof. By direct expansion, we obtain

$$\prod_{\mu=1}^{d}(I-\overline{\mathcal{V}}_\mu\overline{\mathcal{V}}_\mu^*)=\sum_{i\in\{0,1\}^d}\prod_{\mu=1}^{d}(-\overline{\mathcal{V}}_\mu\overline{\mathcal{V}}_\mu^*)^{i_\mu}=\sum_{s=0}^{d}\sum_{\substack{i\in\{0,1\}^d\\|i|=s}}\prod_{\mu=1}^{d}(-\overline{\mathcal{V}}_\mu\overline{\mathcal{V}}_\mu^*)^{i_\mu},$$

where $|i|:=i_1+\cdots+i_d$. We now separate the terms for $s=0$ and $s=1$ from the sum and make use of the fact that $\overline{\mathcal{V}}_\mu\overline{\mathcal{V}}_\mu^*\overline{\mathcal{V}}_\nu\overline{\mathcal{V}}_\nu^*=\mathcal{V}\mathcal{V}^*$ for any $\mu\neq\nu$ as well as $\overline{\mathcal{V}}_\mu\overline{\mathcal{V}}_\mu^*\mathcal{V}\mathcal{V}^*=\mathcal{V}\mathcal{V}^*$. This leads to

$$\prod_{\mu=1}^{d}(I-\overline{\mathcal{V}}_\mu\overline{\mathcal{V}}_\mu^*)=I-\sum_{\mu=1}^{d}\overline{\mathcal{V}}_\mu\overline{\mathcal{V}}_\mu^*+\theta\,\mathcal{V}\mathcal{V}^*,$$

with

$$\theta=\sum_{k=2}^{d}(-1)^k\binom{d}{k}=-1+d+\sum_{k=0}^{d}(-1)^k\binom{d}{k}=-1+d+(1-1)^d=d-1,$$

which concludes the proof. □

Based on the result of Proposition 2.1, we will now represent the residual as a sum of orthogonal terms, which can then be bounded individually. An essential part of this representation are terms $\overline{\mathcal{V}}_\mu^*r$, $\mu=1,\ldots,d$, which we will examine in some more detail before stating the main technical results. Consider the term

$$\overline{\mathcal{V}}_\mu^*r=\overline{\mathcal{V}}_\mu^*(c-\mathcal{A}\mathcal{V}y)=\overline{\mathcal{V}}_\mu^*c-\big(\overline{\mathcal{V}}_\mu^*\mathcal{A}\overline{\mathcal{V}}_\mu\big)\mathcal{V}_\mu y,$$

where we have used $\mathcal{V}=\overline{\mathcal{V}}_\mu\mathcal{V}_\mu$ in the last equality. Analogous to $\mathcal{V}^*\mathcal{A}\mathcal{V}$, the matrix $\overline{\mathcal{V}}_\mu^*\mathcal{A}\overline{\mathcal{V}}_\mu$ inherits the Kronecker structure from $\mathcal{A}$, see (6), but the coefficient matrices A_ν are replaced by

$$\widetilde{A}_\nu:=V_\nu^*A_\nu V_\nu,\quad\text{for}\quad\nu\in\{1,\ldots,d\}\setminus\{\mu\}.\tag{10}$$

This allows us to write $\overline{\mathcal{V}}_\mu^*r\equiv\overline{\mathcal{V}}_\mu^*r(\mathcal{V},A_1,\ldots,A_d,c)$ as

$$\overline{\mathcal{V}}_\mu^*r(\mathcal{V},A_1,\ldots,A_d,c)=r(\mathcal{V}_\mu,\widetilde{A}_1,\ldots,\widetilde{A}_{\mu-1},A_\mu,\widetilde{A}_{\mu+1},\ldots,\widetilde{A}_d,\overline{\mathcal{V}}_\mu^*c),$$

where the latter coincides with the residual of the reduced system $\big(\overline{\mathcal{V}}_\mu^*\mathcal{A}\overline{\mathcal{V}}_\mu\big)\big(\mathcal{V}_\mu y\big)=\overline{\mathcal{V}}_\mu^*c$.

Proposition 2.2. *With the notation introduced above, the following statements hold.*

(a) The residual $r(\mathcal{V},A_1,\ldots,A_d,c)=c-\mathcal{A}\widetilde{x}$ can be represented as

$$r(\mathcal{V},A_1,\ldots,A_d,c)=\sum_{\mu=1}^{d}\overline{\mathcal{V}}_\mu r(\mathcal{V}_\mu,\widetilde{A}_1,\ldots,\widetilde{A}_{\mu-1},A_\mu,\widetilde{A}_{\mu+1},\ldots,\widetilde{A}_d,\overline{\mathcal{V}}_\mu^*c)+\widehat{c},\tag{11}$$

where the remainder term $\widehat{c}:=\Big(\prod_{\mu=1}^{d}(I-\overline{\mathcal{V}}_\mu\overline{\mathcal{V}}_\mu^*)\Big)\,c$ vanishes for $c\in\mathrm{span}(\mathcal{V})$.

(b) The vectors $\widehat{c}$ and $\overline{\mathcal{V}}_\mu\overline{\mathcal{V}}_\mu^*r=\overline{\mathcal{V}}_\mu r(\mathcal{V}_\mu,\widetilde{A}_1,\ldots,\widetilde{A}_{\mu-1},A_\mu,\widetilde{A}_{\mu+1},\ldots,\widetilde{A}_d,\overline{\mathcal{V}}_\mu^*c)$ for $\mu=1,\ldots,d$ are mutually orthogonal. In particular, this implies

$$\|r(\mathcal{V},A_1,\ldots,A_d,c)\|_2^2=\sum_{\mu=1}^{d}\big\|r(\mathcal{V}_\mu,\widetilde{A}_1,\ldots,\widetilde{A}_{\mu-1},A_\mu,\widetilde{A}_{\mu+1},\ldots,\widetilde{A}_d,\overline{\mathcal{V}}_\mu^*c)\big\|_2^2+\|\widehat{c}\|_2^2.$$

Proof. *(a)* Multiplying both sides of the equality (9) with the residual $r \equiv r(\mathcal{V}, A_1, \ldots, A_d, c)$ leads to

$$\prod_{\mu=1}^{d}(I - \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*)r = r - \sum_{\mu=1}^{d} \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^* r + (d-1)\mathcal{V}\mathcal{V}^* r = r - \sum_{\mu=1}^{d} \overline{\mathcal{V}}_\mu \big(\overline{\mathcal{V}}_\mu^* r\big),$$

where we used $\mathcal{V}^* r = 0$ from the Galerkin orthogonality condition (8). It therefore remains to show that $\prod_{\mu=1}^{d}(I - \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*)r = \prod_{\mu=1}^{d}(I - \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*)c = \widehat{c}$, or equivalently, that

$$\prod_{\mu=1}^{d}(I - \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*)\mathcal{A}\mathcal{V}y = 0. \tag{12}$$

Inserting $\mathcal{A} = \sum_{\mu=1}^{d} \mathcal{A}_\mu$ leads to

$$\prod_{\mu=1}^{d}(I - \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*)\mathcal{A}\mathcal{V}y = \sum_{\nu=1}^{d} \Big(\prod_{\substack{\mu=1 \\ \mu\neq\nu}}^{d}(I - \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*)\Big)(I - \overline{\mathcal{V}}_\nu \overline{\mathcal{V}}_\nu^*)\mathcal{A}_\nu \overline{\mathcal{V}}_\nu \mathcal{V}_\nu y.$$

Because $\mathcal{A}_\nu$ and $\overline{\mathcal{V}}_\nu$ commute, we have $(I - \overline{\mathcal{V}}_\nu \overline{\mathcal{V}}_\nu^*)\mathcal{A}_\nu \overline{\mathcal{V}}_\nu = (I - \overline{\mathcal{V}}_\nu \overline{\mathcal{V}}_\nu^*)\overline{\mathcal{V}}_\nu \mathcal{A}_\nu = 0$. This shows (12).

(b) The orthogonality relations follow from (a),

$$\big\langle \overline{\mathcal{V}}_\nu \overline{\mathcal{V}}_\nu^* r, \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^* r \big\rangle = r^* \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^* \overline{\mathcal{V}}_\nu \overline{\mathcal{V}}_\nu^* r = r^* \mathcal{V}\mathcal{V}^* r = \|\mathcal{V}^* r\|_2^2 = 0,$$

for $\mu \neq \nu$, and

$$\big\langle \overline{\mathcal{V}}_\nu \overline{\mathcal{V}}_\nu^* r, \widehat{c} \big\rangle = \Big\langle \overline{\mathcal{V}}_\nu \overline{\mathcal{V}}_\nu^* r, \prod_{\mu=1}^{d}(I - \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*)\, c \Big\rangle = c^* \Big(\prod_{\substack{\mu=1 \\ \mu\neq\nu}}^{d}(I - \overline{\mathcal{V}}_\mu \overline{\mathcal{V}}_\mu^*)\Big)(I - \overline{\mathcal{V}}_\nu \overline{\mathcal{V}}_\nu^*)\overline{\mathcal{V}}_\nu \overline{\mathcal{V}}_\nu^* r = 0$$

for any $\nu = 1, \ldots, d$. □

The partial residuals $\overline{\mathcal{V}}_\mu^* r = r(\mathcal{V}_\mu, \widetilde{A}_1, \ldots, \widetilde{A}_{\mu-1}, A_\mu, \widetilde{A}_{\mu+1}, \ldots, \widetilde{A}_d, \overline{\mathcal{V}}_\mu^* c)$ contributing to the overall residual in Proposition 2.2 (b) can also be interpreted more compactly as the residual of a two-dimensional problem. To see this, let us consider the case $\mu = 1$. Then $\overline{\mathcal{V}}_1^* r = \overline{\mathcal{V}}_1^* c - \overline{\mathcal{V}}_1^* \mathcal{A} \overline{\mathcal{V}}_1 \mathcal{V}_1 y$, with the reduced matrix

$$\overline{\mathcal{V}}_1^* \mathcal{A} \overline{\mathcal{V}}_1 = I_{k_d} \otimes \cdots \otimes I_{k_2} \otimes A_1 + \underbrace{\sum_{\mu=2}^{d} I_{k_d} \otimes \cdots \otimes I_{k_{\mu+1}} \otimes \widetilde{A}_\mu \otimes I_{k_{\mu-1}} \otimes \cdots \otimes I_{k_2}}_{=:B_1} \otimes I_{n_1}.$$

Defining $c^{(1)} := \overline{\mathcal{V}}_1^* c = (V_d^* \otimes \cdots \otimes V_2^* \otimes I)c$, we thus arrive at the compact formula

$$\overline{\mathcal{V}}_1^* r = c^{(1)} - (I \otimes A_1 + B_1 \otimes I)(I \otimes V_1)y = r(I \otimes V_1, A_1, B_1, c^{(1)}) =: r^{(1)}, \tag{13}$$

with suitable sizes of the identity matrices. Note that the matrix $I \otimes A_1 + B_1 \otimes I$ represents the Sylvester operator $X \mapsto A_1 X + X B_1^T$. As a consequence of (13), $r^{(1)}$ can be interpreted as the residual obtained from approximating the solution to the linear system $(I \otimes A_1 + B_1 \otimes I)x = c^{(1)}$

(which corresponds to a Sylvester equation) by applying Galerkin projection with the "1D projector" $I \otimes V_1$. As implicitly demonstrated in [2] for Sylvester equations, the convergence of such 1D projections is significantly easier to study than the convergence of the projection as a whole.

For general μ, a similar formula can be shown for $\overline{\mathcal{V}}_\mu^* r$. Let us define the two-dimensional residual

$$r^{(\mu)} := r(I \otimes V_\mu, A_\mu, B_\mu, c^{(\mu)}) = c^{(\mu)} - (I \otimes A_\mu + B_\mu \otimes I)(I \otimes V_\mu)P^{(\mu)} y, \tag{14}$$

where B_μ is now a matrix of size $k_1 \cdots k_{\mu-1} k_{\mu+1} \cdots k_d$. Moreover, both the right-hand side and the residual undergo a permutation with an appropriately chosen permutation matrix $P^{(\mu)}$:

$$r^{(\mu)} = P^{(\mu)}\big(\overline{\mathcal{V}}_\mu^* r\big), \qquad c^{(\mu)} = P^{(\mu)}\big(\overline{\mathcal{V}}_\mu^* c\big). \tag{15}$$

Since a permutation does not change the norm of a vector, the following result follows directly from Proposition 2.2 (b).

Proposition 2.3. *The norm of the residual satisfies*

$$\|r(\mathcal{V}, A_1, \ldots, A_d, c)\|_2^2 = \sum_{\mu=1}^{d} \|r^{(\mu)}\|_2^2 + \|\widehat{c}\|_2^2, \tag{16}$$

with $r^{(\mu)} \equiv r(I \otimes V_\mu, A_\mu, B_\mu, c^{(\mu)})$.

In Section 3, we will derive bounds for the case that the columns of each V_μ form the basis of a rational Krylov subspace with A_μ. To compute these bounds, it is helpful to reformulate (14) in terms of contour integrals in $\mathbb{C}$.

Proposition 2.4. *For the linear system* (5)*, consider a fixed integer* $\mu \in \{1, \ldots, d\}$ *and let* $A_\mu, B_\mu, c^{(\mu)}, V_\mu \in \mathbb{R}^{n_\mu \times k_\mu}$ *(with* $k_\mu < n_\mu$*) be defined as explained above. Suppose that the linear systems*

$$(I \otimes A_\mu + B_\mu \otimes I)x^{(\mu)} = c^{(\mu)} \tag{17}$$

and

$$(I \otimes \widetilde{A}_\mu + B_\mu \otimes I)y^{(\mu)} = (I \otimes V_\mu^*)c^{(\mu)}, \qquad \text{with } \widetilde{A}_\mu = V_\mu^* A_\mu V_\mu,$$

admit unique solutions.

Then, the residual $r^{(\mu)} = c^{(\mu)} - (I \otimes A_\mu + B_\mu \otimes I)\widetilde{x}_\mu$ *of* (17) *for the approximate solution* $\widetilde{x}_\mu = V_\mu y^{(\mu)}$ *can be written as*

$$r^{(\mu)} = \int_{\Gamma_{B_\mu}} \big(zI + B_\mu\big)^{-1} \otimes \Big[I - (zI - A_\mu)V_\mu(zI - \widetilde{A}_\mu)^{-1}V_\mu^*\Big]c^{(\mu)}\, \frac{dz}{2\pi i}.$$

with a compact curve Γ_{B_μ} *which encircles the spectrum of* $-B_\mu$ *once, but does not encircle the spectrum of* A_μ *and* $\widetilde{A}_\mu$.

Proof. For a compact curve Γ_A encircling the spectrum of A once but not encircling the spectrum of $-B$, where A, B are general square matrices, the unique solution of $(I \otimes A + B \otimes I)x = c$ can be written as

$$x = \int_{\Gamma_A} \Big((zI + B)^{-1} \otimes (zI - A)^{-1}\Big)\, c\, \frac{dz}{2\pi i}\ . \tag{18}$$

This is seen by inserting (18) into $(I \otimes A + B \otimes I)x = c$:

$$\begin{aligned}(I \otimes A + B \otimes I)\, x &= \Big(I \otimes (A - zI) + (zI + B) \otimes I\Big)\, x \\ &= -\int_{\Gamma_A} \Big((zI + B)^{-1} \otimes I\Big)\, c\, \frac{dz}{2\pi i} + \int_{\Gamma_A} \Big(I \otimes (zI - A)^{-1}\Big)\, c\, \frac{dz}{2\pi i} \\ &= 0 + c.\end{aligned}$$

Choosing a contour Γ encircling once the spectrum of both A_μ and $\widetilde{A}_\mu$ but not that of $-B_\mu$, we obtain integral representations of the solution vectors $x^{(\mu)}$ and $y^{(\mu)}$. Hence the 1D projection error is

$$x^{(\mu)} - (I \otimes V_\mu)y^{(\mu)} = \int_\Gamma (zI + B_\mu)^{-1} \otimes \Big((zI - A_\mu)^{-1} - V_\mu (zI - \widetilde{A}_\mu)^{-1} V_\mu^*\Big)\, c^{(\mu)}\, \frac{dz}{2\pi i}$$

together with the residual

$$\begin{aligned}r^{(\mu)} &= (I \otimes A_\mu + B_\mu \otimes I)(x^{(\mu)} - (I \otimes V_\mu)y^{(\mu)}) \\ &= -\Big(I \otimes (zI - A_\mu) + (zI + B_\mu) \otimes I\Big)(x^{(\mu)} - (I \otimes V_\mu)y^{(\mu)}) \\ &= -\int_\Gamma (zI + B_\mu)^{-1} \otimes \Big(I - (zI - A_\mu)V_\mu (zI - \widetilde{A}_\mu)^{-1} V_\mu^*\Big)\, c^{(\mu)}\, \frac{dz}{2\pi i} \\ &\quad + \int_\Gamma I \otimes \Big((zI - A_\mu)^{-1} - V_\mu (zI - \widetilde{A}_\mu)^{-1} V_\mu^*\Big)\, c^{(\mu)}\, \frac{dz}{2\pi i}.\end{aligned}$$

Notice that both integrands have the same expansion $(I - V_\mu V_\mu^*)c^{(\mu)}/z + \mathcal{O}(1/z^2)$ at ∞. By the Cauchy formula, we can switch to a contour integral along Γ_{B_μ}, changing the sign of both integrals. Note that the second integral vanishes. □

3 Rational Krylov subspace projection

In this section, we will concentrate on the projection to rational Krylov subspaces. Specifically, we will assume that c can be written as the Kronecker product of d vectors:

$$c = c_d \otimes c_{d-1} \otimes \cdots \otimes c_1, \qquad c_\mu \in \mathbb{R}^{n_\mu}. \tag{19}$$

This is a rather strong assumption that is rarely satisfied in applications. However, for moderate d and a vector c obtained from the discretization of a smooth d-variate function, it is possible to approximate c by a short sum of vectors having the form (19). By superposition, we can reduce this to the situation (19). Alternatively, one could use a method based on block Krylov subspaces.

For a right-hand side of the form (19), the term $c^{(\mu)}$ from the previous section, see (14)–(15), becomes $c^{(\mu)} = \overline{c}_\mu \otimes c_\mu$, with $\overline{c}_\mu = \widetilde{c}_d \otimes \cdots \otimes \widetilde{c}_{\mu+1} \otimes \widetilde{c}_{\mu-1} \otimes \cdots \widetilde{c}_1$ and $\widetilde{c}_\nu = V_\nu^* c_\nu$. As a consequence, our integral formula of Proposition 2.4 for the partial residual $r^{(\mu)}$ involves the expression

$$\Big[I - (zI - A_\mu)V_\mu (zI - \widetilde{A}_\mu)^{-1} V_\mu^*\Big] c_\mu,$$

which coincides with the residual of the OR (Orthogonal Residual) method for the shifted system $(zI - A_\mu)x = c_\mu$. Provided that z is not an element of the field of values $W(A_\mu)$,

one knows to relate this quantity with the corresponding minimal residual following, e.g., the techniques of [12, Thm 6.2.6]. It will be therefore convenient in what follows to suppose that $0 \notin W(\mathcal{A})$. We will also make use of the fact that

$$W(\mathcal{A}) = W(A_1) + \cdots + W(A_d) = W(A_\mu) - W(-B_\mu),$$

see, for instance, [10, Proof of Thm 4.2]. In particular, this shows that the solvability conditions of Proposition 2.4 are satisfied.

Let the columns of the matrix V_μ represent an orthonormal basis of the rational Krylov subspace

$$\mathcal{K}_{k_\mu}^{(\mu)}(A_\mu, c_\mu) := \{R_\mu(A_\mu)c_\mu : R_\mu \in \mathcal{P}_{k_\mu - 1}/Q_\mu\} \quad \text{for } \mu = 1, \ldots, d.$$

Here, $\mathcal{P}_r$ denotes the set of polynomials of degree at most r, and $Q_\mu \in \mathcal{P}_{k_\mu}$ is a fixed polynomial defined as

$$Q_\mu(z) = \prod_{\substack{i=1 \\ z_{\mu,i} \neq \infty}}^{k_\mu} (z - z_{\mu,i}).$$

For example, with $k_\mu = 2$ and $z_{\mu,1} = \infty$, $z_{\mu,2} \in \mathbb{R}$, the associated Krylov subspace is given by $\mathcal{K}_2^{(\mu)}(A_\mu, c_\mu) = \operatorname{span}\{c_\mu, (A_\mu - z_{\mu,2}I)^{-1}c_\mu\}$. We will further fix the first shift to $z_{\mu,1} = \infty$ for each $\mu = 1, \ldots, d$, which ensures that $c_\mu \in \mathcal{R}(V_\mu)$. It follows that $\widehat{c} = 0$ and we simply have

$$\|r\|_2^2 = \|r^{(1)}\|_2^2 + \|r^{(2)}\|_2^2 + \cdots + \|r^{(d)}\|_2^2$$

from Proposition 2.2. The norm of each partial residual $r^{(\mu)}$ can be seen as the solution of a minimization problem, as described in [2]. The following theorem summarizes these results.

Theorem 3.1. *Suppose that $0 \notin W(\mathcal{A})$. Then the partial residual $r^{(\mu)}$ defined in (14) satisfies*

$$\|r^{(\mu)}\|_2 = \min_{R_\mu \in \mathcal{P}_{k_\mu}/Q_\mu} \left[\|R_\mu(A_\mu)c_\mu\|_2 + g_0\|R_\mu(\widetilde{A}_\mu)\widetilde{c}_\mu\|_2\right] \|R_\mu^{-1}(B_\mu)\overline{c}_\mu\|_2, \tag{20}$$

with the constant g_0 defined as $g_0 = \|\mathcal{A}\|_2/\operatorname{dist}(0, W(\mathcal{A}))$.

Proof. The proof will proceed as follows. In a first step, we prove that, for all $R_\mu \in \mathcal{P}_{k_\mu}/Q_\mu$,

$$\|r^{(\mu)}\|_2 \leq \|R_\mu(A_\mu)c_\mu\|_2\|R_\mu^{-1}(B_\mu)\overline{c}_\mu\|_2 + g_0\|R_\mu(\widetilde{A}_\mu)\widetilde{c}_\mu\|_2\|R_\mu^{-1}(B_\mu)\overline{c}_\mu\|_2. \tag{21}$$

In a second step, we will describe a function $R_\mu^G \in \mathcal{P}_{k_\mu}/Q_\mu$ for which equality holds in the above statement.

Using the exactness property for rational Krylov subspaces [2, Lemma 3.2], the following representation has been derived in [2, Lemma 3.3] for the error of the OR method applied to shifted systems:

$$\begin{aligned}
&(zI - A_\mu)^{-1}c_\mu - V_\mu(zI - \widetilde{A}_\mu)^{-1}V_\mu^* c_\mu \\
&= \frac{R_\mu(A_\mu)}{R_\mu(z)}(zI - A_\mu)^{-1}c_\mu - V_\mu\frac{R_\mu(\widetilde{A}_\mu)}{R_\mu(z)}(zI - \widetilde{A}_\mu)^{-1}V_\mu^* c_\mu
\end{aligned}$$

for any $R_\mu \in \mathcal{P}_{k_\mu}/Q_\mu$. Inserting this relation into the integral representation of $r^{(\mu)}$ from Proposition 2.4 yields

$$\begin{aligned} r^{(\mu)} = &\int_{\Gamma_{B_\mu}} \big(zI + B_\mu\big)^{-1}\overline{c}_\mu \otimes (zI - A_\mu)\frac{R_\mu(A_\mu)}{R_\mu(z)})(zI - A_\mu)^{-1}c_\mu \,\frac{dz}{2\pi i} \\ &- \int_{\Gamma_{B_\mu}} \big(zI + B_\mu\big)^{-1}\overline{c}_\mu \otimes (zI - A_\mu)\Big(V_\mu \frac{R_\mu(\widetilde{A}_\mu)}{R_\mu(z)}(zI - \widetilde{A}_\mu)^{-1}V_\mu^* c_\mu\Big)\,\frac{dz}{2\pi i}. \end{aligned}$$

We will call the two integral terms s_1 and s_2, and start by considering s_1. Using the fact that all the terms containing A_μ commute, we find

$$s_1 = \int_{\Gamma_{B_\mu}} \big(zI + B_\mu\big)^{-1}\overline{c}_\mu \otimes \frac{R_\mu(A_\mu)}{R_\mu(z)}c_\mu \,\frac{dz}{2\pi i} = R_\mu^{-1}(-B_\mu)\overline{c}_\mu \otimes R_\mu(A_\mu)c_\mu$$

and thus $r^{(\mu)} = R_\mu^{-1}(-B_\mu)\overline{c}_\mu \otimes R_\mu(A_\mu)c_\mu - s_2$. It has been shown in [2, p. 2443] that

$$s_2 = (I \otimes A_\mu + B_\mu \otimes I)(I \otimes V_\mu)(I \otimes \widetilde{A}_\mu + B_\mu \otimes I)^{-1}\big(R_\mu^{-1}(-B_\mu)\overline{c}_\mu \otimes R_\mu(\widetilde{A}_\mu)\widetilde{c}_\mu\big).$$

Using that $\|I \otimes A_\mu + B_\mu \otimes I\|_2 = \|\mathcal{A}\|_2$ and

$$\|(I \otimes \widetilde{A}_\mu + B_\mu \otimes I)^{-1}\|_2 \leq \frac{1}{\operatorname{dist}(0, W(I \otimes A_\mu + B_\mu \otimes I))} = \frac{1}{\operatorname{dist}(0, W(\mathcal{A}))},$$

we conclude that

$$\|r^{(\mu)}\|_2 \leq \|R_\mu^{-1}(-B_\mu)\overline{c}_\mu \otimes R_\mu(A_\mu)c_\mu\ \|_2 + g_0\,\|R_\mu^{-1}(-B_\mu)\overline{c}_\mu \otimes R_\mu(\widetilde{A}_\mu)\widetilde{c}_\mu\|_2,$$

as claimed in the first part of the statement. To address the second part, we define the rational function

$$R_\mu^G(z) := \frac{\det(zI - \widetilde{A}_\mu)}{Q_\mu(z)}, \quad \mu = 1, \ldots, d.$$

Note that $R_\mu^G(\widetilde{A}_\mu) = 0$, implying $s_2 = 0$ for this choice of R_μ. Therefore

$$r^{(\mu)} = (R_\mu^G)^{-1}(-B_\mu)\overline{c}_\mu \otimes R_\mu^G(A_\mu)c_\mu.$$

This shows equality in (21) for $R_\mu = R_\mu^G$ and therefore completes the proof. □

Up to this point, an exact representation of the residual norm was given. Now, we aim to give a bound on the residual, involving only the fields of values $W(A_\mu)$ for the matrices $A_1, \ldots, A_d$. First, we note that

$$\begin{aligned} \|r^{(\mu)}\|_2 &= \min_{R_\mu \in \mathcal{P}_{k_\mu}/Q_\mu} \Big[\|R_\mu(A_\mu)c_\mu\|_2 + g_0\|R_\mu(\widetilde{A}_\mu)\widetilde{c}_\mu\|_2\Big]\,\|R_\mu^{-1}(B_\mu)\overline{c}_\mu\|_2 \\ &\leq \|c\|_2 \min_{R_\mu \in \mathcal{P}_{k_\mu}/Q_\mu} \Big[\|R_\mu(A_\mu)\|_2 + g_0\|R_\mu(\widetilde{A}_\mu)\|_2\Big]\,\|R_\mu^{-1}(B_\mu)\|_2 \\ &\leq C_{\text{Crouzeix}}\|c\|_2 \min_{R_\mu \in \mathcal{P}_{k_\mu}/Q_\mu} \max_{z \in W(B_\mu)} \Big[\|R_\mu(A_\mu)\|_2 + g_0\|R_\mu(\widetilde{A}_\mu)\|_2\Big]\,|R_\mu^{-1}(z)|. \end{aligned} \tag{22}$$

Here, the constant C_{Crouzeix} is such that

$$\|f(A)\|_2 \le C_{\text{Crouzeix}} \|f\|_{L^\infty(W(A))}$$

for any matrix A and any function f analytic in $W(A)$. Recently, it was proven that $C_{\text{Crouzeix}} \le 11.08$, and it has been conjectured that $C_{\text{Crouzeix}} = 2$ [8].

Let us now define the Green's function $g_{A_\mu}(\cdot, \zeta)$ of $\overline{\mathbb{C}} \setminus W(A_\mu)$, $\mu = 1, \ldots, d$, with pole $\zeta \in \overline{\mathbb{C}}$, and set

$$u_\mu(z) := \exp\left(- \sum_{j=1}^{k_\mu} g_{A_\mu}(z, z_{\mu,j}) \right), \qquad \mu = 1, \ldots, d. \tag{23}$$

Note that $u_\mu(z)$ can be given explicitly for the case when A_μ is Hermitian positive definite, and thus $W(A_\mu) = [\alpha_\mu, \beta_\mu]$ with $0 < \alpha_\mu < \beta_\mu$. It then takes the form

$$u_\mu(z) = \prod_{j=1}^{k_\mu} \left| \frac{\sqrt{\dfrac{z - \beta_\mu}{z - \alpha_\mu} \dfrac{z_{\mu,j} - \alpha_\mu}{z_{\mu,j} - \beta_\mu}} - 1}{\sqrt{\dfrac{z - \beta_\mu}{z - \alpha_\mu} \dfrac{\overline{z}_{\mu,j} - \alpha_\mu}{\overline{z}_{\mu,j} - \beta_\mu}} + 1} \right|. \tag{24}$$

The following theorem is a straightforward extension of [2, Theorem 2.3] from $d = 2$ to general d.

Theorem 3.2. *Suppose that $0 \notin W(\mathcal{A})$ and define*

$$\gamma_\mu := \max \left\{ u_\mu(-z) : z \in \sum_{\substack{\nu=1 \\ \nu \ne \mu}}^{d} W(A_\nu) \right\}, \qquad \mu = 1, \ldots, d. \tag{25}$$

Then, the residual for the rational Galerkin method described above is bounded by

$$\|r\|_2 \le 2\, C_{\text{Crouzeix}}\, (1 + g_0) \|c\|_2 \sqrt{\sum_{\mu=1}^{d} \left(\frac{\gamma_\mu}{1 - \gamma_\mu} \right)^2}. \tag{26}$$

For the case of Hermitian positive definite matrices A_μ, a tighter bound is given by

$$\|r\|_2 \le 2\, \|c\|_2 \sqrt{\frac{\lambda_{\max}(\mathcal{A})}{\lambda_{\min}(\mathcal{A})}} \sqrt{\sum_{\mu=1}^{d} \gamma_\mu^2}, \tag{27}$$

provided that each set of shifts $\{z_{\mu,1}, \ldots, z_{\mu,k_\mu}\}$, $\mu = 1, \ldots, d$, is closed under complex conjugation.

Proof. According to [2, Theorem 3.4] with the convex set $\mathbb{E} = W(A_\mu) \supset W(\widetilde{A}_\mu)$, there exists a function $R^\# \in \mathcal{P}_{k_\mu}/Q_\mu$ such that

$$\|R^\#(A_\mu)\|_2 \le 2, \quad \|R^\#(\widetilde{A}_\mu)\|_2 \le 2, \quad \frac{1}{|R^\#(z)|} \le \frac{u_\mu(z)}{1 - u_\mu(z)} \quad \forall z \notin W(A_\mu).$$

Applying this result to (22) directly leads to

$$\|r^{(\mu)}\|_2 \leq C_{\text{Crouzeix}}\|c\|_2 \max_{z\in W(B_\mu)} 2(1+g_0)\frac{u_\mu(z)}{1-u_\mu(z)}.$$

Note that $t/(1-t)$ is monotonically increasing for $t \in [0,1)$, and that $u_\mu(z) \in [0,1)$ because $g_{A_\mu}(z,\xi) \geq 0$ for all z, ξ and $g_{A_\mu}(z,\xi) > 0$ for $z, \xi \notin W(A_\mu)$. This directly leads to (26).

For the tighter bound in the case of Hermitian positive definite matrices, we refer to the proof of Theorem 2.3 on pages 2447–2448 of [2]. □

Remark 3.3. For $\mathcal{A}$ being Hermitian positive definite, error bounds in the energy norm, $\|x-\widetilde{x}\|_{\mathcal{A}}$ have been given in [21] for the case of standard Krylov subspaces, that is, $z_{\mu,j} \equiv \infty$.

4 Application to specific examples

In this section, we consider several concrete choices of rational Krylov subspaces, and calculate the convergence bounds resulting from Theorem 3.2. We will focus on Hermitian positive definite matrices A_μ, $\mu = 1, \ldots, d$ and consider the following three choices of subspaces:

(i) standard Krylov subspaces (all shifts $z_{\mu,j} = \infty$);

(ii) extended Krylov subspaces ($z_{\mu,j} \in \{0, \infty\}$ alternatingly), also called Krylov plus inverse Krylov (KPIK);

(iii) modified extended Krylov subspaces ($z_{\mu,j} \in \{\sigma, \infty\}$ alternatingly, with $\sigma \in \mathbb{R}$).

Theorem 3.2 applies to all these cases and in the following we will only specify the parameter γ_μ that appears in the residual bound (27). Recall that, for Hermitian positive definite A_μ, the field of values $W(A_\mu) = [\alpha_\mu, \beta_\mu]$ coincides with the convex hull of the spectrum, and hence $0 < \alpha_\mu < \beta_\mu$. Our convergence bounds will be expressed in terms of

$$\kappa_\mu := \frac{\beta_\mu}{\alpha_\mu}, \quad \kappa_{L,\mu} := \frac{\lambda_{\max}(\mathcal{A})}{\lambda_{\max}(\mathcal{A}) - \beta_\mu + \alpha_\mu}, \quad \kappa_{R,\mu} := 1 + \frac{\beta_\mu - \alpha_\mu}{\lambda_{\min}(\mathcal{A})}. \tag{28}$$

In what follows, the quantities $\kappa_{L,\mu}$ and $\kappa_{R,\mu}$ will be referred to as effective condition numbers. It is immediate to check that the inequalities

$$1 < \kappa_{L,\mu} < \kappa_{R,\mu} < \min\{\kappa_\mu, \frac{\lambda_{\max}(\mathcal{A})}{\lambda_{\min}(\mathcal{A})}\}. \tag{29}$$

hold for $d \geq 2$. Using the substitution $f = \sqrt{\frac{z-\beta_\mu}{z-\alpha_\mu}}$ and combining (25) with (24), we obtain in the Hermitian positive definite case the simplified formula

$$\gamma_\mu = \max_{f\in[\sqrt{\kappa_{L,\mu}}, \sqrt{\kappa_{R,\mu}}]} \prod_{j=1}^{k_\mu} \left|\frac{f-\theta_{\mu,j}}{f+\theta_{\mu,j}}\right|, \quad \theta_{\mu,j} := \sqrt{\frac{z_{\mu,j}-\beta_\mu}{z_{\mu,j}-\alpha_\mu}}, \tag{30}$$

from which it becomes clear that if is sufficient to restrict our attention to poles on the negative real axis $z_{\mu,j} \in [-\infty, 0]$ or, equivalently, $\theta_{\mu,j} \in [1, \sqrt{\kappa_\mu}]$.

We start by giving a bound for standard Krylov subspaces.

Corollary 4.1. *Let A_μ be Hermitian positive definite matrices with $W(A_\mu) = [\alpha_\mu, \beta_\mu]$ for $\mu = 1, \dots, d$. Applying the Galerkin projection method with standard Krylov subspaces*

$$\operatorname{span}\{c_\mu, A_\mu c_\mu, \dots, A_\mu^{k_\mu - 1} c_\mu\},$$

the convergence factor γ_μ satisfies

$$\gamma_\mu = \left(\frac{\sqrt{\kappa_{R,\mu}} - 1}{\sqrt{\kappa_{R,\mu}} + 1} \right)^{k_\mu}, \qquad \text{with} \qquad \kappa_{R,\mu} = 1 + \frac{\beta_\mu - \alpha_\mu}{\lambda_{\min}(\mathcal{A})}.$$

Proof. The choice of standard Krylov subspaces corresponds to $z_{\mu,j} \equiv \infty$, and thus $\theta_{\mu,j} = 1$ for $j = 1, \dots, k_\mu$. Inserting this into (30) and taking into account that $(1, \infty) \ni f \mapsto (f-1)/(f+1)$ is positive and increasing, the assertion follows. □

Note that the convergence factor of Corollary 4.1 matches the one obtained in [21, Corollary 4.4]. However, for the case of extended Krylov subspaces, the approach from this paper gives a substantially better factor compared to [21, Lemma 6.1], especially for $d > 2$.

Corollary 4.2. *Let A_μ be Hermitian positive definite matrices with $W(A_\mu) = [\alpha_\mu, \beta_\mu]$ for $\mu = 1, \dots, d$. Applying the Galerkin projection method with extended Krylov subspaces*

$$\operatorname{span}\{c_\mu, A_\mu^{-1} c_\mu, A_\mu c_\mu, \dots, A_\mu^{k_\mu/2 - 1} c_\mu, A_\mu^{-k_\mu/2} c_\mu\}$$

for even k_μ, the convergence factor γ_μ satisfies

$$\gamma_\mu \le \left(\frac{\sqrt[4]{\kappa_\mu} - 1}{\sqrt[4]{\kappa_\mu} + 1} \right)^{k_\mu}, \qquad \text{with} \qquad \kappa_\mu = \frac{\beta_\mu}{\alpha_\mu}.$$

More specifically, we have the equalities

$$\gamma_\mu = \begin{cases} \left(\frac{\sqrt{\kappa_{L,\mu}} - 1}{\sqrt{\kappa_{L,\mu}} + 1} \frac{\sqrt{\kappa_\mu / \kappa_{L,\mu}} - 1}{\sqrt{\kappa_\mu / \kappa_{L,\mu}} + 1} \right)^{k_\mu/2}, & \text{for } \lambda_{\max}(\mathcal{A}) < \beta_\mu + \sqrt{\beta_\mu \alpha_\mu}, \\ \left(\frac{\sqrt{\kappa_{R,\mu}} - 1}{\sqrt{\kappa_{R,\mu}} + 1} \frac{\sqrt{\kappa_\mu / \kappa_{R,\mu}} - 1}{\sqrt{\kappa_\mu / \kappa_{R,\mu}} + 1} \right)^{k_\mu/2}, & \text{for } \lambda_{\min}(\mathcal{A}) > \alpha_\mu + \sqrt{\beta_\mu \alpha_\mu}, \\ \left(\frac{\sqrt[4]{\kappa_\mu} - 1}{\sqrt[4]{\kappa_\mu} + 1} \right)^{k_\mu}, & \text{otherwise.} \end{cases}$$

Proof. The choice of extended Krylov subspaces corresponds to $z_{\mu,j} = \infty$ (hence, $\theta_{\mu,j} = 1$) for odd j, and $z_{\mu,j} = 0$ (hence, $\theta_{\mu,j} = \sqrt{\beta_\mu / \alpha_\mu} = \sqrt{\kappa_\mu}$) for even j. We need to find

$$\gamma_\mu^{2/k_\mu} = \max_{f \in [\sqrt{\kappa_{L,\mu}}, \sqrt{\kappa_{R,\mu}}]} g(f), \qquad g(f) = \frac{f - 1}{f + 1} \cdot \frac{\sqrt{\kappa_\mu} - f}{f + \sqrt{\kappa_\mu}}.$$

where we have used inequalities (29). Simple elementary calculus shows that the maximum of g on the interval $[1, \sqrt{\kappa_\mu}]$ is attained at $f^* = \sqrt[4]{\kappa_\mu}$, and that

$$\gamma_\mu^{2/k_\mu} = \max_{f \in [\sqrt{\kappa_{L,\mu}}, \sqrt{\kappa_{R,\mu}}]} g(f) \le \max_{f \in [1, \sqrt{\kappa_\mu}]} g(f) = g(f^*) = \left(\frac{\sqrt[4]{\kappa_\mu} - 1}{\sqrt[4]{\kappa_\mu} + 1} \right)^2.$$

This shows the first statement. If $\sqrt[4]{\kappa_\mu} \in [\sqrt{\kappa_{L,\mu}}, \sqrt{\kappa_{R,\mu}}]$, this inequality becomes an equality. Otherwise, the maximum is given by $g(\sqrt{\kappa_{L,\mu}})$ if $f^* < \sqrt{\kappa_{L,\mu}}$ and by $g(\sqrt{\kappa_{R,\mu}})$ if $f^* > \sqrt{\kappa_{R,\mu}}$, as claimed in the second statement. □

Finally, we get the following bound for a rational Krylov subspace method with two shifts, ∞ and $\sigma \in \mathbb{R} \setminus W(A_\mu)$, the latter possibly depending on μ. This is a generalization of extended Krylov subspaces, where $\sigma = 0$, and we will see that it allows for faster convergence, while requiring the same number of linear system solves in the method.

Corollary 4.3. *Let A_μ be Hermitian positive definite matrices with $W(A_\mu) = [\alpha_\mu, \beta_\mu]$ for $\mu = 1, \ldots, d$. Applying the Galerkin projection method with rational Krylov subspaces of the form*

$$\operatorname{span}\{c_\mu, (A_\mu - \sigma I)^{-1} c_\mu, A_\mu c_\mu, \ldots, A_\mu^{k_\mu/2-1} c_\mu, (A_\mu - \sigma I)^{-k_\mu/2} c_\mu\}$$

for even k_μ, the convergence factor γ_μ satisfies

$$\gamma_\mu \leq \max\left\{ \left(\frac{\sqrt{\theta}-1}{\sqrt{\theta}+1}\right)^{k_\mu}, \left(\frac{\sqrt{\kappa_{R,\mu}}-1}{\sqrt{\kappa_{R,\mu}}+1} \cdot \left|\frac{\sqrt{\kappa_{R,\mu}}-\theta}{\sqrt{\kappa_{R,\mu}}+\theta}\right|\right)^{k_\mu/2} \right\}, \tag{31}$$

with $\theta = \sqrt{\frac{\sigma-\beta_\mu}{\sigma-\alpha_\mu}}$ and $\kappa_{R,\mu} = 1 + \frac{\beta_\mu - \alpha_\mu}{\lambda_{\min}(\mathcal{A})}$.

The shift minimizing this bound is given by

$$\sigma_{\text{opt}} = \alpha_\mu(\theta_{\text{opt}}^2 - \kappa_\mu)/(\theta_{\text{opt}}^2 - 1), \tag{32}$$

with $\theta_{\text{opt}} := s^{-1}(\sqrt{\kappa_\mu})$ for $s(\theta) := \left[(\theta+1)^2 + (\theta-1)\sqrt{\theta^2+6\theta+1}\right]/(4\sqrt{\theta})$. When using this shift σ_{opt}, we have

$$\gamma_\mu \leq \left(\frac{\sqrt[6]{4\kappa_{R,\mu}}-1}{\sqrt[6]{4\kappa_{R,\mu}}+1}\right)^{k_\mu}. \tag{33}$$

Proof. The choice of rational Krylov subspaces corresponds to $z_{\mu,j} = \infty$ (hence, $\theta_{\mu,j} = 1$) for odd j, and $z_{\mu,j} = \sigma$ (hence, $\theta_{\mu,j} = \sqrt{\frac{\sigma-\beta_\mu}{\sigma-\alpha_\mu}} =: \theta$) for even j. We need to find

$$\gamma_\mu^{2/k_\mu} = \max_{f \in [\sqrt{\kappa_{L,\mu}}, \sqrt{\kappa_{R,\mu}}]} g_\theta(f), \qquad g_\theta(f) = \frac{f-1}{f+1} \cdot \left|\frac{f-\theta}{f+\theta}\right|.$$

The function $g_\theta(f)$ is continuously differentiable in all points except 1 and θ, see Figure 1. It has a unique local maximum at $\sqrt{\theta}$. Combined with (29), this shows the bound (31):

$$\gamma_\mu^{2/k_\mu} = \max_{f \in [\sqrt{\kappa_{L,\mu}}, \sqrt{\kappa_{R,\mu}}]} g_\theta(f) \leq \max_{f \in [1, \sqrt{\kappa_{R,\mu}}]} g_\theta(f) = \max\left\{g_\theta(\sqrt{\theta}), g_\theta(\sqrt{\kappa_{R,\mu}})\right\}.$$

To prove (32), we need to find $\theta_{\text{opt}} \in (0, \infty)$ which minimizes the function $h(\theta)$ given by

$$\begin{aligned} h(\theta) &:= \max\left\{g_\theta(\sqrt{\theta}), g_\theta(\sqrt{\kappa_{R,\mu}})\right\} = \max\{h_1(\theta), h_2(\theta)\} \\ &\text{with } h_1(\theta) := \left(\frac{\sqrt{\theta}-1}{\sqrt{\theta}+1}\right)^2, \quad h_2(\theta) := \frac{\sqrt{\kappa_{R,\mu}}-1}{\sqrt{\kappa_{R,\mu}}+1}\left|\frac{\sqrt{\kappa_{R,\mu}}-\theta}{\sqrt{\kappa_{R,\mu}}+\theta}\right|. \end{aligned}$$

The function $h_1(\theta)$ is zero in 1, and monotonically increases with $|\theta - 1|$. Similarly, $h_2(\theta)$ is zero in $\sqrt{\kappa_{R,\mu}}$, and monotonically increases with $|\theta - \sqrt{\kappa_{R,\mu}}|$. As both h_1 and h_2 are monotonously decreasing on $(0, 1]$, we clearly have $\theta_{\text{opt}} \geq 1$. Similarly, we find that $\theta_{\text{opt}} \leq \sqrt{\kappa_{R,\mu}}$. Therefore, θ_{opt} is uniquely defined by the relation

$$h_1(\theta_{\text{opt}}) = h_2(\theta_{\text{opt}}), \quad \theta_{\text{opt}} \in [1, \sqrt{\kappa_{R,\mu}}].$$

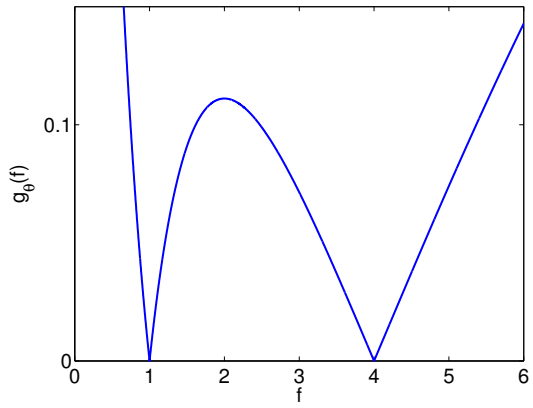

Figure 1: Function $g_\theta(f)$ from the proof of Corollary 4.3 for $\theta = 4$.

Inserting both functions leads to the condition $\sqrt{\kappa_{R,\mu}} = s(\theta_{\text{opt}})$, with the bijective function $s : [1, \sqrt{\kappa_{R,\mu}}] \to [1, s(\sqrt{\kappa_{R,\mu}})]$ defined in the statement of the corollary.

To show (33), it remains to prove

$$\gamma_\mu^{2/k_\mu} = \left(\frac{\sqrt{\theta_{\text{opt}}} - 1}{\sqrt{\theta_{\text{opt}}} + 1}\right)^2 \leq \left(\frac{\sqrt[6]{4\kappa_{R,\mu}} - 1}{\sqrt[6]{4\kappa_{R,\mu}} + 1}\right)^2,$$

which is the case if and only if $\sqrt{\theta_{\text{opt}}} \leq \sqrt[6]{4\kappa_{R,\mu}}$, or, equivalently, $\theta_{\text{opt}}^{3/2}/2 \leq \sqrt{\kappa_{R,\mu}} = s(\theta_{\text{opt}})$. This is easily seen from

$$2\theta^2 \leq (\theta+1)^2 + (\theta-1)^2 \leq (\theta+1)^2 + (\theta-1)\sqrt{\theta^2+6\theta+1}, \qquad \forall q \geq 0. \qquad \square$$

Remark 4.4. The convergence behavior tends to improve when d, the number of dimensions, increases. We illustrate this by considering the case $A_\mu \equiv A$ with $W(A) = [\alpha, \beta]$ and $k_\mu \equiv k$. Then the bound obtained from Corollary 4.1 for standard Krylov subspaces becomes

$$\|r\|_2 \leq 2\sqrt{d}\,\|c\|_2 \sqrt{\frac{\lambda_{\max}(\mathcal{A})}{\lambda_{\min}(\mathcal{A})}} \left(\frac{\sqrt{\kappa_R} - 1}{\sqrt{\kappa_R} + 1}\right)^k,$$

with $\kappa_R = 1 + \frac{\beta-\alpha}{d\alpha} < \kappa = \beta/\alpha$. Clearly, the effective condition number κ_R decreases as d increases. A similar statement holds for the bound obtained from Corollary 4.3 for rational Krylov subspaces with shifts ∞ and σ_{opt}. However, for the case of extended Krylov subspaces, the bound generally only improves if the condition $\lambda_{\min}(\mathcal{A}) > \alpha_\mu + \sqrt{\beta_\mu \alpha_\mu}$ holds, that is, if $\sqrt{\kappa} < d - 1$. Since it is unlikely that such a condition is satisfied, it follows that choosing an optimal shift is essential for achieving improved results in higher dimensions.

Remark 4.5. For different but similar Blaschke-type rational functions, the quantity γ_k of (30) has been determined in [3, Section 6] for various configurations of poles, including those discussed in this section. If it is affordable to solve shifted systems with A_μ for several different shifts, one could imagine to choose $z_{\mu,1} = \infty$ and then cyclic shifts

$$\begin{array}{rclrclcrcl}
z_{\mu,2} & = & \sigma_1, & z_{\mu,3} & = & \sigma_2, & \ldots, & z_{\mu,p+1} & = & \sigma_p, \\
z_{\mu,p+2} & = & \sigma_1, & z_{\mu,p+3} & = & \sigma_2, & \ldots, & z_{\mu,2p+1} & = & \sigma_p, \\
\ldots & & & & & & & & &
\end{array}$$

for some fixed p. Note that $\sigma_1, \ldots, \sigma_p$ will depend on μ. The so-called ADI optimal shifts are obtained by solving the third Zolotarev problem

$$\min_{\theta_{\mu,2},\ldots,\theta_{\mu,p+1}} \max_{f\in[\sqrt{\kappa_{L,\mu}},\sqrt{\kappa_{R,\mu}}]} \prod_{j=2}^{p+1} \left| \frac{f-\theta_{\mu,j}}{f+\theta_{\mu,j}} \right|,$$

see, e.g., [25, Sec. 2.1] and [11]. For instance, in the case $p = 1$ we get the convergence rate $(\sqrt{\kappa_{R,\mu}/\kappa_{L,\mu}} - 1)/(\sqrt{\kappa_{R,\mu}/\kappa_{L,\mu}} + 1)$ for the parameter $\theta_{\mu,2} = \sqrt{\kappa_{R,\mu}\kappa_{L,\mu}}$.

5 Numerical Experiments

In this section, numerical experiments are presented to illustrate the theoretical results on the convergence behavior obtained in Section 4 for Hermitian positive definite matrices.

Remark 5.1. The implementation of the Galerkin projection method requires the solution of the linear system

$$\left(\sum_{\mu=1}^{d} I_{k_d} \otimes \cdots \otimes I_{k_{\mu+1}} \otimes \widetilde{A}_\mu \otimes I_{k_{\mu-1}} \otimes \cdots \otimes I_{k_1} \right) y = \widetilde{c}_d \otimes \cdots \otimes \widetilde{c}_1,$$

at least once in the final step of the method. This system has size $k_1 k_2 \cdots k_d$, which makes a direct approach infeasible for larger d, even when $k_\mu \ll n_\mu$. Instead, we use an approximate solution based on exponential sums:

$$\widetilde{y} = \sum_{j=1}^{R} \omega_j \exp(-\alpha_j \widetilde{A}_d)\widetilde{c}_d \otimes \cdots \otimes \exp(-\alpha_j \widetilde{A}_1)\widetilde{c}_1,$$

see [14, 21]. The vector $\widetilde{y}$ is only stored implicitly through the vectors $\exp(-\alpha_j \widetilde{A}_\mu)\widetilde{c}_\mu$. The coefficients α_j, ω_j, $j = 1, \ldots, R$ are chosen to approximate $1/\xi$ by the exponential sum $\sum_{j=1}^{R} \omega_j e^{-\alpha_j \xi}$ [7, 15, 16]. As the approximation error decreases exponentially depending with R, even moderate values of R lead to an accuracy at the level of about 10^{-8}, which is used in all experiments below.

Experiment 1: Standard Krylov subspaces. In the first example, we set all matrices $A_\mu \equiv A$, where $A \in \mathbb{R}^{200\times 200}$ is the standard finite difference discretization of the one-dimensional Laplace operator on $[0, 1]$ with homogeneous Dirichlet boundary conditions. The right-hand side is set to $c = b \otimes b \otimes \cdots \otimes b$, where the entries of b are random numbers from $\mathcal{N}(0, 1)$ and b is normalized such that $\|b\|_2 = 1$. The obtained results are shown in the left plot of Figure 2. For $d = 2$, observed and predicted convergence rates match quite well. However, as d increases, the observed convergence becomes faster than predicted by Corollary 4.1. This phenomenon seems to depend strongly on the choice of b. To demonstrate this, let

$$b = UD^{-1}e/\|UD^{-1}e\|_2, \tag{34}$$

where $A = UDU^T$ is the eigenvalue decomposition of A and $e = (1, \ldots, 1)^T$. Then the results shown in the right plot of Figure 2 reveal that the observed and predicted convergence rates match very well even for large d. Moreover, both the convergence rate does not improve visibly as d increases, which confirms the observation made in Remark 4.4.

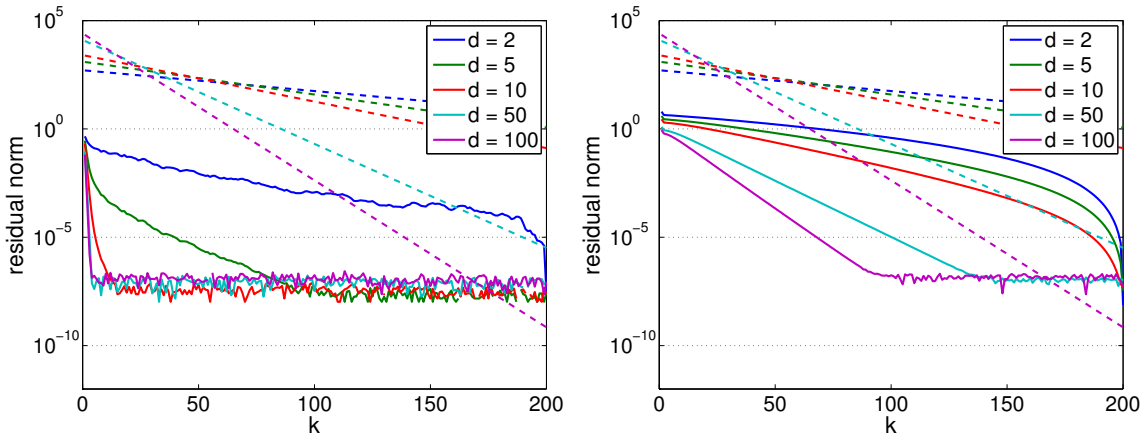

Figure 2: Norm of residuals for standard Krylov subspace method. Solid lines correspond to computed residuals and dashed lines correspond to the bound obtained from Corollary 4.1. **Left:** Randomly chosen right-hand side. **Right:** Particularly constructed right-hand side (34).

Experiment 2: Extended Krylov subspaces. We repeat Experiment 1 for extended Krylov subspaces. The results for a random right-hand side are shown in the left plot of Figure 3. Even for $d = 2$, the observed convergence is much faster than predicted by Corollary 4.2. Moreover, this phenomenon does not disappear with the right-hand side (34).

Inspired by a numerical experiment in [20, Example 4.2], we also consider a diagonal matrix $A \in \mathbb{R}^{n\times n}$, $n = 10^4$, with diagonal elements

$$a_{jj} = \frac{1}{2\sqrt{\kappa}}\Big((\kappa+1) + (\kappa-1)\cos\frac{\pi(j-1)}{n-1}\Big), \quad j = 1,\dots,n, \tag{35}$$

where $\kappa = 2500 = \kappa(A)$. The right-hand side vector is set to $c = b \otimes b \otimes \cdots \otimes b$, with $b = A^{-1}e$. The obtained results are shown in the right plot of Figure 3. Again, the observed and predicted convergence rates match very well even for large d.

Experiment 3: Rational Krylov subspaces with shifts ∞ and σ_{opt}. We repeat Experiment 2 for rational Krylov subspaces with shifts ∞ and σ_{opt} defined in Corollary 4.3. The results are displayed in Figure 4. The observed and predicted convergence rates match well. However, in contrast to Experiment 2, the convergence rates improve as d increases.

Experiment 4: Rational Krylov subspaces with optimal ADI shifts. In this experiment, we apply rational Krylov subspaces with optimal ADI shifts $\sigma_1, \cdots, \sigma_p$ to the matrix (35) and the corresponding right-hand side. The obtained results for $p = 1$ and $p = 3$ optimal ADI shifts are shown in the left and the right plot of Figure 5, respectively.

6 Conclusions

In this paper, we have provided an analysis of Galerkin projection onto tensor products of subspaces for linear systems that can be regarded as d-dimensional analogues of the Sylvester

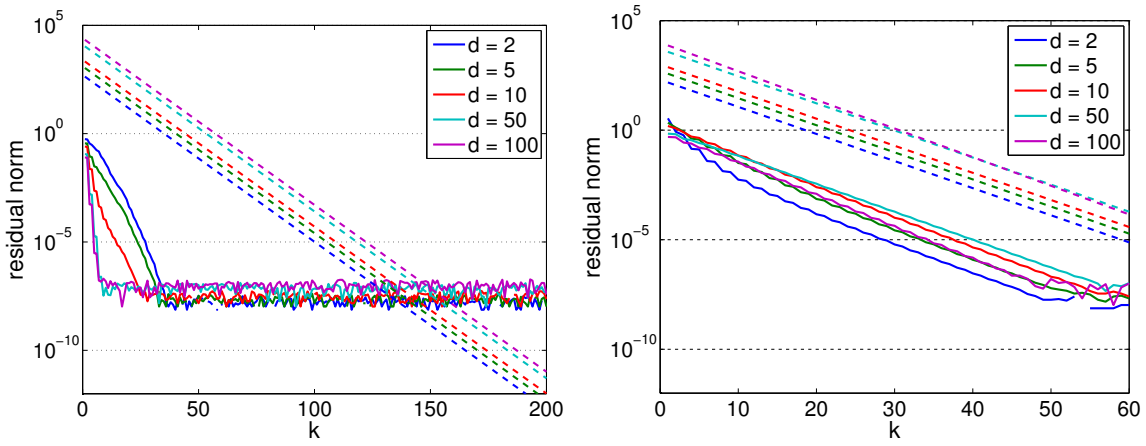

Figure 3: Norm of residuals for extended Krylov subspace method. Solid lines correspond to computed residuals and dashed lines correspond to the bound obtained from Corollary 4.2. **Left:** Discrete Laplace. **Right:** Diagonal matrix (35).

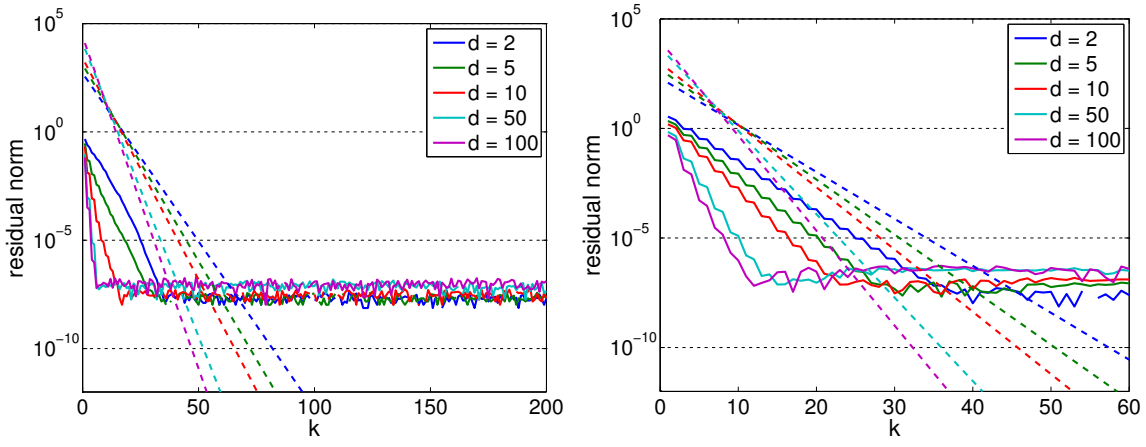

Figure 4: Norm of residuals for rational Krylov subspace method with shifts ∞ and σ_{opt}. Solid lines correspond to computed residuals and dashed lines correspond to the bound obtained from Corollary 4.3. **Left:** Discrete Laplace. **Right:** Diagonal matrix. (35).

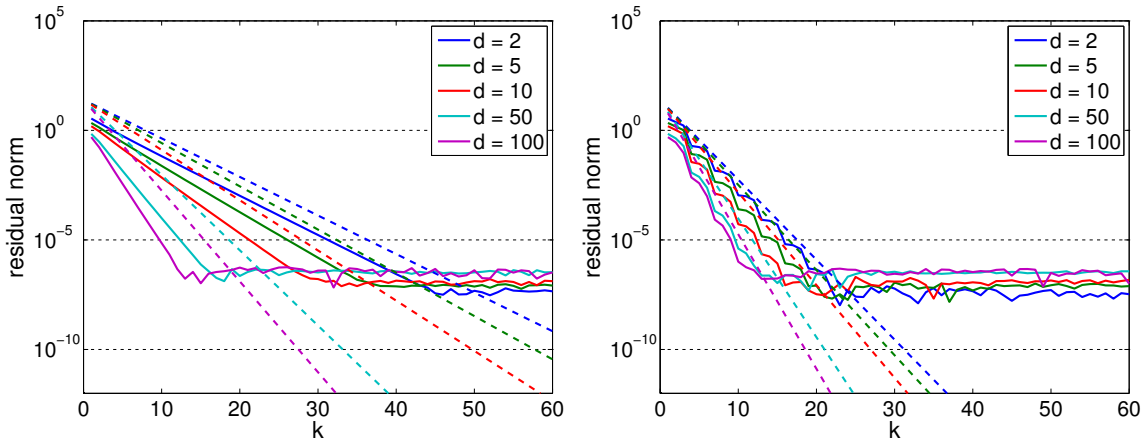

Figure 5: Norm of residuals for rational Krylov subspace method with optimal ADI shifts. Solid lines correspond to computed residuals and dashed lines correspond to the bound from Theorem 3.2. **Left:** One shift. **Right:** Three shifts.

equation. The orthogonal decomposition of the residual derived in Proposition 2.2 is the key observation and allows to decompose the residual into a sum of residuals for simpler 1D projections, see Proposition 2.3. When applied to polynomial and rational Krylov subspaces, this decomposition allows to derive a priori error estimates via extremal problems for univariate rational functions. This contrasts with the analysis in [21], which involves multivariate approximation problems. Numerical experiments demonstrate that the convergence rates derived in this paper are sharp. Moreover, our bounds allow for a better understanding of the dependence of the convergence rates on the shifts in the rational Krylov subspaces. In turn, this can be used to optimize the choice of shifts. Interestingly, the convergence of rational Krylov subspace methods with optimized shifts improves as the dimension d increases.

References

[1] A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. SIAM Publications, Philadelphia, PA, 2005.

[2] B. Beckermann. An error analysis for rational Galerkin projection applied to the Sylvester equation. *SIAM J. Numer. Anal.*, 49(6):2430–2450, 2011.

[3] B. Beckermann and L. Reichel. Error estimates and evaluation of matrix functions via the Faber transform. *SIAM J. Numer. Anal.*, 47(5):3849–3883, 2009.

[4] P. Benner and T. Breiten. Low rank methods for a class of generalized Lyapunov equations and related issues. Technical report, MPI Magdeburg, 2012. Available from `http://www.mpi-magdeburg.mpg.de/preprints/`.

[5] P. Benner, R. Byers, E. S. Quintana-Ortí, and G. Quintana-Ortí. Solving algebraic Riccati equations on parallel computers using Newton's method with exact line search. *Parallel Comput.*, 26(10):1345–1368, 2000.

[6] P. Benner, R.-C. Li, and N. Truhar. On the ADI method for Sylvester equations. *J. Comput. Appl. Math.*, 233(4):1035–1045, 2009.

[7] D. Braess and W. Hackbusch. Approximation of $1/x$ by exponential sums in $[1, \infty)$. *IMA J. Numer. Anal.*, 25(4):685–697, 2005.

[8] M. Crouzeix. Numerical range and functional calculus in hilbert space. *J. Funct. Anal.*, 244(2):668–690, 2007.

[9] J. W. Demmel. Three methods for refining estimates of invariant subspaces. *Computing*, 38:43–57, 1987.

[10] 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(5):1875–1898, 2011.

[11] N. S. Ellner and E. L. Wachspress. Alternating direction implicit iteration for systems with complex spectra. *SIAM J. Numer. Anal.*, 28(3):859–870, 1991.

[12] O. G. Ernst. Minimal and orthogonal residual methods and their generalizations for solving linear operator equations. Habilitation thesis, TU Bergakademie Freiberg, 2001.

[13] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[14] L. Grasedyck. Existence and computation of low Kronecker-rank approximations for large linear systems of tensor product structure. *Computing*, 72(3-4):247–265, 2004.

[15] W. Hackbusch. Approximation of $1/x$ by exponential sums. Available from `http://www.mis.mpg.de/scicomp/EXP_SUM/1_x/tabelle`. Retrieved August 2008.

[16] W. Hackbusch. Entwicklungen nach Exponentialsummen. Technical Report, Max-Planck-Institut für Mathematik in den Naturwissenschaften, 2009. Revised version. See `http://www.mis.mpg.de/preprints/tr/report-0405.pdf`.

[17] I. M. Jaimoukha and E. M. Kasenally. Krylov subspace methods for solving large Lyapunov equations. *SIAM J. Numer. Anal.*, 31:227–251, 1994.

[18] K. Jbilou. ADI preconditioned Krylov methods for large Lyapunov matrix equations. *Linear Algebra Appl.*, 432(10):2473–2485, 2010.

[19] B. N. Khoromskij and Ch. Schwab. Tensor-structured Galerkin approximation of parametric and stochastic elliptic PDEs. *SIAM J. Sci. Comput.*, 33(1):364–385, 2011.

[20] L. Knizhnerman and V. Simoncini. Convergence analysis of the extended Krylov subspace method for the Lyapunov equation. *Numer. Math.*, 118(3):567–586, 2011.

[21] D. Kressner and C. Tobler. Krylov subspace methods for linear systems with tensor product structure. *SIAM J. Matrix Anal. Appl.*, 31(4):1688–1714, 2010.

[22] D. Kressner and C. Tobler. Preconditioned low-rank methods for high-dimensional elliptic PDE eigenvalue problems. *Comput. Methods Appl. Math.*, 11(3):363–381, 2011.

[23] T. Mach and J. Saak. Towards an ADI iteration for tensor structured equations. Technical report, MPI Magdeburg, 2012. Available from `www.mpi-magdeburg.mpg.de/preprints/2011/12/`.

[24] Y. Saad. Numerical solution of large Lyapunov equations. In *Signal processing, scattering and operator theory, and numerical methods (Amsterdam, 1989)*, volume 5 of *Progr. Systems Control Theory*, pages 503–511. Birkhäuser Boston, Boston, MA, 1990.

[25] J. Sabino. *Solution of Large-Scale Lyapunov Equations via the Block Modified Smith Method*. PhD thesis, Rice University, 2006.

[26] V. Simoncini. A new iterative method for solving large-scale Lyapunov matrix equations. *SIAM J. Sci. Comput.*, 29(3):1268–1288, 2007.

[27] V. Simoncini and V. Druskin. Convergence analysis of projection methods for the numerical solution of large Lyapunov equations. *SIAM J. Numer. Anal.*, 47:828–843, 2009.

[28] N. Truhar, Z. Tomljanović, and R.-C. Li. Analysis of the solution of the Sylvester equation using low-rank ADI with exact shifts. *Systems Control Lett.*, 59(3-4):248–257, 2010.