Peter Benner Tobias Breiten

On optimality of interpolation-based low rank approximations of large-scale matrix equations

Max Planck Institute Magdeburg Preprints

MPIMD/11-10 **February 3, 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 will discuss some optimality results for the approximation of large-scale matrix equations. In particular, this will include the special case of Lyapunov and Sylvester equations, respectively. We show a relation between the iterative rational Krylov algorithm and a Riemannian optimization method which recently has been shown to locally minimize a certain energy norm of the underlying Lyapunov operator. Moreover, we extend the results for a more general setting leading to a slight modification of IRKA. By means of some numerical test examples, we will show the efficiency of the proposed methods.

Keywords: matrix equations, low rank approximations, rational Krylov subspaces, $\mathcal{H}_2$-model reduction

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 note, we will discuss some optimality results for low rank approximations of solutions of large-scale matrix equations of the form

$$AXE + FXB + CD = 0, \tag{1}$$

with $A, F \in \mathbb{R}^{n\times n}$, $B, E \in \mathbb{R}^{m\times m}$, $C \in \mathbb{R}^{n\times p}$ and $D \in \mathbb{R}^{p\times m}$. As is well-known, these types of equations play an important role in analyzing the structure of dynamical systems. For example, in case of $A = B^T$, $E = F^T$ and $C = D^T$, the resulting Lyapunov equation characterizes the stability properties of the associated dynamical system

$$E\dot{x}(t) = Ax(t) + Cu(t), \quad y(t) = Dx(t). \tag{2}$$

Throughout the paper we will assume the matrix pencils (A, F) and (B, E) to be stable, i.e. all eigenvalues are included in the open left complex plane. While for small-to-medium scale equations the exact solution can be computed explicitly by means of the Bartels-Stewart algorithm (see [3]) or Hammarlings method (see [16]), for n exceeding $\mathcal{O}(10^5)$, this will no longer be possible. However, in a lot of interesting real-life applications it holds true that $\operatorname{rank}(CD) \ll n, m$ which often induces a strong singular value decay of the solution matrix X. For a detailed discussion on this topic, we refer to, e.g., [2, 14, 21, 24]. The fact that X can be well approximated by low rank matrices, i.e. $X \approx UV^T$ with $U, V \in \mathbb{R}^{n\times k}, k \ll n$, has caused the development of several numerically efficient low rank methods for solving (1). Here, the most prominent ones are the alternating direction implicit (ADI) iteration as well as the Krylov-plus-inverted-Krylov method (KPIK), see e.g. [6, 27, 22]. Recently, for the case of the Lyapunov equation, a rather different approach has been proposed in [25] which relies on Riemannian optimization and minimizes a certain energy norm associated with the underlying Lyapunov operator. The main result of this paper will be to reveal a connection between this method and the iterative rational Krylov algorithm (IRKA), a well-established approach in the context of model order reduction of dynamical systems of the form (2), see [15]. The structure now will be as follows. In Section 2, we will briefly review the problem of $\mathcal{H}_2$-model reduction of large-scale dynamical systems as well as IRKA. In Section 3, we will then discuss the special case of $A = B^T$, $E = F^T$ and $C = D^T$. Starting with symmetric state space systems, we will explain how to generalize the concepts for the unsymmetric case and point out some difficulties. Subsequently, Section 4 deals with more general Sylvester equations of the form (1). Here, we will introduce an appropriate objective function whose minimization will automatically yield certain optimality results concerning low rank approximations. As it will turn out, the approach is a direct extension of IRKA and thus can be implemented in a stable and efficient

projection framework. We will show the efficiency of our results by means of some standard test examples in Section 5 and conclude with a short review as well as some related topics of further research in Section 6. However, it should be noted at this point that all our results are primarily of theoretical interest and hopefully lead to new insight and better understanding of the relations between the currently most popular methods for solving large-scale Lyapunov and Sylvester equations. In a certain sense, the approximations we propose are optimal and we will provide numerical algorithms which will converge to these optimal approximations. Nevertheless, since all of them are of iterative nature, they strongly depend on the convergence rate and the cost per iteration. At this point they will in most cases be not competitive in terms of numerical efficiency when compared to methods like, e.g., the ADI method and rational Krylov, but we expect that our findings will be helpful in further optimizing these methods and possibly arriving at a hybrid method aggregating the best features of them.

Throughout the paper, $I_n \in \mathbb{R}^{n\times n}$ and $I_{\hat{n}} \in \mathbb{R}^{\hat{n}\times\hat{n}}$ will denote the identity. With $\otimes$ we denote the Kronecker product of two matrices and $\mathrm{vec}\,(\cdot)$ will be understood as the vectorization of a matrix into a long vector. Finally, $\mathrm{tr}\,()$ denotes the trace of a matrix, i.e. the sum of its diagonal entries while $\hat{A} \in \mathbb{R}^{\hat{n}\times\hat{n}}$ will always indicate that the matrix is related to a reduced-order model, meaning that $\hat{n} \ll n$.

2 Optimal $\mathcal{H}_2$-model reduction

In this section, we will briefly introduce the topic of model order reduction of a dynamical system. Basically, the goal is to replace a system of the form (2) by another system with the same structure but much fewer states, i.e.

$$\hat{E}\dot{\hat{x}}(t) = \hat{A}\hat{x}(t) + \hat{C}u(t), \quad \hat{y}(t) = \hat{D}\hat{x}(t), \tag{3}$$

where $\hat{E}, \hat{A} \in \mathbb{R}^{\hat{n}\times\hat{n}}, \hat{C} \in \mathbb{R}^{n\times p}, \hat{D} \in \mathbb{R}^{q\times n}$ and $\hat{n} \ll n$. We will abbreviate the above structure by using the notation $\hat{\Sigma} = (\hat{E}; \hat{A}, \hat{C}, \hat{D})$. Since (3) should approximate (2) in some sense, one usually demands $y \approx \hat{y}$. For linear systems, the deviation from a reduced system to the original one can be measured in terms of different system norms. Besides the $\mathcal{H}_\infty$-norm, a common way for this is to use the $\mathcal{H}_2$-norm which, for a dynamical system Σ, is defined as follows:

$$||\Sigma||_{\mathcal{H}_2} := \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \mathrm{tr}\left(H(-i\omega)H^T(i\omega)\right) d\omega\right)^{\frac{1}{2}},$$

where $H(s) = D(sI_n - A)^{-1}C$ is called the *transfer function* of Σ. In the following, it will be important to recall that the $\mathcal{H}_2$-norm alternatively can be characterized in terms of the solution of the Lyapunov equations associated

with the system. In more detail, we have

$$||\Sigma||_{\mathcal{H}_2}^2 = \operatorname{tr}\left(DXD^T\right) = \operatorname{tr}\left(C^TYC\right),$$

where

$$AXE^T + EXA^T + CC^T = 0, \quad A^TYE + E^TYA + D^TD = 0.$$

In order to compute the approximation of a reduced model with respect to the $\mathcal{H}_2$-norm, one can now define an *error system* as

$$\begin{bmatrix} E & 0 \\ 0 & \hat{E} \end{bmatrix}, \quad \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix}, \quad \begin{bmatrix} C \\ \hat{C} \end{bmatrix}, \quad \begin{bmatrix} D & -\hat{D} \end{bmatrix}. \tag{4}$$

Finally, based on the above definitions and properties, it is possible to derive first-order necessary conditions for $\mathcal{H}_2$-optimality of a reduced system. Although this has already been done in [20], until recently in [15], there has not been an efficient way of computing such a reduced order model. For simplicity, recall that in case of a single input and single output (SISO) system, these conditions imply that the reduced transfer function $\hat{H}$ is a Hermite interpolant of the original transfer function H at the mirror images of its own system poles, i.e.

$$\hat{H}(-\lambda_i) = H(-\lambda_i), \quad \hat{H}'(-\lambda_i) = H'(-\lambda_i), \tag{5}$$

with λ_i being the i-th eigenvalue of the pencil $(\hat{A}, \hat{E})$. Though obviously these poles are not known a priori, the iterative rational Krylov algorithm proposed in [15] has largely resolved this problem. A pseudocode for the SISO case is depicted in Algorithm 1.

Algorithm 1 Iterative rational Krylov algorithm (IRKA) for SISO case

Input: A, E, c, d, $\hat{A}$, $\hat{E}$, $\hat{c}$, $\hat{d}$
Output: $\hat{A}^{opt}$, $\hat{E}^{opt}$, $\hat{c}^{opt}$, $\hat{d}^{opt}$

1: **while** (change in $\Lambda > tol$) **do**
2: $\quad \hat{A}Q = \hat{E}Q\Lambda$
3: $\quad V_i = (-\lambda_i E - A)^{-1}c, \quad W_i = (-\lambda_i E - A)^{-T}d^T,$
4: $\quad V = \operatorname{orth}(V), W = \operatorname{orth}(W)$
5: $\quad \hat{A} = W^TAV, \hat{E} = W^TEV, \hat{c} = W^Tc, \hat{d} = dV$
6: **end while**
7: $\hat{A}^{opt} = \hat{A}, \hat{E}^{opt} = \hat{E}, \hat{c}^{opt} = \hat{c}, \hat{d}^{opt} = \hat{d}$

3 The Lyapunov equation

In this section, we will analyze the case of Lyapunov equations. Once more, for the matrix equation (1), this means $A = B^T$ and $E = F^T$. We start with a detailed discussion for the symmetric case including a very brief review of the method proposed in [25] and will then successively transfer some results to the unsymmetric case.

The symmetric case

In the following we will use the setting specified in [25]. Hence, let us assume that $E = E^T \succ 0$ is symmetric positive definite (s.p.d.) and $A = A^T \prec 0$ is symmetric negative definite. Resulting from the vectorization of equation (1) we define $\mathcal{L} := -E \otimes A - A \otimes E$ to be the Kronecker representation of the associated Lyapunov operator. Note that due to the assumption on the pencil (A, E), we automatically have that $\sigma(L) \subset \mathbb{R}_+$. Moreover, since we are dealing with symmetric matrices, it makes sense to define the energy norm

$$|| \cdot ||_{\mathcal{L}} = \sqrt{\langle \cdot, \cdot \rangle_{\mathcal{L}}} \text{ with } \langle u, v \rangle_{\mathcal{L}} = \langle u, \mathcal{L}v \rangle.$$

In [25], the authors construct a method based on Riemannian optimization that computes a low rank approximation X_k by minimizing the objective function

$$f : \mathcal{M} \to \mathbb{R},\ X \mapsto \operatorname{tr}(XAXE) + \operatorname{tr}\left(XCC^T\right)$$

on the manifold $\mathcal{M}$ of symmetric positive semi-definite matrices of rank k,

$$\mathcal{M} = \{X : X \in S_n^{sym}, X \geq 0, \operatorname{rank}(X) = k\}.$$

The specific function f is motivated by the fact that it holds

$$\begin{aligned} || \operatorname{vec}(X - X_k) ||_{\mathcal{L}}^2 &= 2 \operatorname{tr}(X_k E X_k A) + 2 \operatorname{tr}\left(X_k C C^T\right) + 2 \operatorname{tr}(XEXA) \\ &= 2f(X_k) + 2 \operatorname{tr}(XEXA). \end{aligned}$$

Since the second term depends only on the true solution X it is constant and it thus suffices to minimize f. The first step now is to realize that there is a close relationship between the elements in $\mathcal{M}$ and approximants constructed by a projection-based approach. This is seen as follows. Let $X_k \in \mathcal{M}$. Hence it can be written as $V\hat{X}V^T$, where $V \in \mathbb{R}^{n \times k}$ is an orthogonal matrix and $\hat{X} = \hat{X}^T \in \mathbb{R}^{k \times k}$. In order to minimize the objective function f, we compute the derivative of f with respect to $\hat{X}$ and obtain (see [10]):

$$\begin{aligned} \frac{\partial f}{\partial \hat{X}} &= \frac{\partial}{\partial \hat{X}} \left(\operatorname{tr}\left(V\hat{X}V^T E V \hat{X} V^T A\right) + \operatorname{tr}\left(V\hat{X}V^T C C^T\right) \right) \\ &= \hat{A}\hat{X}\hat{E} + \hat{E}\hat{X}\hat{A} + \hat{C}\hat{C}^T. \end{aligned}$$

Consequently, as a necessary optimality condition we obtain that $\hat{X}$ has to be the solution of the Lyapunov equation associated with the projected system matrices $\hat{E} = V^T E V, \hat{A} = V^T A V$ and $\hat{C} = V^T C$. For this reason, instead of using the Riemannian optimization approach, we want to construct an approximation by projecting onto a suitable subspace V. In the following, we will now point out a crucial link between the iterative rational

Krylov algorithm and the Riemannian optimization method. The most important observation is that the energy norm of every low rank approximant is bounded below by the $\mathcal{H}_2$-norm of an associated error system. For this, we need the following result.

Lemma 3.1. *Let $\Sigma = (E; A, C, C^T)$ denote a symmetric dynamical system with E and $-A$ being s.p.d. Further assume that $\hat{\Sigma} = (\hat{E}; \hat{A}, \hat{C}, \hat{C}^T)$ is a reduced order system obtained by a Galerkin-type projection $\mathcal{P} = VV^T$. Then it holds*

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

with equality in case of $\hat{\Sigma}$ being a locally $\mathcal{H}_2$-optimal reduced order system.

Proof. By the definition of the $\mathcal{H}_2$-inner product, we know that it holds:

$$\begin{aligned}\langle \Sigma - \hat{\Sigma}, \Sigma - \hat{\Sigma}\rangle_{\mathcal{H}_2} &= \langle \Sigma, \Sigma\rangle_{\mathcal{H}_2} - 2\langle \Sigma, \hat{\Sigma}\rangle_{\mathcal{H}_2} + \langle \hat{\Sigma}, \hat{\Sigma}\rangle_{\mathcal{H}_2} \\ &= \langle \Sigma, \Sigma\rangle_{\mathcal{H}_2} - 2\langle \Sigma - \hat{\Sigma}, \hat{\Sigma}\rangle_{\mathcal{H}_2} - \langle \hat{\Sigma}, \hat{\Sigma}\rangle_{\mathcal{H}_2}.\end{aligned}$$

Note that we have

$$\langle \Sigma - \hat{\Sigma}, \hat{\Sigma}\rangle_{\mathcal{H}_2} = \operatorname{tr}\left(C_e^T P \hat{C}\right) = \operatorname{vec}\left(C_e \hat{C}^T\right)^T \operatorname{vec}(P),$$

where $C_e = \begin{bmatrix} C^T & -\hat{C}^T \end{bmatrix}$ and P is the solution of the Sylvester equation

$$A_e P \hat{E} + E_e P \hat{A}^T + C_e \hat{C}^T = 0,$$

with $A_e = \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix}$ and $E_e = \begin{bmatrix} E & 0 \\ 0 & \hat{E} \end{bmatrix}$. Using the vectorization of the above line, we obtain

$$\operatorname{vec}(P) = \left(-\hat{E} \otimes A_e - \hat{A} \otimes E_e\right)^{-1} \operatorname{vec}\left(B_e \hat{B}^T\right).$$

However, since $E_e, \hat{E} \succ 0$ and $A_e, \hat{A} \prec 0$, respectively, we can conclude that the previous inverse is s.p.d. Hence, $\langle \Sigma - \hat{\Sigma}, \hat{\Sigma}\rangle_{\mathcal{H}_2} \geq 0$. Assume now that $P := \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$. Then, it follows

$$AP_1\hat{E} + EP_1\hat{A}^T + C\hat{C}^T = 0, \quad \hat{A}P_2\hat{E} + \hat{E}P_2\hat{A}^T + \hat{C}\hat{C}^T = 0$$

and

$$\langle \Sigma - \hat{\Sigma}, \hat{\Sigma}\rangle_{\mathcal{H}_2} = \operatorname{tr}\left(C^T P_1 \hat{C} - \hat{C}^T P_2 \hat{C}\right).$$

Finally, due to the Wilson conditions for $\mathcal{H}_2$-optimality, a locally optimal reduced order model satisfies $C^T P_1 - \hat{C}^T P_2 = 0$ which proves the statement. □

Remark 3.1. *Note that the above statement also follows from the pole-residue expression for the $\mathcal{H}_2$-error and results on the residues of symmetric state space systems as well as on the difference of the transfer functions which have been shown in [12].*

However, it now easily follows that for symmetric state space systems, IRKA yields low rank approximations X_k that minimize the distance of the true solution X and X_k with respect to the $\mathcal{L}$-norm.

Theorem 3.1. *Let $\Sigma = (E; A, C, C^T)$ denote a symmetric dynamical system Σ with E and $-A$ being s.p.d. and V denote a projection matrix corresponding to a reduced order model $\hat{\Sigma}$ with system matrices $\hat{E}, \hat{A}$ and $\hat{C}$. Let further X and $\hat{X}$ denote the solution of the associated Lyapunov equations. Then*

$$\langle X - V\hat{X}V^T, X - V\hat{X}V^T\rangle_{\mathcal{L}} \geq ||\Sigma - \hat{\Sigma}||^2_{\mathcal{H}_2}, \tag{6}$$

with equality in case of $\hat{\Sigma}$ being a locally $\mathcal{H}_2$-optimal reduced order system.

Proof. Let $\tilde{X} = V\hat{X}V^T$. First, note that the vectorized solutions of the original and reduced Lyapunov equations are obtained as follows

$$\begin{aligned}
\operatorname{vec}(X) &= \underbrace{(-E \otimes A - A \otimes E)^{-1}}_{\mathcal{L}^{-1}} \operatorname{vec}\left(CC^T\right), \\
\operatorname{vec}\left(\hat{X}\right) &= \underbrace{(-\hat{E} \otimes \hat{A} - \hat{A} \otimes \hat{E})^{-1}}_{\hat{\mathcal{L}}^{-1}} \operatorname{vec}\left(\hat{C}\hat{C}^T\right).
\end{aligned}$$

Hence, it now subsequently follows

$$\begin{aligned}
&\langle X - \tilde{X}, X - \tilde{X}\rangle_{\mathcal{L}} = \operatorname{vec}\left(X - \tilde{X}\right)^T \mathcal{L} \operatorname{vec}\left(X - \tilde{X}\right) \\
&= ||\Sigma||^2_{\mathcal{H}_2} - \operatorname{vec}\left(\tilde{X}\right)^T \mathcal{L} \operatorname{vec}(X) - \operatorname{vec}(X)^T \mathcal{L} \operatorname{vec}\left(\tilde{X}\right) + \operatorname{vec}\left(\tilde{X}\right)^T \mathcal{L} \operatorname{vec}\left(\tilde{X}\right) \\
&= ||\Sigma||^2_{\mathcal{H}_2} - \operatorname{vec}\left(\hat{X}\right)^T \left(V^T \otimes V^T\right) \mathcal{L} \operatorname{vec}(X) - \operatorname{vec}(X)^T \mathcal{L} \left(V \otimes V\right) \operatorname{vec}\left(\hat{X}\right) \\
&\quad + \operatorname{vec}\left(\hat{X}\right)^T \left(V^T \otimes V^T\right) \mathcal{L} \left(V \otimes V\right) \operatorname{vec}\left(\hat{X}\right) \\
&= ||\Sigma||^2_{\mathcal{H}_2} - \operatorname{vec}\left(\hat{C}\hat{C}^T\right)^T \hat{\mathcal{L}}^{-1} \left(V^T \otimes V^T\right) \mathcal{L} \operatorname{vec}(X) \\
&\quad - \operatorname{vec}(X)^T \mathcal{L} \left(V \otimes V\right) \hat{\mathcal{L}}^{-1} \operatorname{vec}\left(\hat{C}\hat{C}^T\right) + \operatorname{vec}\left(\hat{C}\hat{C}^T\right)^T \hat{\mathcal{L}}^{-1} \hat{\mathcal{L}} \hat{\mathcal{L}}^{-1} \operatorname{vec}\left(\hat{C}\hat{C}^T\right) \\
&= ||\Sigma||^2_{\mathcal{H}_2} - \operatorname{vec}\left(\hat{C}\hat{C}^T\right)^T \hat{\mathcal{L}}^{-1} \left(V^T \otimes V^T\right) \operatorname{vec}\left(CC^T\right) \\
&\quad - \operatorname{vec}\left(CC^T\right)^T \left(V \otimes V\right) \hat{\mathcal{L}}^{-1} \operatorname{vec}\left(\hat{C}\hat{C}^T\right) + \operatorname{vec}\left(\hat{C}\hat{C}^T\right)^T \hat{\mathcal{L}}^{-1} \hat{\mathcal{L}} \hat{\mathcal{L}}^{-1} \operatorname{vec}\left(\hat{C}\hat{C}^T\right) \\
&= ||\Sigma||^2_{\mathcal{H}_2} - ||\hat{\Sigma}||^2_{\mathcal{H}_2}
\end{aligned}$$

The assertion follows with the previous Lemma. □

Remark 3.2. *At this point we want to emphasize the implication of the previous result since it might not be too obvious at a first glance. Theorem 3.1 states that the energy norm error of each low rank approximation obtained by orthogonally prolongating the solution of a reduced Lyapunov equation is bounded below by the $\mathcal{H}_2$-norm of its associated error system. Since for the iterative rational Krylov algorithm this lower bound is not only minimized but at the same time equality is attained, we thus know that the left-hand side of (6) is automatically minimized as well. However, this means that the corresponding low rank approximation to the solution of the Lyapunov equation is optimal w.r.t. the energy norm.*

The unsymmetric case

Next, we want to consider the more difficult case of A being unsymmetric. However, as it will turn out, for the general case of unsymmetric dynamical systems, defining an energy norm will no longer be possible. Thus we begin with the assumption that the system (A, c, d^T) is equivalent to a symmetric state space system, i.e. there exists a state space transformation $x = Tz$ such that $(T; AT, c, d^T T)$ is a symmetric realization as in Theorem 3.1. Let us further assume that instead of one we now want to solve the two dual Lyapunov equations

$$AX + XA^T + cc^T = 0, \quad A^T Y + YA + dd^T = 0, \tag{7}$$

with rank one right hand sides, i.e. $c, d \in \mathbb{R}^n$. A crucial tool now is the symmetrizer of A, i.e. a matrix J which fulfills $J = J^T$ and $AJ = JA^T$. Here, we look for a special choice of J which has the additional property $d^T J = c^T$. As has been shown in [23], under the reasonable assumption of (A, c, d^T) being a stable dynamical system that is reachable and observable, it is always possible to construct J by means of the reachability matrix

$$\mathcal{K}_c := \begin{bmatrix} c & Ac & A^2c & \dots & A^{n-1}c \end{bmatrix} \tag{8}$$

and the observability matrix

$$\mathcal{K}_d := \begin{bmatrix} d & A^T d & (A^T)^2 d & \dots & (A^T)^{n-1} d \end{bmatrix}^T, \tag{9}$$

respectively. For this, simply define $J := \mathcal{K}_c \, \mathcal{K}_d^{-T}$. Due to the above assumption that the system should be equivalent to a symmetric state space realization, we have that $J \succ 0$. Instead of the above Lyapunov equations, the idea now is to slightly change the situation by looking at the transformed equations

$$AJ\bar{X}J + J\bar{X}JA^T + cc^T = 0, \quad JA^T\bar{Y}J + J\bar{Y}AJ + Jdd^TJ = 0. \tag{10}$$

Obviously, the solution $\bar{Y}$ of (10) is also the solution to the standard Lyapunov equation (7). On the other hand, we have $X = J\bar{X}J$ and we thus have

a nice connection between the Lyapunov equations (7) and (10). Note that it also holds $\bar{Y} = \bar{X}$. The main advantage now is that for the transformed equations it again makes sense to define an energy norm by

$$||\cdot||_{\bar{\mathcal{L}}} = \sqrt{\langle \cdot, \cdot \rangle_{\bar{\mathcal{L}}}} \text{ with } \langle u, v \rangle_{\bar{\mathcal{L}}} = \langle u, \bar{\mathcal{L}} v \rangle$$

and

$$\bar{\mathcal{L}} := -J \otimes AJ - AJ \otimes J. \tag{11}$$

A crucial observation is the fact that the unsymmetric iterative rational Krylov algorithm implicitly is related to the problem of finding the optimal rank k approximation to the solution of the symmetrized Lyapunov equations with respect to the above norm. In more detail, we obtain.

Theorem 3.2. *Let $(\hat{A}, \hat{c}, \hat{d}^T)$ be a reduced model constructed by the iterative rational Krylov algorithm applied to the original system (A, c, d^T) which is assumed to be equivalent to a symmetric state space system. Moreover, let V and W denote the projection matrices associated with the final reduction step. Then $\tilde{Y} = W(V^TW)^{-1}\hat{Y}(W^TV)^{-1}W^T$, with $\hat{Y}$ being the solution of $\hat{A}^T\hat{Y} + \hat{Y}\hat{A} + \hat{d}\hat{d}^T = 0$, is a local minimizer of*

$$\min_{Y_k \in \mathcal{M}} \{(\text{vec}\left(\bar{Y} - Y_k\right))^T \bar{\mathcal{L}}\, \text{vec}\left(\bar{Y} - Y_k\right)\}.$$

Furthermore, $\tilde{X} = J^{-1}V\hat{X}V^TJ^{-1}$, with $\hat{X}$ solving $\hat{A}\hat{X} + \hat{X}\hat{A}^T + \hat{c}\hat{c}^T = 0$, is a local minimizer of

$$\min_{X_k \in \mathcal{M}} \{(\text{vec}\left(\bar{X} - X_k\right))^T \bar{\mathcal{L}}\, \text{vec}\left(\bar{X} - X_k\right)\}.$$

Proof. First of all, note the following useful relation between V and W.

$$\begin{aligned}
V &= \left[(\sigma_1 I - A)^{-1}c, \ldots, (\sigma_k I - A)^{-1}c\right] \\
&= \left[(\sigma_1 I - JA^TJ^{-1})^{-1}c, \ldots, (\sigma_k I - JA^TJ^{-1})^{-1}c\right] \\
&= \left[(J(\sigma_1 I - A^T)J^{-1})^{-1}c, \ldots, (J(\sigma_k I - A^T)J^{-1})^{-1}c\right] \\
&= \left[J(\sigma_1 I - A^T)^{-1}J^{-1}c, \ldots, J(\sigma_k I - A^T)^{-1}J^{-1}c\right] \\
&= \left[J(\sigma_1 I - A^T)^{-1}d, \ldots, J(\sigma_k I - A^T)^{-1}d\right] = JW.
\end{aligned}$$

Thus, we obtain $Z = W(V^TW)^{-1} = W\underbrace{(W^TJW)^{-1}}_{\hat{J}^{-1}}$. Now we can proceed

by rewriting $\tilde{Y}$ in vectorized notation, i.e.

$$\begin{aligned}
\operatorname{vec}\left(\tilde{Y}\right) &= (Z \otimes Z)\left(-\hat{A}^T \otimes I - I \otimes \hat{A}^T\right)^{-1}\left(V^T d \otimes V^T d\right) \\
&= (Z \otimes Z)\left(-V^T A^T Z \otimes I - I \otimes V^T A^T Z\right)^{-1}\left(W^T J d \otimes W^T J d\right) \\
&= (Z \otimes Z)\left(-W^T J A^T W \hat{J}^{-1} \otimes I - I \otimes W^T J A^T W \hat{J}^{-1}\right)^{-1}\left(W^T c \otimes W^T c\right) \\
&= \left(W\hat{J}^{-1} \otimes W\hat{J}^{-1}\right)\left(-W^T J A^T W \hat{J}^{-1} \otimes I - I \otimes W^T J A^T W \hat{J}^{-1}\right)^{-1}\left(W^T c \otimes W^T c\right) \\
&= (W \otimes W)\left(-W^T J A^T W \otimes \hat{J} - \hat{J} \otimes W^T J A^T W\right)^{-1}\left(W^T c \otimes W^T c\right).
\end{aligned}$$

Finally, we note that

$$\begin{aligned}
W &= \left[(\sigma_1 I - A^T)^{-1} d, \ldots, (\sigma_k I - A^T)^{-1} d\right] \\
&= \left[(\sigma_1 I - A^T)^{-1} J^{-1} c, \ldots, (\sigma_k I - A^T)^{-1} J^{-1} c\right] \\
&= \left[(\sigma_1 J - J A^T)^{-1} c, \ldots, (\sigma_k J - J A^T)^{-1} c\right].
\end{aligned}$$

Hence, we can conclude that $\tilde{Y}$ is the approximation which one would have obtained after applying the symmetric IRKA to the system $(J; JA^T, c, d^T J)$. However, by Theorem 3.1, we know that this implies that $\tilde{Y}$ is a local minimizer. Similarly, for the vectorized solution $\tilde{X}$, we have

$$\begin{aligned}
\operatorname{vec}\left(\tilde{X}\right) &= \left(J^{-1}V \otimes J^{-1}V\right)\left(-\hat{A} \otimes I - I \otimes \hat{A}\right)^{-1}\left(\hat{c} \otimes \hat{c}\right) \\
&= \left(J^{-1}V \otimes J^{-1}V\right)\left(-Z^T A V \otimes I - I \otimes Z^T A V\right)^{-1}\left(Z^T c \otimes Z^T c\right) \\
&= (W \otimes W)\left(-\hat{J}^{-1} W^T A J W \otimes I - I \otimes \hat{J}^{-1} W^T A J W\right)^{-1}\left(\hat{J}^{-1} W^T c \otimes \hat{J}^{-1} W^T c\right) \\
&= (W \otimes W)\left(-W^T J A^T W \otimes \hat{J} - \hat{J} \otimes W^T J A^T W\right)^{-1}\left(W^T c \otimes W^T c\right).
\end{aligned}$$

□

Remark 3.3. *Note that Theorem 3.2 yields an explanation why one can expect $V\hat{X}V^T$ and $Z\hat{Y}Z^T$ to be good low rank approximations for the solutions to the unsymmetric Lyapunov equations.*

Before we turn our attention to more general unsymmetric dynamical systems, we will now show that the interpolation points resulting from IRKA are also optimal parameters for the ADI iteration w.r.t. the energy norm. For this we need the following result.

Lemma 3.2. *Let $A \in \mathbb{R}^{n\times n}, c \in \mathbb{R}^n$ and $S = \{\sigma_1, \ldots, \sigma_k\} \in \mathbb{C}_+$ denote a set of shift parameters. Let $\mathcal{K}_k(A, c, S)$ be the associated rational Krylov subspace, i.e.,*

$$\mathcal{K}_k(A, c, S) = \left[(\sigma_1 I - A)^{-1} c, \quad \ldots, \quad (\sigma_k I - A)^{-1} c\right].$$

Assume that V is an orthonormal basis for $\mathcal{K}_k(A, c, S)$ and let $Z \in \mathbb{R}^{n\times k}$ be arbitrary with $Z^T V = I_k$. Then

$$\mathcal{K}_k(A, c, S) = \mathcal{K}_k(VZ^T A, VZ^T c, S).$$

Moreover, let Z_{adi} denote the approximation obtained by applying $-S$ in k steps of the LRCF-ADI iteration for (A, c), and let $\tilde{Z}_{adi}$ be the approximation obtained by applying $-S$ in k steps of the LRCF-ADI iteration applied to $(VZ^T A, VZ^T, c)$. Then

$$Z_{adi} = \tilde{Z}_{adi}.$$

Proof. Let $\mathcal{P} = VZ^T$. Hence, we have $\mathcal{P}^2 = \mathcal{P}$ and, since V is an orthonormal basis for $\mathcal{K}_k(A, c, S)$, it follows that $\mathcal{P}(\sigma_i I - A)^{-1}c = (\sigma_i I - A)^{-1}c$. Thus, for each shift σ_i we obtain

$$\begin{aligned}(\sigma_i I - \mathcal{P}A)^{-1}\mathcal{P}c &= (\sigma_i I - \mathcal{P}A)^{-1}\mathcal{P}(\sigma_i I - A)(\sigma_i I - A)^{-1}c \\ &= (\sigma_i I - \mathcal{P}A)^{-1}\mathcal{P}(\sigma_i I - A)\mathcal{P}(\sigma_i I - A)^{-1}c \\ &= (\sigma_i I - \mathcal{P}A)^{-1}(\sigma_i \mathcal{P} - \mathcal{P}A\mathcal{P})(\sigma_i I - A)^{-1}c \\ &= (\sigma_i I - \mathcal{P}A)^{-1}(\sigma_i I - \mathcal{P}A)\mathcal{P}(\sigma_i I - A)^{-1}c \\ &= \mathcal{P}(\sigma_i I - A)^{-1}c = (\sigma_i I - A)^{-1}c\end{aligned}$$

Consequently, $\mathcal{K}_k(A, c, S) = \mathcal{K}_k(\mathcal{P}A, \mathcal{P}c, S)$. The statement on the ADI approximations follows with the exact same argument and the fact that Z_{adi} can be written as $Z_{adi} = V_{adi} M_{adi} D_{adi}$, where

$$V_{adi} = \left[(\sigma_1 I - A)^{-1}c, \quad \ldots, \quad (\sigma_k I - A)^{-1}c\right],$$

and M_{adi} as well as D_{adi} are independent of A and c, see [19]. □

Next, we need the following result from [11, 13]. For a similar statement, we also want to refer to [9].

Theorem 3.3. *Suppose $S = \{\sigma_1, \ldots, \sigma_k\} \subset \mathbb{C}_+$ are points that satisfy $\lambda(V^T AV) = -\{\sigma_1, \ldots, \sigma_k\}$, where V is an orthonormal basis for the rational Krylov subspace $\mathcal{K}_k(A, c, S)$. Let $X_k \in \mathbb{R}^{k\times k}$ solve*

$$V^T AVX_k + X_k V^T A^T V + V^T cc^T V = 0$$

and assume $\tilde{X}_k$ to be computed by employing $-\{\sigma_1, \ldots, \sigma_k\}$ in exactly k steps of the ADI iteration. Then $\tilde{X}_k = VX_k V^T$.

For our main result on the ADI iteration, it is crucial to note that the proof given in [13], does not use that A is invertible. As long as the reduced Lyapunov equation is uniquely solvable, the above equality still holds true. This now allows to show the optimality of the IRKA points applied in the ADI iteration.

Theorem 3.4. *Let $S = \{-\lambda_1, \ldots, -\lambda_k\}$ denote the set of interpolation points obtained by the iterative rational Krylov algorithm applied to the original system (A, c, d^T) which is assumed to be equivalent to a symmetric state space system. Let further $X_{adi} = Z_{adi} Z_{adi}^T$ and $Y_{adi} = \bar{Z}_{adi} \bar{Z}_{adi}^T$ be the approximations obtained by applying $-S$ in exactly k steps of the ADI iteration applied to the system (A, c, d^T). Then $J^{-1} X_{adi} J^{-1}$ and Y_{adi} are local minimizers of*

$$\min_{X_k \in \mathcal{M}} \{(\operatorname{vec}\left(\bar{X} - X_k\right))^T \bar{\mathcal{L}} \operatorname{vec}\left(\bar{X} - X_k\right)\},$$

and

$$\min_{Y_k \in \mathcal{M}} \{(\operatorname{vec}\left(\bar{Y} - Y_k\right))^T \bar{\mathcal{L}} \operatorname{vec}\left(\bar{Y} - Y_k\right)\}.$$

respectively.

Proof. Let us take a closer look at the reduced Lyapunov equation from Theorem 3.2, i.e.,

$$\hat{A}\hat{X} + \hat{X}\hat{A}^T + \hat{c}\hat{c}^T = 0.$$

In more detail, the above means

$$Z^T A V \hat{X} + \hat{X} V^T A^T Z + Z^T c c^T Z = 0,$$

with $Z^T = (W^T V)^{-1} W^T$. Since V is orthonormal, this is equivalent to solving

$$V^T V Z^T A V \hat{X} + \hat{X} V^T A^T Z V^T V + V^T V Z^T c c^T Z V^T V = 0$$

which, using $\tilde{A} = V Z^T A$ and $\tilde{c} = V Z^T c$, can be interpreted as

$$V^T \tilde{A} V \hat{X} + \hat{X} V^T \tilde{A}^T V + \tilde{c}\tilde{c}^T = 0.$$

From Lemma 3.2 we know that the rational Krylov subspace as well as the ADI iteration produce the same results if we use $(\tilde{A}, \tilde{c})$ instead of (A, c). Finally, due to the $\mathcal{H}_2$-optimality conditions, it holds

$$\sigma(V^T A V) = \sigma(V^T V Z^T A V) = \sigma(Z^T A V) = \sigma(\hat{A}) = -S$$

and we can apply Theorem 3.3.

For the dual result, let $\mathcal{Y} = (V^T W)^{-1} \hat{Y} (W^T V)^{-1}$ be defined as in Theorem 3.2, i.e. $\tilde{Y} = W \mathcal{Y} W^T$. A similar reformulation of the reduced Lyapunov equation leads to solving

$$V^T A W \mathcal{Y} (W^T V) + (V^T W) \mathcal{Y} W^T A V + V^T d d^T V = 0,$$

which, by pre- and postmultiplication, can be rewritten as

$$(V^TW)^{-1}V^TA^TW\mathcal{Y} + \mathcal{Y}W^TAV(W^TV)^{-1} + (V^TW)^{-1}V^Tdd^T(W^TV)^{-1} = 0.$$

From here, the procedure is completely analogue to the previous case. One can make use of the orthonormality of W and then simply apply the preceding statements in order to show that $W\mathcal{Y}W^T = Y_{adi}$. □

So far, we have assumed that the system under consideration (A, c, d^T) is at least equivalent to a symmetric state space system. This is an essential property which ensures that the symmetrizer J and thus the Lyapunov operator $\bar{\mathcal{L}}$ is positive definite and we can define an energy norm. However, these assumptions characterize a very limited class of dynamical systems. For example, let us consider the following simple two-dimensional system

$$E = \begin{bmatrix} 2 & 0 \\ 0 & -1 \end{bmatrix}, \quad A = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ 1 \end{bmatrix} = d^T.$$

Although the above system is stable, dissipative and has only real eigenvalues, it can be trivially shown that we can never transform it into a symmetric state space system which allows the definition of an energy norm with equal inputs and outputs. This is due to the fact that the spectra of E and A will always lie on both sides of the imaginary axis and thus $\mathcal{L} = -E \otimes A - A \otimes E$ will be indefinite. Otherwise, if it is transformed into a definite matrix the inputs and outputs will no longer be equal.

Moreover, all systems with complex poles automatically exclude the possibility of an induced energy norm of the form $-E \otimes A - A \otimes E$. This is seen as follows. Assume that an unsymmetric dynamical system (A, c, d^T) is given, with A having complex eigenvalues. Assume now that the system can be transformed into a generalized symmetric state space system of the form $(\tilde{E}; \tilde{A}, \tilde{c}, \tilde{c}^T)$ and that the operator $\tilde{\mathcal{L}} = -\tilde{E} \otimes \tilde{A} - \tilde{A} \otimes \tilde{E}$ is positive definite. Due to the Theorem of Stephanos, see e.g. [18], the eigenvalues of $\tilde{E}$ and $\tilde{A}$ must all have equal or opposite sign since otherwise $\tilde{\mathcal{L}}$ would be indefinite. Next, w.l.o.g we assume that $\sigma(\tilde{E}) \subset \mathbb{C}_+$. This means that $\tilde{E}$ is symmetric positive definite and the eigenvalue problem for the pencil $(\tilde{A}, \tilde{E})$ can be transformed into a symmetric one. However, this would imply that all eigenvalues of $(\tilde{A}, \tilde{E})$ are real. Since the poles of a dynamical system are invariant under state space transformations, this would mean that all eigenvalues of A are real which is a contradiction to our assumption. Thus we cannot define the desired energy norm in a straightforward way.

Nevertheless, it remains the question if low rank Lyapunov approximations obtained by an IRKA reduced order model still can be expected to be accurate even if the underlying dynamical system is unsymmetric and exhibits complex poles. For this, it might be an interesting observation that the $\mathcal{H}_2$-norm of the error system vanishes if and only if the corresponding

Lyapunov approximations that are generated by the reduced system are exact.

Theorem 3.5. *Let (A, C, D) denote a minimal stable dynamical system with $A \in \mathbb{R}^{n\times n}, C \in \mathbb{R}^{n\times p}$ and $D \in \mathbb{R}^{p\times n}$. Assume that a stable reduced order model $(\hat{A}, \hat{C}, \hat{D})$ is constructed by a Petrov-Galerkin projection $\mathcal{P} = V \underbrace{(W^TV)^{-1}W^T}_{Z^T}$ with $V, W \in \mathbb{R}^{n\times k}$ and $V^TV = I$. Let further $X_k = V\hat{X}V^T$ and $Y_k = Z\hat{Y}Z^T$ be obtained by solving the projected Lyapunov equations*

$$\hat{A}\hat{X} + \hat{X}\hat{A}^T + \hat{C}\hat{C}^T = 0, \quad \hat{A}^T\hat{Y} + \hat{Y}\hat{A} + \hat{D}^T\hat{D} = 0.$$

Then the $\mathcal{H}_2$-norm of the error system, cf. (4), is zero if and only if $X_k = X$ and $Y_k = Y$, where X and Y are the exact solutions of the original Lyapunov equations

$$AX + XA^T + CC^T = 0, \quad A^TY + YA + D^TD = 0. \tag{12}$$

Proof. Let us assume that $||\Sigma_e|| = ||\Sigma - \hat{\Sigma}||_{\mathcal{H}_2} = 0$. By the definition of the $\mathcal{H}_2$-norm this means that

$$\int_0^\infty ||D_e e^{A_e t} C_e||_F dt = 0.$$

Hence, since $D_e e^{A_e t} C_e$ is continuous it has to be the constant zero function and thus its derivatives evaluated at zero have to be zero as well, i.e. $D_e A_e^i C_e = 0,\ i \geq 0$. Due to the structure of the error system this means $DA^iC = \hat{D}\hat{A}^i\hat{C}^i,\ i \geq 0$. Thus, the Markov parameters of Σ and $\hat{\Sigma}$ coincide. Since we assumed Σ to be a minimal realization, from [1], it follows that $k = n$. Consequently, the projection matrices V and $Z = W(V^TW)^{-1}$ are orthogonal. Let us now have a look at the transformed Lyapunov equation

$$\hat{A}\hat{X} + \hat{X}\hat{A}^T + \hat{C}\hat{C}^T = 0.$$

Inserting the definition of $\hat{\Sigma}$, we have

$$Z^TAV\hat{X} + \hat{X}V^TA^TZ + Z^TCC^TZ = 0.$$

Multiplying from the left with Z^{-T} and from the right with V^T, we see that X_k solves the original Lyapunov equation. Similarly, one can show that $Y_k = Y$.

Conversely, let us assume that the approximation is exact, i.e. $X_k = X$. As we have seen in the proof of Lemma 3.1, for the $\mathcal{H}_2$-norm of the error system, it holds

$$\langle \Sigma - \hat{\Sigma}, \Sigma - \hat{\Sigma}\rangle_{\mathcal{H}_2} = \langle \Sigma, \Sigma\rangle_{\mathcal{H}_2} - 2\langle \Sigma - \hat{\Sigma}, \hat{\Sigma}\rangle_{\mathcal{H}_2} - \langle \hat{\Sigma}, \hat{\Sigma}\rangle_{\mathcal{H}_2}.$$

Since $X_k = X$, it follows that

$$\langle \hat{\Sigma}, \hat{\Sigma} \rangle_{\mathcal{H}_2} = \hat{D}\hat{X}\hat{D}^T = DV\hat{X}V^T D^T = DX_k D^T = DXD^T = \langle \Sigma, \Sigma \rangle_{\mathcal{H}_2}.$$

Hence, in order to prove the assertion, it remains to show that it holds $\langle \Sigma - \hat{\Sigma}, \hat{\Sigma} \rangle_{\mathcal{H}_2} = 0$. Once again, analog to the proof of Lemma 3.1, we know that

$$\langle \Sigma - \hat{\Sigma}, \hat{\Sigma} \rangle_{\mathcal{H}_2} = \operatorname{tr}\left(DM\hat{D}^T - \hat{D}\hat{X}\hat{D}^T \right),$$

where M is the solution of $AM + M\hat{A}^T + C\hat{C}^T = 0$. Since A and $\hat{A}$ are assumed to be stable, the solution M is unique. However, since X_k is the exact solution of eq. (12), we have

$$AV\hat{X}V^T + V\hat{X}V^T A^T + CC^T = 0.$$

Multiplying from the right with Z, it follows

$$AV\hat{X} + V\hat{X}\hat{A}^T + C\hat{C}^T = 0.$$

Thus, it holds $V\hat{X} = M$ and also $\langle \Sigma - \hat{\Sigma}, \hat{\Sigma} \rangle_{\mathcal{H}_2} = 0$. □

Remark 3.4. *Note that Theorem 3.5 shows that the $\mathcal{H}_2$-norm of the error system is an objective function which is zero if and only if the low rank Lyapunov approximations are the exact solutions. Hence, it seems reasonable to minimize this objective function in order to obtain approximations which are close to the exact solutions. However, this is exactly what the iterative rational Krylov algorithm aims at.*

To sum up, we have seen that the previous results yield a theoretical explanation for the often very accurate low rank approximations obtained by orthogonally prolongating the solution corresponding to an IRKA reduced order model. More precisely, for the special case of systems that are equivalent to symmetric state space systems, we have shown that locally $\mathcal{H}_2$-optimal reduced order models lead to approximations that implicitly minimize the energy norm induced by a symmetrized Lyapunov operator of the form (11), which is given by the controllability and observability matrices (8) and (9), respectively. Moreover, we have seen that in this case we can obtain the exact same result by employing the corresponding interpolation points within the ADI iteration, generalizing the existing results for one-sided projections specified in [9, 11, 13]. Since this equivalence does not require that the poles of the system are real, we know that $\mathcal{H}_2$-optimal interpolation points are in a certain sense also optimal parameters for the ADI iteration which, in contrast to the parameters derived in [26], can be computed even if the matrices exhibit complex spectra. Thus, we can finally state that if one uses appropriate shifts, the ADI iteration, the rational Krylov framework and, at least for state space symmetric systems, the Riemannian optimization approach from [25], are all equivalent and minimize the naturally induced energy norm.

4 Sylvester equations

In this section, we will generalize the situation and study Sylvester equations of the form (1). As in the previous section, we begin with a detailed analysis of the symmetric case before we briefly sketch the extension to the unsymmetric case which is quite similar to what we have presented before.

The symmetric case

We will now subsequently develop a more general approach which relies on the idea of the iterative rational Krylov algorithm, but is not restricted to linear dynamical systems and hence Lyapunov equations. Again, we assume that all involved square matrices are symmetric and have eigenvalues either in $\mathbb{C}_-$ or in $\mathbb{C}_+$. To be more precise, we want to have $A = A^T \prec 0, B = B^T \prec 0, E = E^T \succ 0$ and $F = F^T \succ 0$. This allows us to consider an energy norm based on the following symmetric positive definite matrix

$$\mathcal{L}_S = -E \otimes A - B \otimes F.$$

We now seek for approximations of the form $\tilde{X} = V\hat{X}W^T$ of rank k which minimize the $\mathcal{L}_S$-norm between the original solution X and $\tilde{X}$, i.e.

$$|| \operatorname{vec}\left(X - \tilde{X}\right) ||^2_{\mathcal{L}_S} = \left(\operatorname{vec}\left(X - \tilde{X}\right)\right)^T \mathcal{L}_S \operatorname{vec}\left(X - \tilde{X}\right).$$

Here, $\hat{X}$ will again be determined by solving a reduced Sylvester equation

$$\hat{A}\hat{X}\hat{E} + \hat{F}\hat{X}\hat{B} + \hat{C}\hat{D} = 0,$$

while V and W denote projection matrices with $V^TV = W^TW = I$. From now on, let

$$\Sigma = (A, B, C, D, E, F)$$

denote an abbreviation for an associated Sylvester equation of the form (1). Furthermore, let us consider the following objective function

$$f(\Sigma) = \operatorname{tr}\left(C^T X D^T\right), \tag{13}$$

with X fulfilling (1). As is easily seen, this function results from a slight modification of the $\mathcal{H}_2$-norm of a dynamical system and thus can be computed as

$$f(\Sigma) = \left(\operatorname{vec}\left(I_p\right)\right)^T \left(D \otimes C^T\right) \left(-E \otimes A - B \otimes F\right)^{-1} \left(D^T \otimes C\right) \operatorname{vec}\left(I_p\right). \tag{14}$$

Assume now that we have constructed a reduced set of matrices

$$\hat{\Sigma} = (\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{E}, \hat{F})$$

by the following projection, i.e.

$$\hat{A} = V^T A V, \quad \hat{B} = W^T B W, \quad \hat{C} = V^T C,$$
$$\hat{D} = DW, \quad \hat{E} = W^T E W, \quad \hat{F} = V^T F V.$$

Next, for Σ and $\hat{\Sigma}$ we define the corresponding error set

$$\Sigma_{err} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}, \mathcal{F})$$

with

$$\mathcal{A} = \begin{bmatrix} -A & 0 \\ 0 & \hat{A} \end{bmatrix}, \quad \mathcal{B} = \begin{bmatrix} -B & 0 \\ 0 & \hat{B} \end{bmatrix}, \quad \mathcal{C} = \begin{bmatrix} C \\ \hat{C} \end{bmatrix}$$
$$\mathcal{D} = \begin{bmatrix} D & \hat{D} \end{bmatrix}, \quad \mathcal{E} = \begin{bmatrix} -E & 0 \\ 0 & \hat{E} \end{bmatrix}, \quad \mathcal{F} = \begin{bmatrix} -F & 0 \\ 0 & \hat{F} \end{bmatrix}.$$

The following result will later on be crucial for constructing optimal low rank approximations.

Lemma 4.1. *Let Σ and $\hat{\Sigma}$ denote two sets of matrices associated with large and reduced Sylvester equations of the form (1), respectively. Then for the associated error set Σ_{err}, it holds*

$$f(\Sigma_{err}) \leq f(\Sigma) - f(\hat{\Sigma}).$$

Proof. We begin with analyzing the solution $\mathcal{X}$ of the Sylvester equation for the error set $\Sigma_{err} = (\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}, \mathcal{F})$. What we obtain is

$$\begin{bmatrix} -A & 0 \\ 0 & \hat{A} \end{bmatrix} \underbrace{\begin{bmatrix} X & Y \\ Z & \hat{X} \end{bmatrix}}_{\mathcal{X}} \begin{bmatrix} -E & 0 \\ 0 & \hat{E} \end{bmatrix} + \begin{bmatrix} -F & 0 \\ 0 & \hat{F} \end{bmatrix} \underbrace{\begin{bmatrix} X & Y \\ Z & \hat{X} \end{bmatrix}}_{\mathcal{X}} \begin{bmatrix} -B & 0 \\ 0 & \hat{B} \end{bmatrix}$$
$$+ \begin{bmatrix} C \\ \hat{C} \end{bmatrix} \begin{bmatrix} D & \hat{D} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Hence, X and $\hat{X}$ are the solutions to the Sylvester equations

$$AXE + FXB + CD = 0 \quad \text{and} \quad \hat{A}\hat{X}\hat{E} + \hat{F}\hat{X}\hat{B} + \hat{C}\hat{D} = 0.$$

On the other hand, Y and Z satisfy

$$-AY\hat{E} - FY\hat{B} + C\hat{D} = 0 \quad \text{and} \quad -\hat{A}ZE - \hat{F}ZB + \hat{C}D = 0.$$

Next, let us have a look at the generalized Sylvester equations

$$\mathcal{A}\mathcal{Y}\hat{E} + \mathcal{F}\mathcal{Y}\hat{B} + \mathcal{C}\hat{D} = 0 \quad \text{and} \quad \hat{A}\mathcal{Z}\mathcal{E} + \hat{F}\mathcal{Z}\mathcal{B} + \hat{C}\mathcal{D} = 0.$$

Comparing the structure of the error set, it trivially follows that $\mathcal{Y} = \begin{bmatrix} Y \\ \hat{X} \end{bmatrix}$ and $\mathcal{Z} = \begin{bmatrix} Z & \hat{X} \end{bmatrix}$. Consequently, we can derive the objective function for the error set as

$$\begin{aligned} f(\Sigma_{err}) &= \operatorname{tr}\left(\mathcal{C}^T \mathcal{X} \mathcal{D}^T\right) = \operatorname{tr}\left(\begin{bmatrix} C^T & \hat{C}^T \end{bmatrix} \begin{bmatrix} X & Y \\ Z & \hat{X} \end{bmatrix} \begin{bmatrix} D^T \\ \hat{D}^T \end{bmatrix}\right) \\ &= \operatorname{tr}\left(C^T X D^T + \hat{C}^T Z D^T + C^T Y \hat{D}^T + \hat{C}^T \hat{X} \hat{D}^T\right) \\ &= \operatorname{tr}\left(C^T X D^T + \mathcal{C}^T \mathcal{Y} \hat{D}^T + \hat{C}^T \mathcal{Z} \mathcal{D}^T - \hat{C}^T \hat{X}^T \hat{D}^T\right). \end{aligned}$$

In order to compare the above identity with $f(\Sigma) - f(\hat{\Sigma})$, let us analyze the term $\operatorname{tr}\left(\mathcal{C}^T \mathcal{Y} \hat{D}^T\right)$, for which we obtain

$$\begin{aligned} \operatorname{tr}\left(\mathcal{C}^T \mathcal{Y} \hat{D}^T\right) &= \operatorname{vec}(I_p)^T \operatorname{vec}\left(\mathcal{C}^T \mathcal{Y} \hat{D}^T\right) = \operatorname{vec}(I_p)^T \left(\hat{D} \otimes \mathcal{C}^T\right) \operatorname{vec}(\mathcal{Y}) \\ &= \operatorname{vec}(I_p)^T \left(\hat{D} \otimes \mathcal{C}^T\right) \left(-\hat{E} \otimes \mathcal{A} - \hat{B} \otimes \mathcal{F}\right)^{-1} \left(\hat{D}^T \otimes \mathcal{C}\right) \operatorname{vec}(I_p). \end{aligned}$$

For the following step, we refer to [4], where by means of a special permutation matrix, the evaluation of Kronecker products of structured matrices of the above form can be simplified as

$$\begin{aligned} \operatorname{tr}\left(\mathcal{C}^T \mathcal{Y} \hat{D}^T\right) &= \operatorname{vec}(I_p) \left(\hat{D} \otimes C^T\right) \left(\hat{E} \otimes A + \hat{B} \otimes F\right)^{-1} \left(\hat{D}^T \otimes C\right) \operatorname{vec}(I_p) \\ &\quad - \operatorname{vec}(I_p) \left(\hat{D} \otimes \hat{C}^T\right) \left(\hat{E} \otimes \hat{A} + \hat{B} \otimes \hat{F}\right)^{-1} \left(\hat{D}^T \otimes \hat{C}\right) \operatorname{vec}(I_p). \end{aligned}$$

Note that the structure of the above terms is

$$x^T M^{-1} x - x^T \mathcal{V} \left(\mathcal{V}^T M \mathcal{V}\right)^{-1} \mathcal{V}^T x,$$

where M is a symmetric negative definite matrix. Moreover,

$$M^{-1} - \mathcal{V} \left(\mathcal{V}^T M \mathcal{V}\right)^{-1} \mathcal{V}^T$$

is the Schur complement of $\mathcal{S} = \begin{bmatrix} \mathcal{V}^T M \mathcal{V} & \mathcal{V}^T \\ \mathcal{V} & M^{-1} \end{bmatrix}$ in M^{-1}. Let $s = \begin{bmatrix} y \\ z \end{bmatrix}$ now be an arbitrary vector. Then it holds

$$s^T \mathcal{S} s = y^T \mathcal{V}^T M \mathcal{V} y + y^T \mathcal{V}^T z + z^T \mathcal{V} y + z^T M^{-1} z.$$

Defining $q := M \mathcal{V} y$, it follows

$$s^T \mathcal{S} s = \left(q^T + z^T\right) M^{-1} (q + z) \leq 0.$$

However, this means that $\mathcal{S}$ as well as its Schur complement are negative semi-definite. Finally, this shows that $\operatorname{tr}\left(\mathcal{C}^T\mathcal{Y}\hat{D}^T\right) \leq 0$. A completely analogue argumentation leads to $\hat{C}^T\mathcal{Z}\mathcal{D}^T \leq 0$, finishing the proof. □

Similar to the Lyapunov case, our goal will be locally minimizing $f(\Sigma_{err})$ simultaneously leading to a minimizer of $||\operatorname{vec}\left(X - \tilde{X}\right)||^2_{\mathcal{L}_S}$. For this, we will derive first order necessary conditions on $\hat{\Sigma} = (\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{E}, \hat{F})$. In the following, we make use of the spectral decomposition of the pencil $(\hat{A}, \hat{F})$, i.e. $\hat{A} = \hat{F}Q\Lambda Q^{-1}$, where Q contains the eigenvectors corresponding to the eigenvalues of $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_k)$. Since a similar derivation with equivalent Kronecker structures can be found in [4], we only state the most important aspects. First, note that it holds

$$\begin{aligned}
f(\Sigma_{err}) = \operatorname{vec}(I_p)^T &\left(\begin{bmatrix} D & \hat{D}\end{bmatrix} \otimes \begin{bmatrix} C^T & \hat{C}^T\end{bmatrix}\right) \times \\
&\left(\begin{bmatrix} -E & 0 \\ 0 & \hat{E}\end{bmatrix} \otimes \begin{bmatrix} -A & 0 \\ 0 & \hat{A}\end{bmatrix} + \begin{bmatrix} -F & 0 \\ 0 & \hat{F}\end{bmatrix} \otimes \begin{bmatrix} -B & 0 \\ 0 & \hat{B}\end{bmatrix}\right)^{-1} \times \\
&\left(-\begin{bmatrix} D^T \\ \hat{D}^T\end{bmatrix} \otimes \begin{bmatrix} C \\ \hat{C}\end{bmatrix}\right) \operatorname{vec}(I_p) \\
= \operatorname{vec}(I_p)^T &\left(\begin{bmatrix} D & \hat{D}\end{bmatrix} \otimes \begin{bmatrix} C^T & \hat{C}^T\end{bmatrix}\right) \left(\begin{bmatrix} I & 0 \\ 0 & \hat{I}\end{bmatrix} \otimes \begin{bmatrix} I & 0 \\ 0 & Q\end{bmatrix}\right) \times \\
&\left(\begin{bmatrix} -E & 0 \\ 0 & \hat{E}\end{bmatrix} \otimes \begin{bmatrix} -A & 0 \\ 0 & \Lambda\end{bmatrix} + \begin{bmatrix} -F & 0 \\ 0 & \hat{F}\end{bmatrix} \otimes \begin{bmatrix} -B & 0 \\ 0 & \hat{B}\end{bmatrix}\right)^{-1} \times \\
&\left(\begin{bmatrix} I & 0 \\ 0 & \hat{I}\end{bmatrix} \otimes \begin{bmatrix} I & 0 \\ 0 & Q^{-1}\hat{F}^{-1}\end{bmatrix}\right) \left(-\begin{bmatrix} D^T \\ \hat{D}^T\end{bmatrix} \otimes \begin{bmatrix} C \\ \hat{C}\end{bmatrix}\right) \operatorname{vec}(I_p)\,.
\end{aligned}$$

Since $\hat{A}$ and $\hat{F}$ are symmetric matrices, we can compute a set of $\hat{F}$-orthonormal eigenvectors Q, i.e. $Q^T\hat{F}Q = I$. Hence, we have

$$\begin{aligned}
f(\Sigma_{err}) = \operatorname{vec}(I_p)^T &\left(\begin{bmatrix} D & \hat{D}\end{bmatrix} \otimes \begin{bmatrix} C^T & \hat{C}^T\end{bmatrix}\right) \left(\begin{bmatrix} I & 0 \\ 0 & \hat{I}\end{bmatrix} \otimes \begin{bmatrix} I & 0 \\ 0 & Q\end{bmatrix}\right) \times \\
&\left(\begin{bmatrix} -E & 0 \\ 0 & \hat{E}\end{bmatrix} \otimes \begin{bmatrix} -A & 0 \\ 0 & \Lambda\end{bmatrix} + \begin{bmatrix} -F & 0 \\ 0 & \hat{F}\end{bmatrix} \otimes \begin{bmatrix} -B & 0 \\ 0 & \hat{B}\end{bmatrix}\right)^{-1} \times \\
&\left(\begin{bmatrix} I & 0 \\ 0 & \hat{I}\end{bmatrix} \otimes \begin{bmatrix} I & 0 \\ 0 & Q^T\end{bmatrix}\right) \left(-\begin{bmatrix} D^T \\ \hat{D}^T\end{bmatrix} \otimes \begin{bmatrix} C \\ \hat{C}\end{bmatrix}\right) \operatorname{vec}(I_p)\,.
\end{aligned}$$

Next, for the derivative w.r.t. λ_i, we make use of the product rule for

Kronecker products (see [4]) in order to obtain

$$\begin{aligned}\frac{\partial f}{\partial \lambda_i} = 2 \cdot \operatorname{vec}(I_p)^T \left(\begin{bmatrix} D & \hat{D} \end{bmatrix} \otimes \begin{bmatrix} C^T & \hat{C}^T \end{bmatrix}\right) \left(\begin{bmatrix} I & 0 \\ 0 & \hat{I} \end{bmatrix} \otimes \begin{bmatrix} I & 0 \\ 0 & Q \end{bmatrix}\right) \times \\ \left(\begin{bmatrix} -E & 0 \\ 0 & \hat{E} \end{bmatrix} \otimes \begin{bmatrix} -A & 0 \\ 0 & \Lambda \end{bmatrix} + \begin{bmatrix} -F & 0 \\ 0 & \hat{F} \end{bmatrix} \otimes \begin{bmatrix} -B & 0 \\ 0 & \hat{B} \end{bmatrix}\right)^{-1} \times \\ \left(\begin{bmatrix} -E & 0 \\ 0 & \hat{E} \end{bmatrix} \otimes \begin{bmatrix} 0 & 0 \\ 0 & e_i e_i^T \end{bmatrix}\right) \times \\ \left(\begin{bmatrix} -E & 0 \\ 0 & \hat{E} \end{bmatrix} \otimes \begin{bmatrix} -A & 0 \\ 0 & \Lambda \end{bmatrix} + \begin{bmatrix} -F & 0 \\ 0 & \hat{F} \end{bmatrix} \otimes \begin{bmatrix} -B & 0 \\ 0 & \hat{B} \end{bmatrix}\right)^{-1} \times \\ \left(\begin{bmatrix} I & 0 \\ 0 & \hat{I} \end{bmatrix} \otimes \begin{bmatrix} I & 0 \\ 0 & Q^T \end{bmatrix}\right) \left(\begin{bmatrix} D^T \\ \hat{D}^T \end{bmatrix} \otimes \begin{bmatrix} C \\ \hat{C} \end{bmatrix}\right) \operatorname{vec}(I_p).\end{aligned}$$

By setting the last expression equal to zero and carefully analyzing the structure, it turns out that for optimality we have to require

$$\begin{aligned}&\operatorname{vec}(I_p)^T \left(D \otimes \hat{C}^T Q\right) \left(-E \otimes \Lambda - B \otimes \hat{I}\right)^{-1} \left(-E \otimes e_i e_i^T\right) \times \\ &\left(-E \otimes \Lambda - B \otimes \hat{I}\right)^{-1} \left(D^T \otimes Q^T \hat{C}\right) \operatorname{vec}(I_p) \\ &= \operatorname{vec}(I_p)^T \left(\hat{D} \otimes \hat{C}^T Q\right) \left(-\hat{E} \otimes \Lambda - \hat{B} \otimes \hat{I}\right)^{-1} \left(-\hat{E} \otimes e_i e_i^T\right) \times \\ &\left(-\hat{E} \otimes \Lambda - \hat{B} \otimes \hat{I}\right)^{-1} \left(\hat{D}^T \otimes Q^T \hat{C}\right) \operatorname{vec}(I_p).\end{aligned}$$

This now can be simplified as follows

$$\begin{aligned}&\tilde{C}_i^T D \left(-\lambda_i E - B\right)^{-1} E \left(-\lambda_i E - B\right)^{-1} D^T \tilde{C}_i \\ &= \tilde{C}_i^T D \left(-\lambda_i E - B\right)^{-1} E \left(-\lambda_i E - B\right)^{-1} D^T \tilde{C}_i,\end{aligned}$$

with $\tilde{C} = \hat{C}^T Q$. Furthermore, if we introduce a *transfer function* $H(s) = D\left(sE - B\right)^{-1} D^T$, the above means

$$\tilde{C}_i^T H'(-\lambda_i) \tilde{C}_i = \tilde{C}_i^T \hat{H}'(-\lambda_i) \tilde{C}_i,$$

i.e. the derivative of the reduced *transfer function* has to tangentially interpolate the derivative of the original *transfer function* at the mirror images of the reduced system poles of the pencil $(\hat{A}, \hat{F})$. Using $\tilde{C}$ as optimization parameter determining one part of the reduced set of matrices $(\hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{E}, \hat{F})$, for the derivative of f w.r.t. $\tilde{C}_i$, one will obtain

$$\begin{aligned}H(-\lambda_i) \tilde{C}_i &= \hat{H}(-\lambda_i) \tilde{C}_i, \\ \tilde{C}_i H(-\lambda_i) &= \tilde{C}_i \hat{H}(-\lambda_i).\end{aligned}$$

Similarly, introducing the *dual transfer function* as $G(s) = C^T(sF - A)^{-1}C$, and deriving necessary conditions as well, we finally arrive at

$$\begin{aligned} G(-\mu_i)\tilde{D}_i &= \hat{G}(-\mu_i)\tilde{D}_i, \\ \tilde{D}_i G(-\mu_i) &= \tilde{D}_i \hat{G}(-\mu_i), \\ \tilde{D}_i G'(-\mu_i)\tilde{D}_i &= \tilde{D}_i \hat{G}'(-\mu_i)\tilde{D}_i, \end{aligned}$$

where $\tilde{D} = \hat{D}R$ and $\hat{B} = \hat{E}R\Theta R^{-1}$, $\Theta = \mathrm{diag}(\mu_1, \ldots, \mu_k)$. Due to the fact that the tangential directions as well as the interpolation points are determined by the reduced pair of dual matrices there is obviously a close connection to the case of optimal $\mathcal{H}_2$-model reduction. Hence, one can adapt the iterative rational Krylov algorithm by means of appropriately changing the shift strategy. This then leads to Algorithm 2 which converges to a reduced set of matrices fulfilling these optimality conditions.

Algorithm 2 Sylvester IRKA (Symmetric)

Input: A, B, C, D, E, F, $\hat{A}$, $\hat{B}$, $\hat{C}$, $\hat{D}$, $\hat{E}$, $\hat{F}$
Output: X^{opt}

1: **while** (change in $\Lambda, \Theta > 0$) **do**
2: $\hat{A}Q = \hat{F}Q\Lambda, \hat{B}R = \hat{E}R\Theta, Q^T\hat{F}Q = R^T\hat{E}R = I, \tilde{C} = \hat{C}^TQ, \tilde{D} = \hat{D}R$
3: $V_i = (-\mu_i F - A)^{-1}C\tilde{D}_i \quad W_i = (-\lambda_i E - B)^{-1}D^T\tilde{C}_i$
4: $V = \mathrm{orth}\,(V)$, $W = \mathrm{orth}\,(W)$
5: $\hat{A} = V^TAV$, $\hat{B} = W^TBW$, $\hat{C} = V^TC$, $\hat{D} = DW$, $\hat{E} = W^TEW$, $\hat{F} = V^TFV$
6: **end while**
7: $\hat{A}^{opt} = \hat{A}$, $\hat{B}^{opt} = \hat{B}$, $\hat{C}^{opt} = \hat{C}$, $\hat{D}^{opt} = \hat{D}$, $\hat{E}^{opt} = \hat{E}$, $\hat{F}^{opt} = \hat{F}$
8: Solve $\hat{A}^{opt}\hat{X}\hat{E}^{opt} + \hat{F}^{opt}\hat{X}\hat{B}^{opt} + \hat{C}^{opt}\hat{D}^{opt} = 0$.
9: $X^{opt} = V\hat{X}W^T$

Remark 4.1. *Due to the connection to optimal $\mathcal{H}_2$-model reduction, it should be mentioned that instead of Step 3 of Algorithm 2, one can alternatively solve two reduced Sylvester equations of the form*

$$\begin{aligned} AV\hat{E} + FV\hat{B} + C\hat{D} &= 0, \\ EW\hat{A} + BW\hat{F} + D^T\hat{C}^T &= 0. \end{aligned}$$

For a robust solver for this type of equations, we refer to e.g. [5].

It remains to show that, in case of convergence of Algorithm 2, it holds $f(\Sigma_{err}) = f(\Sigma) - f(\hat{\Sigma})$. However, from the proof of Lemma 4.1, recall that this is equivalent to showing $\mathrm{tr}\left(\mathcal{C}^T\mathcal{Y}\hat{D}^T\right) = 0$ and $\mathrm{tr}\left(\hat{C}^T\mathcal{Z}\mathcal{D}^T\right) = 0$. This

is easily seen as follows. We know that

$$\begin{aligned}\operatorname{tr}\left(\mathcal{C}^T\mathcal{Y}\hat{D}^T\right) = \operatorname{vec}(I_p)\left(\hat{D}\otimes C^T\right)\left(\hat{E}\otimes A+\hat{B}\otimes F\right)^{-1}\left(\hat{D}^T\otimes C\right)\operatorname{vec}(I_p)\\ -\operatorname{vec}(I_p)\left(\hat{D}\otimes \hat{C}^T\right)\left(\hat{E}\otimes \hat{A}+\hat{B}\otimes \hat{F}\right)^{-1}\left(\hat{D}^T\otimes \hat{C}\right)\operatorname{vec}(I_p).\end{aligned}$$

From [4], we know that V and W can also be computed using vectorized notation, i.e.

$$\begin{aligned}\operatorname{vec}(V) &= \left(-\Theta\otimes F-\hat{I}\otimes A\right)^{-1}\left(\tilde{D}^T\otimes C\right)\operatorname{vec}(I_p),\\ \operatorname{vec}(W) &= \left(-\Lambda\otimes E-\hat{I}\otimes B\right)^{-1}\left(\tilde{C}^T\otimes D^T\right)\operatorname{vec}(I_p).\end{aligned}$$

Moreover, for $z\subset \operatorname{span}(\operatorname{vec}(V))$, it holds $(\hat{I}\otimes VV^T)z=z$. Next, let us have a look at

$$\operatorname{vec}(I_p)\left(\hat{D}\otimes \hat{C}^T\right)\left(\hat{E}\otimes \hat{A}+\hat{B}\otimes \hat{F}\right)^{-1}\left(\hat{D}^T\otimes \hat{C}\right)\operatorname{vec}(I_p).$$

Using the spectral decomposition from Algorithm 2, this expression is equivalent to

$$\operatorname{vec}(I_p)\left(\tilde{D}\otimes \hat{C}^T\right)\left(\hat{I}\otimes \hat{A}+\Theta\otimes \hat{F}\right)^{-1}\left(\tilde{D}^T\otimes \hat{C}\right)\operatorname{vec}(I_p). \tag{15}$$

Finally, due to the mentioned properties of the projection matrix V, (15) can be transformed into

$$\operatorname{vec}(I_p)\left(\tilde{D}\otimes C^T\right)\left(\hat{I}\otimes A+\Theta\otimes F\right)^{-1}\left(\tilde{D}^T\otimes C\right)\operatorname{vec}(I_p)$$

which in turn yields $\operatorname{tr}\left(\mathcal{C}^T\mathcal{Y}\hat{D}^T\right)=0$. Similarly, we obtain $\operatorname{tr}\left(\hat{C}^T\mathcal{Z}\mathcal{D}^T\right)=0$. Analog to the proof of Theorem 3.1, one can eventually show that

$$\operatorname{vec}\left(X-V\hat{X}W^T\right)^T\mathcal{L}_S\operatorname{vec}\left(X-V\hat{X}W^T\right)=f(\Sigma)-f(\hat{\Sigma}).$$

Altogether, we have proven our main result.

Theorem 4.1. *Let $\Sigma=(A,B,C,D,E,F)$ denote a set of matrices determining a Sylvester equation as in (1) with solution X. Assume that $A=A^T\prec 0, B=B^T\prec 0, E=E^T\succ 0$ and $F=F^T\succ 0$. Let further X^{opt} be computed by Algorithm 2. Then X^{opt} is a local minimizer of*

$$\min_{X_k\in\mathcal{M}}\{(\operatorname{vec}(X-X_k))^T\mathcal{L}_S\operatorname{vec}(X-X_k)\}.$$

The unsymmetric case

For sake of completeness, let us briefly analyze the case

$$AX + XB + cd^T = 0,$$

where $A \neq A^T \in \mathbb{R}^{n\times n}, B \neq B^T \in \mathbb{R}^{m\times m}, c \in \mathbb{R}^n$ and $d \in \mathbb{R}^m$. Similar to the Lyapunov equations, in order to ensure the possibility of defining an appropriate norm, we will have to compute symmetrizers $J_1 = J_1^T$ and $J_2 = J_2^T$ s.t. $J_1A = A^TJ_1$ and $BJ_2 = J_2B^T$. For this, we assume that (A, c, c^T) and (B, d, d^T) are dynamical systems that are at least equivalent to state space symmetric systems. Following the approach of the previous section, we can then always find matrices which additionally satisfy $J_1c = c$ and $d^TJ_2 = d^T$. As a consequence, we are thus faced with the transformed equation

$$J_1AXJ_2 + J_1XBJ_2 + J_1cd^TJ_2 = 0,$$

for which we again define an energy norm as

$$||\cdot||_{\bar{\mathcal{L}}_S} = \sqrt{\langle\cdot,\cdot\rangle_{\bar{\mathcal{L}}_S}} \quad \text{with} \quad \langle u, v\rangle_{\bar{\mathcal{L}}_S} = \langle u, \bar{\mathcal{L}_S} v\rangle$$

and

$$\bar{\mathcal{L}}_S := -J_2 \otimes J_1A - BJ_2 \otimes J_1.$$

Although by means of Algorithm 2, we now would be able to compute a locally optimal low rank approximation to X, in practice we obviously want to avoid the ill-conditioned computation of the symmetrizers J_1 and J_2. Instead, the goal is to directly operate on the original matrices A and B. Completely analog to the $\mathcal{H}_2$-optimal model reduction problem, this will yield the necessity of using oblique projections for the reduction. Note that the first order necessary conditions on $f(\Sigma_{err})$ will no longer include tangential directions. However, we will still have to make sure that

$$\begin{aligned} H(-\lambda_i) &= \hat{H}(-\lambda_i), & H'(-\lambda_i) &= \hat{H}'(-\lambda_i), \\ G(-\mu_i) &= \hat{G}(-\mu_i), & G'(-\mu_i) &= \hat{G}'(-\mu_i), \end{aligned}$$

where $H(s) = d^T(sI_m - B)^{-1}d$ and $G(s) = c^T(sI_n - A)^{-1}c$. A slight modification of Algorithm 2 then leads to Algorithm 3.

Algorithm 3 Sylvester IRKA (Unsymmetric)

Input: A, B, c, d, $\hat{A}$, $\hat{B}$, $\hat{c}$, $\hat{d}$
Output: X^{opt}

1: **while** (change in $\Lambda, \Theta > 0$) **do**
2: $\hat{A}Q = Q\Lambda, \hat{B}R = R\Theta$
3: $U_i = (-\mu_i I_n - A)^{-1}c, \quad V_i = (-\mu_i I_n - A^T)^{-1}c,$
4: $W_i = (-\lambda_i I_m - B)^{-1}d, \quad Z_i = (-\lambda_i I_m - B^T)^{-1}d$
5: $U = \text{orth}\,(U)$, $V = \text{orth}\,(V)$, $W = \text{orth}\,(W)$, $Z = \text{orth}\,(Z)$
6: $\hat{A} = (V^TU)^{-1}V^TAU$, $\hat{B} = (Z^TW)^{-1}Z^TBW$,
$\hat{c} = (V^TU)^{-1}V^Tc$, $\hat{d} = W^Td$
7: **end while**
8: $\hat{A}^{opt} = \hat{A}$, $\hat{B}^{opt} = \hat{B}$, $\hat{c}^{opt} = \hat{c}$, $\hat{d}^{opt} = \hat{d}$
9: Solve $\hat{A}^{opt}\hat{X} + \hat{X}\hat{B}^{opt} + \hat{c}^{opt}(\hat{d}^{opt})^T = 0.$
10: $X^{opt} = U\hat{X}(Z^TW)^{-1}Z^T$

Though it might not be obvious at a first glance how to construct the approximations, a careful analysis justifies the following result.

Corollary 4.1. *Let X^{opt} be constructed by Algorithm 3 applied to the original matrices (A, B, c, d). Let further (A, c, c^T) and (B, d, d^T) be dynamical systems that can be transformed into state space symmetric realizations. Then $X^{opt} = U\hat{X}(Z^TW)^{-1}Z^T$ is a local minimizer of*

$$\min_{X_k \in \mathcal{M}} \{(\text{vec}\,(X - X_k))^T \bar{\mathcal{L}}_S \,\text{vec}\,(X - X_k)\}.$$

Remark 4.2. *As in the Lyapunov case, it is possible to show equivalence between the rational Krylov framework and the ADI iteration for Sylvester equations. For the assumptions and a more detailed discussion we refer to [13].*

5 Numerical examples

In this section, we will study the performance of the proposed algorithms by means of some standard numerical test examples. As stopping criterion for the iterative algorithms we always use a relative residual of 10^{-8}. All simulations were generated on an Intel® Dual-Core CPU E5400, 2 MB cache, 3 GB RAM, Ubuntu Linux 10.04 (x86_64), MATLAB® Version 7.11.0 (R2010b) 64-bit (glnxa64).

Lyapunov equations

The first example is a semi-discretized heat transfer problem from the Oberwolfach benchmark collection.[1] Here, we used the coarsest discretization

[1] `http://portal.uni-freiburg.de/imteksimulation/downloads/benchmark`

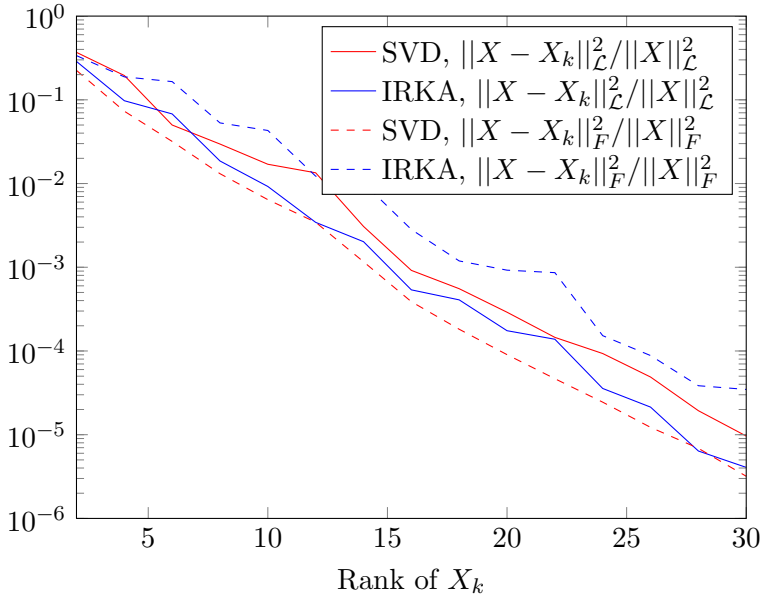

Figure 1: Steel profile

leading to symmetric matrices $E, A \in \mathbb{R}^{1357\times 1357}$ together with the input matrix $C \in \mathbb{R}^{1357\times 6}$, see [7]. In Figure 5, we present a comparison between the singular value decomposition based approximation of the exact solution with the approximation given by the rational Krylov space obtained by IRKA. As expected, the relative error in the Frobenius norm is obviously better when the SVD approximation is used. However, the slope of the IRKA approximation is almost parallel to that and for an approximation of rank 30, the relative error is only one order of magnitude smaller than the best approximation given by the SVD. On the other hand, we see that IRKA outperforms the SVD for nearly every rank k when the relative error is measured in terms of the $\mathcal{L}-$norm introduced in [25]. The few values of k where this is not the case may be explained by the fact that IRKA only is able to find a local minimum of the underlying $\mathcal{H}_2$-model reduction problem.

The second example is also quite common in the context of model order reduction. The symmetric system matrices $E, A \in \mathbb{R}^{1668\times 1668}$ and $C \in \mathbb{R}^{1668\times 5}$ stem from the finite element discretization of a thermal model of a filter device and thus are sparse, see [17]. Similar to the previous example, from Figure 5 we can again conclude that the SVD approximation dominates the performance with respect to the Frobenius norm while the IRKA approach performs better when the $\mathcal{L}$-norm is taken as a basis for judgment. However, here one can observe a constant improvement of the SVD approximation for ranks larger than 30.

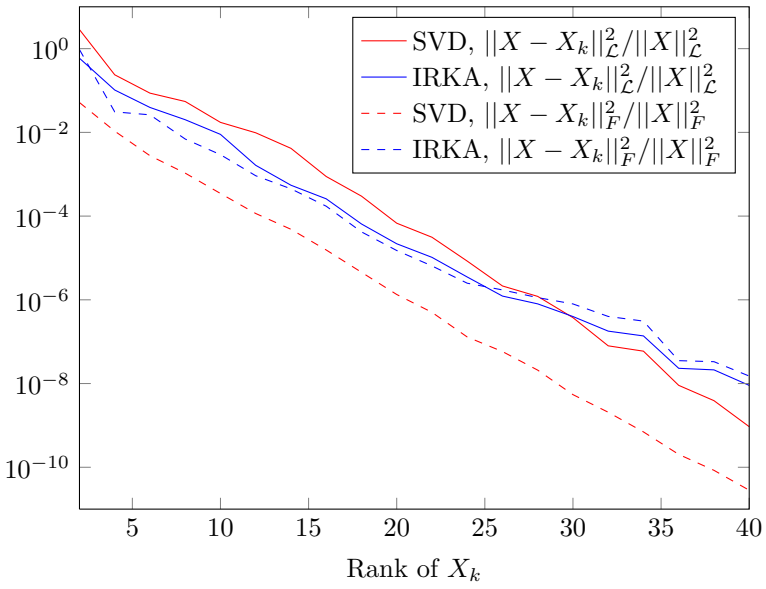

Figure 2: Tunable optical filter

Our final example concerning the Lyapunov equations is a clamped beam model taken from the Slicot model reduction benchmark collection. Though the matrices $A \in \mathbb{R}^{348\times 348}, c, d \in \mathbb{R}^{348\times 1}$ are smaller than in the previous cases we are now faced with an unsymmetric matrix $A \neq A^T$, see [8].[2] Recall from Section 3 that the two-sided IRKA still implicitly yields an optimal approximation subspace. However, it is obvious that the computation of the observability and controllability matrices $\mathcal{K}_c$ and $\mathcal{K}_d$ is infeasible due to numerical instabilities caused by the ill-conditionedness. For this reason, in Figure 5, we show the results for the approximations of the two Lyapunov equations

$$AX + XA^T + cc^T = 0 \quad \text{and} \quad A^TY + YA + dd^T = 0$$

by means of the relative error measured in the Frobenius norm. Despite the fact that the IRKA approximants lead to larger relative errors than those resulting from the SVD of the true solution, we observe a monotone decrease with satisfying relative errors in the range of 10^{-9} for $k = 20$. Moreover, the approach clearly outperforms the approximations of same rank computed by KPIK. However, recall that KPIK is computationally far more attractive than IRKA which has to be taken into account when computing low rank approximations.

Since in Section 3 we have seen that the $\mathcal{H}_2$-norm of the error system only vanishes if the corresponding approximation of the Lyapunov equation

[2] `http://www.slicot.org`

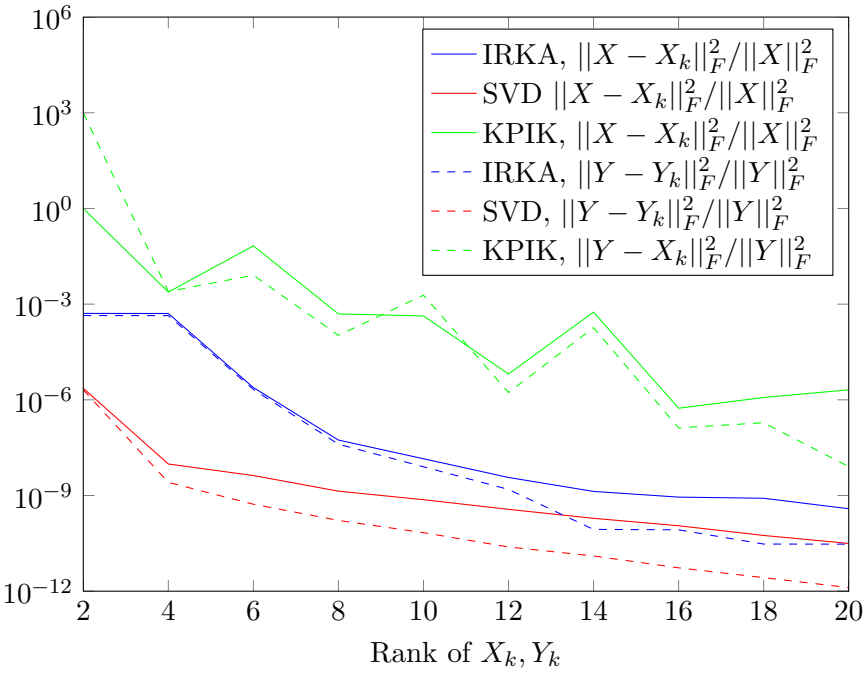

Figure 3: Clamped beam

is exact, in Figure 5 we plot a comparison between the $\mathcal{H}_2$-norm of the error system corresponding to a reduced order model and the error of the associated Lyapunov approximation $X_k = V\hat{X}V^T$ measured in the Frobenius norm. The results belong to four completely random stable SISO systems of dimension $n = 200$ with complex poles. Each reduced model is obtained by a one-sided interpolatory projection $\mathcal{P} = VV^T$ with interpolation points randomly chosen within the interval $[0, 1000]$. The solution of the original Lyapunov equation is denoted with X. As Figure 5 indicates, the $\mathcal{H}_2$-error behaves in a similar way as the Frobenius error $X - X_k$ indicating that a locally $\mathcal{H}_2$-reduced model might yield accurate low rank Lyapunov approximations as well.

Sylvester equations

Let us now finally draw our attention to the more general case specified in 2. Analogue to the Lyapunov case, our first example is given by the process of optimal cooling of steel profiles. In order to end up with a general Sylvester equation including different matrix dimensions, we have used matrix sets (A, B, C, D, E, F), where A, E, C is as specified above while B, F, D is obtained by a finer resolution with mesh size $m = 5177$. In Figure 5, we see a comparison between the rational Krylov subspace approximation computed

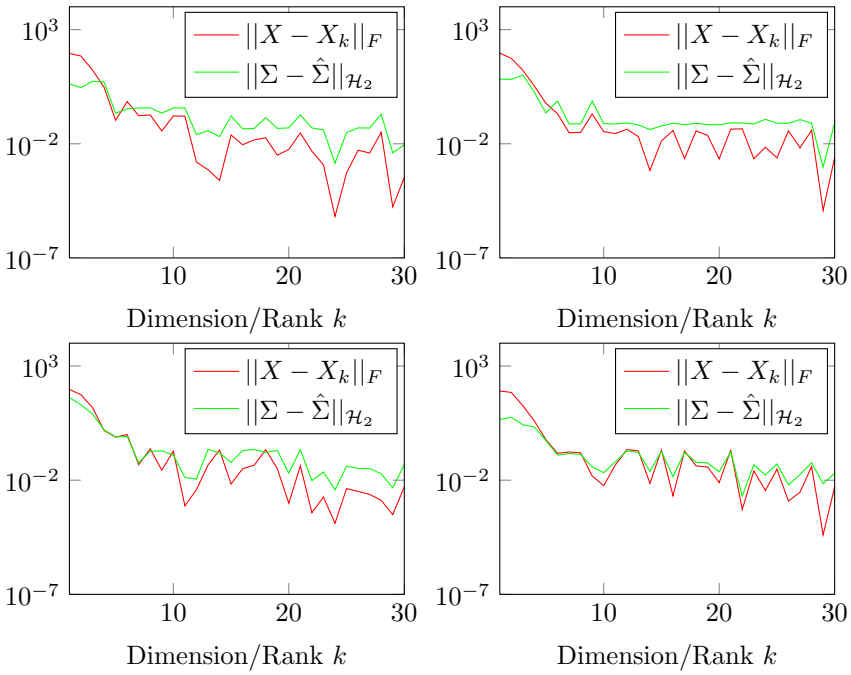

Figure 4: Frobenius error and $\mathcal{H}_2$-error

by Algorithm 2 which is abbreviated with SIRKA and the SVD-based approximation. Due to the lack of an exact solver for the original Sylvester equation, we used our new method with an approximation of rank 250 for reference values. It should be mentioned that the relative residual for this approach was smaller than 10^{-13} and thus should be sufficient for comparison. Again, we see that SIRKA is dominated by the SVD approximation if the quality is measured in terms of the Frobenius norm while it performs constantly better if we use the $\mathcal{L}_S$-norm from Section 4.

For reasons of completeness, we also present one final example for an unsymmetric Sylvester equation. The matrices are obtained by combination of a continuous heat transfer model as well as the clamped beam. Both models can be found in the Slicot benchmark collection and exhibit unsymmetric matrices $A \neq A^T \in \mathbb{R}^{200\times 200}, C \in \mathbb{R}^{200\times 1}$ and B, D as above. Of course, the use of such a Sylvster equation with completely unrelated models may be questionable. However, here they should only serve as realistic test matrices and for our purposes thus are sufficient. In Figure 5, we present the results for the approximations obtained by Algorithm 3, again abbreviated with SIRKA and compare them with the SVD approximation of the true solution.

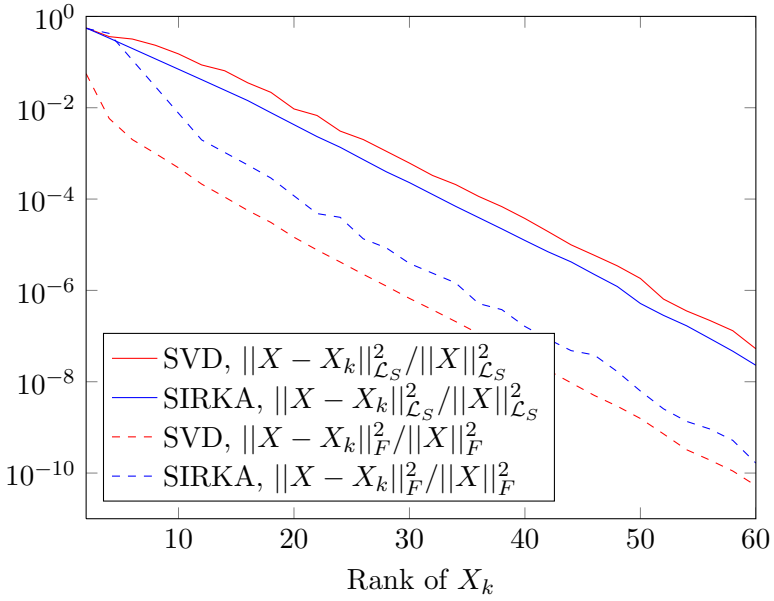

Figure 5: Steel profiles with different mesh sizes

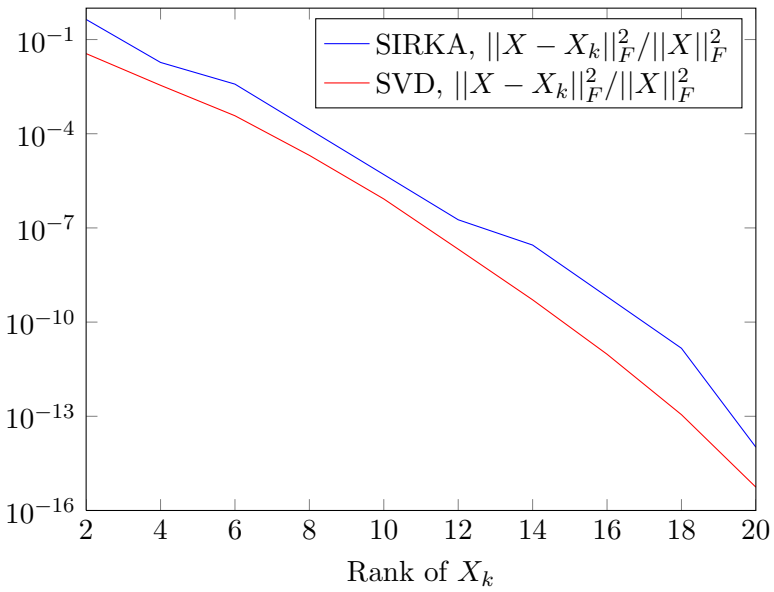

Figure 6: Clamped beam and heat equation

6 Conclusions and Outlook

In this paper we have studied low rank approximations of large-scale matrix equations including the special case of Lyapunov equations arising in the context of dynamical systems. For symmetric systems, we have shown that the rational Krylov subspaces that lead to local minimizers of the underlying $\mathcal{H}_2$-model reduction problem additionally yield optimal low rank approximations with respect to a certain energy norm. Hence, we could show a close connection to the recently proposed solvers based on Riemannian optimization, see [25]. We further extended our results to the case of unsymmetric Lyapunov equations with rank one right hand side. Here, under the assumption that the systems can be transformed into a state space symmetric realization, we have shown that the two-sided IRKA implicitly minimizes a symmetrized system obtained by means of the controllability and observability matrix, respectively. Moreover, we have proven that the $\mathcal{H}_2$-norm of the error system arising in model order reduction is an objective function which vanishes if and only if the corresponding low rank Lyapunov approximations are the exact solutions and thus might be used as a reasonable error measure. Furthermore, we extended our discussion to the case of more general Sylvester equations. For this, we introduced an appropriate extension of the iterative rational Krylov algorithm which also minimizes the corresponding energy norm induced by the Kronecker notation of the associated Sylvester operator.

As an outlook for further research, we want to draw our attention to the case of more general matrix equations of the form

$$AXE + FXB + \sum_{j=1}^{m} N_j X M_j + CD = 0,$$

which have been shown to possess possible applications in the context of model reduction of bilinear and linear stochastic control systems.

Acknowledgements

We thank Bart Vandereycken from EPF Lausanne who pointed out a mistake in the initial version of this manuscript. Furthermore, we thank Garret Flagg from Virginia Tech for some interesting discussions that helped to improve the presentation of the paper.

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] R.H. Bartels and G.W. Stewart. Solution of the matrix equation $AX + XB = C$: Algorithm 432. *Comm. ACM*, 15:820–826, 1972.

[4] P. Benner and T. Breiten. Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems, 2011. Preprint MPIMD/11–02.

[5] P. Benner, M. Köhler, and J. Saak. Sparse-dense sylvester equations in $\mathcal{H}_2$-model order reduction, 2011. Preprint MPIMD/11–11.

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

[7] P. Benner and J. Saak. Efficient numerical solution of the LQR-problem for the heat equation. *Proc. Appl. Math. Mech.*, 4(1):648–649, 2004.

[8] Y. Chahlaoui and P. Van Dooren. A collection of benchmark examples for model reduction of linear time invariant dynamical systems. SLICOT Working Note 2002–2, 2002.

[9] V. Druskin, L. Knizhnerman, and V. Simoncini. Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation. *SIAM J. Numer. Anal.*, 49:1875–1898, 2011.

[10] E.V. Dulov and N.A. Andrianova. On differentiability of the matrix trace operator and its applications. *Korean J. Comput. Appl. Math.*, 8:97–109, 2001.

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

[12] G. Flagg, C.A. Beattie, and S. Gugercin. Convergence of the Iterative Rational Krylov Algorithm. Technical report, 2011. submitted, available as arXiv:1107.5363v1.

[13] 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.

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

[15] S. Gugercin, A.C. Antoulas, and S. Beattie. $\mathcal{H}_2$ Model Reduction for large-scale dynamical systems. *SIAM J. Matrix Anal. Appl.*, 30(2):609–638, 2008.

[16] S.J. Hammarling. Numerical solution of the stable, non-negative definite Lyapunov equation. *IMA J. Numer. Anal.*, 2:303–323, 1982.

[17] D. Hohlfeld and H. Zappe. An all-dielectric tunable optical filter based on the thermo-optic effect. *Journal of Optics A: Pure and Applied Optics*, 6:504–511, 2004.

[18] P. Lancaster and M. Tismenetsky. *The Theory of Matrices*. Academic Press, Orlando, 2nd edition, 1985.

[19] J.-R. Li. *Model Reduction of Large Linear Systems via Low Rank System Gramians*. PhD thesis, Massachusettes Institute of Technology, September 2000.

[20] L. Meier and D.G. Luenberger. Approximation of linear constant systems. *IEEE Trans. Automat. Control*, 12(5):585–588, 1967.

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

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

[23] D.C. Sorensen and A.C. Antoulas. The Sylvester equation and approximate balanced reduction. *Linear Algebra Appl.*, 351–352:671–700, 2002.

[24] D.C. Sorensen and Y. Zhou. Bounds on eigenvalue decay rates and sensitivity of solutions of lyapunov equations. Technical Report 7, Rice University, 2002.

[25] B. Vandereycken and S. Vandewalle. A Riemannian optimization approach for computing low-rank solutions of Lyapunov equations. *SIAM J. Matrix Anal. Appl.*, 31(5):2553–2579, 2010.

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

[27] Y. Zhou and D. Sorensen. Approximate implicit subspace iteration with alternating directions for LTI system model reduction. *Numer. Lin. Alg. Appl.*, 15:873–886, 2008.