Residual-based iterations for the generalized Lyapunov equation

Tobias Breiten¹ · Emil Ringh²

Received: 14 September 2018 / Accepted: 5 June 2019 / Published online: 15 June 2019
© The Author(s) 2019

Abstract

This paper treats iterative solution methods for the generalized Lyapunov equation. Specifically, a residual-based generalized rational-Krylov-type subspace is proposed. Furthermore, the existing theoretical justification for the alternating linear scheme (ALS) is extended from the stable Lyapunov equation to the stable generalized Lyapunov equation. Further insights are gained by connecting the energy-norm minimization in ALS to the theory of H₂-optimality of an associated bilinear control system. Moreover it is shown that the ALS-based iteration can be understood as iteratively constructing rank-1 model reduction subspaces for bilinear control systems associated with the residual. Similar to the ALS-based iteration, the fixed-point iteration can also be seen as a residual-based method minimizing an upper bound of the associated energy norm.

Keywords Generalized Lyapunov equation · H₂-optimal model reduction · Bilinear control systems · Alternating linear scheme · Projection methods · Matrix equations · Rational Krylov

Mathematics Subject Classification 65F10 · 58E25 · 65F30 · 65F35

1 Introduction

This paper concerns iterative ways to compute approximate solutions to what has become known as the *generalized Lyapunov equation*,

$$\mathcal{L}(X) + \Pi(X) + BB^T = 0, \quad (1)$$

Communicated by Michiel E. Hochstenbach.

✉ Emil Ringh
eringh@kth.se

Extended author information available on the last page of the article

where $X \in \mathbb{R}^{n \times n}$ is unknown, $B \in \mathbb{R}^{n \times r}$ is given, and the operators $\mathcal{L}, \Pi : \mathbb{R}^{n \times n} \rightarrow \mathbb{R}^{n \times n}$ are defined as

$$\mathcal{L}(X) := AX + XA^T \quad (2)$$

$$\Pi(X) := \sum_{i=1}^m N_i X N_i^T, \quad (3)$$

with $A, N_i \in \mathbb{R}^{n \times n}$ for $i = 1, \dots, m$ given. The operator $\mathcal{L}$ is commonly known as the *Lyapunov operator*, and Π is sometimes called a *correction*. We further assume that A is *stable*, i.e., A has all its eigenvalues in the left-half plane, which implies that $\mathcal{L}$ is invertible [23, Theorem 4.4.6]. Moreover, we assume that $\rho(\mathcal{L}^{-1} \Pi) < 1$, where ρ denotes the (operator) spectral radius. The assumption on the spectral radius implies that (1) has a unique solution [24, Theorem 2.1]. Furthermore, the definition of Π in (3) implies that it is non-negative, in the sense that $\Pi(X)$ is positive semidefinite when X is positive semidefinite. Thus one can assert that, for all positive definite right-hand-sides, the unique solution X is indeed positive definite [9, Theorem 3.9] [12, Theorem 4.1]. Under these assumptions we prove that the alternating linear scheme (ALS) presented by Kressner and Sirković in [25] computes search directions which at each step fulfill a first order necessary condition for being $\mathcal{H}_2$ -optimal. Moreover, we show an equivalence between the bilinear iterative rational Krylov (BIRKA) method [5, 19] and the ALS-iteration for the generalized Lyapunov equation. The established equivalence leads to that the ALS-iteration for the generalized Lyapunov equation can be understood as iteratively computing model reduction spaces of dimension 1 for a sequence of bilinear control systems associated with the residual of the generalized Lyapunov equation (Sect. 3). We also present a residual-based generalized rational-Krylov-type subspace adapted for solving the generalized Lyapunov equation (Sect. 5). A further result regards the fixed-point iteration, a residual-based iteration which we show minimizes an upper bound of the energy-norm (Sect. 4).

The standard Lyapunov equation, $AX + XA^T + BB^T = 0$, has been well studied for a long time and considerable research effort has been, and is still, put into finding efficient algorithms for computing the solution and approximations thereof. For large and sparse problems it is typical to look for low-rank approximations since algorithms can be adapted to exploit the low-rank format, reducing computational effort and storage requirement. One such algorithm is the Riemannian optimization method from [40] which computes a low-rank approximation by minimizing an associated cost function over the manifold of rank- k matrices, where $k \ll n$. The Lyapunov equation has a close connection to control theory. Hence methods such as the iterative rational Krylov algorithm (IRKA) [18, 21], which computes subspaces for locally $\mathcal{H}_2$ -optimal reduced order linear systems, provide good approximation spaces for low-rank approximations. Related research is presented in a series of papers [13–15], where Druskin and co-authors develop a strategy to choose shifts for the rational Krylov subspace for efficient subspace reduction when solving PDEs [13, 14], as well as for model reduction of linear single-input-single-output (SISO) systems and solutions to Lyapunov equations [15]. Instead of computing full spaces iteratively with a method such as IRKA, the idea is to construct an infinite sequence with asymptotically optimal con-

vergence speed [13]. Then the subspace can be dynamically extended as needed, until required precision is achieved. The idea is also further developed by using tangential directions, proving especially useful for situations where the right-hand-side is not of particularly low rank [16], e.g., multiple-input-multiple-output (MIMO) systems. For a more complete overview of results and techniques for Lyapunov equations see the review article [38].

The generalized Lyapunov equation has received increased attention over the last decade. Results on low-rank approximability have emerged [6,24]. More precisely, similarly to the standard Lyapunov equation one can in certain cases when the right-hand-side B is of low rank, $r \ll n$, expect the singular values of the solution to decay rapidly even for the generalized Lyapunov equation. The result [6, Theorem 1] is applicable when the matrices N_i for $i = 1, \dots, m$ have low rank, and the result [24, Theorem 2] when $\rho(\mathcal{L}^{-1} \Pi) < 1$. Examples of algorithms exploiting low-rank structures are a Bilinear ADI method [6], specializations of Krylov methods for matrix equations [24], as well as greedy low-rank methods [25], and exploitations of the fixed-point iteration [37]. Through the connection with bilinear control systems there is an extension of IRKA, known as bilinear iterative rational Krylov (BIRKA) [5,19]. There are also methods based on Lyapunov and ADI-preconditioned GMRES and BICGSTAB [12], and in general for problems with tensor product structure [26]. In the context of stochastic steady-state diffusion equations, rational Krylov subspace methods for generalized Sylvester equations have also been analyzed in [32]. The suggested search space is based on a union of rational Krylov subspaces, as well as combinations of rational functions, generated by the coefficient matrices defining the generalized Sylvester operator. We also mention that for the case when the correction Π has low operator-rank, there is a specialization of the Sherman-Morrison-Woodbury formula to the linear matrix equation; see [33] or [12, Section 3]. The result has been exploited in works such as [6,28,34]. Recently, the generalized Lyapunov equation has also been considered on an infinite-dimensional Hilbert space, see [4]. In particular, the authors show ([4, Proposition 1.1]) that the Gramians solving the generalized linear operator equations can be approximated by truncated Gramians that are associated to a sequence of standard operator Lyapunov equations.

2 Preliminaries

2.1 Generalized matrix equations and approximations

We recall some basic definitions and results that will be used later in the paper. In general we will think of $\hat{X}_k \in \mathbb{R}^{n \times n}$ as an approximation of the solution to (1), where k is typically an iteration count. Connected with an approximation $\hat{X}_k$ is the corresponding *error*

$$X_k^e := X - \hat{X}_k, \quad (4)$$

where X is the exact solution to (1), and the *residual*,

$$\mathcal{R}_k := \mathcal{L}(\hat{X}_k) + \Pi(\hat{X}_k) + BB^T. \quad (5)$$

The goal is to find an $\hat{X}_k$ such that $\|X_k^e\|$ is small for some norm. Since $\|X_k^e\|$ is usually not available in practice, one instead aims at a small residual norm $\|\mathcal{R}_k\|$. To discuss projection methods and make the results precise, we make the following (standard) definition.

Definition 1 (*The Galerkin approximation*) Let $\mathcal{K}_k \subseteq \mathbb{R}^n$ be an $n_k \leq n$ dimensional subspace for $k = 0, 1, \dots$, and let $V_k \in \mathbb{R}^{n \times n_k}$ be a matrix containing an orthogonal basis of $\mathcal{K}_k$. We call $\hat{X}_k$ the *Galerkin approximation* to (1), in $\mathcal{K}_k$, if $\hat{X}_k = V_k Y_k V_k^T$ and Y_k is determined by the condition

$$V_k^T \left(\mathcal{L}(\hat{X}_k) + \Pi(\hat{X}_k) + BB^T \right) V_k = 0. \quad (6)$$

For the generalized Lyapunov equation there are certain sufficient conditions for the Galerkin approximation to exist and be unique, e.g., the criteria in [9, Theorem 3.9], [12, Theorem 4.1] or [24, Proposition 3.2]. Related to the Galerkin approximation there is also the (standard) definition of the Galerkin residual.

Definition 2 (*The Galerkin residual*) We call $\mathcal{R}_k$ from (5) the *Galerkin residual* if $\hat{X}_k$ is the Galerkin approximation.

The condition (6) is known as both the *projected problem* and the *Galerkin condition*, and it states that $V_k^T \mathcal{R}_k V_k = 0$ for the Galerkin residual. Some of the results and arguments presented below are valid for a (generic) residual and others, more specialized, only for the Galerkin residual. However, it will be clear from context and the Galerkin residual will always be referenced as such.

The following fundamental result from linear algebra will be important for us. The specialization for the Lyapunov equation was presented already by Smith in [39]. For generalized matrix equations cf. [12, Section 4.2], and [25, Algorithm 2]; and an analogy for the algebraic Riccati equation in [29].

Proposition 3 (Residual equation) *Consider Eq. (1). Let $\hat{X}_k$ be an approximation of the solution, $\mathcal{R}_k$ be the residual (5), and X_k^e be the error (4). Then*

$$\mathcal{L}(X_k^e) + \Pi(X_k^e) + \mathcal{R}_k = 0.$$

One strategy for computing updates to the current approximation is to compute approximations of the error. Proposition 3 allows such iterations by connecting the error with the known, or computable, quantities $\mathcal{L}$, Π and $\mathcal{R}_k$. The idea is well established in the literature and is, e.g., analogous to the defect correction method [29] and the RADI method [8] for the algebraic Riccati equation, as well as the iterative improvement [20, Section 3.5.3] for a general linear system. For future reference we also need the following basic definition.

Definition 4 (*Symmetric generalized Lyapunov equation*) The generalized Lyapunov equation

$$AX + XA^T + \sum_{i=1}^m N_i X N_i^T + BB^T = 0, \quad (7)$$

is called *symmetric* if $A = A^T$ and $N_i = N_i^T$ for $i = 1, \dots, m$.

2.2 Bilinear systems

We recall some control theoretic concepts for *bilinear control systems* of the form

$$\Sigma \begin{cases} \dot{x}(t) = Ax(t) + \sum_{i=1}^m N_i x(t) w_i(t) + Bu(t) \\ y(t) = Cx(t), \end{cases} \quad (8)$$

with $A, N_i \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times r}$ and $C \in \mathbb{R}^{p \times n}$ and control inputs $u(t) \in \mathbb{R}^r$ and $w(t) \in \mathbb{R}^m$.

Remark 5 Note that the bilinear system (8) differs from the notation frequently used in the literature, e.g., [1,2,5,9,12,19,41]. The formulation (8) is convenient since it allows for $m \neq r$. However, the system Σ can be put into the usual form by considering the input vector $[w(t)^T, u(t)^T]^T$, adding m zero-columns to the beginning of B , i.e., $[0, B]$, and considering the matrices $N_{m+1} = 0, \dots, N_{m+r} = 0$. The system Σ can also be compared to systems from applications, e.g., [30, Equation (2)].

As in [2], for a MIMO bilinear system (8), we define the controllability and observability Gramians as follows.

Definition 6 ([2], *Bilinear Gramians*) Consider the bilinear system (8) and let A be stable. Moreover, let $P_1(t_1) := e^{At_1} B$, $P_j(t_1, \dots, t_j) := e^{At_j} [N_1 P_{j-1}, \dots, N_m P_{j-1}]$ for $j = 2, 3, \dots$, $Q_1(t_1) := C e^{At_1}$, and $Q_j(t_1, \dots, t_j) := [N_1^T Q_{j-1}^T, \dots, N_m^T Q_{j-1}^T]^T e^{At_j}$ for $j = 2, 3, \dots$. We define the controllability and observability Gramian respectively as

$$P := \sum_{j=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} P_j P_j^T dt_1 \cdots dt_j$$

$$Q := \sum_{j=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} Q_j^T Q_j dt_1 \cdots dt_j.$$

It is possible that the generalized Gramians from Definition 6 do not exist; sufficient conditions are given in, e.g., [41, Theorem 2]. However, if the Gramians exist we also

know that they satisfy the following matrix equations

$$\begin{aligned} AP + PA^T + \sum_{i=1}^m N_i P N_i^T + BB^T &= 0 \\ A^T Q + QA + \sum_{i=1}^m N_i^T Q N_i + C^T C &= 0. \end{aligned} \quad (9)$$

In relation to the generalized controllability and observability Gramians, one might also define a generalized cross Gramian similar to the SISO case discussed in [36]. Consider the symmetric generalized Lyapunov equation (7), and an approximation $\hat{X}_k$ with related error X_k^e , and residual $\mathcal{R}_k$. One can easily verify that for the auxiliary system

$$\Sigma^e = \begin{cases} \dot{x}(t) = Ax(t) + \sum_{i=1}^m N_i x(t) w_i(t) + B_{\mathcal{R}_k} u(t) \\ y(t) = C_{\mathcal{R}_k} x(t), \end{cases}$$

with $B_{\mathcal{R}_k} = US^{1/2}$ and $C_{\mathcal{R}_k} = S^{1/2}V^T$, where $\mathcal{R}_k = USV^T$ is a singular value decomposition of $\mathcal{R}_k$, the associated cross Gramian coincides with the error X_k^e . In the special case where $\mathcal{R}_k = \mathcal{R}_k^T \succeq 0$, it is easy to show the following result.

Proposition 7 *Consider a symmetric generalized Lyapunov equation (7). Let $\hat{X}_k$ be an approximation such that the residual $\mathcal{R}_k = \mathcal{R}_k^T \succeq 0$. Then one can choose $B_{\mathcal{R}_k} = C_{\mathcal{R}_k}^T$ and the error X_k^e is the controllability and observability Gramian of the system Σ^e .*

For what follows, we recall the definition of the $\mathcal{H}_2$ -norm for bilinear systems that was introduced by Zhang and Lam in [41].

Definition 8 ([41], *Bilinear $\mathcal{H}_2$ -norm*) Consider the bilinear system Σ from (8). We define the $\mathcal{H}_2$ -norm of Σ as

$$\|\Sigma\|_{\mathcal{H}_2}^2 := \text{trace} \left(\sum_{j=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \sum_{\ell_1, \dots, \ell_{j-1}=1}^m \sum_{\ell_j=1}^r g_j^{(\ell_1, \dots, \ell_j)} (g_j^{(\ell_1, \dots, \ell_j)})^T ds_1 \cdots ds_j \right),$$

with $g_j^{(\ell_1, \dots, \ell_j)}(s_1, \dots, s_j) := Ce^{As_j} N_{\ell_1} e^{As_{j-1}} N_{\ell_2} \cdots e^{As_1} b_{\ell_j}$.

It has been shown [41, Theorem 6] that if the Gramians from Definition 6 exist, then

$$\|\Sigma\|_{\mathcal{H}_2}^2 = \text{trace} \left(CPC^T \right) = \text{trace} \left(B^T QB \right).$$

3 ALS and $\mathcal{H}_2$ -optimal model reduction for bilinear systems

In this section, we discuss a low-rank approximation method proposed by Kressner and Sirković in [25]. We show that several results can be generalized from the case of the standard Lyapunov equation to the more general form (1). Moreover, we show that in the symmetric case the method allows for an interpretation in terms of $\mathcal{H}_2$ -optimal model reduction for bilinear control systems. With this in mind, we assume that we have a symmetric generalized Lyapunov equation (7). If additionally $A \prec 0$ and $\rho(\mathcal{L}^{-1}\Pi) < 1$, then the operator $\mathcal{M}(X) := -\mathcal{L}(X) - \Pi(X)$ is positive definite and allows us to define a weighted inner product via

$$\begin{aligned} \langle \cdot, \cdot \rangle_{\mathcal{M}} : \mathbb{R}^{n \times n} \times \mathbb{R}^{n \times n} &\rightarrow \mathbb{R} \\ \langle X, Y \rangle_{\mathcal{M}} = \langle X, \mathcal{M}(Y) \rangle &= \text{trace} \left(X^T \mathcal{M}(Y) \right), \end{aligned}$$

with a corresponding induced $\mathcal{M}$ -norm, also known as *energy norm*,

$$\|X\|_{\mathcal{M}}^2 = \langle X, X \rangle_{\mathcal{M}}.$$

3.1 ALS for the generalized Lyapunov equation

In [25], it is suggested to construct iterative approximations $\hat{X}_k$ by rank-1 updates that are locally optimal with respect to the $\mathcal{M}$ -norm. To be more precise, assume that X is a solution to the symmetric Lyapunov equation (7), i.e., $AX + XA + \sum_{i=1}^m N_i X N_i + BB^T = 0$. Given an approximation $\hat{X}_k$, we consider the minimization problem

$$\min_{v, w \in \mathbb{R}^n} \|X - \hat{X}_k - vw^T\|_{\mathcal{M}}^2 = \langle X - \hat{X}_k - vw^T, X - \hat{X}_k - vw^T \rangle_{\mathcal{M}}.$$

Since the minimization involves the constant term $\|X - \hat{X}_k\|_{\mathcal{M}}^2$, it suffices to focus on

$$J(v, w) := \langle vw^T, vw^T \rangle_{\mathcal{M}} - 2 \text{trace} \left(wv^T \mathcal{R}_k \right), \quad (10)$$

where $\mathcal{R}_k$ is the current residual, i.e., (5). Locally optimal vectors v_k and w_k are then (approximately) determined via an *alternating linear scheme* (ALS). The main step is to fix one of the two vectors, e.g., v and then minimize the strictly convex objective function to obtain an update for w . A pseudocode is given in Algorithm 1.

In view of Proposition 3 the ALS-based approach for computing new subspace extensions can be seen as searching for an approximation to X_k^c of the form $v_k w_k^T$ by iterating $(\mathcal{L}(v_k w_k^T) + \Pi(v_k w_k^T) + \mathcal{R}_k)w_k = 0$ when determining v_k and $v_k^T(\mathcal{L}(v_k w_k^T) + \Pi(v_k w_k^T) + \mathcal{R}_k) = 0$ when determining w_k . This is to say that the error is approximated by a rank-1 matrix, and at convergence this would result in the new residual, $\mathcal{R}_{k+1}$, being left-orthogonal to v_k and right-orthogonal to w_k . In the symmetric case, local minimizers of (10) are necessarily symmetric positive semidefinite. This yields the following extension of [25, Lemma 2.3].

Algorithm 1: ALS for the generalized Lyapunov equation [25, Algorithm 1]**input** : System: $A, N_1, \dots, N_m \in \mathbb{R}^{n \times n}$ $\mathcal{R}_k \in \mathbb{R}^{n \times n}$, tol Initial guess: $v, w \in \mathbb{R}^n$ **output**: Approximation vectors: $v^a, w^a \in \mathbb{R}^n$ **while** Change in $(\frac{v^T A v}{\|v\|^2} + \frac{w^T A^T w}{\|w\|^2})/2$ larger than tol **do**

- 1 Normalize $w = w/\|w\|$
- 2 $\hat{A}_1 = A + I(w^T A w) + \sum_{i=1}^m N_i(w^T N_i w)$
- 3 Solve $\hat{A}_1 v = -\mathcal{R}_k w$
- 4 Normalize $v = v/\|v\|$
- 5 $\hat{A}_2 = A + I(v^T A v) + \sum_{i=1}^m (v^T N_i v) N_i$
- 6 Solve $\hat{A}_2 w = -\mathcal{R}_k^T v$

end7 Normalize such that $\|w\| = \|v\|$ **return** $v^a = v, w^a = w$

Lemma 9 Consider the symmetric generalized Lyapunov equation (7) and assume that $A \prec 0$, $\rho(\mathcal{L}^{-1} \Pi) < 1$, and $\mathcal{R}_k = \mathcal{R}_k^T \geq 0$. Let J be as in (10). Then every local minimum (v_*, w_*) of J is such that $v_* w_*^T$ is symmetric positive semidefinite.

Proof The proof naturally follows along the lines of [25, Lemma 2.3], and hence without loss of generality we assume that $v_* \neq 0$, $w_* \neq 0$, and $\|v_*\| = \|w_*\|$. Thus $v_* w_*^T$ is positive semidefinite if and only if $v_* = w_*$. The proof is by contradiction and we assume that $v_* \neq w_*$. Then, since $J(v_*, w)$ is strictly convex in w and $J(v, w_*)$ is strictly convex in v , it follows that

$$2J(v_*, w_*) < J(v_*, v_*) + J(w_*, w_*).$$

Simplifying the left-hand-side we get

$$2J(v_*, w_*) = -2v_*^T \mathcal{L}(v_* w_*^T) w_* - 2v_*^T \Pi(v_* w_*^T) w_* - 4v_*^T \mathcal{R}_k w_*,$$

and similarly the right-hand-side gives

$$\begin{aligned} J(v_*, v_*) + J(w_*, w_*) &= -v_*^T \mathcal{L}(v_* v_*^T) v_* - v_*^T \Pi(v_* v_*^T) v_* - 2v_*^T \mathcal{R}_k v_* \\ &\quad - w_*^T \mathcal{L}(w_* w_*^T) w_* - w_*^T \Pi(w_* w_*^T) w_* - 2w_*^T \mathcal{R}_k w_*. \end{aligned}$$

Collecting the terms involving the $\mathcal{L}$ -operator we observe that

$$\begin{aligned} &-2v_*^T \mathcal{L}(v_* w_*^T) w_* + v_*^T \mathcal{L}(v_* v_*^T) v_* + w_*^T \mathcal{L}(w_* w_*^T) w_* \\ &= 2(v_*^T v_*)(w_*^T A w_* - v_*^T A v_*) + 2w_*^T w_*(v_*^T A v_* - w_*^T A w_*) = 0, \end{aligned}$$

Thus by collecting the terms involving the Π -operator to the left, and the residual to the right, the inequality reduces to

$$-2v_*^T \Pi(v_* w_*^T) w_* + v_*^T \Pi(v_* v_*^T) v_* + w_*^T \Pi(w_* w_*^T) w_* < -2(v_* - w_*)^T \mathcal{R}_k (v_* - w_*).$$

The argument is now concluded by showing that

$$-2v_*^T \Pi(v_* w_*^T) w_* + v_*^T \Pi(v_* v_*^T) v_* + w_*^T \Pi(w_* w_*^T) w_* \geq 0,$$

since this implies that $-2(v_* - w_*)^T \mathcal{R}_k(v_* - w_*) > 0$ in contradiction to the positive semidefiniteness of $\mathcal{R}_k$. We can without loss of generality consider $m = 1$, i.e., only one N -matrix, since the following argument can be applied to all terms in the sum independently. We observe that

$$-2v_*^T N v_* w_*^T N w_* + v_*^T N v_* v_*^T N v_* + w_*^T N w_* w_*^T N w_* = (v_*^T N v_* - w_*^T N w_*)^2 \geq 0,$$

which shows the desired inequality and thus concludes the proof. $\square$

Algorithm 1 and the argument in Lemma 9 are based on a residual. However, if $\hat{X}_k = 0$, then $\mathcal{R}_k = B B^T$, and hence the result is applicable directly to any symmetric generalized Lyapunov equation. The focus on the residual in the previous results is natural since it leads to the following extension of [25, Theorem 2.4] to the case of the symmetric generalized Lyapunov equation.

Theorem 10 *Consider the symmetric generalized Lyapunov equation (7) with the additional assumptions that $A \prec 0$ and $\rho(\mathcal{L}^{-1} \Pi) < 1$. Moreover, consider the sequence of approximations constructed as*

$$\begin{aligned} \hat{X}_0 &= 0 \\ \hat{X}_{k+1} &= \hat{X}_k + v_{k+1} v_{k+1}^T, \quad k = 0, 1, \dots, \end{aligned} \quad (11)$$

where v_{k+1} is a locally optimal vector computed with ALS (Algorithm 1). Then $\mathcal{R}_{k+1} = \mathcal{R}_{k+1}^T \geq 0$ for all $k \geq -1$.

Proof We show the assertion by induction. It clearly holds that $\mathcal{R}_0 = \mathcal{R}_0^T \geq 0$. Now assume that this is the case for some k . From Lemma 9 the local minimizers of (10) are symmetric and hence $\hat{X}_{k+1}$ is reasonably defined in (11). Moreover, since $\hat{X}_{k+1}$ and the operators in (1) are symmetric it follows that $\mathcal{R}_{k+1}$ is symmetric. Thus what is left to show is that $\mathcal{R}_{k+1} \geq 0$, which is true if and only if $y^T \mathcal{R}_{k+1} y \geq 0$ for all $y \in \mathbb{R}^n$. Hence take an arbitrary $y \in \mathbb{R}^n$ and consider $y^T \mathcal{R}_{k+1} y$. We derive properties similar to [25, equations (12)–(14)]:

Since (v_{k+1}, v_{k+1}) is a (local) minimizer of $J(v, w)$, it also follows that v_{k+1} is a (global) minimizer of the (convex) cost function

$$J_w(v) := J(v, w) = \langle v w^T, v w^T \rangle_{\mathcal{M}} - 2 \operatorname{trace}(w v^T \mathcal{R}_k),$$

where $w = v_{k+1}$. Note that the gradient $\nabla_v J_w$ of J_w with respect to v is given by

$$(\nabla_v J_w)_i = 2 \langle e_i w^T, v w^T \rangle_{\mathcal{M}} - 2 e_i^T \mathcal{R}_k w.$$

Due to the optimality of v_{k+1} with respect to $J_{v_{k+1}}$, first order optimality conditions then imply that

$$-Av_{k+1}v_{k+1}^T v_{k+1} - v_{k+1}v_{k+1}^T Av_{k+1} - \sum_{i=1}^m N_i v_{k+1}v_{k+1}^T N_i v_{k+1} = \mathcal{R}_k v_{k+1}. \quad (12)$$

Striking this equality with v_{k+1}^T from the left implies that

$$2v_{k+1}^T Av_{k+1} \|v_{k+1}\|^2 = -v_{k+1}^T \mathcal{R}_k v_{k+1} - \sum_{i=1}^m (v_{k+1}^T N_i v_{k+1})^2. \quad (13)$$

Based on (12) and its transpose, and by exploiting the symmetry of the involved matrices, we can write the residual as

$$\begin{aligned} y^T \mathcal{R}_{k+1} y &= y^T \mathcal{R}_k y + y^T \left(Av_{k+1}v_{k+1}^T + v_{k+1}v_{k+1}^T A + \sum_{i=1}^m N_i v_{k+1}v_{k+1}^T N_i \right) y \\ &= y^T \mathcal{R}_k y + \sum_{i=1}^m y^T N_i v_{k+1}v_{k+1}^T N_i y + \frac{1}{\|v_{k+1}\|^2} y^T (U_{k+1} + U_{k+1}^T) y, \end{aligned}$$

with $U_{k+1} := -\mathcal{R}_k v_{k+1}v_{k+1}^T - (v_{k+1}^T A v_{k+1})v_{k+1}v_{k+1}^T - \sum_{i=1}^m N_i v_{k+1}z_{i,k+1}^T$, and where $z_{i,k+1} := (v_{k+1}^T N_i v_{k+1})v_{k+1}$. We rearrange, identify the term $-2(v_{k+1}^T A v_{k+1})v_{k+1}v_{k+1}^T$ and insert (13) to get

$$\begin{aligned} y^T \mathcal{R}_{k+1} y &= y^T \mathcal{R}_k y \\ &\quad + \frac{1}{\|v_{k+1}\|^2} y^T \left(-\mathcal{R}_k v_{k+1}v_{k+1}^T - v_{k+1}v_{k+1}^T \mathcal{R}_k + \frac{1}{\|v_{k+1}\|^2} v_{k+1}^T \mathcal{R}_k v_{k+1}v_{k+1}^T \right) y \\ &\quad + \frac{1}{\|v_{k+1}\|^2} y^T \left(\sum_{i=1}^m N_i v_{k+1}v_{k+1}^T N_i \|v_{k+1}\|^2 + \frac{1}{\|v_{k+1}\|^2} \sum_{i=1}^m z_{i,k+1} z_{i,k+1}^T \right) y \\ &\quad + \frac{1}{\|v_{k+1}\|^2} y^T \left(-\sum_{i=1}^m N_i v_{k+1} z_{i,k+1}^T - \sum_{i=1}^m z_{i,k+1} v_{k+1}^T N_i \right) y \\ &= y^T \mathcal{R}_k y + \frac{1}{\|v_{k+1}\|^2} \left(-2(y^T \mathcal{R}_k v_{k+1})(v_{k+1}^T y) + \frac{1}{\|v_{k+1}\|^2} (v_{k+1}^T \mathcal{R}_k v_{k+1})(v_{k+1}^T y)^2 \right) \\ &\quad + \frac{1}{\|v_{k+1}\|^2} \left(\sum_{i=1}^m (y^T N_i v_{k+1})^2 \|v_{k+1}\|^2 + \frac{1}{\|v_{k+1}\|^2} (z_{i,k+1}^T y)^2 - 2(y^T N_i v_{k+1})(z_{i,k+1}^T y) \right) \\ &= (y - v_{k+1} \frac{v_{k+1}^T y}{\|v_{k+1}\|^2})^T \mathcal{R}_k (y - v_{k+1} \frac{v_{k+1}^T y}{\|v_{k+1}\|^2}) \\ &\quad + \frac{1}{\|v_{k+1}\|^2} \sum_{i=1}^m \left(\|v_{k+1}\| (y^T N_i v_{k+1}) - \frac{1}{\|v_{k+1}\|} (z_{i,k+1}^T y) \right)^2 \geq 0. \end{aligned}$$

This asserts the inductive step and hence concludes the proof. $\square$

Corollary 11 *The iteration (11) produces an increasing sequence of approximations $0 = \hat{X}_0 \preceq \hat{X}_1 \preceq \dots \preceq X$.*

3.2 $\mathcal{H}_2$ -optimal model reduction for symmetric state space systems

For the standard Lyapunov equation it has been shown, in [7], that minimization of the energy norm induced by the Lyapunov operator (see [40]) is related to $\mathcal{H}_2$ -optimal model reduction for linear control systems. We show that a similar conclusion can be drawn for the minimization of the cost functional (10) and $\mathcal{H}_2$ -optimal model reduction for symmetric bilinear control systems. In this regard, let us briefly summarize the most important concepts from bilinear model reduction. Given a bilinear system Σ as in (8) with $\dim(\Sigma) = n$, the goal of model reduction is to construct a surrogate model $\hat{\Sigma}$ of the form

$$\hat{\Sigma}: \begin{cases} \dot{\hat{x}}(t) = \hat{A}\hat{x}(t) + \sum_{i=1}^m \hat{N}_i \hat{x}(t) w_i(t) + \hat{B}u(t) \\ \hat{y}(t) = \hat{C}\hat{x}(t), \end{cases} \quad (14)$$

with $\hat{A}, \hat{N}_i \in \mathbb{R}^{k \times k}$, $\hat{B} \in \mathbb{R}^{k \times r}$, $\hat{C} \in \mathbb{R}^{r \times k}$ and control inputs $u(t) \in \mathbb{R}^r$ and $w(t) \in \mathbb{R}^m$. In particular, the reduced system should satisfy $k \ll n$ and $\hat{y}(t) \approx y(t)$ in some norm. In [5, 19] the authors have suggested an algorithm, BIRKA, that iteratively tries to compute a reduced model satisfying first order necessary conditions for $\mathcal{H}_2$ -optimality, for the bilinear $\mathcal{H}_2$ -norm as defined in Definition 8. A corresponding pseudocode is given in Algorithm 2.

Algorithm 2: BIRKA [5, Algorithm 2] and [19, Algorithm 5]

input : System: $A, N_1, \dots, N_m \in \mathbb{R}^{n \times n}$ $B \in \mathbb{R}^{n \times r}$, $C \in \mathbb{R}^{r \times n}$, tol
 Initial guess: $\tilde{V}, \tilde{W} \in \mathbb{R}^{n \times k}$
output: Approximation spaces: $\tilde{V}^{\text{opt}}, \tilde{W}^{\text{opt}} \in \mathbb{R}^{n \times k}$ satisfying necessary conditions for $\mathcal{H}_2$ -optimality

while *Change in eigenvalues of $\tilde{A}$ larger than tol* **do**

- 1 | Update guess $\tilde{A} = (\tilde{W}^T \tilde{V})^{-1} \tilde{W}^T A \tilde{V}$,
 $\tilde{N}_1 = (\tilde{W}^T \tilde{V})^{-1} \tilde{W}^T N_1 \tilde{V}, \dots, \tilde{N}_m = (\tilde{W}^T \tilde{V})^{-1} \tilde{W}^T N_m \tilde{V}$, $\tilde{B} = (\tilde{W}^T \tilde{V})^{-1} \tilde{W}^T B$, $\tilde{C} = C \tilde{V}$
- 2 | Decompose $R \tilde{A} R^{-1} = \tilde{A}$
- 3 | Compute $\hat{B} = R^{-1} \tilde{B}$, $\hat{C} = \tilde{C} R$, $\hat{N}_1 = R^{-1} \tilde{N}_1 R, \dots, \hat{N}_m = R^{-1} \tilde{N}_m R$
- 4 | Solve $\tilde{V} \tilde{A} + A \tilde{V} + \sum_{i=1}^m N_i \tilde{V} \tilde{N}_i^T + B \hat{B}^T = 0$ for $\tilde{V}$
- 5 | Solve $\tilde{W} \tilde{A} + A^T \tilde{W} + \sum_{i=1}^m N_i^T \tilde{W} \tilde{N}_i + C^T \hat{C} = 0$ for $\tilde{W}$
- 6 | Orthogonalize $\tilde{V} = \text{orth}(\tilde{V})$, $\tilde{W} = \text{orth}(\tilde{W})$

end
return $\tilde{V}^{\text{opt}} = \tilde{V}$, $\tilde{W}^{\text{opt}} = \tilde{W}$

To establish the connection we introduce the following generalizations of the operator $\mathcal{M}$:

$$\begin{aligned}\widetilde{\mathcal{M}}: \mathbb{R}^{n \times k} &\rightarrow \mathbb{R}^{n \times k}, \quad \widetilde{\mathcal{M}}(X) := -AX - X\hat{A}^T - \sum_{i=1}^m N_i X \hat{N}_i^T, \\ \widehat{\mathcal{M}}: \mathbb{R}^{k \times k} &\rightarrow \mathbb{R}^{k \times k}, \quad \widehat{\mathcal{M}}(X) := -\hat{A}X - X\hat{A}^T - \sum_{i=1}^m \hat{N}_i X \hat{N}_i^T,\end{aligned}$$

where $\hat{A} = V^T A V$, $\hat{N}_i = V^T N_i V$ for $i = 1, \dots, m$, and $V \in \mathbb{R}^{n \times k}$ is orthogonal. Our first result is concerned with the invertibility of the operators $\widetilde{\mathcal{M}}$ and $\widehat{\mathcal{M}}$, respectively.

Proposition 12 *If $\sigma(\mathcal{M}) = -\sigma(\mathcal{L} + \Pi) \subset \mathbb{C}_+$ then $\sigma(\widetilde{\mathcal{M}}) \subset \mathbb{C}_+$ and $\sigma(\widehat{\mathcal{M}}) \subset \mathbb{C}_+$.*

Proof Note that $\sigma(\widetilde{\mathcal{M}})$ is determined by the eigenvalues of the matrix

$$\widetilde{\mathbf{M}} := -I \otimes A - \hat{A} \otimes I - \sum_{i=1}^m \hat{N}_i \otimes N_i. \quad (15)$$

Similarly, we obtain $\sigma(\mathcal{M})$ by computing the eigenvalues of the matrix

$$\mathbf{M} := -I \otimes A - A \otimes I - \sum_{i=1}^m N_i \otimes N_i. \quad (16)$$

Since A and N_i are assumed to be symmetric, we conclude that $\mathbf{M} = \mathbf{M}^T > 0$. Let us then define the orthogonal matrix $\mathbf{V} = V \otimes I$. It follows that $\widetilde{\mathbf{M}} = \mathbf{V}^T \mathbf{M} \mathbf{V}$ and, consequently, $\widetilde{\mathbf{M}} = \widetilde{\mathbf{M}}^T > 0$. A similar argument with $\mathbf{V} = V \otimes V$ can be applied to show the second assertion. $\square$

Given a reduced bilinear system, we naturally obtain an approximate solution to the generalized Lyapunov equation. Moreover, the error with respect to the $\mathcal{M}$ -inner product is given by the $\mathcal{H}_2$ -norms of the original and reduced system, respectively.

Proposition 13 *Let Σ denote a bilinear system (8) and let $A = A^T < 0$, $N_i = N_i^T$ for $i = 1, \dots, m$, and $B = C^T$. Assume that $\rho(\mathcal{L}^{-1} \Pi) < 1$. Given an orthogonal $V \in \mathbb{R}^{n \times k}$, $k < n$, define $\widehat{\Sigma}$, the reduced bilinear system (14), via $\hat{A} = V^T A V$, $\hat{N}_i = V^T N_i V$ and $\hat{B} = V^T B = \hat{C}^T$. Let X be the solution to $\mathcal{M}(X) = B B^T$, and let $\hat{X}$ be the solution to $\widehat{\mathcal{M}}(\hat{X}) = \hat{B} \hat{B}^T$. Then*

$$\|X - V \hat{X} V^T\|_{\mathcal{M}}^2 = \|\Sigma\|_{\mathcal{H}_2}^2 - \|\widehat{\Sigma}\|_{\mathcal{H}_2}^2.$$

Proof By assumption it holds that $\mathcal{M}$ and $\widehat{\mathcal{M}}$ are invertible and the controllability Gramians X and $\hat{X}$ exist. We observe that $\|X\|_{\mathcal{M}}^2 = \text{trace}(X B B^T) = \|\Sigma\|_{\mathcal{H}_2}^2$ and that $\langle V \hat{X} V^T, X \rangle_{\mathcal{M}} = \text{trace}(V \hat{X} V^T B B^T) = \|\widehat{\Sigma}\|_{\mathcal{H}_2}^2$. Moreover, for the reduced system we obtain

$$\widehat{\mathcal{M}}(\hat{X}) = -V^T (A V \hat{X} V^T + V \hat{X} V^T A + \sum_{i=1}^m N_i V \hat{X} V^T N_i) V = V^T \mathcal{M}(V \hat{X} V^T) V,$$

which implies that $\|V\hat{X}V^T\|_{\mathcal{M}}^2 = \text{trace}(\hat{X}\widehat{\mathcal{M}}(\hat{X})) = \|\widehat{\Sigma}\|_{\mathcal{H}_2}^2$. Hence, we obtain

$$\|X - V\hat{X}V^T\|_{\mathcal{M}}^2 = \|X\|_{\mathcal{M}}^2 + \|V\hat{X}V^T\|_{\mathcal{M}}^2 - 2\langle V\hat{X}V^T, X \rangle_{\mathcal{M}} = \|\Sigma\|_{\mathcal{H}_2}^2 - \|\widehat{\Sigma}\|_{\mathcal{H}_2}^2.$$

□

Extending the results from [7], we obtain a lower bound for the previous terms by the $\mathcal{H}_2$ -norm of the error system $\Sigma - \widehat{\Sigma}$.

Proposition 14 *Let Σ denote a bilinear system (8) and let $A = A^T \prec 0$, $N_i = N_i^T$ for $i = 1, \dots, m$, and $B = C^T$. Assume that $\rho(\mathcal{L}^{-1}\Pi) < 1$. Given an orthogonal $V \in \mathbb{R}^{n \times k}$, $k < n$, define $\widehat{\Sigma}$, the reduced bilinear system (14), via $\hat{A} = V^T A V$, $\hat{N}_i = V^T N_i V$ and $\hat{B} = V^T B = \hat{C}^T$. Then, it holds*

$$\|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2}^2 \leq \|\Sigma\|_{\mathcal{H}_2}^2 - \|\widehat{\Sigma}\|_{\mathcal{H}_2}^2,$$

with equality if $\widehat{\Sigma}$ is locally $\mathcal{H}_2$ -optimal.

Proof The proof follows by arguments similar to those used in [7, Lemma 3.1]. By definition of the $\mathcal{H}_2$ -norm for bilinear systems

$$\|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2}^2 = \text{trace} \left([B^T - \hat{B}^T] X_e \begin{bmatrix} B \\ -\hat{B} \end{bmatrix} \right),$$

where $X_e = \begin{bmatrix} X & Y \\ Y^T & \hat{X} \end{bmatrix}$ is the solution of

$$\begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} X_e + X_e \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} + \sum_{i=1}^m \begin{bmatrix} N_i & 0 \\ 0 & \hat{N}_i \end{bmatrix} X_e \begin{bmatrix} N_i & 0 \\ 0 & \hat{N}_i \end{bmatrix} + \begin{bmatrix} B \\ \hat{B} \end{bmatrix} [B^T \ \hat{B}^T] = 0.$$

Analyzing the block structure of X_e , adding and subtracting $\|\widehat{\Sigma}\|_{\mathcal{H}_2}^2 = \text{trace}(\hat{B}^T \hat{X} \hat{B})$, we find the equivalent expression

$$\|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2}^2 = \|\Sigma\|_{\mathcal{H}_2}^2 - \|\widehat{\Sigma}\|_{\mathcal{H}_2}^2 - 2 \left(\text{trace}(B^T Y \hat{B}) - \text{trace}(\hat{B}^T \hat{X} \hat{B}) \right).$$

We claim that $\text{trace}(B^T Y \hat{B}) - \text{trace}(\hat{B}^T \hat{X} \hat{B}) \geq 0$ which then shows the first assertion. In fact, Y and $\hat{X}$ are the solutions of $\widehat{\mathcal{M}}(Y) = B\hat{B}^T$ and $\widehat{\mathcal{M}}(\hat{X}) = \hat{B}\hat{B}^T$. With the operators introduced in (15) and (16), we obtain

$$\begin{aligned} \text{trace}(B^T Y \hat{B}) - \text{trace}(\hat{B}^T \hat{X} \hat{B}) &= \tilde{\mathbf{b}}^T \text{vec}(Y) - \tilde{\mathbf{b}}^T \text{vec}(\hat{X}) = \tilde{\mathbf{b}}^T \tilde{\mathbf{M}}^{-1} \tilde{\mathbf{b}}^T - \tilde{\mathbf{b}}^T \tilde{\mathbf{M}}^{-1} \tilde{\mathbf{b}} \\ &= \tilde{\mathbf{b}}^T \left(\tilde{\mathbf{M}}^{-1} - \mathbf{V}(\mathbf{V}^T \tilde{\mathbf{M}} \mathbf{V})^{-1} \mathbf{V}^T \right) \tilde{\mathbf{b}}, \end{aligned}$$

where $\tilde{\mathbf{b}} = \text{vec}(B\hat{B}^T)$ and $\hat{\mathbf{b}} = \text{vec}(\hat{B}\hat{B}^T)$. As in [7, Lemma 3.1], it follows that the previous expression contains the Schur complement of $\tilde{\mathbf{M}}^{-1}$ in $\mathbf{S} = \begin{bmatrix} \mathbf{V}^T \tilde{\mathbf{M}} \mathbf{V} & \mathbf{V}^T \\ \mathbf{V} & \tilde{\mathbf{M}}^{-1} \end{bmatrix}$ which can be shown to be positive semidefinite. We omit the details and refer to [7].

Assume now that $\hat{\Sigma}$ is locally $\mathcal{H}_2$ -optimal. From [41], we have the following first-order necessary optimality conditions

$$\begin{aligned} Y^T Z + \hat{X} \hat{Z} &= 0, & Z^T N_i Y + \hat{X} \hat{N}_i \hat{Z} &= 0, \quad i = 1, \dots, m, \\ Z^T B + \hat{Z} \hat{B} &= 0, & CY + \hat{C} \hat{X} &= 0, \end{aligned}$$

where $Y, \hat{X}$ are as before and $Z, \hat{Z}$ satisfy

$$A^T Z + Z \hat{A} + \sum_{i=1}^m N_i^T Z \hat{N}_i - C^T \hat{C} = 0, \quad \hat{A}^T \hat{Z} + \hat{Z} \hat{A} + \sum_{i=1}^m \hat{N}_i^T \hat{Z} \hat{N}_i + \hat{C}^T \hat{C} = 0.$$

From the symmetry of $A, \hat{A}, N_i$ and $\hat{N}_i$ as well as the fact that $B = C^T$ and $\hat{B} = \hat{C}^T$, we conclude that $\hat{Z} = \hat{X}$ and $Z = -Y$. Hence, from the optimality conditions, we obtain

$$0 = Z^T B + \hat{Z} \hat{B} = -Y^T B + \hat{X} \hat{B}$$

which in particular implies that

$$\text{trace}(B^T Y \hat{B}) - \text{trace}(\hat{B}^T \hat{X} \hat{B}) = \text{trace}(\hat{B}^T (Y^T B - \hat{X} \hat{B})) = 0.$$

This shows the second assertion. $\square$

As a consequence of Propositions 13 and 14, we obtain the following result.

Theorem 15 *Let Σ denote a bilinear system (8) and let $A = A^T \prec 0, N_i = N_i^T$ for $i = 1, \dots, m$ and $B = C^T$. Assume that $\rho(\mathcal{L}^{-1} \Pi) < 1$. Given an orthogonal $V \in \mathbb{R}^{n \times k}, k < n$, define $\hat{\Sigma}$, the reduced bilinear system (14), via $\hat{A} = V^T A V, \hat{N}_i = V^T N_i V$ and $\hat{B} = V^T B = \hat{C}^T$. Assume that $\hat{X}$ solves $\widehat{\mathcal{M}}(\hat{X}) = \hat{B} \hat{B}^T$. If $\hat{\Sigma}$ is locally $\mathcal{H}_2$ -optimal, then $V \hat{X} V^T$ is locally optimal with respect to the $\mathcal{M}$ -norm.*

3.3 Equivalence of ALS and rank-1 BIRKA

So far we have shown that a subspace producing a locally $\mathcal{H}_2$ -optimal model reduction is also a subspace for which the Galerkin approximation is locally optimal in the $\mathcal{M}$ -norm. In this part we, algorithmically, establish an equivalence between BIRKA and ALS. More precisely, for the symmetric case the equivalence is between BIRKA applied with the target model reduction subspace of dimension 1 for (8), and ALS applied to (1). The proof is based on the following lemmas.

Lemma 16 Consider using BIRKA (Algorithm 2) with $k = 1$, i.e., both the initial guesses and the output are vectors. Then $\tilde{A} \in \mathbb{R}$ is a scalar and hence we can take $\tilde{A} = \tilde{A}$ and $R = 1$ in Step 2. Thus $\hat{B} = \tilde{B}$, $\hat{C} = \tilde{C}$, $\hat{N}_1 = \tilde{N}_1, \dots, \hat{N}_m = \tilde{N}_m$, and hence Steps 2–3 are redundant. Moreover, since $\tilde{V}$ and $\tilde{W}$ are vectors, Step 6, is redundant.

Proof The result follows from direct computation. $\square$

When speaking about *redundant* steps and operations we mean that the entities assigned in that step are exactly equal to another, existing, entity. In such a situation the algorithm can be rewritten, by simply changing the notation, in a way that skips the redundant step and still produces the same result.

Lemma 17 Consider a symmetric generalized Lyapunov equation (7) and let $v, w \in \mathbb{R}^n$ be two given vectors. Let $v_{\text{BIRKA}}, w_{\text{BIRKA}} \in \mathbb{R}^n$ be the approximations obtained by applying BIRKA (Algorithm 2) to (1) with $C = B^T$ and initial guesses v and w . If $v = w$, then $v_{\text{BIRKA}} = w_{\text{BIRKA}}$.

Proof The proof is by induction, and it suffices to show that if $\tilde{V} = \tilde{W}$ at the beginning of a loop, the same holds at the end of the loop. Thus assume $\tilde{V} = \tilde{W}$. Then $\tilde{N}_i = (\tilde{W}^T \tilde{V})^{-1} \tilde{W}^T N_i \tilde{V} = \tilde{V}^T N_i \tilde{V} / \|\tilde{V}\|^2 = \tilde{V}^T N_i^T \tilde{V} / \|\tilde{V}\|^2 = \hat{N}_i^T$ for $i = 1, \dots, m$, and $\tilde{C} = C \tilde{V} = B^T \tilde{W} = \tilde{B}^T$. By Lemma 16 we do not need to consider Steps 2–3. We can now conclude that Step 4 and Step 5 are equal, and thus at the end of the iteration we still have $\tilde{V} = \tilde{W}$. $\square$

Lemma 18 Consider a symmetric generalized Lyapunov equation (7) and let $v, w \in \mathbb{R}^n$ be two given vectors. Let $v_{\text{ALS}}, w_{\text{ALS}} \in \mathbb{R}^n$ be the approximations obtained by applying the ALS algorithm (Algorithm 1) to (1) with initial guesses v and w . If $v = w$, then $v_{\text{ALS}} = w_{\text{ALS}}$.

Proof Similar to the proof of Lemma 17 it is enough to show that if $v = w$ at the beginning of a loop then it also holds at the end of the loop. Hence we assume that $v = w$. Then $\hat{A}_1 = \hat{A}_2$ follows by direct calculations. Moreover, by assumption $\mathcal{R}_k = \mathcal{R}_k^T$. Thus Step 3 and Step 6 are equal, and hence at the end of the iteration we still have that $v = w$. $\square$

Theorem 19 Consider a symmetric generalized Lyapunov equation (7) and let $v \in \mathbb{R}^n$ be a given vector. Let $v_{\text{BIRKA}} \in \mathbb{R}^n$ be the approximation obtained by applying BIRKA (Algorithm 2) to (1) with $C = B^T$ and initial guess v . Moreover, let $v_{\text{ALS}} \in \mathbb{R}^n$ be the approximation obtained by applying the ALS algorithm (Algorithm 1) to (1) with initial guess v . Then $v_{\text{BIRKA}} = v_{\text{ALS}}$.

Proof First, Lemma 17 and Lemma 18 makes it reasonable to assess the algorithms with only a single initial guess as well as a single output. Moreover, Step 5 in BIRKA as well as Steps 2–4 in ALS are redundant. Furthermore, it follows from Lemma 16 that in this situation Steps 2, 3, and 6 of BIRKA are also redundant. Hence we need to compare the procedure consisting of Steps 1 and 4 from BIRKA, with the procedure consisting of Steps 1, 5, and 6 from ALS. It can be observed that the computations

are equivalent and thus the asserted equality holds if they stop after an equal amount of iterations. We hence consider the stopping criteria and note that they are the same, since $(v^T A^T v + v^T A v)/2\|v\|^2 = v^T A v/\|v\|^2 = \tilde{A} \in \mathbb{R}$. $\square$

Corollary 20 *Theorem 10 is applicable with ALS changed to BIRKA, using subspaces of dimension 1.*

Remark 21 Note that ALS can be generalized such that the optimization is computing rank- ℓ corrections, see [25, Remark 2.2]. With similar arguments as above, one can show that for symmetric systems this can equivalently be achieved by BIRKA. From a theoretical point of view, this will yield more accurate approximations. However, the computational complexity increases quickly since each ALS or BIRKA step then requires solving a generalized Sylvester equation of dimension $n \times \ell$.

4 Fixed-point iteration and approximative $\mathcal{M}$ -norm minimization

In the previous section we showed that the ALS-based iteration (11) locally minimizes the error in the $\mathcal{M}$ -norm with rank-1 updates. In contrast we here show that the fixed-point iteration minimizes an upper bound for the $\mathcal{M}$ -norm, but with no rank constraint on the minimizer.

Recall the fixed-point iteration for the generalized Lyapunov equation (1),

$$\mathcal{L}(\hat{X}_{k+1}) = -\Pi(\hat{X}_k) - BB^T, \quad k = 0, 1, \dots, \quad (17)$$

with $\hat{X}_0 = 0$. Under the assumption that $\rho(\mathcal{L}^{-1}\Pi) < 1$ the iteration is a convergent splitting and has been presented in, e.g., [12, Equation (12)], [41, Equation (12)], and [37, Equation (4)]. The fixed-point iteration is a residual-based iteration since (17) is known to be equivalent to

$$\hat{X}_{k+1} = \hat{X}_k - \mathcal{L}^{-1}(\mathcal{R}_k), \quad k = 0, 1, \dots, \quad (18)$$

with $\hat{X}_0 = 0$. To relate the fixed-point iteration to the $\mathcal{M}$ -norm minimization problem we consider the problem

$$\min_{\Delta = \Delta^T \succeq 0} \|X - \hat{X}_k - \Delta\|_{\mathcal{M}}^2.$$

The minimization is restricted to symmetric positive semidefinite matrices since we know that the solution $X = X^T \succeq 0$. Hence it is desired to have that if $\hat{X}_k = \hat{X}_k^T \succeq 0$, then the new iterate $\hat{X}_{k+1} = \hat{X}_k + \Delta$ also fulfills $\hat{X}_{k+1} = \hat{X}_{k+1}^T \succeq 0$; specifically then $\hat{X}_{k+1} \succeq \hat{X}_k$. Proposition 3 gives us the solution in just one step. However, the computation is as difficult as the original problem and hence the goal is to minimize an upper bound on the error. As before we disregard the constant term $\|X - \hat{X}_k\|_{\mathcal{M}}^2$ in the minimization and consider

$$\begin{aligned}
\min_{\substack{\Delta \\ \Delta = \Delta^T \geq 0}} \langle \Delta, \Delta \rangle_{\mathcal{M}} - 2 \operatorname{trace}(\Delta^T \mathcal{R}_k) &= \min_{\substack{\Delta \\ \Delta = \Delta^T \geq 0}} \operatorname{trace}(\Delta^T (-\mathcal{L}(\Delta) - 2\mathcal{R}_k) - \Delta^T \Pi(\Delta)) \\
&\leq \min_{\substack{\Delta \\ \Delta = \Delta^T \geq 0}} \operatorname{trace}(\Delta^T (-\mathcal{L}(\Delta) - 2\mathcal{R}_k)),
\end{aligned}$$

where the inequality is a consequence of the linearity of the trace and the positive semidefiniteness of Δ^T and $\Pi(\Delta)$. Hence the trace is non-negative [31]. The last expression is minimized by $\Delta = -\mathcal{L}^{-1}(\mathcal{R}_k)$, if $\mathcal{R}_k$ is symmetric and positive definite. The latter is asserted in the following theorem.

Theorem 22 *Consider the symmetric generalized Lyapunov equation (7) with the additional assumptions that $A < 0$ and $\rho(\mathcal{L}^{-1} \Pi) < 1$. Moreover, consider the sequence of approximations constructed by (18) where $\mathcal{R}_k$ is the residual associated with $\hat{X}_k$. Then $\hat{X}_k = \hat{X}_k^T \geq 0$ and $\mathcal{R}_k = \mathcal{R}_k^T \geq 0$, for all $k \geq 0$.*

Proof The proof is by induction and similar to that of Theorem 10. It holds that $X_0 = X_0^T \geq 0$ and $\mathcal{R}_0 = \mathcal{R}_0^T \geq 0$. Now assume that this is the case for some k . Then $\Delta = -\mathcal{L}^{-1}(\mathcal{R}_k)$ is symmetric and positive semidefinite, and hence $\hat{X}_{k+1}$ is symmetric and positive semidefinite. Moreover, since $\hat{X}_{k+1}$ and the operators in (1) are symmetric it follows that $\mathcal{R}_{k+1}$ is symmetric. Thus what is left to show is $\mathcal{R}_{k+1} \geq 0$, which is true if and only if $y^T \mathcal{R}_{k+1} y \geq 0$ for all $y \in \mathbb{R}^n$. Hence take an arbitrary $y \in \mathbb{R}^n$ and consider

$$y^T \mathcal{R}_{k+1} y = y^T \mathcal{R}_k y + y^T (\mathcal{L}(\Delta) + \Pi(\Delta)) y = y^T (\Pi(\Delta)) y \geq 0.$$

The last inequality holds since Δ is symmetric and positive semidefinite and Π is a symmetric operator. $\square$

Corollary 23 *The fixed-point iteration (17) produces an increasing sequence of approximations $0 = \hat{X}_0 \leq \hat{X}_1 \leq \dots \leq X$.*

Remark 24 One could consider creating a subspace iteration from (18), by computing a few singular vectors of $\mathcal{L}^{-1}(\mathcal{R}_k)$ and adding these to the basis. The method seems to have nice convergence properties per iteration in the symmetric case, but not in the non-symmetric case. However, the (naïve) computations are prohibitively expensive. See [37] for a computationally more efficient way of exploiting the fixed-point iteration.

5 A residual-based rational Krylov generalization

A viable technique for designing iterative methods for the generalized Lyapunov equation seems to be working with the residual; see the discussion in connection to Proposition 3, and in Sects. 3 and 4. In [25, Section 4] it is suggested that, so called, preconditioned residuals can be used to expand the search space. It is further suggested that one such preconditioner could be a one-step-ADI preconditioner $P_{\text{ADI}}^{-1} = (A - \sigma I)^{-1} \otimes (A - \sigma I)^{-1}$, for a suitable choice of the shift. We present a method along those lines, and show that it can be seen as a generalization of the rational Krylov subspace method.

5.1 Suggested search space

For the generalized Lyapunov equation (1), we suggest the following search space:

$$\mathcal{H}_k := \text{Span}\{B, (A - \sigma_1 I)^{-1}u_0, (A - \sigma_2 I)^{-1}u_1, \dots, (A - \sigma_k I)^{-1}u_{k-1}\}, \quad (19)$$

where u_{k-1} is the most dominant left singular vector of the Galerkin residual $\mathcal{R}_{k-1}$ of $\mathcal{H}_{k-1}$, and $\{\sigma_\ell\}_{\ell=1}^k$ is a sequence of shifts that needs to be chosen. In analogy with the discussion in [15], we suggest that the shifts are chosen according to the largest approximation error along the current direction. More precisely,

$$\sigma_k := \arg \max_{\sigma \in [\sigma_{\min}, \sigma_{\max}]} \left(\left\| u_{k-1} - (A - \sigma I)V_{k-1}(\hat{A}_{k-1} - \sigma I)^{-1}V_{k-1}^T u_{k-1} \right\| \right), \quad (20)$$

where V_{k-1} is a matrix with orthogonal columns containing a basis of $\mathcal{H}_{k-1}$, the matrix $\hat{A}_{k-1} = V_{k-1}^T A V_{k-1}$, and $[\sigma_{\min}, \sigma_{\max}]$ is a search interval. Typically for a stable matrix A we let $\sigma_{\min}$ be the negative real part of the eigenvalue of A with largest real part (closest to 0). Correspondingly we let $\sigma_{\max}$ the negative real part of the eigenvalue of A with smallest real part. Equations (19) and (20) can be straightforwardly incorporated in a Galerkin method for the generalized Lyapunov equation; the pseudocode is presented in Algorithm 3.

Algorithm 3: Residual-based rational-Krylov-type solver

input : $A, N_1, \dots, N_m \in \mathbb{R}^{n \times n}$ $B \in \mathbb{R}^{n \times r}$, tol
output: $\hat{X}$

```

1  $V_0 = \emptyset, v_1 = \text{orth}(B)$ 
for  $k = 1, 2, \dots$  until convergence do
2    $V_k = [V_{k-1}, v_k]$ 
3   Compute the projected matrices:  $\hat{A}_k = V_k^T A V_k$ , and  $\hat{N}_{i,k} = V_k^T N_i V_k$  for  $i = 1, 2, \dots, m$ , and
    $\hat{B}_k = V_k^T B$ 
4   Solve the projected problem:  $\hat{A}_k Y_k + Y_k \hat{A}_k^T + \sum_{i=1}^m \hat{N}_{i,k} Y_i \hat{N}_{i,k}^T + \hat{B}_k \hat{B}_k^T = 0$ 
5   Construct the (Galerkin) approximation:  $\hat{X}_k = V_k Y_k V_k^T$ 
6   Compute the residual:  $\mathcal{R}_k = A \hat{X}_k + \hat{X}_k A^T + \sum_{i=1}^m N_i \hat{X}_k N_i^T + B B^T$ 
7   if  $\|\mathcal{R}_k\| < \text{tol}$  then
   | Break
   end
8    $u_k \leftarrow$  the most dominant left singular vector of  $\mathcal{R}_k$ 
9   Select shift  $\sigma_{k+1}$  according to (20)
10   $v_{k+1} = (A - \sigma_{k+1} I)^{-1} u_k$ 
11   $v_{k+1} \leftarrow$  orthogonalize  $v_{k+1}$  with respect to  $V_k$ 
end
12 return  $\hat{X} = \hat{X}_k$ 
```

Remark 25 In practice the computation of the left singular vector can typically be done approximatively in an iterative fashion. This would also remove the need of computing the approximative solution $\hat{X}_k$ in Step 5 and the residual in Step 6 explicitly, since the matrix vector product can be implemented as $\mathcal{R}_k v = A V_k Y_k V_k^T v +$

$V_k Y_k V_k^T A^T v + \sum_{i=1}^m N_i V_k Y_k V_k^T N_i^T v + B B^T v$. However, such computations may introduce inexactness which can present a difficulty in a subspace method.

Remark 26 Heuristically the dynamic shift-search in Step 9 can be changed to an analogue of [15, (2.4) and (2.2)]. We suggest

$$\sigma_k := \arg \max_{\sigma \in \partial S} \frac{1}{|\tau_{k-1}(z)|}, \quad (21)$$

where S approximates the mirrored spectrum of A and ∂S is the boundary of S , and

$$\tau_{k-1}(z) := \frac{\prod_{j=1}^{\dim(\mathcal{K}_{k-1})} (z - \lambda_j^{(k-1)})}{\prod_{\ell=1}^{k-1} (z - \sigma_\ell)},$$

with $\lambda_j^{(k-1)}$ being the Ritz values of $\hat{A}_{k-1}$. Typically S is approximated at each step using the convex hull of the Ritz values of $\hat{A}_{k-1}$. It has been observed efficient in experiments since the maximization of (21) is computationally faster compared to (20). See Sect. 6 for a practical comparison of convergence properties.

Remark 27 The steps 8–9 in Algorithm 3 can be changed for a tangential-direction approach according to [16]. One practical way, although a heuristic, is to do the shift search according to either (20) or (21), and then compute the principal direction(s) according to [16, Section 3], i.e., through a singular value decomposition of $\mathcal{R}_{k-1} - (A - \sigma_k I) V_{k-1} (\hat{A}_{k-1} - \sigma_k I)^{-1} V_{k-1}^T \mathcal{R}_{k-1}$. It has been observed in experiments that such an approach tends to speed up the convergence, in terms of computation time, since the computation of the residual is costly.

Remark 28 It is (sometimes) desirable to allow for complex conjugate shifts σ_k and $\bar{\sigma}_k$, although, for reasons of computations and model interpretation one wants to keep the basis real. This goal is achievable using the same idea as in [35]. More precisely, one can utilize the relation $\text{Span} \{(A - \sigma_k I)^{-1} u_{k-1}, (A - \bar{\sigma}_k I)^{-1} u_{k-1}\} = \text{Span} \{\text{Re}((A - \sigma_k I)^{-1} u_{k-1}), \text{Im}((A - \sigma_k I)^{-1} u_{k-1})\}$. Although it requires two shifts to be used together with the vector u_{k-1} .

5.2 Analogies to the linear case

To give further insight to the suggested subspace in (19), we draw parallels with the (standard) rational Krylov subspace for the (standard) Lyapunov equation,

$$\mathcal{L}(X) + B B^T = 0, \quad (22)$$

where $\mathcal{L}$ is defined by (2) and $B \in \mathbb{R}^{n \times r}$. The idea is to show that the suggested space (19), reduces to something well known in this case. As a technical note we observe that Definitions 1 and 2 are analogous for the (standard) Lyapunov equation (22), but with $\Pi = 0$. The reasoning in this section can be compared to that of [3, Section 2]. To prove the main result of this section we need the following lemma.

Lemma 29 Let $A \in \mathbb{R}^{n \times n}$ and $\sigma_a \in \mathbb{R}$ be any scalar such that $(A - \sigma_a I)$ is nonsingular. Moreover, let $V \in \mathbb{R}^{n \times k}$, $k \leq n$, be orthogonal, i.e., $V^T V = I$, and let $\mathcal{R} \in \mathbb{R}^{n \times n}$ be such that $\text{Range}((A - \sigma_a I)^{-1} \mathcal{R}) \subseteq \text{Span}(V)$. Then $\mathcal{R} = (A - \sigma_a I)V(V^T AV - \sigma_a I)^{-1} V^T \mathcal{R}$.

Proof We introduce the notation $S := (A - \sigma_a I)$ and $\hat{S} := (V^T AV - \sigma_a I)$. To prove the statement we consider the right-hand-side of the asserted equality,

$$SV\hat{S}^{-1}V^T \mathcal{R} = SV\hat{S}^{-1}V^T S S^{-1} \mathcal{R} = SV\hat{S}^{-1}V^T S V V^T S^{-1} \mathcal{R},$$

where the second equality follows from the assumption $\text{Range}(S^{-1} \mathcal{R}) \subseteq \text{Span}(V)$. By observing that $\hat{S}^{-1} V^T S V = I$ the expression can be further simplified as

$$SV\hat{S}^{-1}V^T \mathcal{R} = S V V^T S^{-1} \mathcal{R} = S S^{-1} \mathcal{R} = \mathcal{R},$$

where, again, the second equality follows from $\text{Range}(S^{-1} \mathcal{R}) \subseteq \text{Span}(V)$. $\square$

Theorem 30 Let $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times r}$, and let $\{\sigma_\ell\}_{\ell=1}^{k+1}$ be a sequence of shifts such that $A - \sigma_\ell I$ is nonsingular for $\ell = 1, \dots, k+1$. Define the space $\mathcal{K}_k := \text{Span}\{B, (A - \sigma_1 I)^{-1} B, \dots, \prod_{\ell=1}^k (A - \sigma_\ell I)^{-1} B\}$, and $\mathcal{K}_{k+1}$ analogously. Let V_k be an orthogonal basis of $\mathcal{K}_k$, V_{k+1} an orthogonal basis of $\mathcal{K}_{k+1}$, and let $v_{k+1} \in \mathbb{R}^{n \times r}$ be such that $V_{k+1} = [V_k, v_{k+1}]$.¹ Moreover, let $\mathcal{R}_k \in \mathbb{R}^{n \times n}$ be the Galerkin residual with respect to (22). Then each column of $(A - \sigma_{k+1} I)^{-1} \mathcal{R}_k$ is in $\text{Span}(V_{k+1})$, i.e., $\text{Range}((A - \sigma_{k+1} I)^{-1} \mathcal{R}_k) \subseteq \text{Span}(V_{k+1})$. Furthermore, if $\text{Range}((A - \sigma_{k+1} I)^{-1} \mathcal{R}_k) \subseteq \text{Span}(V_k)$, then $\mathcal{R}_k = 0$.

Proof We introduce the notation $S_{k+1} := (A - \sigma_{k+1} I)$ and $\hat{S}_{k+1} := (V^T AV - \sigma_{k+1} I)$.

From existing results on rational Krylov subspaces, see, e.g., [27, Proposition 2.2], there exists $\alpha \in \mathbb{R}^{r \times n}$ such that

$$\begin{aligned} \mathcal{R}_k &= A V_k Y_k V_k^T + V_k Y_k V_k^T A^T + B B^T \\ &= \sigma_{k+1} v_{k+1} \alpha - (I - V_k V_k^T) A v_{k+1} \alpha + V_k T_k Y_k V_k^T + V_k Y_k V_k^T A^T + V_k V_k^T B B^T \\ &= -S_{k+1} v_{k+1} \alpha + V_k \beta \end{aligned}$$

for a suitable $\beta \in \mathbb{R}^{(k+1)r \times n}$. This shows the first claim.

To prove the second claim we assume that $\text{Range}(S_{k+1}^{-1} \mathcal{R}_k) \subseteq \mathcal{K}_k = \text{Span}(V_k)$. Under this assumption we can use Lemma 29 and the fact that $\mathcal{R}_k = \mathcal{R}_k^T$ to get

$$\mathcal{R}_k = S_{k+1} V_k \hat{S}_{k+1}^{-1} V_k^T \mathcal{R}_k = S_{k+1} V_k \hat{S}_{k+1}^{-1} V_k^T \mathcal{R}_k V_k \hat{S}_{k+1}^{-1} V_k^T S_{k+1} = 0,$$

since $\mathcal{R}_k$ is the Galerkin residual and thus $V_k^T \mathcal{R}_k V_k = 0$. $\square$

¹ Here we have, implicitly, assumed that the dimension of the $\mathcal{K}_{k+1}$ is $n \times (k+2)r$, i.e., all the columns in the definition of the space are linearly independent.

Remark 31 The interpretation of Theorem 30 is easiest in the case when $B = b \in \mathbb{R}^n$. Consider the two spaces $\mathcal{K}_k := \text{Span}\{b, (A - \sigma_1 I)^{-1}b, \dots, \prod_{\ell=1}^k (A - \sigma_\ell I)^{-1}b\}$ and $\hat{\mathcal{K}}_k := \text{Span}\{\mathcal{R}_{-1}, (A - \sigma_1 I)^{-1}\mathcal{R}_0, \dots, (A - \sigma_k I)^{-1}\mathcal{R}_{k-1}\}$, where $\mathcal{R}_{-1} = b$ and $\mathcal{R}_j$ is the Galerkin residual in space $\mathcal{K}_j$, with $j = 0, 1, \dots, k-1$. Then for all relevant cases, i.e., $\mathcal{R}_j \neq 0$ for $j = -1, 0, \dots, k-1$, we have that $\mathcal{K}_k = \hat{\mathcal{K}}_k$. In this sense the suggested subspace in (19) can be seen as a natural generalization of a rational Krylov subspace for linear matrix equations.

6 Numerical examples

We now numerically compare different methods discussed in the paper. All algorithms are treated in a subspace fashion² and we compare practically achieved approximation properties as a function of subspace dimension. Since the paper focuses on the symmetric problem we use Galerkin projection in the tested methods, except BIRKA. However, to (numerically) investigate the domain of application we test the methods on problems with varying degree of symmetry.

For small and moderate sized problems there are algorithms for computing the full solution, see [24, Algorithm 2], cf. [41, equation (12)]. Although costly, this nevertheless allows for inspection of the *relative error*, i.e.,

$$\|X - \hat{X}_k\|/\|X\|.$$

Moreover, it also allows comparison with the (in the Frobenius norm) optimal low-rank approximation based on the SVD.

We summarize some of the implementation details. Specifically, BIRKA is implemented as described in Algorithm 2, with a maximum allowed number of iterations equal to 100. Convergence tolerance is implemented as relative norm difference of the vector of sorted eigenvalues and was set to 10^{-3} . Each subspace is computed independently from a random initial guess. We emphasize that the method based on ALS is a subspace method, and not an iteratively updated method as described in (11). Because of the structure of the generalized Lyapunov equation, the solution is symmetric even if the coefficient matrices are not, we use a symmetric version of ALS even for the non-symmetric examples. More precisely, a symmetrized version of Algorithm 1, cf. Lemma 18, was used as an inner iteration, the resulting vector was used in an outer iteration to expand the search space, and the approximation was found using Galerkin projection. The maximum allowed number of iterations in ALS (inner iteration) was set to 20, and the tolerance to 10^{-2} . With reference to [25] we note that preconditioned residuals were not used, although it may accelerate convergence. Regarding the rational-Krylov-type methods we compare the following methods, which we give short labels for the legends further down:

- A: $\mathcal{K}_k$ as in (19), according to Algorithm 3

² The technique of turning an iterative method, such as, e.g., ALS, into a subspace method is known as *Galerkin acceleration*. The idea is nicely explained in [25, Section 3].

- B: Algorithm 3 but with tangential directions according to Remark 27, though with shifts according to (20)
- C: Algorithm 3 but with shifts according to (21)
- D: Algorithm 3 but with tangential directions according to Remark 27 and shifts according to (21)
- E: Standard rational Krylov. More precisely, similar to Algorithm 3, but instead of using u_{k-1} we use the right-hand-side B in both (19) and (20)
- F: $\mathcal{K}_k$ as in (19), but with on-beforehand-prescribed shifts given as the recycling of mirrored eigenvalues from a size-10-BIRKA (convergence tolerance set to 10^{-3}). Mirrored eigenvalues are potentially complex, with positive real part, and taken in ascending order according to the real parts.

For methods C, D, and F the shifts may be complex-valued, and the complex arithmetic is avoided by creating the space in accordance with Remark 28. For methods A–E, the shift-search-boundaries were set to $\sigma_{\max} = -1.01 \cdot \min_{\lambda \in \sigma(A)} \lambda$ and $\sigma_{\min} = -0.99 \cdot \max_{\lambda \in \sigma(A)} \lambda$, as to slightly enlarge the region. For methods A, B, and E, the shifts are taken as approximations to (20). The approximation is computed by discretizing the interval $[\sigma_{\min}, \sigma_{\max}]$ in 30 equidistant points and comparing the value of the target function. Orthogonalization of the basis is implemented using MATLAB built-in QR factorization and keeping vectors only if the corresponding diagonal element in R is large enough. Implementations for the methods A–F are available online.³ The simulations were done in MATLAB R2018a (9.4.0.813654) on a computer with four 1.6 GHz processors and 16 GB of RAM.

We test the algorithms on three different problems. All examples are bilinear control systems and we approximate the associated controllability Gramian, as in (9). The examples all have stable Lyapunov operators. The first example is symmetric, the second is non-symmetric but symmetrizable, and the third example is non-symmetric.

6.1 Heat equation

The first example is motivated by an optimal control problem for selective cooling of steel profiles, see [17]. In this example, the state variable w models the evolution of a temperature and is described by a two-dimensional heat equation,

$$\frac{\partial}{\partial t} w(x, y, t) = \Delta w(x, y, t) \quad (x, y, t) \in (0, 1) \times (0, 1) \times (0, T),$$

where a control $u(t)$ enters bilinearly from the left through a Robin condition,

$$-\frac{\partial}{\partial x} w(0, y, t) = 0.5(w(0, y, t) - 1)u(t) \quad (y, t) \in (0, 1) \times (0, T).$$

The control can be interpreted as the spraying intensity of a cooling fluid. The other spatial boundaries satisfy homogeneous Dirichlet conditions, and at $t = 0$ an initial temperature profile is specified. The equation is discretized in space using centered

³ https://people.kth.se/~eringh/software/res_rat_Kry_type/.

Fig. 1 Cross-algorithm comparison for the heat equation. Relative residual norm (left), relative error in Frobenius norm (right)

Fig. 2 Rational-Krylov-type method comparison for the heat equation. Relative residual norm (left), relative error in Frobenius norm (right). Compare with Fig. 1 as the lines for method A are the same in the two figures. For a description of the labels, see the beginning of this section

finite difference, which yields a bilinear system with $A \in \mathbb{R}^{5041 \times 5041}$, $B \in \mathbb{R}^{5041}$, $m = 1$, and $N_1 = N \in \mathbb{R}^{5041 \times 5041}$. It can be further noted that, $A = A^T < 0$ and $N = N^T$, and hence the theory of $\mathcal{H}_2$ -optimality and the definition of the $\mathcal{M}$ -norm is applicable.

We compare different methods discussed in the paper, both the relative residual norm and the relative error. For readability the plots have been split in different figures. Hence in Fig. 1 we compare across different classes of methods, and in Fig. 2 we compare between different flavors the rational-Krylov-type methods. It can be observed, see Fig. 1, that for this example BIRKA has extremely good performance, even outperforming the SVD in relative residual norm. Nevertheless, the larger BIRKA subspaces can be rather costly to compute. In comparison ALS shows good performance compared to the rational-Krylov-type subspace, and is rather cheap to compute. When comparing the different rational-Krylov-type methods, see Fig. 2, we see that stan-

Fig. 3 The cross-algorithm comparison (left) and rational-Krylov-type methods (right) for the heat equation. The relative error is measured in the $\mathcal{M}$ -norm. The lines for method A are the same in the two plots

dard rational-Krylov (E) has the problem that the convergence stagnates. The methods A, C, and F have similar performance. In comparison, B and D are only slightly worse in the error per subspace dimension comparison but are practically sometimes faster to compute.

Since the $\mathcal{M}$ -norm is defined for this example we compare the relative error also in this norm, see Fig. 3. The trend is similar as in the Frobenius norm, although it can be observed that in general the error is smaller and BIRKA has best performance, even compared to the SVD.

6.2 1D Fokker–Planck

The second example is from quantum physics, where a one-dimensional Fokker–Planck equation is used to describe the evolution of a probability density function, ρ , of a particle affected by a potential. Parts of the potential can be manipulated by a so-called optical tweezer, which constitutes the control. For further details of the problem see [22]. More precisely we consider

$$\begin{aligned} \frac{\partial}{\partial t} \rho(x, t) &= \nu \frac{\partial^2}{\partial x^2} \rho(x, t) + \frac{\partial}{\partial x} \left(\rho(x, t) \frac{\partial}{\partial x} V(x, t) \right) & (x, t) &\in (-6, 6) \times (0, T) \\ \rho(x, 0) &= \rho_0(x) & x &\in (-6, 6) \\ \nu \frac{\partial}{\partial x} \rho(x, t) &= -\rho(x, t) \frac{\partial}{\partial x} V(x, t) & (x, t) &\in \{-6, 6\} \times (0, T), \end{aligned}$$

where the potential is $V(x, t) = W(x) + \alpha(x)u(t)$, with the ground (fixed) potential being $W(x) = (((0.5x^2 - 15)x^2 + 199)x^2 + 28x + 50)/200$, and $\alpha(x)$ is an approximately linear control shape function; for more details see [11]. In a weighted inner product, the dynamics can be described by self-adjoint operators. However, here we employ an upwinding type finite difference scheme with 5000 grid points, leading to a non-symmetric system. As has been pointed out in [22], the system matrix A is not

Fig. 4 Cross-algorithm comparison for 1D Fokker–Planck equation. Relative residual norm (left) and relative error (right)

Fig. 5 Rational-Krylov-type method comparison for 1D Fokker–Planck equation. Compare with Fig. 4 as the lines for method A are the same in the two figures. For a description of the labels, see the beginning of this section

asymptotically stable due to a simple zero eigenvalue associated with the stationary probability distribution. Using a projection-based decoupling, it is however possible to work with an asymptotically stable system of dimension $n = 4999$. Similar to the first example, the control variable is a scalar and, consequently, we only obtain a single bilinear coupling matrix $N_1 = N$. Since the system is non-symmetric, the operator $\mathcal{M}$ is generally indefinite and hence we make no comparisons in the $\mathcal{M}$ -norm.

The plots in Figs. 4 and 5 are analogous to the plots in Figs. 1 and 2 respectively. However, for this example the direct solver stagnated at a relative residual of about 10^{-8} , which can be seen in the stagnation of the SVD approximation in the left of Fig. 4. As a result, the comparisons of relative error performance, the right of Figs. 4 and 5, show an artificial stagnation. At a certain level the convergence stagnates since it measures the discrepancy between the method approximations and the inexact reference solution, rather than the true error of the method approximations. Nevertheless

we believe the comparisons to be fair more or less up to the point of stagnation, which is justified by the relative residual plots showing similar behavior. However, the relative residual indicates stagnation around 10^{-8} for the other methods as well, although not quite as clear as for the SVD.

From Fig. 4 we see the BIRKA performs well for this example. However, the subspaces of dimension 28 and 29 did not converge in a 100 iterations and hence for clarity these are left out of the plots. This illustrates a drawback of the method. The performance difference between ALS and the rational-Krylov-type method is slightly smaller compared to the previous example. Among the rational-Krylov-type methods A, B, and F seems to have similar performance, whereas C is clearly worse. Method E is competitive for about 10 iterations and then the convergence is significantly slower. However, method D ends up with an insufficient subspace.

6.3 Burgers' equation

In the third example we consider an approximation to the one-dimensional viscous Burgers' equation

$$\begin{aligned} \frac{\partial}{\partial t} w(x, t) + w(x, t) \frac{\partial}{\partial x} w(x, t) &= \nu \frac{\partial^2}{\partial x^2} w(x, t) & (x, t) \in (0, 1) \times (0, T) \\ w(x, 0) &= w_0(x) & x \in (0, 1) \end{aligned}$$

where $\nu = 0.1$ is constant. The spatial boundary conditions are Dirichlet conditions. More specifically, $w(1, t) = 0$ and $w(0, t) = u(t)$, where $u(t)$ is an applied control input. The solution $w(x, t)$ can be interpreted as a velocity and the equation occurs in, e.g., modeling of gas or traffic flow. The problem is discretized in space using centered finite differences with 71 uniformly distributed grid points. Using a second order Carleman bilinearization, we obtain a bilinear control system approximation with $A, N \in \mathbb{R}^{5112 \times 5112}$ and $B \in \mathbb{R}^{5112}$; see [10] for further details. Note that in this case A is an asymptotically stable but non-symmetric matrix. To ensure the positive semidefiniteness of the Gramian, we scale the control matrices N and B with a factor $\alpha = 0.25$. We emphasize that the control law is scaled proportionally with $\frac{1}{\alpha}$ such that the dynamics remain unchanged, for further discussion see [9, Section 3.4].

The comparison is similar to the previous examples and the Figs. 6 and 7 are analogous to the respective Figs. 1 and 2. The problem is difficult in the sense that the singular values of the solution decay slowly. Moreover, the direct method stagnates at a relative residual norm of 5×10^{-6} . This is, however, less visible compared to the previous example since in general the convergence is slower.

For this example the performance of BIRKA is not significantly better than other methods, which is not surprising since the theoretical justifications for the method are not valid. ALS shows faster convergence in relative residual norm but slower convergence in relative error, as well as indications of stagnation. However, the theoretical justifications for ALS are also not valid for this example and the result is in line with the results in [25]. Regarding the rational-Krylov-type methods it seems as if method

Fig. 6 Cross-algorithm comparison for Burgers' equation. Relative residual norm (left) and relative error (right)

Fig. 7 Rational-Krylov-type method comparison for Burgers' equation. Compare with Fig. 6 as the lines for method A are the same in the two figures. For a description of the labels, see the beginning of this section

D and B has the best performance. However, method E does not provide a useful subspace for this example.

6.4 Execution time experiment

We conclude the numerical examples with a small experiment comparing the execution time of different methods considered. The problems are the same as above, i.e., the heat equation, the 1D Fokker–Planck equation, and the Burger's equation. For all these we generate a BIRKA subspace of dimension 30, an ALS subspace of dimension 60, and a subspace of type A of dimension 60. The approximation properties of these spaces are similar for the heat equation, see Fig. 1. The cumulative CPU time as a function of iteration count, in the respective method, is plotted in Fig. 8. Note that for ALS and method A the iteration count corresponds to increasing the dimension of the subspace

Fig. 8 Cumulative time as function of number of iterations. The plots are for the: Heat equation (left), Fokker–Planck (middle), and Burgers’ equation (right)

with one, since the right-hand-side is rank one. However, for BIRKA the dimension of the subspace is fixed on beforehand and hence there is an irregular number of iterations, corresponding to the convergence of the fixed-point problem rather than the size of the subspace. It was, for example, mentioned above that the BIRKA iterations for subspaces of dimension 28 and 29, for the Fokker–Planck equation, did not converge to the specified tolerance in the allowed 100 iterations.

In this situation, and for the chosen parameters, BIRKA is faster for the heat equation, and slower for the Fokker–Planck equation. In the case of the Burger’s equation it seems as if BIRKA is faster. However, if we take the approximation properties into account we find, by looking at Fig. 6, that a more fair comparison with method A is to consider the latter only up to iteration 30. Moreover, fixing the subspace dimension, rather than the tolerance, is (likely) advantageous for BIRKA.

7 Conclusions and outlooks

We have proposed a rational-Krylov-type subspace for solving the generalized Lyapunov equation. Simulations indicate competitive performance, at least in the non-symmetric case where optimality statements for the other methods are no longer valid. Simulations show that methods A and F perform well for all three examples. The ALS iteration, as well as results from the literature, cf. [1], seems to indicate that subspaces of the type $(A - \sigma I - \mu N_i)^{-1} B$ could be useful. Although we have not been able to exploit this efficiently. Another generalization of the rational Krylov subspace, for general linear matrix equations, is presented in [32]. It is suggested to use subspaces of the type $(A - \sigma I)^{-1} v$, and $(N_i - \sigma I)^{-1} v$, where v is a vector from the previous space. We see that more research is needed to understand the theoretical aspects of the suggested, and related, spaces.

Common for all methods studied is that they use the current residual in the iterations. Computing the residual can in itself be costly for a truly large scale problem, although approximate dominant directions can be computed in an iterative fashion, resulting

in an inner-outer-type iteration. However, more research is needed to understand the consequences of such inexact subspaces.

Acknowledgements We wish to thank the anonymous referees, who's comments helped improve the manuscript. The authors also wish to thank Elias Jarlebring (KTH) for support and discussions. This research started when the second author visited the first author at the Karl-Franzens-Universität in Graz; the kind hospitality was greatly appreciated. The visit was made possible due to the generous support from the European Model Reduction Network (COST action TD1307, STSM Grant 38025).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Ahmad, M., Baur, U., Benner, P.: Implicit Volterra series interpolation for model reduction of bilinear systems. *J. Comput. Appl. Math.* **316**(Supplement C), 15–28 (2017)
2. Al-Baiyat, S.A., Bettayeb, M.: A new model reduction scheme for k-power bilinear systems. In: *Proceedings of 32nd IEEE Conference on Decision and Control*, vol. 1, pp. 22–27 (1993)
3. Baars, S., Viebahn, J., Mulder, T., Kuehn, C., Wubs, F., Dijkstra, H.: Continuation of probability density functions using a generalized Lyapunov approach. *J. Comput. Phys.* **336**, 627–643 (2017)
4. Becker, S., Hartmann, C.: Infinite-dimensional bilinear and stochastic balanced truncation with error bounds. Technical report. [arXiv:1806.05322](https://arxiv.org/abs/1806.05322) (2018)
5. Benner, P., Breiten, T.: Interpolation-based $\mathcal{H}_2$ -model reduction of bilinear control systems. *SIAM J. Matrix Anal. Appl.* **33**(3), 859–885 (2012)
6. Benner, P., Breiten, T.: Low rank methods for a class of generalized Lyapunov equations and related issues. *Numer. Math.* **124**(3), 441–470 (2013)
7. Benner, P., Breiten, T.: On optimality of approximate low rank solutions of large-scale matrix equations. *Syst. Control Lett.* **67**, 55–64 (2014)
8. Benner, P., Bujanović, Z., Kürschner, P., Saak, J.: RADI: a low-rank ADI-type algorithm for large scale algebraic Riccati equations. *Numer. Math.* **138**(2), 301–330 (2018)
9. Benner, P., Damm, T.: Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems. *SIAM J. Control Optim.* **49**(2), 686–711 (2011)
10. Breiten, T., Damm, T.: Krylov subspace methods for model order reduction of bilinear control systems. *Syst. Control Lett.* **59**(8), 443–450 (2010)
11. Breiten, T., Kunisch, K., Pfeiffer, L.: Numerical study of polynomial feedback laws for a bilinear control problem. *Math. Control Relat. Fields* **8**(3&4), 557–582 (2018)
12. Damm, T.: Direct methods and ADI-preconditioned Krylov subspace methods for generalized Lyapunov equations. *Numer. Linear Algebra Appl.* **15**(9), 853–871 (2008)
13. Druskin, V., Knizhnerman, L., Zaslavsky, M.: Solution of large scale evolutionary problems using rational Krylov subspaces with optimized shifts. *SIAM J. Sci. Comput.* **31**(5), 3760–3780 (2009)
14. Druskin, V., Lieberman, C., Zaslavsky, M.: On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems. *SIAM J. Sci. Comput.* **32**(5), 2485–2496 (2010)
15. Druskin, V., Simoncini, V.: Adaptive rational Krylov subspaces for large-scale dynamical systems. *Syst. Control Lett.* **60**(8), 546–560 (2011)
16. Druskin, V., Simoncini, V., Zaslavsky, M.: Adaptive tangential interpolation in rational Krylov subspaces for MIMO dynamical systems. *SIAM J. Matrix Anal. Appl.* **35**(2), 476–498 (2014)
17. Eppler, K., Tröltzsch, F.: Fast optimization methods in the selective cooling of steel. In: Grötschel, M., Krumke, S., Rambau, J. (eds.) *Online Optimization of Large Scale Systems*, pp. 185–204. Springer, Berlin (2001)
18. Flagg, G., Beattie, C., Gugercin, S.: Convergence of the iterative rational Krylov algorithm. *Syst. Control Lett.* **61**(6), 688–691 (2012)
19. Flagg, G., Gugercin, S.: Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems. *SIAM J. Matrix Anal. Appl.* **36**(2), 549–579 (2015)

20. Golub, G., Van Loan, C.: *Matrix Computations*, 4th edn. The Johns Hopkins University Press, Baltimore (2013)
21. Gugercin, S., Antoulas, A., Beattie, C.: $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appl.* **30**(2), 609–638 (2008)
22. Hartmann, C., Schäfer-Bung, B., Thöns-Zueva, A.: Balanced averaging of bilinear systems with applications to stochastic control. *SIAM J. Control Optim.* **51**(3), 2356–2378 (2013)
23. Horn, R., Johnson, C.: *Topics in Matrix Analysis*. Cambridge University Press, Cambridge (1991)
24. Jarlebring, E., Mele, G., Palitta, D., Ringh, E.: Krylov methods for low-rank commuting generalized Sylvester equations. *Numer. Linear Algebra Appl.* **25**(6), e2176 (2018)
25. Kressner, D., Sirković, P.: Truncated low-rank methods for solving general linear matrix equations. *Numer. Linear Algebra Appl.* **22**(3), 564–583 (2015)
26. Kressner, D., Tobler, C.: Krylov subspace methods for linear systems with tensor product structure. *SIAM J. Matrix Anal. Appl.* **31**(4), 1688–1714 (2010)
27. Lin, Y., Simoncini, V.: Minimal residual methods for large scale Lyapunov equations. *Appl. Numer. Math.* **72**, 52–71 (2013)
28. Massei, S., Palitta, D., Robol, L.: Solving rank-structured Sylvester and Lyapunov equations. *SIAM J. Matrix Anal. Appl.* **39**(4), 1564–1590 (2018)
29. Mehrmann, V., Tan, E.: Defect correction method for the solution of algebraic Riccati equations. *IEEE Trans. Autom. Control* **33**(7), 695–698 (1988)
30. Mohler, R.R., Kolodziej, W.J.: An overview of bilinear system theory and applications. *IEEE Trans. Syst. Man Cybern.* **10**(10), 683–688 (1980)
31. Neudecker, H.: A matrix trace inequality. *J. Math. Anal. Appl.* **166**(1), 302–303 (1992)
32. Powell, C.E., Silvester, D., Simoncini, V.: An efficient reduced basis solver for stochastic Galerkin matrix equations. *SIAM J. Sci. Comput.* **39**(1), A141–A163 (2017)
33. Richter, S., Davis, L.D., Collins Jr., E.G.: Efficient computation of the solutions to modified Lyapunov equations. *SIAM J. Matrix Anal. Appl.* **14**(2), 420–431 (1993)
34. Ringh, E., Mele, G., Karlsson, J., Jarlebring, E.: Sylvester-based preconditioning for the waveguide eigenvalue problem. *Linear Algebra Appl.* **542**, 441–463 (2018). *Proceedings of the 20th ILAS Conference*, p. 2016. Belgium, Leuven
35. Ruhe, A.: The rational Krylov algorithm for nonsymmetric eigenvalue problems. III: complex shifts for real matrices. *BIT* **34**(1), 165–176 (1994)
36. Shaker, H.R., Tahavori, M.: Control configuration selection for bilinear systems via generalised Hankel interaction index array. *Int. J. Control* **88**(1), 30–37 (2015)
37. Shank, S.D., Simoncini, V., Szyld, D.B.: Efficient low-rank solution of generalized Lyapunov equations. *Numer. Math.* **134**(2), 327–342 (2016)
38. Simoncini, V.: Computational methods for linear matrix equations. *SIAM Rev.* **58**(3), 377–441 (2016)
39. Smith, R.: Matrix equation $XA + BX = C$. *SIAM J. Appl. Math.* **16**(1), 198–201 (1968)
40. Vandereycken, B., Vandewalle, S.: A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations. *SIAM J. Matrix Anal. Appl.* **31**(5), 2553–2579 (2010)
41. Zhang, L., Lam, J.: On H_2 model reduction of bilinear systems. *Autom. J. IFAC* **38**(2), 205–216 (2002)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Tobias Breiten¹ · Emil Ringh²

Tobias Breiten
tobias.breiten@uni-graz.at

¹ Institute for Mathematics and Scientific Computing, Karl-Franzens-Universität, 8010 Graz, Austria

² Department of Mathematics, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden