Peter Benner Tobias Breiten

Low rank methods for a class of generalized Lyapunov equations and related issues

Max Planck Institute Magdeburg Preprints

MPIMD/12-03 **February 21, 2012**

Impressum:

Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg

Publisher:
Max Planck Institute for Dynamics of Complex Technical Systems

Address:
Max Planck Institute for Dynamics of Complex Technical Systems
Sandtorstr. 1
39106 Magdeburg

www.mpi-magdeburg.mpg.de/preprints

Abstract

In this paper, we study possible low rank solution methods for generalized Lyapunov equations arising in bilinear and stochastic control. We show that under certain assumptions one can expect a strong singular value decay in the solution matrix allowing for low rank approximations. Since the theoretical tools strongly make use of a connection to the standard linear Lyapunov equation, we can even extend the result to the d-dimensional case described by a tensorized linear system of equations. We further provide some reasonable extensions of some of the most frequently used linear low rank solution techniques such as the alternating directions implicit (ADI) iteration and the Krlyov-plus-inverse-Krylov (KPIK) method. By means of some standard numerical examples used in the area of bilinear model order reduction, we will show the efficiency of the new methods.

Keywords: Lyapunov equations, bilinear systems, ADI iteration, low rank approximations

Author's addresses:

Peter Benner
Computational Methods in Systems and Control Theory,
Max Planck Institute for Dynamics of Complex Technical Systems,
Sandtorstr. 1,
39106 Magdeburg
Germany
(benner@mpi-magdeburg.mpg.de)

Tobias Breiten
Computational Methods in Systems and Control Theory,
Max Planck Institute for Dynamics of Complex Technical Systems,
Sandtorstr. 1,
39106 Magdeburg
Germany
(breiten@mpi-magdeburg.mpg.de)

1 Introduction

In this paper, we want to study a certain class of generalized Lyapunov equations of the form

$$AX + XA^T + \sum_{j=1}^{m} N_j X N_j^T + BB^T = 0, \tag{1}$$

where $A, N_j \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$. This type of equation has been shown to arise in the context of bilinear and special linear stochastic differential equations, see e.g. [6, 12, 20]. For the latter type of systems, the matrices N_j are associated with uncertainties described by independent zero mean real Wiener processes and usually are small compared to the system matrix A. In more detail, there is a direct relation between the solution X of (1) being positive definite and mean-square-stability of the system, see e.g. [1]. Moreover, one can define certain energy functionals which can be characterized by means of X and the solution of a dual Lyapunov equation. This allows to determine the importance of the system states w.r.t. an input-output map and thus opens up the way to generalize model reduction methods such as balanced truncation for linear systems. For bilinear control systems, control concepts as well as the meaning of X is far less obvious. For example, stability and controllability have to be defined locally and X might become indefinite although the system is locally stable. Nevertheless, there has recently been given an interpretation which also makes use of energy functionals, see [6]. Furthermore, the so-called generalized reachability Gramian X has been extensively used in the context of model order reduction for bilinear systems as well, see e.g. [6, 10, 20]. Note that, although bilinear control systems belong to the class of nonlinear control systems, the reachability Gramian X is defined as the solution of a linear matrix equation and thus one might extend some well-known results for linear systems. Hence, let us for a moment consider the more prominent case given by the standard Lyapunov equation

$$AX + XA^T + BB^T = 0, \tag{2}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, which arises e.g. in the stability analysis of linear continuous time-invariant control systems. In the following, whenever we speak of the *standard case*, we will refer to eq. (2). Here, if the pair (A, B) is controllable, there is a one-to-one-correspondence between the location of the eigenvalues of A and the property of the solution matrix $X = X^T \in \mathbb{R}^{n\times n}$ being positive definite.

If the number of inputs is small, i.e. $m \ll n$, the singular values of X often tend to decay exponentially fast, see e.g. [2, 19, 30], meaning that there exists a matrix $L \in \mathbb{R}^{n\times r}$ with $r \ll n$ s.t. $X \approx \hat{X} = LL^T$. Over the last years, this has caused the development of several different iterative methods like e.g. KPIK (see [33]) and the ADI iteration (see e.g. [7, 29, 30]) that work solely on the low rank factor L and therefore allow to compute solutions for dimensions up to $n \sim 10^6$. Other low rank techniques rely on considering the explicit system of linear equations corresponding to (2). To be more precise, for $m = 1$, we obtain the following tensor product structure

$$(I \otimes A + A \otimes I)\,\mathrm{vec}\,(X) = -B \otimes B. \tag{3}$$

As has been shown in [19, 26], the main advantage now is that most of the low rank approaches dealing with the above structure can be even generalized to the d-dimensional case with additional mass matrices appearing within the tensor structure

$$\left(\sum_{i=1}^{d} E_1 \otimes \cdots \otimes E_{i-1} \otimes A_i \otimes E_{i+1} \otimes \cdots \otimes E_d\right) \operatorname{vec}(X) = \bigotimes_{i=1}^{d} b_i. \tag{4}$$

Since the special structure of (4) allows to diagonalize the left-hand side by a matrix of tensor rank 1, the approximation procedure amounts to approximating the function

$$f(x_1, \ldots, x_d) = \frac{1}{x_1 + \cdots + x_d}.$$

Our intention now is to generalize the above ideas to (1). Hence, let us come back to the more general situation. Except in the case that the N_j commute with A, little is known about the singular value decay of X. However, as already observed in [6], the solution X still seems to exhibit similar properties as in the *standard case*. Moreover, in the context of high-dimensional eigenvalue problems, in [28], the authors already have proposed some methods for the d-dimensional case that expect the solution X to possess good low rank approximations. The goal of this paper is to give a theoretical explanation for this phenomenon in case that some restrictions are imposed on the matrices. This will justify the generalization of some linear low rank solution methods which will be suggested in this paper as well. Moreover, we will briefly address the d-dimensional tensor product case, i.e.

$$\left(\sum_{i=1}^{d} E_1 \otimes \cdots \otimes E_{i-1} \otimes A_i \otimes E_{i+1} \otimes \cdots \otimes E_d + \sum_{j=1}^{k} N_{j_1} \otimes \cdots \otimes N_{j_d}\right) \operatorname{vec}(X) = \bigotimes_{i=1}^{d} b_i. \tag{5}$$

In more detail, we will now proceed as follows. In Section 2, we will briefly review the basic results known from the *standard case*. This will include a short discussion on the tensor rank of a vectorized matrix, which will be an essential tool throughout the rest of the paper. Subsequently, in Section 3, we will motivate the need for efficient numerical treatment of large-scale generalized Lyapunov equations by means of giving some detailed background on model order reduction of very large-scale bilinear and stochastic linear control systems. In Section 4, we will make use of the Sherman-Morrison-Woodbury formula to show the desired exponential decay of the singular values of the solution matrix X. Starting with the special case $d = 2$, we will transfer the ideas to the multidimensional case. Based on these results, several low rank solution methods for $d = 2$, will be proposed in Section 5. Here, this will lead to a generalization of some rational Krylov subspace methods as well as the low rank Cholesky-factor ADI iteration. Finally, in Section 6, the new methods will be evaluated by means of several numerical examples reaching up to dimensions $n = 500\ 000$. We will conclude with a short summary and a discussion of future research perspectives in Section 7.

2 Preliminaries

We will start with the following definitions of the vec ()-operator as well as the Kronecker product of two matrices.

Definition 2.1. *([22]) Let $X = [x_1, \ldots, x_m] \in \mathbb{R}^{n\times m}$ and $Y \in \mathbb{R}^{p\times q}$. Then*

$$\operatorname{vec}(X) := \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \in \mathbb{R}^{n\cdot m\times 1}, \quad X \otimes Y = \begin{bmatrix} x_{11}Y & \ldots & x_{1m}Y \\ \vdots & & \vdots \\ x_{n1}Y & \ldots & x_{nm}Y \end{bmatrix} \in \mathbb{R}^{n\cdot p\times m\cdot q}.$$

Note the very useful link between the vectorization of matrix products and the Kronecker product operation

$$\operatorname{vec}(ABC) = (C^T \otimes A)\operatorname{vec}(B). \tag{6}$$

These properties immediately yield a proof for the equivalence of equations (2) and (3). Furthermore, throughout the rest of this paper, we will make use of the tensor rank of a vectorized matrix.

Definition 2.2. *([25]) Let $x = \operatorname{vec}(X) \in \mathbb{R}^{n^2}$. Then the minimal number k s.t.*

$$x = \sum_{i=1}^{k} u_i \otimes v_i,$$

where $u_i, v_i \in \mathbb{R}^n$, is called the tensor rank of the vector x.

Remark 2.1. *Due to the properties of the Kronecker product, it is easily seen that the tensor rank of a vectorized matrix X coincides with* $\operatorname{rank}(X)$.

Let us now return to the tensor product representation highlighted in (4). For linear systems with such a structure, in [19] it is shown that there exists a vector x_k of tensor rank k that fulfills a profitable error bound. The basic idea is to make use of the integral representation of the inverse of the matrix in (4) for which the use of a certain quadrature formula leads to the following lemma.

Lemma 2.1. *[19] Let $\mathcal{A}$ denote a matrix of tensor product structure as in* (4) *with tensor right-hand side $\mathcal{B}$. Assume that the sum of the spectra of the $E_i^{-1}A_i$ is contained in the strip $\Omega := [\lambda_{\min}, \lambda_{\max}] \oplus i[-\mu, \mu] \subseteq \mathbb{C}_-$. Let $k \in \mathbb{N}$ and define the following quadrature weights and points*

$$h_{st} := \frac{\pi}{\sqrt{k}}, \tag{7}$$

$$t_j := \log\left(\exp(jh_{st}) + \sqrt{1+\exp(2jh_{st})}\right), \tag{8}$$

$$w_j := \frac{h_{st}}{\sqrt{1+\exp(-2jh_{st})}}. \tag{9}$$

Then $\exists\, C_{st}$ s.t. the solution x to $\mathcal{A}x = \mathcal{B}$ can be approximated by

$$\tilde{x} := \sum_{j=-k}^{k} \frac{2w_j}{\lambda_{\max}} \bigotimes_{i=1}^{d} \exp\left(\frac{2t_j}{\lambda_{\max}} E_i^{-1} A_i\right) E_i^{-1} b_i, \tag{10}$$

with approximation error

$$||x - \tilde{x}|| \leq \frac{C_{st}}{\pi \lambda_{\max}} \exp\left(\frac{2\mu\lambda_{\max}^{-1} + 1}{\pi} - \pi\sqrt{k}\right) \oint_{\Gamma} ||\lambda I - 2\frac{\mathcal{A}}{\lambda_{\max}}||\, \mathrm{d}_{\Gamma}\lambda \;\; ||\bigotimes_{i=1}^{d} E^{-1} b_i||.$$

Obviously, in the special case $d = 2$, the above statement immediately reveals that the solution to the Lyapunov equation

$$AXE^T + EXA^T + BB^T = 0,$$

can be approximated by a low rank matrix $\tilde{X} = LL^T,\ L \in \mathbb{R}^{n\times k}$, with exponentially decreasing approximation error $||X - \tilde{X}||$. The basic ideas for proving the assertion are, on the one hand, the exponential character of the solution matrix $\mathcal{A}^{-1}$ corresponding to a system of linear equations $\mathcal{A}x = \mathcal{B}$ as well as the Dunford-Cauchy representation of the underlying matrix exponential. On the other hand, one can exploit the special tensor structure which allows to decompose the approximant $\tilde{x}$ and thus leads to the above tensor structure. However, for a more detailed analysis, we refer to [19].

Remark 2.2. *The quadrature weights and points from Lemma 2.1 go back to the quadrature formula of Stenger, see e.g. [35]. Note that the constant C_{st} is independent of the individual problem and has been experimentally determined as $C_{st} \approx 2.75$, see [26].*

Remark 2.3. *As has been shown in [26], at least for the symmetric and supersymmetric case, respectively, one can construct even better approximations $\tilde{x}$ that, although depending on the condition number of $\mathcal{A}$, exhibit an exponentially decreasing approximation error which is not slowed down by a square root term.*

A somewhat different explanation for the singular value decay that is dedicated to the 2-dimensional case makes use of the error expression of the ADI iteration and additional properties of Cauchy matrices. Although the corresponding theory is interesting as well, we want to stick with the tensor product formulation and therefore only refer to [2, 30, 34] for these alternative error expressions.

3 Generalized Lyapunov Equations Arising in Bilinear Model Reduction

Now that we have reviewed the main theoretical concepts, we want to discuss the origin of generalized Lyapunov equations of the form (1). For this, let us have a look at an interesting subclass of nonlinear control systems which naturally appear in certain boundary controlled dynamic processes, see e.g. [6, 9]. These so-called bilinear control

systems have already been studied for some years and have the following state-space representation

$$\Sigma : \begin{cases} \dot{x}(t) = Ax(t) + \sum_{j=1}^{m} N_j x(t) u_j(t) + Bu(t), \\ y(t) = Cx(t), \quad x(0) = x_0, \end{cases} \tag{11}$$

where $A,\ N_j \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$. The concepts of reachability and observability, respectively, which are known from linear system theory have been shown (see e.g. [11, 13]) to possess bilinear analogs that can be constructed iteratively as follows. Let

$$\begin{aligned} P_1(t_1) &= e^{At_1}B, \\ P_i(t_1,\ldots,t_i) &= e^{At_i}\begin{bmatrix} N_1P_{i-1} & \ldots & N_mP_{i-1} \end{bmatrix}, \quad i = 2,3,\ldots \end{aligned}$$

Then the reachability Gramian corresponding to (11) is defined as

$$P = \sum_{i=1}^{\infty} \int_0^{\infty} \ldots \int_0^{\infty} P_i P_i^T \mathrm{d}t_1 \cdots \mathrm{d}t_i \tag{12}$$

if it exists. Moreover, P then satisfies the bilinear Lyapunov equation specified in (1). Similarly, the observability Gramian of (11) satisfies the dual equation

$$A^TQ + QA + \sum_{j=1}^{m} N_j^T Q N_j + C^TC = 0.$$

A nice property of these equations is given by their close relation to the problem of model order reduction. As was shown in [6, 12], under certain reasonable assumptions, the solutions P and Q are symmetric positive-definite matrices, offering the possibility of computing Cholesky decompositions $P = UU^*$ and $Q = LL^*$, respectively. By means of a singular value decomposition of the product of the two Cholesky factors $U^*L = ZSY^*$ it is possible to construct a state-space transformation $T = S^{\frac{1}{2}}Z^*U^{-1} = S^{-\frac{1}{2}}Y^*L^{-1}$ which converts the system into a balanced realization, allowing to neglect states that are hard to reach and, at the same time, difficult to observe, without influencing the systems transfer behavior significantly. However, due to their rather complicated structure, solutions of those Lyapunov equations so far can be computed only up to an order of $n \sim 10^3$ – 10^4. Nevertheless, it is often the case that the singular values of P and Q decay rather rapidly, raising the question if there exists a theoretical explanation for this phenomenon.

Although belonging to a quite different class of control systems, Itô-type stochastic linear systems

$$dx = Ax\ dt + \sum_{j=1}^{m} A_j x\ dw_j + Bu\ dt, \quad y = Cx,$$

with $w_j = w_j(t)$ denoting independent zero mean real Wiener processes on a probability space $(\Omega, \mathcal{F}, \mu)$, lead to the same generalized Lyapunov equations, see e.g. [12, 21]. Finally, since one can interpret linear parameter-varying systems as a special class of bilinear systems (see [4]) and hence construct reduced balanced realizations as well, there is certainly a wide range of areas where these Lyapunov equations play a crucial role.

4 Existence of Low Rank Approximations

In order to get a better understanding for the problems that occur in showing the existence of low rank approximations to the solution of equations of the form

$$\left(I \otimes A + A \otimes I + \sum_{j=1}^{m} N_j \otimes N_j \right) \operatorname{vec}(X) = -\operatorname{vec}\left(BB^T\right), \tag{13}$$

let us at first again have a look at the main aspects used in the *standard case*. As we already mentioned in the beginning, one way of constructing low rank approximations is based on the possibility of alternatively considering the approximation of the function

$$f(x_1, x_2) = \frac{1}{x_1 + x_2}.$$

This equivalence is easily seen as follows. Let $A = Q\Lambda Q^{-1}$ be the eigenvalue decomposition of A. Then for the linear Lyapunov equation we have

$$(I \otimes A + A \otimes I) \operatorname{vec}(X) = -\operatorname{vec}\left(BB^T\right)$$

which is the same as

$$(Q \otimes Q)(I \otimes \Lambda + \Lambda \otimes I)\left(Q^{-1} \otimes Q^{-1}\right) \operatorname{vec}(X) = -\operatorname{vec}\left(BB^T\right).$$

However, this means that we can solve the transformed linear system of equations

$$(I \otimes \Lambda + \Lambda \otimes I) \operatorname{vec}\left(\tilde{X}\right) = -\operatorname{vec}\left(\tilde{B}\tilde{B}^T\right), \tag{14}$$

with $\operatorname{vec}\left(\tilde{X}\right) = \left(Q^{-1} \otimes Q^{-1}\right) \operatorname{vec}(X)$ and $\tilde{B} = Q^{-1}B$. In (14), we have to invert a diagonal matrix leading to expressions of the form $\frac{1}{\lambda_i + \lambda_j}$.

Obviously, to obtain an at least similar structure in the bilinear case, one has to impose severe restrictions on the matrices A and N_j. Indeed, what one needs is a simultaneous diagonalization as $A = Q\Lambda Q^{-1}$ and $N_j = Q\Gamma_j Q^{-1}$. As is well-known, see e.g. [22], this means that A and N_j must commute which in practice is almost never the case.

Hence, let us consider what happens if we want to make use of the integral representation of the solution of a system of linear equations. If we denote the coefficient matrix in (5) with M, following [19], the inverse

$$M^{-1} = \int_0^\infty \exp(tM) \mathrm{d}t,$$

can be approximated by

$$\sum_{i=-k}^{k} w_i \exp(t_i M), \tag{15}$$

with the quadrature points t_i and weights w_i as specified in Section 2. Once more, in the *standard case* the computation of the above matrix exponentials (see [22]) boils down to

$$\exp(t_i(I \otimes A + A \otimes I)) = \exp(t_i A) \otimes \exp(t_i A).$$

This in turn means that the approximate inverse of the matrix M is of tensor rank $2k+1$, leading to an approximative solution $\mathrm{vec}(X)$ of tensor rank or, equivalently, of column rank $(2k+1) \cdot m$, where m is the number of columns of B. Again, for the bilinear case there arise some problems. Here, we end up with expressions of the form

$$\exp\left(t_i\left(I \otimes A + A \otimes I + \sum_{j=1}^{m} N_j \otimes N_j\right)\right), \tag{16}$$

where we can neither make an assertion on their tensor ranks nor on the column rank of the solution X. As we can see, the crucial point is that the matrix exponential cannot be split up into its components if the matrices do not commute, i.e.

$$\exp\left(t_i\left(I \otimes A + A \otimes + \sum_{j=1}^{m} N_j \otimes N_j\right)\right) \neq (\exp(t_i A) \otimes \exp(t_i A)) \exp\left(t_i\left(\sum_{j=1}^{m} N_j \otimes N_j\right)\right).$$

However, in case of commutativity and additional low rank structure of the matrices N_j, we obtain a first simple result.

Proposition 4.1. *Let A, $N_j \in \mathbb{R}^{n \times n}$ be diagonalizable and assume they commute. Further assume that $r = \sum_{j=1}^{m} \underbrace{\mathrm{rank}(N_j)}_{r_j} < n$. Then the inverse of*

$$M = I \otimes A + A \otimes I + \sum_{j=1}^{m} N_j \otimes N_j$$

can be approximated by a matrix of tensor rank not greater than $(2k+1) \cdot (r+1)$ with approximation error decreasing proportionally to $\exp(-\pi\sqrt{k})$.

Proof. Due to commutativity, the matrix exponentials given in (15) simplify according to the aforementioned splitting. Thus, we only need to check the exponential term of the N_j summands. However, since we assumed commutativity, all $N_j = T D_j T^{-1}$ can be

diagonalized simultaneously, leading to

$$\exp\left(t_i\Big(\sum_{j=1}^{m} N_j \otimes N_j\Big)\right) = (T\otimes T)\exp\left(t_i\Big(\sum_{j=1}^{m} D_j \otimes D_j\Big)\right)(T\otimes T)^{-1}$$
$$= (T\otimes T)\exp\left(t_i\Big(\sum_{j=1}^{m}\sum_{k=1}^{r_{n_j}} d_{j_{kk}}\, e_{j_{kk}} e_{j_{kk}}^T \otimes D_j\Big)\right)(T\otimes T)^{-1},$$

with j_{kk} denoting the index of the k-th nonzero diagonal entry of D_j. The assertion now trivially follows by the definition of the matrix exponential and the fact that $e_{j_{kk}} e_{j_{kk}}^T$ is an idempotent matrix. □

Remark 4.1. *Note that the idea of the proof is not influenced if we replace the terms $N_j \otimes N_j$ by $N_j \otimes R_j$, where R_j is a matrix of full rank. Moreover, by inspection we observe that an equivalent assertion is true if N_j has full rank and R_j is a low rank matrix.*

Although we already discussed the absence of commutative matrices in practice, Proposition 4.1 not only explains the singular value decay of the solution P of the generalized Lyapunov equation (1), but yields an approximation of low tensor rank to the inverse M^{-1} as well. Obviously, in general this is more complicated than showing the singular value decay of P. However, for our purposes it will be sufficient to show the property for P. Let us now assume that the matrices N_j have a low rank representation given by matrices $U_j,\ V_j \in \mathbb{R}^{n\times r_j}$ s.t. $N_j = U_j V_j^T$. As discussed in [12], we can make use of the splitting

$$\underbrace{I\otimes A + A\otimes I}_{\mathcal{L}} + \sum_{j=1}^{m} N_j \otimes N_j$$

in order to apply the Sherman-Morrison-Woodbury formula which will help us to prove our main result.

Theorem 4.1. *Let $\mathcal{A}$ denote a matrix of tensor product structure as in* (13) *with right-hand side $\mathcal{B} = -\operatorname{vec}\left(BB^T\right)$. Assume that the spectrum of $\mathcal{A}$ is contained in the strip $\Omega := [\lambda_{\min}, \lambda_{\max}] \oplus i[-\mu,\mu] \subseteq \mathbb{C}_-$. Let further $N_j = U_j V_j^T$, with $U_j, V_j \in \mathbb{R}^{n\times r_j}$ and let $r = \sum_{j=1}^{m} r_j$ and $U = \begin{bmatrix} U_1\otimes U_1 & \dots & U_m \otimes U_m\end{bmatrix}$ and $V = \begin{bmatrix} V_1\otimes V_1 & \dots & V_m\otimes V_m\end{bmatrix}$. Then the solution x to $\mathcal{A}x = \mathcal{B}$ can be approximated by a vector of tensor rank $(2\cdot k+1)\cdot(m+r)$ of the form*

$$\tilde{x} := \sum_{\ell=-k}^{k} \frac{2w_\ell}{\lambda_{\max}}\left(\exp\Big(\frac{2t_\ell}{\lambda_{\max}}A\Big)\otimes\exp\Big(\frac{2t_\ell}{\lambda_{\max}}A\Big)\right)\begin{bmatrix}\mathcal{B} & -U\mathcal{Y}\end{bmatrix}, \tag{17}$$

where $\mathcal{Y}$ is the solution of

$$\left(I_{r^2} + V^T\mathcal{L}^{-1}U\right)\mathcal{Y} = V^T\mathcal{L}^{-1}\mathcal{B} \tag{18}$$

and w_ℓ, t_ℓ are the quadrature weights and points from Lemma 2.1. The corresponding approximation error is given as

$$\begin{aligned} ||x-\tilde{x}||_2 \leq \frac{C_{st}}{\pi\lambda_{\max}} \exp\left(\frac{2\mu\lambda_{\max}^{-1}+1}{\pi} - \pi\sqrt{k}\right) \oint_\Gamma ||\lambda I - 2\frac{\mathcal{A}}{\lambda_{\max}}||_F \, \mathrm{d}_\Gamma\lambda \\ \times ||BB^T + \sum_{j=1}^{m} U_j \operatorname{vec}^{-1}\left(\mathcal{Y}_{r_j}\right) U_j^T||_F, \end{aligned} \tag{19}$$

where $\mathcal{Y}_{r_j}$ denotes the r_j^2 elements ranging from $\sum_{i=1}^{j-1} r_i^2 + 1$ to $\sum_{i=1}^{j} r_i^2$.

Proof. Let us consider the tensor structure

$$\Big(\underbrace{I\otimes A + A\otimes I}_{\mathcal{L}} + \underbrace{\sum_{j=1}^{m} N_j \otimes N_j}_{UV^T}\Big)x = \mathcal{B}.$$

Making use of the low rank structure and the Sherman-Morrison-Woodbury formula, the computation of the inverse of $\mathcal{A}$ simplifies to

$$\mathcal{A}^{-1} = \mathcal{L}^{-1} - \mathcal{L}^{-1}U\left(I_{r^2} + V^T\mathcal{L}^{-1}U\right)^{-1}V^T\mathcal{L}^{-1}.$$

Hence, solving $\mathcal{A}x = \mathcal{B}$ is equivalent to solving

$$(I\otimes A + A\otimes I)\,x = \mathcal{B} - U\underbrace{\left(I_{r^2} + V^T\mathcal{L}^{-1}U\right)^{-1}V^T\mathcal{L}^{-1}\mathcal{B}}_{\mathcal{Y}}.$$

However, the last equation is a standard Lyapunov equation for which we can apply the results from Lemma 2.1. Nevertheless, for the assertion on the tensor rank of $\tilde{x}$, it remains to show that the tensor rank of $\mathcal{B} - U\mathcal{Y}$ is $m + r$. This is easily seen by the definition of $U = \begin{bmatrix} U_1\otimes U_1 & \dots & U_m\otimes U_m \end{bmatrix}$. In fact, what we obtain is

$$\begin{aligned} U\mathcal{Y} &= \begin{bmatrix} U_1\otimes U_1 & \dots & U_m\otimes U_m \end{bmatrix}\mathcal{Y} = \sum_{j=1}^{m} U_j \underbrace{\operatorname{vec}^{-1}\left(\mathcal{Y}_{r_j}\right)U_j^T}_{:=Y_j^T} \\ &= \sum_{j=1}^{m}\sum_{i=1}^{r_j} U_{j,i}Y_{j,i}^T = \sum_{j=1}^{m}\sum_{i=1}^{r_j} Y_{j,i}\otimes U_{j,i}. \end{aligned}$$

In the last line, the second subscript i denotes the i-th column of the matrices. By assumption the r_j sum up to r, leading to a tensor rank of $(2\cdot k+1)\cdot(m+r)$. The approximation error trivially follows by the same inversion of the vec ()-operator and applying well-known linear results, e.g. [2, 19, 30], for a modified right-hand side $-\operatorname{vec}\left(BB^T\right) - U\mathcal{Y}$. □

Remark 4.2. *Obviously, there exist special cases where the N_j are full-rank matrices and we still can expect a strong singular value decay of the solution X. Here, one might think of*

$$AX + XA^T + AXA^T + BB^T = 0,$$

or the even easier case

$$AX + XA^T + X + BB^T = 0.$$

Both of the above equations reduce to a modified linear Lyapunov equation with right-hand side of rank m. However, note that each time N commutes with A. Nevertheless, so far it remains an open question if it is possible to extend decay results for a more general setting as well. The numerical results shown in Section 6 indicate that there seem to be conditions for low rank properties also in this case.

Although for the higher dimensional case the tensor rank will increase exponentially with the dimensions, it might be worth noting that we can still expect low rank approximations as stated in the following corollary. For this, let

$$\mathcal{L}_d = \sum_{i=1}^{d} E_1 \otimes \cdots \otimes E_{i-1} \otimes A_i \otimes E_{i+1} \otimes \cdots \otimes E_d.$$

Corollary 4.1. *Let $\mathcal{A}$ denote a matrix of tensor product structure as in (5) with tensor right-hand side $\mathcal{B} = \bigotimes_{i=1}^{d} b_i$ and $N_{j_\ell} = N_j$, with $\operatorname{rank}(N_j) = r_j$. Assume that the sum of the spectra of the $E_i^{-1}A_i$ is contained in the strip $\Omega := [\lambda_{\min}, \lambda_{\max}] \oplus i[-\mu, \mu] \subseteq \mathbb{C}_-$. Let further $N_j = U_j V_j^T$, with $U_j, V_j \in \mathbb{R}^{n \times r_j}$ and let $r = \sum_{j=1}^{m} r_j$ and $U = \begin{bmatrix} \bigotimes_{i=1}^{d} U_1 & \dots & \bigotimes_{i=1}^{d} U_m \end{bmatrix}$ and $V = \begin{bmatrix} \bigotimes_{i=1}^{d} V_1 & \dots & \bigotimes_{i=1}^{d} V_m \end{bmatrix}$. Then the solution x to $\mathcal{A}x = \mathcal{B}$ can be approximated by a vector of tensor rank $(2 \cdot k + 1) \cdot (m + r^{d-1})$ of the form*

$$\tilde{x} := \sum_{\ell=-k}^{k} \frac{2w_\ell}{\lambda_{\max}} \bigotimes_{i=1}^{d} \exp\left(\frac{2t_\ell}{\lambda_{\max}} E_i^{-1} A_i \right) \begin{bmatrix} \mathcal{B} & -U\mathcal{Y} \end{bmatrix}, \tag{20}$$

where $\mathcal{Y}$ is the solution of

$$\left(I_{r^d} + V^T \mathcal{L}_d^{-1} U \right) \mathcal{Y} = V^T \mathcal{L}_d^{-1} \mathcal{B} \tag{21}$$

and w_ℓ, t_ℓ are the weights from Lemma 2.1. The corresponding approximation error is given as

$$\begin{aligned} ||x - \tilde{x}||_2 \leq \frac{C_{st}}{\pi \lambda_{\max}} \exp\left(\frac{2\mu \lambda_{\max}^{-1} + 1}{\pi} - \pi\sqrt{k} \right) \oint_\Gamma ||\lambda I - 2\frac{\mathcal{A}}{\lambda_{\max}}||_F \, \mathrm{d}_\Gamma \lambda \\ \times ||\mathcal{B} + \sum_{j=1}^{m} \left(\bigotimes_{i=1}^{d} U_j \right) \mathcal{Y}||_2. \end{aligned} \tag{22}$$

Proof. The assertion on the tensor rank easily follows by iteratively applying the procedure from the proof of Theorem 4.1 to the terms $\left(\bigotimes_{i=1}^{d} U_j\right) \mathcal{Y}$, e.g. for $d = 3$, we obtain

$$(U_j \otimes U_j \otimes U_j)\, \mathcal{Y} = \left[U_{j_1} \otimes (U_j \otimes U_j)\, \mathcal{Y}_1 \quad \ldots \quad U_{j_r} \otimes (U_j \otimes U_j)\, \mathcal{Y}_r\right].$$

Since each of the terms $(U_j \otimes U_j)\, \mathcal{Y}_i$ is of tensor rank r, it is clear that $(U_j \otimes U_j \otimes U_j)\, \mathcal{Y}$ is of tensor rank r^2. All other results can be proved analogously as before. □

Remark 4.3. *Though the rank of the approximation increases exponentially with d, so does the maximum possible tensor rank which is n^{d-1}. Hence, the ratio between full and approximate solution is $\sim \left(\frac{r}{n}\right)^{d-1}$.*

5 Low Rank Solution Methods

Now that we have seen that we indeed can expect a singular value decay of the solution matrix X of (1), we want to discuss possible extensions of existing linear low rank Lyapunov solvers that have been proven to yield accurate low rank approximations $LL^T \approx X$. Here, we will point out the LRCF-ADI iteration, KPIK together with the more general rational Krylov framework and finally approaches that rely on solving the explicit linear system in tensorized form by iterative solvers like e.g. BiCGstab. As has been pointed out in [12], for the generalized Lyapunov equation

$$\underbrace{AX + XA^T}_{\mathcal{L}} + \underbrace{\sum_{j=1}^{m} N_j X N_j^T}_{\Pi} + BB^T = 0,$$

it makes sense to demand that the spectral radius $\rho\left(\mathcal{L}^{-1}\Pi\right) < 1$ since we otherwise cannot ensure that X is positive semi-definite. However, at least in the bilinear case there exist a lot of interesting applications that lead to indefinite solution matrices X and we therefore will address problems that might occur in these cases.

5.1 The Low Rank ADI iteration

Let us now focus on the low rank version of the alternating directions implicit iteration. We will take for granted that the reader is familiar with the main concepts in the *standard case* which go back to solving elliptic and parabolic partial differential equations (see [36, 37]). In general, the main idea is that for any parameter $p > 0$, the Lyapunov operator $\mathcal{L}$ can be shifted according to

$$AX + XA^T = \frac{1}{2p}\left((A + pI)X(A + pI)^T - (A - pI)X(A - pI)^T\right). \tag{23}$$

In [12], for a given set of shift parameters $\{p_0, p_1, \dots\}$, this circumstance was used to solve (1) via the following fixed-point iteration

$$\begin{aligned} X_{k+1} = (A - p_k I)^{-1}(A + p_k I)X_k(A + p_k I)^T(A - p_k I)^{-T} \\ + 2p_k(A - p_k I)^{-1}\left(\sum_{j=1}^{m} N_j X_k N_j^T + BB^T\right)(A - p_k I)^{-T}. \end{aligned}$$

However, for dimensions n larger than 10^3 the above scheme will not be feasible since in each step we have to solve a linear system with a matrix right-hand side which might easily become too expensive. Moreover, for even larger dimensions, the simple storing of the generally dense matrix X_k will cause serious memory problems. On the other hand, we can take advantage of the fact that we know the solution matrix X is symmetric and, according to the previous section, tends to have a strong singular value decay as well. For this reason, as in the *standard case*, suggested in [7, 29, 30], instead of the full-rank version, it is reasonable to start with a symmetric initial guess, e.g. $X_0 = BB^T$, and then only compute the low rank factors Z_k according to

$$Z_{k+1} = \begin{bmatrix} (A - p_k I)^{-1}(A + p_k I)Z_k & \sqrt{2p_k}(A - p_k I)^{-1}N_j Z_k & \sqrt{2p_k}(A - p_k I)^{-1}B \end{bmatrix}.$$

Obviously, the advantage is that we now only have to solve $2 + m$ systems of linear equations with low rank right-hand side. In the *standard case*, it has been shown that the iteration can be rewritten in such a way that $Z_{k+1} = \begin{bmatrix} Z_k & V_k \end{bmatrix}$, with $V_k \in \mathbb{R}^{n \times m}$, making an appropriate algorithm much cheaper to evaluate. Unfortunately, due to non-commutativity of A and N_j, in our case this is not possible. If we assume that the iterate Z_k consists of r columns, at least theoretically Z_{k+1} consists of $(m+1) \cdot r + m$ columns. However, we often obtain a deflation in the column spaces such that a so-called column compression can prevent a too strong column increase. Another problem might arise in case of the already mentioned absence of a convergent splitting which is quite common for real-life examples of bilinear control systems. Here, it should be noted that the ADI iteration will not converge and we therefore recommend the use of one of the other low rank solvers which we will discuss in the next subsections. Let us now briefly take a look at the choice of the shift parameters p_k. While for linear systems the search for a set of q optimal parameters is equivalent to the rational min-max problem

$$\min_{\{p_1,\dots,p_q\}} \max_{\lambda \in \sigma(A)} \prod_{\ell=1}^{q} \left| \frac{\lambda - p_\ell}{\lambda + p_\ell} \right|,$$

for the generalized version from above, we have to take into account that the additional N_j matrices will influence the speed of convergence. In more detail, let us consider the case $n = 1$ and $j = 1$, s.t. we are looking at the scalar equation

$$AX + XA^T + NXN^T + BB^T = (2A + N^2)X + B^2 = 0.$$

In order to obtain convergence of the ADI iteration after the first step, we want to have that $X_1 = X$. However, this means that we have to compute a p which fulfills

$$2p(A - pI)^{-1}BB^T(A - pI)^{-T} = -\frac{B^2}{2A + N^2}.$$

Solving this equation for p, the optimal parameter p_{opt} can be easily derived to satisfy

$$|p_{opt} + (A + N^2)| = N\sqrt{2A + N^2}.$$

Hence, once more we have to realize that the non-commutativity of A and N_j does not allow to generalize linear concepts like the min-max problem in an obvious way. However, for linear systems it has recently been observed that the interpolation points which minimize the $\mathcal{H}_2$-norm between the original and a reduced-order system often turn out to be good shifts for the ADI iteration as well, see e.g. [14, 17, 18]. A more detailed explanation for this phenomenon is given in [5], where it is actually shown that for symmetric state space systems these points are optimal w.r.t. a certain energy norm naturally induced by the Lyapunov operator. Since the $\mathcal{H}_2$-interpolation problem was shown to possess a bilinear analog (see [3]), in Section 6, we will show that the corresponding interpolation points have a positive effect on the convergence rate of the bilinear ADI iteration as well.

5.2 Low Rank Solutions by Projection

A somewhat different approach for obtaining low rank approximate solutions of (2) is based on Krylov subspace methods. Following [32], one constructs a projection matrix $V \in \mathbb{R}^{n\times q}$ and, due to a Galerkin condition on the residual, solves a reduced Lyapunov equation of much smaller size which is determined by

$$V^T A V \hat{P} + \hat{P} V^T A^T V + V^T B B^T V = 0,$$

with $\hat{P} \in \mathbb{R}^{q\times q}$. Assuming that A is dissipative, i.e. $x^T(A + A^T)x < 0$, one can show that the reduced Lyapunov equation possesses a unique symmetric positive definite solution $\hat{P} = \hat{P}^T$ and the approximation to the full solution then is given by $P \approx V\hat{P}V^T$. The question arises how to choose the subspace V which should contain as much information of the solution subspace as possible. For linear systems, as a fast and reliable method, one should certainly mention the Krylov-plus-inverse-Krylov (KPIK) which has been introduced in [33]. Here, one computes the two (block)-Krylov subspaces

$$\mathcal{K}_q(A, B), \quad \mathcal{K}_q(A^{-1}, A^{-1}B)$$

and then constructs V as an orthogonal basis of the union of the corresponding column spaces. Alternatively, this may be achieved by the following iterative procedure

$$\begin{aligned} V_1 &= [B, A^{-1}B], \\ V_i &= [AV_{i-1}, A^{-1}V_{i-1}], \quad i \le q. \end{aligned}$$

Usually, the above subspaces are generated by a modified Gram-Schmidt process which leads to orthonormal bases in each step. In order to extend the approach to our generalized setting, we suggest to proceed as follows

$$\begin{aligned} V_1 &= [B, A^{-1}B], \\ V_i &= [AV_{i-1}, A^{-1}V_{i-1}, N_jV_{i-1}], \quad i \le q,\ j = 1, \ldots, m. \end{aligned}$$

Again, the Galerkin condition demands an orthogonal V, such that we have $V := \operatorname{orth}(V_q)$. Moreover, similar to the ADI iteration discussed in the previous section, one should perform a column compression which keeps the rank increase in each step at a compatible level. Analog to the discussions on the standard case given in [23, 24, 32, 33], one can use the nestedness of the subspaces generated during the process to prove the following useful result that allows to replace the computation of the residual $R_i \in \mathbb{R}^{n\times n}$ by a matrix of smaller dimension.

Theorem 5.1. *Let* $R_i := AX_i + X_iA^T + \sum_{j=1}^m N_jX_iN_j^T + BB^T$ *denote the residual associated with the approximate solution* $X_i = V_i\hat{X}_iV_i^T$, *where* $\hat{X}_i$ *is the solution of the reduced Lyapunov equation*

$$V_i^TAV_i\hat{X}_i + \hat{X}_iV_i^TA^TV_i + \sum_{j=1}^m V_i^TN_jV_i\hat{X}_iV_i^TN_j^TV_i + V_i^TBB^TV_i = 0.$$

Then it holds $\operatorname{range}(R_i) = \operatorname{range}(V_{i+1})$ *and* $||R_i|| = ||V_{i+1}^TR_iV_{i+1}||$.

Proof. The first assertion follows from the fact that, due to the iterative construction of V_{i+1}, we have

$$V_i \subset V_{i+1},\ AV_i \subset V_{i+1},\ N_jV_i \subset V_{i+1}.$$

Moreover, with the same argument and the orthonormality of V_{i+1}, it holds

$$R_i = V_{i+1}V_{i+1}^TR_iV_{i+1}V_{i+1}^T.$$

This implies $||R_i|| = ||V_{i+1}^TR_iV_{i+1}||$. □

Note that in contrast to the *standard case* it seems to be impossible to further simplify the expression for the residual. Here, the problem is that the Hessenberg structure of the projected system matrix $T = V_i^TAV_i$ is lost.

Another possibility is to project onto a rational Krylov subspace

$$V = [(\sigma_1I - A)^{-1}B, \ldots, (\sigma_qI - A)^{-1}B],$$

where $\sigma_1, \ldots, \sigma_q$ are prespecified interpolation points. Although the computational costs are higher, this method allows to speed up convergence rates significantly, provided a clever choice of the interpolation points is known. For the *standard case*, a detailed analysis on this topic can be found in [14]. As has been shown in [3], the rational Krylov framework can be extended to the bilinear case by making use of the idea of tangential interpolation. To be more specific, if interpolation points $\sigma_1, \ldots, \sigma_q$ together with tangential directions $\tilde{B} \in \mathbb{R}^{q\times m}$ and $\tilde{N}_j \in \mathbb{R}^{q\times q}$ are chosen, the projection matrix V is given by

$$\operatorname{vec}(V) = \left(\operatorname{diag}(\sigma_1, \ldots, \sigma_q) \otimes I_n - I_q \otimes A - \sum_{j=1}^m \tilde{N}_j \otimes N_j\right)^{-1} \left(\tilde{B} \otimes B\right) \operatorname{vec}(I_m).$$

However, so far there is no theory about a priori choices of good interpolation points and at this point we thus leave it as a topic of further research. Before we proceed, we want to point out that the previous method does not require $\sigma\left(\mathcal{L}^{-1}\Pi\right) < 1$. This is due to the fact that the reduced Lyapunov equation is of much smaller size and thus might be solved by forming the explicit system of linear equations.

5.3 Iterative Linear Solvers

Finally, in this section we want to address the possibility of efficiently solving the tensorized linear system of equations (13) by iterative solvers like CG (symmetric case) or BiCGstab (unsymmetric case). Here, the crucial point is to note that we can incorporate the to-expected low rank structure of P into the algorithm which will allow to reduce the complexity significantly.

The Symmetric Case

Since a quite similar discussion for more general tensorized linear systems can be found in [27], we will follow the notations therein and only briefly discuss how to adapt the main concepts to our purposes. Assuming that the matrices A and N_j are symmetric, we can modify the preconditioned CG method. For this, let us have a look at Algorithm 1 which has already been studied in [27] in the context of solving equations of the form (4).

Algorithm 1 Preconditioned CG method

Input: Matrix functions $\mathcal{A}, \mathcal{M} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, low rank factor B of right-hand side $\mathcal{B} = -BB^T$. Truncation operator $\mathcal{T}$ w.r.t. relative accuracy ϵ_{rel}.

Output: Low rank approximation $X = LDL^T$ with $||\mathcal{A}(X) - \mathcal{B}||_F \leq$ tol.

$X_0 = 0$, $R_0 = \mathcal{B}$, $Z_0 = \mathcal{M}^{-1}(R_0)$, $P_0 = Z_0$, $Q_0 = \mathcal{A}(P_0)$, $\xi_0 = \langle P_0, Q_0\rangle$, $k = 0$
while $||R_k||_F >$ tol **do**
  $\omega_k = \frac{\langle R_k, P_k\rangle}{\xi_k}$
  $X_{k+1} = X_k + \omega_k P_k$, $\quad X_{k+1} \leftarrow \mathcal{T}(X_{k+1})$
  $R_{k+1} = \mathcal{B} - \mathcal{A}(X_{k+1})$, $\quad$ *Optionally:* $R_{k+1} \leftarrow \mathcal{T}(R_{k+1})$
  $Z_{k+1} = \mathcal{M}^{-1}(R_{k+1})$
  $\beta_k = -\frac{\langle Z_{k+1}, Q_k\rangle}{\xi_k}$
  $P_{k+1} = Z_{k+1} + \beta_k P_k$, $\quad P_{k+1} \leftarrow \mathcal{T}(P_{k+1})$
  $Q_{k+1} = \mathcal{A}(P_{k+1})$, $\quad$ *Optionally:* $Q_{k+1} \leftarrow \mathcal{T}(Q_{k+1})$
  $\xi_{k+1} = \langle P_{k+1}, Q_{k+1}\rangle$
  $k = k + 1$
end while
$X = X_k$

The application of the matrix function $\mathcal{A}$ to a matrix X here should denote the operation $AX + XA^T + \sum_{j=1}^m N_j X N_j^T$. As a preconditioner $\mathcal{M}^{-1}$ we will use the low rank version of the bilinear ADI iteration which we studied in Subsection 5.1, whereas the

truncation operator $\mathcal{T}$ should be understood as a simple column compression as described in e.g. [27]. The only point to clarify is that we indeed can ensure a decomposition $X_k = L_k D_k L_k^T$, with diagonal matrix D, in each step of the algorithm. We start with $R_0 = \mathcal{B} = -BB^T$ which obviously can be decomposed into $R_0 = L_{R_0} D_{R_0} L_{R_0}^T$ by setting $L_{R_0} = B$ and $D_{R_0} = -I_m$. Next, we note that the bilinear ADI iteration is not restricted to a factorization of the form ZZ^T but can also be applied to low rank decompositions LDL^T, see [8]. This is easily seen as follows. Recalling the iteration procedure, we formally assume that $Z_k = L_k\sqrt{D_k}$ and obtain the new iterate

$$Z_{k+1} = (A - p_k I)^{-1} \left[(A + p_k I) L_k \sqrt{D_k} \quad \sqrt{2p_k} N_j L_k \sqrt{D_k} \quad \sqrt{2p_k} L \sqrt{D} \right],$$

where $L\sqrt{D}$ is the initial input to the ADI iteration. However, forming the product $Z_{k+1} Z_{k+1}^T$, it is clear that we can replace the step by setting

$$\begin{aligned} L_{k+1} &= (A - p_k I)^{-1} \left[(A + p_k I) L_k \quad \sqrt{2p_k} N_j L_k \quad \sqrt{2p_k} L \right], \\ D_{k+1} &= \begin{bmatrix} D_k & 0 & 0 \\ 0 & D_k & 0 \\ 0 & 0 & D \end{bmatrix}. \end{aligned}$$

Now we only have to check for a possible decomposition of the matrix which is returned after applying the matrix function $\mathcal{A}$ to a factorized matrix LDL^T. By the definition of $\mathcal{A}$, it follows

$$\begin{aligned} \mathcal{A}(LDL^T) &= ALDL^T + LDL^T A^T + \sum_{j=1}^{m} N_j L D L^T N_j^T \\ &= \underbrace{\begin{bmatrix} AL & L & N_j L \end{bmatrix}}_{\hat{L}} \underbrace{\begin{bmatrix} 0 & D & 0 \\ D & 0 & 0 \\ 0 & 0 & D \end{bmatrix}}_{\hat{D}} \underbrace{\begin{bmatrix} AL & L & N_j L \end{bmatrix}^T}_{\hat{L}^T}. \end{aligned}$$

Even though $\hat{D}$ is not a diagonal matrix any more, it is a symmetric monomial matrix. However, in this case $\hat{L}\hat{D}\hat{L}^T$ is also symmetric and thus can be factorized by $\tilde{L}\tilde{D}\tilde{L}^T$, where $\tilde{D}$ again is diagonal. All other computations in Algorithm 1 do not influence the diagonal structure of D allowing to preserve the desired factorization and solely operate on the low rank factors L and D, respectively.

The Unsymmetric Case

Similarly, one might implement more sophisticated algorithms, which are also applicable in the case that A and N_j are unsymmetric. Obviously, there are numerous possible iterative solvers which can be used. However, in this paper we will restrict ourselves to the BiCGstab algorithm. Again, we refer to [27], for a similar discussion on Algorithm 2. Once more, note that the only difference is that our version here is dedicated to solving equations of the form (1) which has to be taken care of in evaluating $\mathcal{A}$ and the

special preconditioner $\mathcal{M}^{-1}$ given by the bilinear ADI iteration. As has been discussed in [15, 16] for the standard case, unsymmetric matrices might also be tackled by a low rank variant of the GMRES method together with a suitable preconditioning technique.

Just as solving the Lyapunov equation by a projection onto a smaller subspace, the use of an iterative linear solver has the advantage that we do not need the assumption $\sigma\left(\mathcal{L}^{-1}\Pi\right) < 1$ as long as we refrain from preconditioning with the bilinear ADI iteration which in case of $\sigma\left(\mathcal{L}^{-1}\Pi\right) \geq 1$ will not converge. If this is the case, we can still precondition with a number of linear ADI iterations which we assume to be at least a rough approximation to the inverse of the bilinear Lyapunov operator, see also the discussion in [12].

Algorithm 2 Preconditioned BiCGstab method

Input: Matrix functions $\mathcal{A}, \mathcal{M} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, low rank factor B of right-hand side $\mathcal{B} = -BB^T$. Truncation operator $\mathcal{T}$ w.r.t. relative accuracy ϵ_{rel}.

Output: Low rank approximation $X = LDL^T$ with $||\mathcal{A}(X) - \mathcal{B}||_F \leq$ tol.

$X_0 = 0$, $R_0 = \mathcal{B}$, $\tilde{R} = \mathcal{B}$, $\rho_0 = \langle \tilde{R}, R_0 \rangle$, $P_0 = R_0$, $\hat{P}_0 = \mathcal{M}^{-1}(P_0)$, $V_0 = \mathcal{A}(\hat{P}_0)$, $k = 0$
while $||R_k||_F >$ tol **do**
  $\omega_k = \frac{\langle \tilde{R}, R_k \rangle}{\langle \tilde{R}, V_k \rangle}$,
  $S_k = R_k - \omega_k V_k$ — *Optionally:* $S_k \leftarrow \mathcal{T}(S_k)$
  $\hat{S}_k = \mathcal{M}^{-1}(S_k)$, — *Optionally:* $\hat{S}_k \leftarrow \mathcal{T}(\hat{S}_k)$
  $T_k = \mathcal{A}(\hat{S}_k)$, — *Optionally:* $T_k \leftarrow \mathcal{T}(T_k)$
  if $||S_k||_F \leq$tol **then**
    $X = X_k + \omega_k \hat{P}_k$,
    return,
  end if
  $\xi_k = \frac{\langle T_k, S_k \rangle}{\langle T_k, T_k \rangle}$,
  $X_{k+1} = X_k + \omega_k \hat{P}_k + \xi_k \hat{S}_k$, — $X_{k+1} \leftarrow \mathcal{T}(X_{k+1})$
  $R_{k+1} = \mathcal{B} - \mathcal{A}(X_{k+1})$, — *Optionally:* $R_{k+1} \leftarrow \mathcal{T}(R_{k+1})$
  if $||R_k||_F \leq$tol **then**
    $X = X_k$,
    return,
  end if
  $\rho_{k+1} = \langle \tilde{R}, R_{k+1} \rangle$,
  $\beta_k = \frac{\rho_{k+1}}{\rho_k} \frac{\omega_k}{\xi_k}$,
  $P_{k+1} = R_{k+1} + \beta_k (P_k - \xi_k V_k)$, — $P_{k+1} \leftarrow \mathcal{T}(P_{k+1})$
  $\hat{P}_{k+1} = \mathcal{M}^{-1}(P_{k+1})$, — *Optionally:* $\hat{P}_{k+1} \leftarrow \mathcal{T}(\hat{P}_{k+1})$
  $V_{k+1} = \mathcal{A}(\hat{P}_{k+1})$, — *Optionally:* $V_{k+1} \leftarrow \mathcal{T}(V_{k+1})$
  $k = k + 1$
end while

6 Numerical Examples

In this section, we will now study the performance of the proposed methods by means of some standard numerical test examples. The first and the second benchmark fulfill the assumptions stated in Theorem 4.1, meaning that the bilinear coupling matrix N is of low rank compared to the system dimension n. Hence, we know that we can indeed expect low rank approximations of the generalized Lyapunov equations as well. However, the third benchmark contains a coupling matrix N which has full rank. Nevertheless, we show that there still seems to be a significant singular value decay in the solution matrix X which allows for low rank approximations. All simulations were generated on an Intel®Xeon®Westmere X5650 with 2.66GHz, 48GB DDR3 RAM and MATLAB® Version 7.11.0.584 (R2010b) 64-bit (glnxa64).

Heat equation

The first example we want to discuss is the heat equation subject to the following mixed boundary conditions

$$\begin{aligned} x_t &= \Delta x && \text{in } \Omega = (0,1) \times (0,1), \\ n \cdot \nabla x &= 0.5 \cdot u(x-1) && \text{on } \Gamma_1, \\ x &= 0 && \text{on } \Gamma_2, \Gamma_3, \Gamma_4, \end{aligned}$$

where $\Gamma_1, \Gamma_2, \Gamma_3$ and Γ_4 denote the boundaries of Ω. Since the above benchmark has been studied a couple of times in the literature, we omit a detailed description of the model and instead only refer to e.g. [6, 9, 12]. As is shown in Figure 1, we solve the generalized Lyapunov equation up to a system dimension of $n = 562\,500$, corresponding to a grid consisting of 750 grid points in each direction. For the bilinear extension of the ADI iteration, we test several choices for the shift parameters. As previously indicated, we compare the interpolation points resulting from a locally $\mathcal{H}_2$-optimal reduced order model obtained by the procedure proposed in [3] with the optimal shifts for the *standard case* derived by Wachspress, see [37]. While for a smaller number of parameters, the convergence rate is worse than for the optimal parameters for the linear problem, for an increasing number of $\mathcal{H}_2$-interpolation points, the convergence improves significantly. Note that this effect has already been observed for the *standard* Lyapunov equations, see [5]. Furthermore, in Figure 1, we see that the approximations obtained by using a low rank implementation of the CG method perform the best for this specific example. Note that the rank of the final iterate is only 63, while the corresponding relative residual is smaller than 10^{-9}. On the other hand, the extension of the KPIK method stagnates at a relative residual of the order 10^{-2}.

A nonlinear RC circuit

Our second example has also been studied several times in the literature, especially in the context of nonlinear model order reduction. The system is a scalable RC ladder with k resistors whose voltage-current dependency is given by an exponential term.

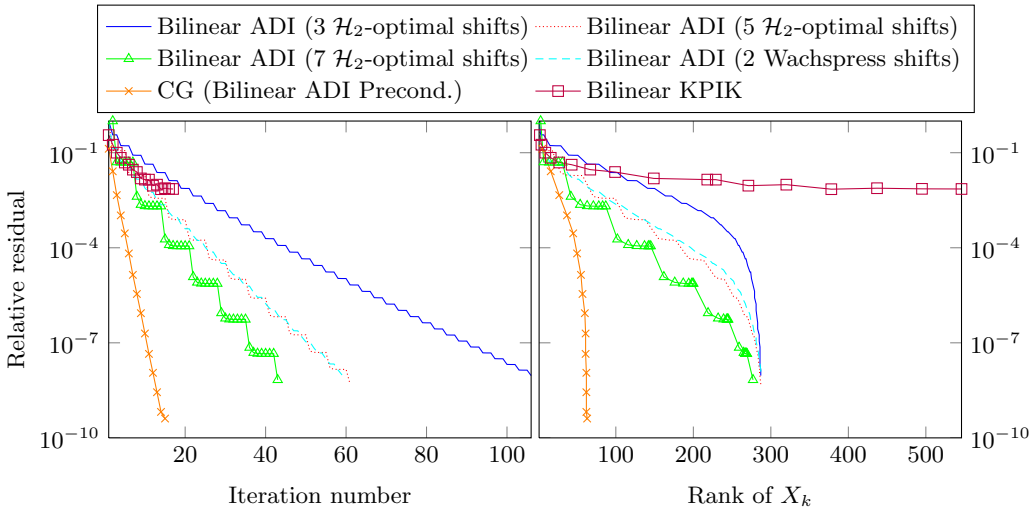

Figure 1: **Heat equation**. Comparison of low rank solution methods for $n = 562\,500$.

Since the original system is in fact nonlinear, we perform a second order Carleman bilinearization, see also [6, 31], leading to a system dimension $n = k + k^2$. Here, the bilinearization process leads to a bilinear coupling matrix N which is only of rank k. The computations were done for $k = 500$ and consequently $n = 250\,000$. Moreover, we scale the matrix N by a factor of 0.5 in order to ensure a positive semi-definite solution P of the associated generalized Lyapunov equation. In Figure 2, we again compare the performance of the bilinear ADI iteration for two different sets of shift parameters. In contrast to the previous example, here the optimal linear parameters proposed by Wachspress clearly outperform a larger set of $\mathcal{H}_2$-optimal shifts. Although the latter ones lead to a fast decrease of the residual within the first steps of the algorithm, we then obtain an almost stagnating residual curve which after 300 steps does not reach a standard stopping criterion of 10^{-8}. Moreover, in Figure 2, we see the results for two different preconditioners for the low rank implementation of the BiCGstab method. The first one is the low rank version of the bilinear ADI iteration which we have previously discussed in detail and for which we only compute the first two iterates in each step of the BiCGstab algorithm. The same is done for the standard low rank ADI iteration which we expect to approximate only the inverse of the standard Lyapunov operator. However, as can be seen in Figure 2, there is no visible advantage which might allow recommending the first method. In fact, using the standard ADI iteration as a preconditioner results in a much smaller final low rank approximation which at least in this example should thus be preferred. Nevertheless, since both variants converged to a relative residual of 10^{-8} after 9 iterations, they seem to be reasonable choices for preconditioning. Finally, the extension of the KPIK method converged to a relative residual of 10^{-7}. It is interesting to note that the ranks of the approximations decrease at the later stage of the algorithm. This is due to the fact that we solved the reduced Lyapunov equation by means of the bilinear ADI iteration as well so that in some cases the ranks of the solutions can be

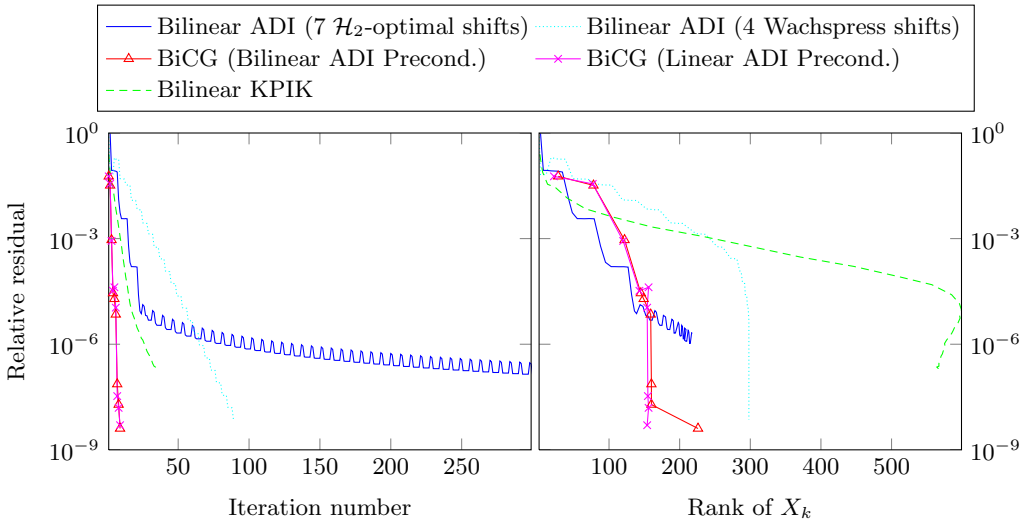

Figure 2: **RC circuit**. Comparison of low rank solution methods for $n = 250\,000$.

further reduced.

Fokker-Planck equation

In order to show that in some cases one might obtain a fast singular value decay even if the bilinear coupling matrix is of full rank, as a final example we consider an application from stochastic control which was already studied in [21]. There the authors discuss a model for a dragged Brownian particle whose one-dimensional motion is described by the stochastic differential equation

$$dX_t = -\nabla V(X_t, t)dt + \sqrt{2\sigma}dW_t,$$

with $V(x, u) = W(x, t) + \Phi(x, u_t) = (x^2 - 1)^2 - xu - x$. Here, we use $\sigma = \frac{1}{2}$ and spatially discretize the underlying probability distribution function with $n = 10\,000$ points. As shown in [21], this setting leads to a bilinear matrix N of rank 10 000. In Figure 3, we see the convergence history for the bilinear ADI iteration using only 2 Wachspress shifts. The relative residual of the final iterate is 10^{-10} while the rank is only 31, indicating that the full solution X indeed exhibits a very strong singular value decay. However, at least for this example, using $\mathcal{H}_2$-optimal shifts for the iteration does not lead to satisfying convergence results and, hence, are neglected in the figure. On the other hand, the low rank implementation of the BiCGstab method as well as the bilinear KPIK variant converged to relative residuals of 10^{-10}, although the approximation of the latter one resulted in having the largest rank.

Remarks on the Computational Complexity

Based on the above results, at a first glance it seems reasonable to recommend the use of an iterative linear solver since the number of iterations as well as the rank of the final

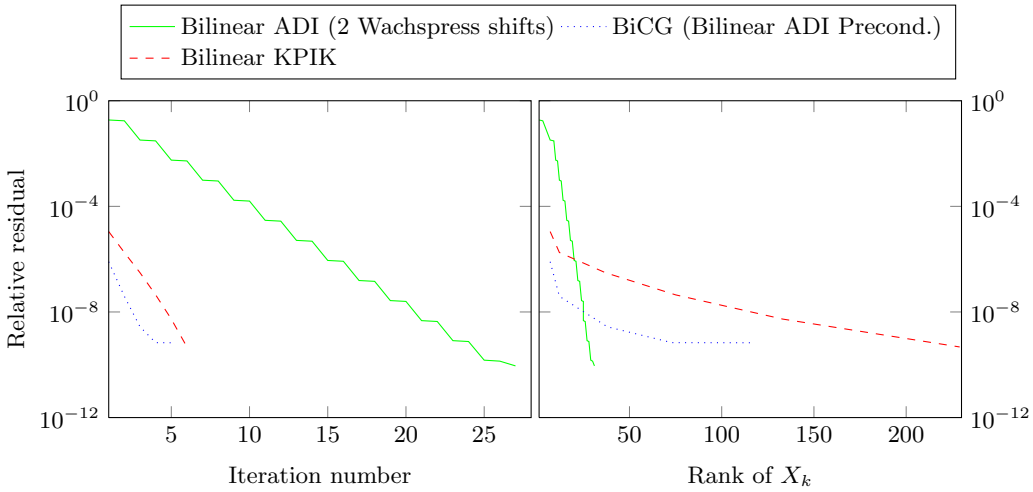

Figure 3: **Fokker-Planck**. Comparison of low rank solution methods for $n = 10\,000$.

approximation often is the smallest. However, when choosing a numerical algorithm, the computational complexity clearly has to be taken into account. Unfortunately, a rigorous complexity analysis of our algorithms is hardly possible. This is due to the fact that if the theoretical costs were actually reached, all our algorithms would no longer be feasible. Let us for example consider the bilinear ADI iteration from Section 5. We have already seen that in each step we have to solve $(2 + m)$ systems of linear equations. The point is that the corresponding right-hand side theoretically grows from size k up to size $(m + 1) \cdot k + m$, where m is the number of inputs. Hence, performing the truncation operation, that keeps the growth of the low rank approximation at a descent level, becomes more and more expensive. Nevertheless, the actual growth of the iterates cannot be specified in general and usually is much smaller than the theoretical expectation. Further, the computation of good shift parameters is even more complicated than in the standard case such that the total costs often might exceed those of the other methods, depending on the speed of convergence.

Regarding the costs of an iterative solver like CG or BiCGstab, one has to keep in mind that using an appropriate preconditioner is essential for obtaining a small iteration number. Since here we proposed to precondition with one step of the bilinear ADI iteration, the complexity also depends on the rank of the current iterate. To be more specific, we can record that the major costs result from the truncation operator and, in case of the projection-based approach, from the necessary orthogonalization by a modified Gram-Schmidt process of the generated Krylov subspaces.

7 Conclusions

In this paper, we have studied a class of generalized Lyapunov equations which naturally arise in the context of model order reduction of bilinear control systems and linear

parameter-varying systems as well as for the stability analysis of linear stochastic differential equations. Under certain low rank assumptions on the involved matrices, we have shown that one can expect a rapid decrease of the singular values of the solutions, justifying the construction of low rank approximations of the form $X = ZZ^T$ and $X = LDL^T$, respectively. We have further proposed some extensions of successful linear low rank approximation procedures and have investigated their usefulness by means of certain large-scale numerical test examples which to some extent fulfill the low rank properties we needed to prove our main results. While the performance is quite good and allows for solving generalized Lyapunov equations of up to the order 562 500, some problems are still open. Here, we think of the generalization of the rational Zolotarev problem which in the *standard case* leads to optimal shift parameters for the ADI iteration. Moreover, it seems to be an interesting topic of further research to give an explanation for the observed fast singular value decay of the solution matrix X although the bilinear coupling matrix N is of full rank.

References

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

[2] A.C. Antoulas, D.C. Sorensen, and Y. Zhou. On the decay rate of Hankel singular values and related issues. *Sys. Control Lett.*, 46(5):323–342, 2002.

[3] P. Benner and T. Breiten. Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems. MPI Magdeburg Preprints MPIMD/11-02, 2011. Available from `http://www.mpi-magdeburg.mpg.de/preprints/abstract.php?nr=11-02&year=2011`.

[4] P. Benner and T. Breiten. On $\mathcal{H}_2$-model reduction of linear parameter-varying systems. In *Proceedings in Applied Mathematics and Mechanics*, volume 11, pages 805–806. WILEY-VCH Verlag, 2011.

[5] P. Benner and T. Breiten. On optimality of interpolation-based low-rank approximations of large-scale matrix equations. MPI Magdeburg Preprints MPIMD/11-10, 2011. Available from `http://www.mpi-magdeburg.mpg.de/preprints/abstract.php?nr=11-10&year=2011`.

[6] P. Benner and T. Damm. Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems. *SIAM J. Cont. Optim.*, 49(2):686–711, 2011.

[7] P. Benner, J.-R. Li, and T. Penzl. Numerical solution of large Lyapunov equations, Riccati equations, and linear-quadratic control problems. *Numer. Lin. Alg. Appl.*, 15(9):755–777, 2008.

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

[9] P. Benner and J. Saak. Linear-quadratic regulator design for optimal cooling of steel profiles. Technical Report SFB393/05-05, Sonderforschungsbereich 393 *Parallele Numerische Simulation für Physik und Kontinuumsmechanik*, TU Chemnitz, 09107 Chemnitz, FRG, 2005. Available from `http://www.tu-chemnitz.de/sfb393`.

[10] M. Condon and R. Ivanov. Krylov subspaces from bilinear representations of nonlinear systems. *COMPEL*, 26:11–26, 2007.

[11] P. D'Alessandro, A. Isidori, and A. Ruberti. Realization and structure theory of bilinear dynamical systems. *SIAM J. Cont. Optim.*, 12(3):517–535, 1974.

[12] T. Damm. Direct methods and ADI-preconditioned Krylov subspace methods for generalized Lyapunov equations. *Numer. Lin. Alg. Appl.*, 15(9):853–871, 2008.

[13] H.T. Dorissen. Canonical forms for bilinear systems. *Sys. Control Lett.*, 13(2):153–160, 1989.

[14] V. Druskin and V. Simoncini. Adaptive rational Krylov subspaces for large-scale dynamical systems. *Sys. Control Lett.*, 60:546–560, 2011.

[15] A. Eppler and M. Bollhöfer. An alternative way of solving large Lyapunov equations. In *Proceedings in Applied Mathematics and Mechanics*, volume 10, pages 547–548. WILEY-VCH Verlag, 2010.

[16] A. Eppler and M. Bollhöfer. Structure-preserving GMRES methods for solving large Lyapunov equations. In *Progress in Industrial Mathematics at ECMI 2010*, volume 17. Springer, 2012. to appear.

[17] G. Flagg. $\mathcal{H}_2$-optimal interpolation: New properties and applications, 2010. Talk given at the 2010 SIAM Annual Meeting, Pittsburgh (PA).

[18] G. Flagg and S. Gugercin. On the ADI method for the Sylvester Equation and the optimal-H2 points. Technical report, 2012. submitted, available as arXiv:1201.4779.

[19] 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.

[20] W.S. Gray and J. Mesko. Energy functions and algebraic Gramians for bilinear systems. In *Preprints of the 4th IFAC Nonlinear Control Systems Design Symposium*, pages 103–108, 1998.

[21] C. Hartmann, A. Zueva, and B. Schäfer-Bung. Balanced model reduction of bilinear systems with applications to positive systems. *SIAM J. Control Optim.*, 2010. submitted.

[22] R.A. Horn and C.R. Johnson. *Matrix analysis*. Cambridge University Press, 1990.

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

[24] K. Jbilou and A. J. Riquet. Projection methods for large Lyapunov matrix equations. *Linear Algebra Appl.*, 415(2-3):344–358, 2006.

[25] T.G. Kolda and B.W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

[26] 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.

[27] D. Kressner and C. Tobler. Low-rank tensor Krylov subspace methods for parametrized linear systems. *SIAM J. Matrix Anal. Appl.*, 32(4):1288–1316, 2011.

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

[29] J.-R. Li and J. White. Low rank solution of Lyapunov equations. *SIAM J. Matrix Anal. Appl.*, 24(1):260–280, 2002.

[30] T. Penzl. Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case. *Sys. Control Lett.*, 40(2):139–144, 2000.

[31] W.J. Rugh. *Nonlinear System Theory.* The John Hopkins University Press, 1982.

[32] Y. Saad. Numerical solution of large Lyapunov equation. In M. A. Kaashoek, J. H. van Schuppen, and A. C. M. Ran, editors, *Signal Processing, Scattering, Operator Theory and Numerical Methods*, pages 503–511. 1990.

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

[34] D.C. Sorensen and Y. Zhou. Bounds on eigenvalue decay rates and sensitivity of solutions to Lyapunov equations. Technical Report TR02-07, Dept. of Comp. Appl. Math., Rice University, Houston, TX, June 2002. Available online from `http://www.caam.rice.edu/caam/trs/tr02.html#TR02`.

[35] F. Stenger. *Numerical methods based on Sinc and analytic functions*, volume 20 of *Springer Series in Computational Mathematics.* Springer-Verlag, New York, 1993.

[36] E.L. Wachspress. Iterative solution of the Lyapunov matrix equation. *Appl. Math. Letters*, 107:87–90, 1988.

[37] E.L. Wachspress. The ADI model problem, 1995. Available from the author.