Peter Benner Tobias Breiten

Interpolation-Based $\mathcal{H}_2$-Model Reduction of Bilinear Control Systems

Max Planck Institute Magdeburg Preprints

MPIMD/10-02 **June 8, 2011**

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 the problem of optimal model order reduction of bilinear control systems with respect to the generalization of the well-known $\mathcal{H}_2$-norm for linear systems. We revisit existing first order necessary conditions for $\mathcal{H}_2$-optimality based on the solutions of generalized Lyapunov equations arising in bilinear system theory and present an iterative algorithm which, upon convergence, will yield a reduced system fulfilling these conditions. While this approach relies on the solution of certain generalized Sylvester equations, we will establish a connection to another method based on generalized rational interpolation. This will lead to another way of computing the $\mathcal{H}_2$-norm of a bilinear system and will extend the pole-residue optimality conditions for linear systems, also allowing for an adaption of the successful iterative rational Krylov algorithm (IRKA) to bilinear systems. By means of several numerical examples, we will then demonstrate that the new techniques outperform the method of balanced truncation for bilinear systems with regard to the relative $\mathcal{H}_2$-error.

Keywords: model order reduction, bilinear systems, $\mathcal{H}_2$-optimality, Sylvester equations

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

The need for efficient numerical treatment of complex dynamical processes often leads to the problem of model order reduction, i.e. the approximation of large-scale systems resulting from e.g., partial differential equations, by significantly smaller ones. Since model reduction of linear systems has been studied for several years now, there exists a well established theory including error bounds and structure-preserving properties fulfilled by a reduced-order model. However, although there are still a lot of open and worthwhile problems, recently more and more attention has been paid to nonlinear systems which are inevitably more complicated. As a first step into this direction, the class of bilinear systems has been pointed out to be an interesting interface between fully nonlinear and linear control systems. More precisely, these special systems are of the form

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

with $A, N_k \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $u(t) = \begin{bmatrix} u_1(t) & \dots & u_m(t) \end{bmatrix}^T \in \mathbb{R}^m$, $y(t) \in \mathbb{R}^p$. Due to their structure, which is obviously closely related to the state space representation of linear systems, many concepts known from linear model order reduction have been shown to possess bilinear analogues. As was already discussed in [9, 22], a variety of biological, physical and economical phenomena naturally result in bilinear models. Here, models for nuclear fission, mechanical brakes or biological species can be mentioned as typical examples. Interestingly enough, a completely similar structure is obtained for a certain type of linear stochastic differential equations. Some interesting applications like, e.g., the Fokker-Planck equation, are discussed in [19]. Coming back to the actual reduction problem, let us recall that we are formally aiming at the construction of another bilinear system

$$\hat{\Sigma} : \begin{cases} \dot{\hat{x}}(t) = \hat{A}\hat{x}(t) + \sum_{k=1}^{m} \hat{N}_k \hat{x}(t) u_k(t) + \hat{B}u(t), \\ \hat{y}(t) = \hat{C}\hat{x}(t), \quad \hat{x}(0) = \hat{x}_0, \end{cases} \tag{2}$$

with $\hat{A}, \hat{N}_k \in \mathbb{R}^{\hat{n}\times\hat{n}}$, $\hat{B} \in \mathbb{R}^{\hat{n}\times m}$, $\hat{C} \in \mathbb{R}^{p\times\hat{n}}$. Since $\hat{\Sigma}$ should approximate Σ in some sense, we certainly expect $\hat{y} \approx y$ for all admissible inputs $u \in L^2[0, \infty[$. Moreover, in order to ensure a significant speed-up in numerical simulations, we demand $\hat{n} \ll n$. There are different ways of achieving this goal. Similar to linear system theory, there exist SVD-based approaches leading to a reasonable generalization of the method of balanced truncation, see [6, 32]. While these methods have been proven to perform very well, they require the solution of two generalized Lyapunov equations which cause serious memory problems already for medium-sized systems. On the other hand, several interpolation-based ideas have evolved that try to approximate generalized transfer functions by projecting the

original model on appropriate Krylov subspaces, see [4, 5, 8, 12, 15, 25, 26]. Despite the fact that a memory efficient implementation is possible, the worse approximation quality compared to the method of balanced truncation make these approaches unfavorable. Moreover, while the choice of optimal interpolation points with respect to a certain norm has been solved for the linear case, see [11, 18], this is still an open question for bilinear system theory. The goal of this paper now is to reveal an appropriate generalized interpolation framework for bilinear systems that allows to propose two different iterative algorithms that aim at finding a local $\mathcal{H}_2$-minimum of the so-called error system. For the first one, we will have to study certain generalized Sylvester equations. The second approach extends the iterative rational Krylov algorithm (IRKA/MIRIAm), see [11, 18], to the bilinear case. We will now proceed as follows. In the subsequent section, we will give a brief review on optimal $\mathcal{H}_2$-model reduction for linear systems. This will include a recapitulation of first order necessary conditions as well as a discussion on the solution provided by IRKA. In Section 3, we will focus on the $\mathcal{H}_2$-norm for bilinear systems, initially introduced in [32]. Here, we present an alternative computation of the norm of the error system which, in Section 4, will enable us to derive first order necessary conditions that extend the ones known from the linear case. Finally, we will study several numerical examples which will underline the superiority of the methods proposed in Section 5 and conclude with a short summary.

2 $\mathcal{H}_2$-Optimal Model Reduction for Linear Systems

Since we will later on extend the concepts from linear $\mathcal{H}_2$-model reduction, we briefly review the existing theory for linear continuous time-invariant systems, i.e.

$$\Sigma_\ell : \begin{cases} \dot{x}(t) = A_\ell x(t) + B_\ell u(t), \\ y(t) = C_\ell x(t), \quad x(0) = x_0, \end{cases} \tag{3}$$

with dimensions as defined in (1) and transfer function $H_\ell(s) = C_\ell \left(sI_n - A_\ell\right)^{-1} B_\ell$. So far, we did not further specify criteria which allow to measure the quality of a reduced-order system. Here, we want to deal with the problem of finding a reduced-order model which approximates the original system as accurately as possible with respect to the $\mathcal{H}_2$-norm. Recall that for linear systems, this norm is defined as

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

where tr denotes the trace of a matrix. As is well-known, there exist two alternative computations for this norm. The first relies on the solution of the Lyapunov equations corresponding to the system, i.e.

$$A_\ell P_\ell + P_\ell A_\ell^T + B_\ell B_\ell^T = 0, \quad A_\ell^T Q_\ell + Q_\ell A_\ell + C_\ell^T C_\ell = 0.$$

It can be shown that it holds

$$||\Sigma_\ell||_{\mathcal{H}_2}^2 = \operatorname{tr}\left(C_\ell P_\ell C_\ell^T \right) = \operatorname{tr}\left(B_\ell^T Q_\ell B_\ell \right).$$

Rather recently, in [3], Antoulas provides a new derivation based on the poles and residues of the transfer function:

$$||\Sigma_\ell||^2_{\mathcal{H}_2} = \sum_{k=1}^{n} \operatorname{tr}\left(\operatorname{res}\left[H_\ell(-s)H_\ell^T(s), \lambda_k\right]\right),$$

where λ_k denotes the eigenvalues of the system matrix A_ℓ and

$$\operatorname{res}\left[H_\ell(-s)H_\ell^T(s), \lambda_k\right] = \lim_{s\to\lambda_k} H_\ell(-s)H_\ell^T(s)(s-\lambda_k).$$

Based on these expressions, it is possible to derive first order necessary conditions for $\mathcal{H}_2$-optimality, i.e. for locally minimizing the norm of the error system $||\Sigma_\ell - \hat{\Sigma}_\ell||_{\mathcal{H}_2}$, see e.g. [18, 21, 31]. On the one hand, the Lyapunov-based norm computation leads to the Wilson conditions

$$P_{12}^T Q_{12} + P_{22}Q_{22} = 0, \quad Q_{12}^T B + Q_{22}\hat{B} = 0, \quad \hat{C}P_{22} - CP_{12} = 0, \tag{4}$$

where

$$P^{err} = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix}, \quad Q^{err} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^T & Q_{22} \end{bmatrix},$$

are the solutions of the Lyapunov equations of the error system

$$A^{err} = \begin{bmatrix} A_\ell & 0 \\ 0 & \hat{A} \end{bmatrix}, \quad B^{err} = \begin{bmatrix} B_\ell \\ \hat{B} \end{bmatrix}, \quad C^{err} = \begin{bmatrix} C_\ell & -\hat{C} \end{bmatrix}.$$

Equivalently, it is possible to characterize the optimality via interpolation-based conditions. Initially derived in [21] and picked up again in [18, 10, 30], the reduced systems' transfer function has to tangentially interpolate the transfer function of the original system at the mirror images of its own poles, i.e. for $1 \le k \le \hat{n}$

$$\tilde{C}_k^T \hat{H}(-\hat{\lambda}_k) = \tilde{C}_k^T H(-\hat{\lambda}_k), \tag{5}$$

$$\hat{H}(-\hat{\lambda}_k)\tilde{B}_k = H(-\hat{\lambda}_k)\tilde{B}_k, \tag{6}$$

$$\tilde{C}_k^T \hat{H}'(-\hat{\lambda}_k)\tilde{B}_k = \tilde{C}_k^T H'(-\hat{\lambda}_k)\tilde{B}_k, \tag{7}$$

where $R\Lambda R^{-1} = \hat{A}$ is the spectral decomposition of $\hat{A}$ with $\Lambda = \operatorname{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{n}})$, $\tilde{B} = \hat{B}^T R^{-T}$, $\tilde{C} = \hat{C}R$ and the subscript k denotes the k-th column of a matrix. For later purposes, it is important to note that there is another way of writing down the above conditions. For this, we will make use of the Kronecker product notation and some simple properties of the vec operator:

$$\operatorname{tr}\left(X^T Y\right) = \operatorname{vec}(X)^T \operatorname{vec}(Y), \quad \operatorname{vec}(XYZ) = (Z^T \otimes X)\operatorname{vec}(Y). \tag{8}$$

Note that the right hand side of equation (5) consists of m columns. Considering now the j-th of those, we obtain:

$$\begin{aligned}
&\tilde{C}_k^T C_\ell \left(-\hat{\lambda}_k I_n - A_\ell\right)^{-1} B_j \\
&= \begin{pmatrix} \tilde{C}_1^T C_\ell & \dots & \tilde{C}_{\hat{n}}^T C_\ell \end{pmatrix} \begin{bmatrix} -\hat{\lambda}_1 I_n - A_\ell & & \\ & \ddots & \\ & & -\hat{\lambda}_{\hat{n}} I_n - A_\ell \end{bmatrix}^{-1} (e_k \otimes B_j) \\
&= \operatorname{vec}(C_\ell^T \tilde{C})^T \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A_\ell\right)^{-1} \left(e_k e_j^T \otimes B_\ell\right) \operatorname{vec}(I_m) \\
&= \operatorname{vec}(I_p)^T \left(\tilde{C} \otimes C_\ell\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A_\ell\right)^{-1} \left(e_k e_j^T \otimes B_\ell\right) \operatorname{vec}(I_m).
\end{aligned}$$

Hence, condition (5) is the same as requiring

$$\begin{aligned}
&\operatorname{vec}(I_p)^T \left(\tilde{C} \otimes \hat{C}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A}\right)^{-1} \left(e_k e_j^T \otimes \hat{B}\right) \operatorname{vec}(I_m) \\
&\quad = \operatorname{vec}(I_p)^T \left(\tilde{C} \otimes C_\ell\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A_\ell\right)^{-1} \left(e_k e_j^T \otimes B_\ell\right) \operatorname{vec}(I_m),
\end{aligned} \tag{9}$$

for $k = 1, \dots, \hat{n}$ and $j = 1, \dots, m$. Similarly, we can derive conditions equivalent to equations (6) and (7):

$$\begin{aligned}
&\operatorname{vec}(I_p)^T \left(e_j e_k^T \otimes \hat{C}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A}\right)^{-1} \left(\tilde{B}^T \otimes \hat{B}\right) \operatorname{vec}(I_m) \\
&\quad = \operatorname{vec}(I_p)^T \left(e_j e_k^T \otimes C_\ell\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A_\ell\right)^{-1} \left(\tilde{B}^T \otimes B_\ell\right) \operatorname{vec}(I_m),
\end{aligned} \tag{10}$$

$$\begin{aligned}
&\operatorname{vec}(I_p)^T \left(\tilde{C} \otimes \hat{C}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A}\right)^{-1} \left(e_k e_k^T \otimes I_{\hat{n}}\right) \times \\
&\qquad \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A}\right)^{-1} \left(\tilde{B}^T \otimes \hat{B}\right) \operatorname{vec}(I_m) \\
&\quad = \operatorname{vec}(I_p)^T \left(\tilde{C} \otimes C_\ell\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A_\ell\right)^{-1} \left(e_k e_k^T \otimes I_n\right) \times \\
&\qquad \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A_\ell\right)^{-1} \left(\tilde{B}^T \otimes B_\ell\right) \operatorname{vec}(I_m).
\end{aligned} \tag{11}$$

Based on these conditions, in [18, 10, 30], the authors have proposed iterative rational Krylov algorithms (IRKA/MIRIAm) which, upon convergence, yield a locally $\mathcal{H}_2$-optimal reduced system. Here, the crucial observation is that if we construct the reduced system by the Petrov-Galerkin projection $\mathcal{P} = VW^T$, i.e.

$$\hat{A} = W^T A_\ell V,\ \hat{B} = W^T B_\ell,\ \hat{C} = C_\ell V,$$

with $V = \begin{bmatrix} V_1 & \dots & V_{\hat{n}} \end{bmatrix}$ and $W = \begin{bmatrix} W_1 & \dots & W_{\hat{n}} \end{bmatrix}$ given as

$$V_i = \left(\sigma_i I_n - A_\ell\right)^{-1} B_\ell \tilde{B}_i, \tag{12}$$

$$W_i = \left(\sigma_i I_n - A_\ell^T\right)^{-1} C_\ell^T \tilde{C}_i, \tag{13}$$

we can guarantee that the transfer function of $\hat{\Sigma}_\ell$ tangentially interpolates the values and first derivatives of the original systems' transfer function at the points σ_i. Again, for later purposes it will be important to note that (12) and (13) can be rewritten by using a vectorized notation:

$$\operatorname{vec}(V) = (\operatorname{diag}(\sigma_1, \dots, \sigma_{\hat{n}}) \otimes I_n - I_{\hat{n}} \otimes A_\ell)^{-1} (\tilde{B}^T \otimes B_\ell) \operatorname{vec}(I_m), \tag{14}$$

$$\operatorname{vec}(W) = \left(\operatorname{diag}(\sigma_1, \dots, \sigma_{\hat{n}}) \otimes I_n - I_{\hat{n}} \otimes A_\ell^T\right)^{-1} (\tilde{C}^T \otimes C_\ell^T) \operatorname{vec}(I_p). \tag{15}$$

3 $\mathcal{H}_2$-Norm for Bilinear Systems

In this section, we will review a possible generalization of the $\mathcal{H}_2$-norm for bilinear systems introduced in [32].

Definition 3.1. *We define the $\mathcal{H}_2$-norm for bilinear systems as*

$$||\Sigma||_{\mathcal{H}_2}^2 = \operatorname{tr}\left(\sum_{k=1}^{\infty} \int_0^\infty \dots \int_0^\infty \sum_{\ell_1,\dots,\ell_k=1}^{m} g_k^{(\ell_1,\dots,\ell_k)} (g_k^{(\ell_1,\dots,\ell_k)})^T ds_1 \dots ds_k\right),$$

with $g_k^{(\ell_1,\dots,\ell_k)}(s_1, \dots, s_k) = Ce^{As_k} N_{\ell_1} e^{As_{k-1}} N_{\ell_2} \cdots e^{As_1} b_{\ell_k}$.

It has been shown that the above definition makes sense in case of the existence of certain generalized observability and reachability Gramians associated with bilinear systems. These, in turn, satisfy the generalized Lyapunov equations

$$AP + PA^T + \sum_{k=1}^{m} N_k P N_k^T + BB^T = 0, \tag{16}$$

$$A^T Q + QA^T + \sum_{k=1}^{m} N_k^T Q N_k + C^T C = 0, \tag{17}$$

and can be computed via the limit of an infinite series of linear Lyapunov equations. Basically, these assumptions are closely related to the notion of stability of Σ. For a more detailed insight, we refer to [32]. Hence, in the following we will always assume that the original system Σ is stable, meaning that the eigenvalues of the system matrix A lie in the open left complex plane and, moreover, the matrices N_k are sufficiently bounded. More precisely, we state the following result on bounded-input-bounded-output (BIBO) stability of bilinear systems, initially obtained in [29].

Theorem 3.1. *Let a bilinear system Σ be given and assume that A is asymptotically stable, i.e. there exist real scalars $\beta > 0$ and $0 < \alpha \leq -\max_i(\operatorname{Re}(\lambda_i(A)))$, such that*

$$||e^{At}|| \leq \beta e^{-\alpha t}, \quad t \geq 0.$$

Further assume that $||u(t)|| = \sqrt{\sum_{k=1}^m |u_k(t)|^2} \leq M$ uniformly on $[0, \infty[$ with $M > 0$, and denote $\Gamma = \sum_{k=1}^m ||N_k||$. Then Σ is BIBO stable, i.e. the corresponding Volterra series of the solution $x(t)$ uniformly converges on the interval $[0, \infty[$, if $\Gamma < \frac{\alpha}{M\beta}$.

Our stability assumption is motivated by the explicit solution formulas for equations (16) and (17) and the demand of having positive semi-definite solutions P and Q, respectively:

$$\operatorname{vec}(P) = -\left(A \otimes I_n + I_n \otimes A + \sum_{k=1}^{m} N_k \otimes N_k\right)^{-1} \operatorname{vec}(BB^T), \tag{18}$$

$$\operatorname{vec}(Q) = -\left(A^T \otimes I_n + I_n \otimes A^T + \sum_{k=1}^{m} N_k^T \otimes N_k^T\right)^{-1} \operatorname{vec}(C^T C). \tag{19}$$

Similarly to the linear case, the $\mathcal{H}_2$-norm now can be computed with the help of the solutions P and Q, see [32].

Proposition 3.1. *Let Σ be a bilinear system. Assume that A is asymptotically stable and the reachability Gramian P and the observability Gramian Q exist. Then it holds*

$$||\Sigma||_{\mathcal{H}_2}^2 = \operatorname{tr}\left(CPC^T\right) = \operatorname{tr}\left(B^T Q B\right).$$

Since in the subsequent section, we want to derive first order necessary conditions for $\mathcal{H}_2$-optimality that extend the interpolation conditions (5), (6) and (7) for linear systems, we propose the following alternative derivation.

Theorem 3.2. *Let Σ be a stable bilinear system. Then it holds*

$$||\Sigma||_{\mathcal{H}_2}^2 = \left(\operatorname{vec}(I_p)\right)^T (C \otimes C) \left(-A \otimes I_n - I_n \otimes A - \sum_{k=1}^{m} N_k \otimes N_k\right)^{-1} (B \otimes B) \operatorname{vec}(I_m).$$

Proof. For the proof, recall the properties from (8), together with the results from Proposition 3.1 and the solution formulas (18) and (19), respectively.

$$\begin{aligned}
||\Sigma||_{\mathcal{H}_2}^2 &= \operatorname{tr}\left(CPC^T\right) = \operatorname{vec}(C^T)^T \operatorname{vec}(PC^T) = \operatorname{vec}(C^T)^T (C \otimes I) \operatorname{vec}(P) \\
&= \operatorname{vec}(C^T)^T (C \otimes I) \left(-A \otimes I - I \otimes A - \sum_{k=1}^{m} N_k \otimes N_k\right)^{-1} \operatorname{vec}(BB^T) \\
&= \left(\left(C^T \otimes I\right) \operatorname{vec}(C^T)\right)^T \left(-A \otimes I - I \otimes A - \sum_{k=1}^{m} N_k \otimes N_k\right)^{-1} (B \otimes B) \operatorname{vec}(I_m) \\
&= \left(\operatorname{vec}(C^T C)\right)^T \left(-A \otimes I - I \otimes A - \sum_{k=1}^{m} N_k \otimes N_k\right)^{-1} (B \otimes B) \operatorname{vec}(I_m) \\
&= \left(\operatorname{vec}(C^T I_p C)\right)^T \left(-A \otimes I - I \otimes A - \sum_{k=1}^{m} N_k \otimes N_k\right)^{-1} (B \otimes B) \operatorname{vec}(I_m) \\
&= \left((C^T \otimes C^T) \operatorname{vec}(I_p)\right)^T \left(-A \otimes I - I \otimes A - \sum_{k=1}^{m} N_k \otimes N_k\right)^{-1} (B \otimes B) \operatorname{vec}(I_m)
\end{aligned}$$

$$= (\mathrm{vec}(I_p))^T (C \otimes C) \left(-A \otimes I - I \otimes A - \sum_{k=1}^{m} N_k \otimes N_k \right)^{-1} (B \otimes B)\, \mathrm{vec}(I_m).$$

□

4 $\mathcal{H}_2$-Optimality Conditions for Bilinear Systems

Next, we want to discuss necessary conditions for $\mathcal{H}_2$-optimality. As in the linear case, for this we have to consider the norm of the error system $\Sigma^{err} := \Sigma - \hat{\Sigma}$, which is defined as follows:

$$A^{err} = \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix}, \quad N_k^{err} = \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix}, \quad B^{err} = \begin{bmatrix} B \\ \hat{B} \end{bmatrix}, \quad C^{err} = \begin{bmatrix} C & -\hat{C} \end{bmatrix}.$$

Based on the assertions from Proposition 3.1, in [32], it is shown that the reduced system matrices have to fulfill conditions that extend the Wilson conditions to the bilinear case:

$$\begin{aligned} Q_{12}^T P_{12} + Q_{22} P_{22} = 0, \quad & Q_{22} \hat{N}_k P_{22} + Q_{12}^T N_k P_{12} = 0, \\ Q_{12}^T B + Q_{22} \hat{B} = 0, \quad & \hat{C} P_{22} - C P_{12} = 0, \end{aligned} \tag{20}$$

where

$$P^{err} = \begin{bmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{bmatrix}, \quad Q^{err} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^T & Q_{22} \end{bmatrix}, \tag{21}$$

are the solutions of the generalized Lyapunov equations

$$A^{err} P^{err} + P^{err} (A^{err})^T + \sum_{k=1}^{m} N_k^{err} P^{err} (N_k^{err})^T + B^{err} (B^{err})^T = 0, \tag{22}$$

$$(A^{err})^T Q^{err} + Q^{err} A^{err} + \sum_{k=1}^{m} (N_k^{err})^T Q^{err} N_k^{err} + (C^{err})^T C^{err} = 0. \tag{23}$$

Since we are heading for a generalization of the iterative rational Krylov algorithm, next we want to derive necessary conditions based on the computation formula from Theorem 3.2. A simple analysis of the structure of the error system leads to the following expression for the error functional E.

Corollary 4.1. *Let Σ and $\hat{\Sigma}$ be the original and reduced bilinear systems, respectively. Then*

$$\begin{aligned} E^2 := ||\Sigma^{err}||_{\mathcal{H}_2}^2 := ||\Sigma - \hat{\Sigma}||_{\mathcal{H}_2}^2 = (\mathrm{vec}(I_{2p}))^T \left(\begin{bmatrix} C & -\tilde{C} \end{bmatrix} \otimes \begin{bmatrix} C & -\hat{C} \end{bmatrix} \right) \times \\ \left(- \begin{bmatrix} A & 0 \\ 0 & \Lambda \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} - \sum_{k=1}^{m} \begin{bmatrix} N_k & 0 \\ 0 & \tilde{N}_k^T \end{bmatrix} \otimes \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix} \right)^{-1} \times \\ \left(\begin{bmatrix} B \\ \tilde{B}^T \end{bmatrix} \otimes \begin{bmatrix} B \\ \hat{B} \end{bmatrix} \right) \mathrm{vec}(I_{2m}), \end{aligned}$$

where $R \Lambda R^{-1} = \hat{A}$ is the spectral decomposition of $\hat{A}$ and $\tilde{B} = \hat{B}^T R^{-T}$, $\tilde{C} = \hat{C} R$, $\tilde{N}_k = R^T \hat{N}_k^T R^{-T}$.

Here, we assume the reduced system matrices left of the Kronecker products to be given by their eigenvalue decomposition. This is motivated by the demand of having optimization parameters Λ, $\tilde{N}_k$, $\tilde{B}$, and $\tilde{C}$ that can be chosen to minimize $||\Sigma - \hat{\Sigma}||^2_{\mathcal{H}_2}$, at least locally. Before we proceed, let us introduce a specific permutation matrix

$$M = \begin{bmatrix} I_{\hat{n}} \otimes \begin{bmatrix} I_n \\ \mathbf{0} \end{bmatrix} & I_{\hat{n}} \otimes \begin{bmatrix} \mathbf{0}^T \\ I_{\hat{n}} \end{bmatrix} \end{bmatrix},$$

which will simplify the computation of Kronecker products for certain block matrices. For this, consider one of the block structures appearing in Corollary 4.1 for which we can show:

$$\begin{aligned}
&M^T \left(\tilde{N}_k^T \otimes \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix} \right) M \\
&= \begin{bmatrix} I_{\hat{n}} \otimes \begin{bmatrix} I_n & \mathbf{0}^T \end{bmatrix} & I_{\hat{n}} \otimes \begin{bmatrix} \mathbf{0} & I_{\hat{n}} \end{bmatrix} \end{bmatrix} \left(\tilde{N}_k^T \otimes \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix} \right) \begin{bmatrix} I_{\hat{n}} \otimes \begin{bmatrix} I_n \\ \mathbf{0} \end{bmatrix} & I_{\hat{n}} \otimes \begin{bmatrix} \mathbf{0}^T \\ I_{\hat{n}} \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} I_{\hat{n}} \otimes \begin{bmatrix} I_n & \mathbf{0}^T \end{bmatrix} & I_{\hat{n}} \otimes \begin{bmatrix} \mathbf{0} & I_{\hat{n}} \end{bmatrix} \end{bmatrix} \begin{bmatrix} \tilde{N}_k^T \otimes \begin{bmatrix} N_k \\ \mathbf{0} \end{bmatrix} & \tilde{N}_k^T \otimes \begin{bmatrix} \mathbf{0}^T \\ \hat{N}_k \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} \tilde{N}_k^T \otimes N_k & 0 \\ 0 & \tilde{N}_k^T \otimes \hat{N}_k \end{bmatrix}.
\end{aligned}$$

If we now differentiate with respect to the optimization parameters, we obtain:

$$\begin{aligned}
\frac{\partial E^2}{\partial \tilde{C}_{ij}} &= (\mathrm{vec}(I_{2p}))^T \left(\begin{bmatrix} 0 & -e_i e_j^T \end{bmatrix} \otimes \begin{bmatrix} C & -\hat{C} \end{bmatrix} \right) \times \\
&\left(- \begin{bmatrix} A & 0 \\ 0 & \Lambda \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} - \sum_{k=1}^{m} \begin{bmatrix} N_k & 0 \\ 0 & \tilde{N}_k^T \end{bmatrix} \otimes \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix} \right)^{-1} \times \\
&\left(\begin{bmatrix} B \\ \tilde{B}^T \end{bmatrix} \otimes \begin{bmatrix} B \\ \hat{B} \end{bmatrix} \right) \mathrm{vec}(I_{2m}) \\
&= (\mathrm{vec}(I_p))^T \left(-e_i e_j^T \otimes \begin{bmatrix} C & -\hat{C} \end{bmatrix} \right) \times \\
&\left(-\Lambda \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} - I_{\hat{n}} \otimes \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix} \right)^{-1} \left(\tilde{B}^T \otimes \begin{bmatrix} B \\ \hat{B} \end{bmatrix} \right) \mathrm{vec}(I_m) \\
&= (\mathrm{vec}(I_p))^T \left(-e_i e_j^T \otimes \begin{bmatrix} C & -\hat{C} \end{bmatrix} \right) \times \\
&\left(M \left(- \begin{bmatrix} \Lambda \otimes I_n & 0 \\ 0 & \Lambda \otimes I_{\hat{n}} \end{bmatrix} - \begin{bmatrix} I_{\hat{n}} \otimes A & 0 \\ 0 & I_{\hat{n}} \otimes \hat{A} \end{bmatrix} - \sum_{k=1}^{m} \begin{bmatrix} \tilde{N}_k^T \otimes N_k & 0 \\ 0 & \tilde{N}_k^T \otimes \hat{N}_k \end{bmatrix} \right) M^T \right)^{-1} \times \\
&\left(\tilde{B}^T \otimes \begin{bmatrix} B \\ \hat{B} \end{bmatrix} \right) \mathrm{vec}(I_m)
\end{aligned}$$

$$= -(\mathrm{vec}(I_p))^T \left(e_i e_j^T \otimes C\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \left(\tilde{B}^T \otimes B\right) \mathrm{vec}(I_m)$$
$$+ (\mathrm{vec}(I_p))^T \left(e_i e_j^T \otimes \hat{C}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k\right)^{-1} \left(\tilde{B}^T \otimes \hat{B}\right) \mathrm{vec}(I_m).$$

Here, the last step is justified by the fact that M is a permutation matrix and, thus, $M^T M = I$ and the identities:

$$\left(-e_i e_j^T \otimes \begin{bmatrix} C & -\hat{C} \end{bmatrix}\right) M = \begin{bmatrix} -e_i e_j^T \otimes C & e_i e_j^T \otimes \hat{C} \end{bmatrix}, \quad M^T \left(\tilde{B} \otimes \begin{bmatrix} B \\ \hat{B} \end{bmatrix}\right) = \begin{bmatrix} \tilde{B} \otimes B \\ \tilde{B} \otimes \hat{B} \end{bmatrix}.$$

Setting the gained expression equal to zero reveals that $\hat{\Sigma}$ has to satisfy:

$$\begin{aligned} &(\mathrm{vec}(I_p))^T \left(e_i e_j^T \otimes C\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \left(\tilde{B}^T \otimes B\right) \mathrm{vec}(I_m) \\ &= (\mathrm{vec}(I_p))^T \left(e_i e_j^T \otimes \hat{C}\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes \hat{A} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k\right)^{-1} \left(\tilde{B}^T \otimes \hat{B}\right) \mathrm{vec}(I_m). \end{aligned} \tag{24}$$

In view of equation (6) in the form of (10), we see that this demand naturally extends the interpolation-based condition known from the linear case. For the differentiation with respect to the poles of $\hat{A}$, recall that we have $\frac{d}{d\hat{A}_{ij}} \left(\hat{A}^{-1}\right) = -\hat{A}^{-1} \frac{d\hat{A}}{d\hat{A}_{ij}} \hat{A}^{-1}$. Hence, we end up with

$$\begin{aligned} \frac{\partial E^2}{\partial \lambda_i} &= \mathrm{vec}(I_{2p})^T \left(\begin{bmatrix} C - \tilde{C} \end{bmatrix} \otimes \begin{bmatrix} C & -\hat{C} \end{bmatrix}\right) \times \\ &\left(\begin{bmatrix} A & 0 \\ 0 & \Lambda \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} + \sum_{k=1}^{m} \begin{bmatrix} N_k & 0 \\ 0 & \tilde{N}_k^T \end{bmatrix} \otimes \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix}\right)^{-1} \times \\ &\left(\begin{bmatrix} 0 & 0 \\ 0 & e_i e_i^T \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix}\right) \times \\ &\left(\begin{bmatrix} A & 0 \\ 0 & \Lambda \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} + \sum_{k=1}^{m} \begin{bmatrix} N_k & 0 \\ 0 & \tilde{N}_k^T \end{bmatrix} \otimes \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix}\right)^{-1} \times \\ &\left(\begin{bmatrix} B \\ \tilde{B}^T \end{bmatrix} \otimes \begin{bmatrix} B \\ \hat{B} \end{bmatrix}\right) \mathrm{vec}(I_{2m}) \\ &= \mathrm{vec}(I_p)^T \left(-\tilde{C} \otimes \begin{bmatrix} C & -\hat{C} \end{bmatrix}\right) \left(\Lambda \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} + I_{\hat{n}} \otimes \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} + \sum_{k=1}^{m} \tilde{N}^T \otimes \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix}\right)^{-1} \times \\ &\left(e_i e_i^T \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix}\right) \left(\Lambda \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_{\hat{n}} \end{bmatrix} + I_{\hat{n}} \otimes \begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} + \sum_{k=1}^{m} \tilde{N}_k^T \otimes \begin{bmatrix} N_k & 0 \\ 0 & \hat{N}_k \end{bmatrix}\right)^{-1} \times \\ &\left(\tilde{B}^T \otimes \begin{bmatrix} B \\ \hat{B} \end{bmatrix}\right) \mathrm{vec}(I_m) \end{aligned}$$

$$
\begin{aligned}
&= -\operatorname{vec}(I_p)^T \left(\tilde{C} \otimes C\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \times \\
&\quad \left(e_i e_i^T \otimes I_n\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \left(\tilde{B}^T \otimes B\right) \operatorname{vec}(I_m) \\
&\quad + \operatorname{vec}(I_p)^T \left(\tilde{C} \otimes \hat{C}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k\right)^{-1} \times \\
&\quad \left(e_i e_i^T \otimes I_{\hat{n}}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k\right)^{-1} \left(\tilde{B}^T \otimes \hat{B}\right) \operatorname{vec}(I_m).
\end{aligned}
$$

Once more, we find an interpolation-based condition generalizing (7) in the form of (11) if we set the last expression equal to zero:

$$
\begin{aligned}
&(\operatorname{vec}(I_p))^T \left(\tilde{C} \otimes C\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \times \\
&\quad \left(e_i e_i^T \otimes I_n\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \left(\tilde{B}^T \otimes B\right) \operatorname{vec}(I_m) \\
&= (\operatorname{vec}(I_p))^T \left(\tilde{C} \otimes \hat{C}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k\right)^{-1} \times \\
&\quad \left(e_i e_i^T \otimes I_{\hat{n}}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k\right)^{-1} \left(\tilde{B}^T \otimes \hat{B}\right) \operatorname{vec}(I_m).
\end{aligned}
\tag{25}
$$

Finally, as a matter of careful analysis, we obtain similar optimality conditions when differentiating with respect to $\tilde{B}$ and $\tilde{N}_k$, respectively:

$$
\begin{aligned}
&\operatorname{vec}(I_p)^T \left(\tilde{C} \otimes C\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \left(e_j e_i^T \otimes B\right) \operatorname{vec}(I_m) \\
&= \operatorname{vec}(I_p)^T \left(\tilde{C} \otimes \hat{C}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k\right)^{-1} \left(e_j e_i^T \otimes \hat{B}\right) \operatorname{vec}(I_m),
\end{aligned}
\tag{26}
$$

$$
\begin{aligned}
&\operatorname{vec}(I_p)^T \left(\tilde{C} \otimes C\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \times \\
&\quad \left(e_j e_i^T \otimes N\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \left(\tilde{B}^T \otimes B\right) \operatorname{vec}(I_m) \\
&= \operatorname{vec}(I_p)^T \left(\tilde{C} \otimes \hat{C}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k\right)^{-1} \times \\
&\quad \left(e_j e_i^T \otimes \hat{N}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k\right)^{-1} \left(\tilde{B}^T \otimes \hat{B}\right) \operatorname{vec}(I_m).
\end{aligned} \tag{27}
$$

Hence, the previous derivations can be summarized in the following theorem.

Theorem 4.1. *Let Σ denote a BIBO stable bilinear system. Assume that $\hat{\Sigma}$ is a reduced bilinear system of dimension $\hat{n}$, minimizing the $\mathcal{H}_2$-norm of the error system among all bilinear systems of dimension $\hat{n}$. Then $\hat{\Sigma}$ fulfills equations (24) - (27).*

5 Generalized Sylvester Equations and Bilinear IRKA

Now that we have specified first order necessary conditions for $\mathcal{H}_2$-optimality, in this section we will propose two algorithms that iteratively construct a reduced-order system which locally minimizes the $\mathcal{H}_2$-error. We will start with a procedure based on certain generalized Sylvester equations which in the linear case reduces to the concept discussed in [30]. For this, let us consider the following two matrix equations:

$$
A X + X \hat{A}^T + \sum_{k=1}^{m} N_k X \hat{N}_k^T + B \hat{B}^T = 0, \tag{28}
$$

$$
A^T Y + Y \hat{A} + \sum_{k=1}^{m} N_k^T Y \hat{N}_k - C^T \hat{C} = 0. \tag{29}
$$

Obviously, the solutions $X, Y \in \mathbb{R}^{n \times \hat{n}}$ can be explicitly computed by vectorizing both sides and making use of the vec-operator. However, this requires solving two linear systems of equations:

$$
\left(-I_{\hat{n}} \otimes A - \hat{A} \otimes I_n - \sum_{k=1}^{m} \hat{N}_k \otimes N_k\right) \operatorname{vec}(X) = \operatorname{vec}(B \hat{B}^T),
$$

$$
\left(I_{\hat{n}} \otimes A^T + \hat{A}^T \otimes I_n + \sum_{k=1}^{m} \hat{N}_k^T \otimes N_k^T\right) \operatorname{vec}(Y) = \operatorname{vec}(C^T \hat{C}).
$$

Throughout the rest of the paper, we will assume that there exist unique solutions satisfying these Sylvester equations. Due to the properties of the eigenvalue computation

of Kronecker products, this certainly is satisfied if the eigenvalues of $\hat{A}$ are located in $\mathbb{C}_-$ and the norms of $\hat{N}_k$ are sufficiently bounded. However, in view of Theorem 3.1 we have already mentioned that this basically characterizes a stable bilinear system. Although in general this cannot be ensured by our proposed algorithms, we did not observe unstable reduced-order systems so far. For a similar discussion for the linear case we refer to [18]. For the sake of completeness, we want to mention that under appropriate assumptions X and Y can be computed as the limit of an infinite series of linear Sylvester equations.

Lemma 5.1. *Let $\mathcal{L}, \Pi : \mathbb{R}^{n\times\hat{n}} \to \mathbb{R}^{n\times\hat{n}}$ denote two linear operators defined by the bilinear systems Σ and $\hat{\Sigma}$, with $\mathcal{L}(X) := AX + X\hat{A}^T$ and $\Pi(X) := \sum_{k=1}^m N_k X \hat{N}_k^T$. If the spectral radius $\rho(\mathcal{L}^{-1}\Pi) < 1$, then the solution X of the generalized Sylvester equation (28) is given as $X = \lim_{j\to\infty} X_j$, with:*

$$AX_1 + X_1\hat{A}^T + B\hat{B}^T = 0, \quad AX_j + X_j\hat{A}^T + \sum_{k=1}^m N_k X_{j-1}\hat{N}_k^T + B\hat{B}^T = 0, \quad j > 1.$$

A dual statement obviously is true for equation (29). Since the statement is a direct consequence of the theory of convergent splittings, we dispense with the proof and instead refer to [13] for an equivalent discussion on bilinear Lyapunov equations. Let us now focus on Algorithm 1 which in each step constructs a reduced system $\hat{\Sigma}$ by a Petrov-Galerkin type projection $\mathcal{P} = V(W^TV)^{-1}W^T$, determined by the solutions of the generalized Sylvester equations associated with the preceding system matrices.

Algorithm 1 Generalized Sylvester iteration

Input: A, N_k, B, C, $\hat{A}$, $\hat{N}_k$, $\hat{B}$, $\hat{C}$
Output: $\hat{A}^{opt}$, $\hat{N}_k^{opt}$, $\hat{B}^{opt}$, $\hat{C}^{opt}$

1: **while** (change in $\sigma(\hat{A}) > 0$) **do**
2: Solve $AX + X\hat{A}^T + \sum_{k=1}^m N_k X \hat{N}_k^T + B\hat{B}^T = 0$.
3: Solve $A^TY + Y\hat{A} + \sum_{k=1}^m N_k^T Y \hat{N}_k - C^T\hat{C} = 0$.
4: $V = \mathrm{orth}(X)$, $W = \mathrm{orth}(Y)$
5: $\hat{A} = \left(W^TV\right)^{-1} W^TAV$, $\hat{N}_k = \left(W^TV\right)^{-1} W^TN_kV$, $\hat{B} = \left(W^TV\right)^{-1} W^TB$, $\hat{C} = CV$
6: **end while**
7: $\hat{A}^{opt} = \hat{A}$, $\hat{N}_k^{opt} = \hat{N}_k$, $\hat{B}^{opt} = \hat{B}$, $\hat{C}^{opt} = \hat{C}$

Finally, we are ready to prove one of our two main results.

Theorem 5.1. *Assume Algorithm 1 converges. Then, $\hat{A}^{opt}$, $\hat{N}_k^{opt}$, $\hat{B}^{opt}$, $\hat{C}^{opt}$ fulfill the necessary $\mathcal{H}_2$-optimality conditions (20).*

Proof. Let $\bar{A}$, $\bar{N}_k$, $\bar{B}$, $\bar{C}$ denote the matrices corresponding to the next to last step in the while loop. Due to convergence, $\hat{\Sigma}^{opt}$ is a state space transformation of $\bar{\Sigma}$, i.e. $\exists T \in \mathbb{R}^{\hat{n}\times\hat{n}}$

nonsingular, such that

$$\bar{A} = T^{-1}\hat{A}^{opt}T,\ \bar{N}_k = T^{-1}\hat{N}_k^{opt}T,\ \bar{B} = T^{-1}\hat{B}^{opt},\ \bar{C} = \hat{C}^{opt}T.$$

Furthermore, according to step 4, we have

$$V^{opt} = X^{opt}F, \quad W^{opt} = Y^{opt}G,$$

with $F, G \in \mathbb{R}^{\hat{n}\times\hat{n}}$ nonsingular. Thus, it holds

$$\left((W^{opt})^T V^{opt}\right)^{-1}(W^{opt})^T = \left(G^T(Y^{opt})^T X^{opt}F\right)^{-1} G^T(Y^{opt})^T = F^{-1}\left((Y^{opt})^T X^{opt}\right)^{-1}(Y^{opt})^T.$$

From step 2, it follows

$$AX^{opt} + X^{opt}\bar{A}^T + \sum_{k=1}^{m} N_k X^{opt}\bar{N}_k^T + B\bar{B}^T = 0.$$

Hence,

$$\begin{aligned}
&\underbrace{F^{-1}\left(Y^{opt^T}X^{opt}\right)^{-1}(Y^{opt})^T}_{((W^{opt})^T V^{opt})^{-1}(W^{opt})^T} A \underbrace{X^{opt}F}_{V^{opt}} \\
&+ F^{-1}\bar{A}^T X^{opt}F + \sum_{k=1}^{m} F^{-1}\left((Y^{opt})^T X^{opt}\right)^{-1}(Y^{opt})^T N_k X^{opt}\bar{N}_k^T X^{opt}F \\
&\quad + F^{-1}\left((Y^{opt})^T X^{opt}\right)^{-1}(Y^{opt})^T B\bar{B}^T X^{opt}F = 0,
\end{aligned}$$

which implies

$$\hat{A}^{opt} + F^{-1}T^T(\hat{A}^{opt})^T T^{-T}F + \sum_{k=1}^{m}\hat{N}_k^{opt}F^{-1}T^T(\hat{N}_k^{opt})^T T^{-T}F + \hat{B}^{opt}(\hat{B}^{opt})^T T^{-T}F = 0.$$

Finally, we end up with

$$\hat{A}^{opt}F^{-1}T^T + F^{-1}T^T(\hat{A}^{opt})^T + \sum_{k=1}^{m}\hat{N}_k^{opt}F^{-1}T^T(\hat{N}_k^{opt})^T + \hat{B}^{opt}(\hat{B}^{opt})^T = 0.$$

From the last line and the fact that we assumed the reduced system to be stable, the solution of the generalized Lyapunov equation is unique and we conclude that $P_{22} = F^{-1}T^T$, were P_{22} is the lower right block from the partitioning in (21). Similarly, we obtain

$$A^T Y^{opt} + Y^{opt}\bar{A} + \sum_{k=1}^{m} N_k^T Y^{opt}\bar{N}_k - C^T\bar{C} = 0.$$

This leads to

$$F^T(X^{opt})^T A^T Y^{opt}((X^{opt})^T Y^{opt})^{-1}F^{-T} + F^T(X^{opt})^T Y^{opt}\bar{A}((X^{opt})^T Y^{opt})^{-1}F^{-T}$$
$$+\sum_{k=1}^{m} F^T(X^{opt})^T N_k^T Y^{opt}\bar{N}_k((X^{opt})^T Y^{opt})^{-1}F^{-T} - F^T(X^{opt})^T C^T\bar{C}((X^{opt})^T Y^{opt})^{-1}F^{-T} = 0,$$

which can be transformed into

$$(\hat{A}^{opt})^T + F^T(X^{opt})^T Y^{opt}T^{-T}\hat{A}^{opt}T((X^{opt})^T Y^{opt})^{-1}F^{-T}$$
$$+\sum_{k=1}^{m} F^T(X^{opt})^T N^T Y^{opt}((X^{opt})^T Y^{opt})^{-1}F^{-T}F^T(X^{opt})^T Y^{opt}\bar{N}_k((X^{opt})^T Y^{opt})^{-1}F^{-T}$$
$$- F^T(X^{opt})^T C^T\hat{C}^{opt}T((X^{opt})^T Y^{opt})^{-1}F^{-T} = 0.$$

Thus it follows

$$(\hat{A}^{opt})^T + F^T(X^{opt})^T Y^{opt}T^{-T}\hat{A}^{opt}T((X^{opt})^T Y^{opt})^{-1}F^{-T}$$
$$+\sum_{k=1}^{m}(\hat{N}_k^{opt})^T F^T X^{opt^T} Y^{opt}T^{-1}\hat{N}_k^{opt}T((X^{opt})^T Y^{opt})^{-1}F^{-T}$$
$$-(\hat{C}^{opt})^T\hat{C}^{opt}T((X^{opt})^T Y^{opt})^{-1}F^{-T} = 0,$$

and, subsequently,

$$-(\hat{A}^{opt})^T F^T(X^{opt})^T Y^{opt}T^{-1} - F^T(X^{opt})^T Y^{opt}T^{-1}(\hat{A}^{opt})^T$$
$$-\sum_{k=1}^{m}(\hat{N}_k^{opt})^T F^T(X^{opt})^T Y^{opt}T^{-1}\hat{N}_k^{opt} + (\hat{C}^{opt})^T\hat{C}^{opt} = 0.$$

Again, the unique solution of the generalized Lyapunov equation of the reduced system satisfies $Q_{22} = -F^T(X^{opt})^T Y^{opt}T^{-1}$, with Q_{22} as defined in (20). Moreover, due to symmetry of the solution, it follows $Q_{22} = -T^{-T}(Y^{opt})^T X^{opt}F$. Finally, we will need the solutions of the generalized Sylvester equations arising in (22). However, it holds that

$$AX^{opt} + X^{opt}\bar{A}^T + \sum_{k=1}^{m} N_k X^{opt}\bar{N}_k^T + B\bar{B}^T = 0$$

is equivalent to

$$AX^{opt} + X^{opt}T^T(\hat{A}^{opt})^T T^{-T} + \sum_{k=1}^{m} N_k X^{opt}T^T(\hat{N}_k^{opt})^T T^{-T} + B(\hat{B}^{opt})^T T^{-T} = 0,$$

yielding

$$AX^{opt}T^T + X^{opt}T^T(\hat{A}^{opt})^T + \sum_{k=1}^{m} N_k X^{opt}T^T(\hat{N}_k^{opt})^T + B(\hat{B}^{opt})^T = 0.$$

Here, we make use of the unique solution of the generalized Sylvester equation. Thus, it follows that $P_{12} = X^{opt}T^T$. Since the argumentation for the dual Sylvester equation is completely analogous, we will skip the derivation that leads to $Q_{12} = Y^{opt}T^{-1}$. Let us now show the optimality conditions (20):

$$Q_{12}^T P_{12} + Q_{22}P_{22} = T^{-T}(Y^{opt})^T X^{opt}T^T - T^{-T}(Y^{opt})^T X^{opt}FF^{-1}T^T = 0,$$

$$\begin{aligned}
Q_{22}\hat{N}_k^{opt}P_{22} + Q_{12}^T N_k P_{12} &= -T^{-T}(Y^{opt})^T X^{opt}F\hat{N}_k^{opt}F^{-1}T^T + T^{-T}(Y^{opt})^T N_k X^{opt}T^T \\
&= -T^{-T}(Y^{opt})^T X^{opt}F\left((W^{opt})^T V^{opt}\right)^{-1}(W^{opt})^T N_k V^{opt}F^{-1}T^T \\
&\quad + T^{-1}(Y^{opt})^T N_k X^{opt}T^T \\
&= -T^{-T}(Y^{opt})^T XFF^{-1}\left((Y^{opt})^T X^{opt}\right)^{-1}(Y^{opt})^T N_k V^{opt}FF^{-1}T^T \\
&\quad + T^{-T}(Y^{opt})^T N_k X^{opt}T^T \\
&= 0,
\end{aligned}$$

$$\begin{aligned}
Q_{12}^T B + Q_{22}\hat{B}^{opt} &= T^{-T}(Y^{opt})^T B - T^{-T}(Y^{opt})^T X^{opt}F\hat{B}^{opt} \\
&= T^{-T}(Y^{opt})^T B - T^{-T}(Y^{opt})^T X^{opt}F\left((W^{opt})^T V^{opt}\right)^{-1}(W^{opt})^T B \\
&= T^{-T}(Y^{opt})^T B - T^{-T}(Y^{opt})^T X^{opt}FF^{-1}\left((Y^{opt})^T X^{opt}\right)^{-1}(Y^{opt})^T B = 0,
\end{aligned}$$

$$\begin{aligned}
\hat{C}^{opt}P_{22} - CP_{12} &= \hat{C}^{opt}F^{-1}T^T - CX^{opt}T^T = CV^{opt}F^{-1}T^T - CX^{opt}T^T \\
&= CX^{opt}FF^{-1}T^T - CX^{opt}T^T = 0.
\end{aligned}$$

□

Remark 5.1. *It should be mentioned that the convergence criterion will only be achieved in exact arithmetic. Nevertheless, in practice, stopping the algorithm whenever the relative change of the eigenvalues is less than a user specified tolerance ϵ will be sufficient for numerical simulations.*

Remark 5.2. *Note that Algorithm 1 generalizes a Sylvester equation based algorithm for $\mathcal{H}_2$-optimality (see [17]) and thus does not require diagonalizability of $\hat{A}$.*

We will now turn our attention to an interpolation-based approach that can be directly derived from Algorithm 1. For a similar derivation in the linear case, see e.g. [17]. Again, let $\hat{A} = R\Lambda R^{-1}$ denote the eigenvalue decomposition of the reduced system. As already mentioned before, the explicit solution for equation (28) in vectorized form reads:

$$\begin{aligned}
\operatorname{vec}(X) &= \left(-I_{\hat{n}} \otimes A - \hat{A} \otimes I_n - \sum_{k=1}^{m} \hat{N}_k \otimes N_k\right)^{-1} \operatorname{vec}(B\hat{B}^T) \\
&= \left(-I_{\hat{n}}A - \hat{A} \otimes I_n - \sum_{k=1}^{m} \hat{N}_k \otimes N_k\right)^{-1} \left(\hat{B} \otimes B\right) \operatorname{vec}(I_m)
\end{aligned}$$

$$
\begin{aligned}
&= \left[(R \otimes I_n) \left(-I_{\hat{n}} \otimes A - \Lambda \otimes I_n - \sum_{k=1}^{m} R^{-1} \hat{N}_k R \otimes N_k \right) \left(R^{-1} \otimes I_n \right) \right]^{-1} \left(\hat{B} \otimes B \right) \operatorname{vec}(I_m) \\
&= (R \otimes I_n) \underbrace{\left(-I_{\hat{n}} \otimes A - \Lambda \otimes I_n - \sum_{k=1}^{m} R^{-1} \hat{N}_k R \otimes N_k \right)^{-1} \left(R^{-1} \hat{B} \otimes B \right) \operatorname{vec}(I_m)}_{\operatorname{vec}(V)}.
\end{aligned}
$$

From the last line, we can now conclude that

$$
(R \otimes I_n)^{-1} \operatorname{vec}(X) = \operatorname{vec}(V) \text{ and hence } X R^{-T} = V.
$$

Similarly, starting from equation (29), we obtain:

$$
\begin{aligned}
\operatorname{vec}(Y) &= \left(I_{\hat{n}} \otimes A^T + \hat{A}^T \otimes I_n + \sum_{k=1}^{m} \hat{N}_k^T \otimes N_k^T \right)^{-1} \operatorname{vec}(C^T \hat{C}) \\
&= \left(I_{\hat{n}} \otimes A^T + \hat{A}^T \otimes I_n + \sum_{k=1}^{m} \hat{N}_k^T \otimes N_k^T \right)^{-1} \left(\hat{C}^T \otimes C^T \right) \operatorname{vec}(I_p) \\
&= \left[\left(R^{-T} \otimes I_n \right) \left(-I_{\hat{n}} \otimes A - \Lambda \otimes I_n - \sum_{k=1}^{m} R^T \hat{N}_k^T R^{-T} \otimes N_k^T \right) \left(-R^T \otimes I_n \right) \right]^{-1} \times \\
&\quad \left(\hat{C}^T \otimes C^T \right) \operatorname{vec}(I_p) \\
&= \left(-R^{-T} \otimes I_n \right) \operatorname{vec}(W).
\end{aligned}
$$

Once again, this leads to

$$
\left(-R^{-T} \otimes I_n \right)^{-1} \operatorname{vec}(Y) = \operatorname{vec}(W) \text{ and } Y(-R) = W,
$$

where

$$
\operatorname{vec}(W) := \left(-I_{\hat{n}} \otimes A - \Lambda \otimes I_n - \sum_{k=1}^{m} R^T \hat{N}_k^T R^{-T} \otimes N_k^T \right)^{-1} \left(R^T \hat{C}^T \otimes C^T \right) \operatorname{vec}(I_p).
$$

According to the proof of Theorem 5.1, as long as $\operatorname{span}\{X\} \subset V$ and $\operatorname{span}\{Y\} \subset W$, we can ensure that the reduced system satisfies the necessary $\mathcal{H}_2$-optimality conditions. Hence, we have found an equivalent method which obviously extends IRKA to the bilinear case, see Algorithm 2.

Algorithm 2 Bilinear IRKA (BIRKA)

Input: A, N_k, B, C, $\hat{A}$, $\hat{N}_k$, $\hat{B}$, $\hat{C}$
Output: $\hat{A}^{opt}$, $\hat{N}_k^{opt}$, $\hat{B}^{opt}$, $\hat{C}^{opt}$

1: **while** (change in $\Lambda > 0$) **do**
2: $R\Lambda R^{-1} = \hat{A}$, $\tilde{B} = \hat{B}^T R^{-T}$, $\tilde{C} = \hat{C}R$, $\tilde{N}_k = R^T \hat{N}_k R^{-T}$
3: $\operatorname{vec}(V) = \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N_k\right)^{-1} \left(\tilde{B}^T \otimes B\right) \operatorname{vec}(I_m)$
4: $\operatorname{vec}(W) = \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A^T - \sum_{k=1}^{m} \tilde{N}_k \otimes N_k^T\right)^{-1} \left(\tilde{C}^T \otimes C^T\right) \operatorname{vec}(I_p)$
5: $V = \operatorname{orth}(V)$, $W = \operatorname{orth}(W)$
6: $\hat{A} = \left(W^T V\right)^{-1} W^T A V$, $\hat{N}_k = \left(W^T V\right)^{-1} W^T N_k V$, $\hat{B} = \left(W^T V\right)^{-1} W^T B$, $\hat{C} = CV$
7: **end while**
8: $\hat{A}^{opt} = \hat{A}$, $\hat{N}_k^{opt} = \hat{N}_k$, $\hat{B}^{opt} = \hat{B}$, $\hat{C}^{opt} = \hat{C}$

Finally, we want to point out the equivalence between the optimality conditions (20) and (24). For this, we need the following projection-based identity.

Lemma 5.2. *Let $V, W \in \mathbb{R}^{n \times \hat{n}}$ be matrices of full rank $\hat{n}$.*
a) Let $z \subset \operatorname{span}\{\operatorname{vec}(V)\}$. Then $\left(I_{\hat{n}} \otimes V(W^TV)^{-1}W^T\right) z = z$.
b) Let $z \subset \operatorname{span}\{\operatorname{vec}(W)\}$. Then $z^T \left(I_{\hat{n}} \otimes V(W^TV)^{-1}W^T\right) = z^T$.

Proof. By assumption, there exists $x \in \mathbb{R}^{n \cdot \hat{n}}$ s.t.

$$\begin{aligned}\left(I_{\hat{n}} \otimes V(W^TV)^{-1}W^T\right) z &= \left(I_{\hat{n}} \otimes V(W^TV)^{-1}W^T\right) \operatorname{vec}(V)x \\ &= \operatorname{vec}(V(W^TV)^{-1}W^TV)x = \operatorname{vec}(V)x = z.\end{aligned}$$

The proof of the second statement is based on the exact same arguments. □

Theorem 5.2. *Assume Algorithm 2 converges. Then $\hat{A}^{opt}$, $\hat{N}_k^{opt}$, $\hat{B}^{opt}$, $\hat{C}^{opt}$ fulfill the necessary interpolation-based $\mathcal{H}_2$-optimality conditions.*

Proof. Since the only difference in proving conditions (24) – (27) lies in using statement b) of Lemma 5.2 and the combination of both a) and b), respectively, we will restrict ourselves to showing optimality condition (24).

$$\begin{aligned}&\operatorname{vec}(I_p)^T \left(e_i e_j^T \otimes \hat{C}^{opt}\right) \left(-\Lambda \otimes I_{\hat{n}} - I_{\hat{n}} \otimes \hat{A}^{opt} - \sum_{k=1}^{m} \tilde{N}_k^T \otimes \hat{N}_k^{opt}\right)^{-1} \left(\tilde{B}^T \otimes \hat{B}^{opt}\right) \operatorname{vec}(I_m) \\ &= \operatorname{vec}(I_p)^T \left(e_i e_j^T \otimes CV\right) \times \\ &\quad \left[\left(I_{\hat{n}} \otimes (W^TV)^{-1}W^T\right) \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N\right) (I_{\hat{n}} \otimes V)\right]^{-1} \times \\ &\quad \left(\tilde{B}^T \otimes (W^TV)^{-1}W^TB\right) \operatorname{vec}(I_m)\end{aligned}$$

$$
\begin{aligned}
&= \operatorname{vec}(I_p)^T \left(e_i e_j^T \otimes CV\right) \times \\
&\quad \left[\left(I_{\hat{n}} \otimes (W^TV)^{-1}W^T\right)\left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N\right)\left(I_{\hat{n}} \otimes V\right)\right]^{-1} \times \\
&\quad \left(I_{\hat{n}} \otimes (W^TV)^{-1}W^T\right) \times \\
&\quad \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N\right)\left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N\right)^{-1} \times \\
&\quad \left(\tilde{B}^T \otimes B\right) \operatorname{vec}(I_m) \\
&\overset{(5.2a)}{=} \operatorname{vec}(I_p)^T \left(e_i e_j^T \otimes CV\right) \times \\
&\quad \left[\left(I_{\hat{n}} \otimes (W^TV)^{-1}W^T\right)\left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N\right)\left(I_{\hat{n}} \otimes V\right)\right]^{-1} \times \\
&\quad \left(I_{\hat{n}} \otimes (W^TV)^{-1}W^T\right)\left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N\right)\left(I_n \otimes V(W^TV)^{-1}W^T\right) \times \\
&\quad \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N\right)^{-1}\left(\tilde{B}^T \otimes B\right) \operatorname{vec}(I_m) \\
&= \operatorname{vec}(I_p)^T \left(e_i e_j^T \otimes CV\right)\left(I_n \otimes (W^TV)^{-1}W^T\right) \times \\
&\quad \left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N\right)^{-1}\left(\tilde{B}^T \otimes B\right) \operatorname{vec}(I_m) \\
&= \operatorname{vec}(I_p)^T \left(e_i e_j^T \otimes C\right)\left(-\Lambda \otimes I_n - I_{\hat{n}} \otimes A - \sum_{k=1}^{m} \tilde{N}_k^T \otimes N\right)^{-1}\left(\tilde{B}^T \otimes B\right) \operatorname{vec}(I_m).
\end{aligned}
$$

□

Remark 5.3. *Note that analogously to the case of solving generalized Sylvester and Lyapunov equations, respectively, it is also possible to construct the matrices appearing in Algorithm 2 as the limit of an infinite series of linear IRKA type computations. For this, in each iteration, one starts with*

$$V_i^1 = (-\lambda_i I - A)^{-1} B\tilde{B}_i,$$

and continues with

$$V_i^j = (-\lambda_i I - A)^{-1}\left(\sum_{k=1}^{m} N_k V_i^{j-1}(\tilde{N}_k)_i\right).$$

The actual projection matrix V then is given as $V = \sum_{j=1}^{\infty} V^j$. A dual derivation obviously yields the projection matrix W.

6 Numerical Examples

In this section, we will now study several applications of bilinear control systems and discuss the performance of the approaches proposed above. As we already mentioned, the method of balanced truncation for bilinear systems is connected to generalized controllability and reachability Gramians of the underlying system, respectively. Hence, similar to the linear case, we expect this method to yield reduced models with small relative $\mathcal{H}_2$-error as well and we will thus use it for a comparison with our algorithms. However, due to the theoretical equivalence of Algorithm 1 and Algorithm 2, we will only report the results for the latter case. Nevertheless, we want to remark that in numerical simulations, there might occur differences with respect to robustness and speed of convergence which might be subject to further studies. Furthermore, in Algorithm 2 we computed the projection matrices V and W by solving the large systems of linear equations explicitly instead of using more sophisticated iterative techniques which might be further investigated as well. Finally, all Lyapunov equations were solved by the method proposed in [13] which allows for solving medium-sized systems. All simulations were generated on an Intel® Core$^{\text{TM}}$i7 CPU 920, 8 MB cache, 12 GB RAM, openSUSE Linux 11.1 (x86_64), MATLAB® Version 7.11.0.584 (R2010b) 64-bit (glnxa64).

6.1 An interconnected power system

The first application is a model for two interconnected power systems which can be described by a bilinear system of state dimension 17. The hydro unit as well as the steam unit each can be controlled by two input variations resulting in a system with 4 inputs and 3 outputs. Since we are only interested in the reduction process, we refer to [2] where a detailed derivation of the dynamics can be found. We have successively reduced the original model to systems varying from $\hat{n} = 1, \ldots, 16$ state variables. A comparison of the associated relative $\mathcal{H}_2$-norm of the error system between our approach and the method of balanced truncation is shown in Figure 1.

Except for the cases $\hat{n} = 2$ and $\hat{n} = 12$, we always obtain better results with the new technique. The initialization of Algorithm 2 is done completely at random, using arbitrary interpolations points and tangential directions, respectively. As indicated for system dimensions $\hat{n} = 5, 10, 14$, the algorithm converges in a few steps, see Figure 2. However, for $\hat{n} = 2, 6, 12$, the stopping criterion which is chosen to be that the relative change of the norm of the eigenvalues of the reduced system becomes smaller than $\sqrt{\epsilon}$, where ϵ denotes machine precision, is not fulfilled. This might explain the mentioned superiority of balanced truncation.

6.2 Fokker-Planck equation

The second example is an application from stochastic control and was already discussed in [19]. Let us consider a dragged Brownian particle whose one-dimensional motion is described by the stochastic differential equation

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

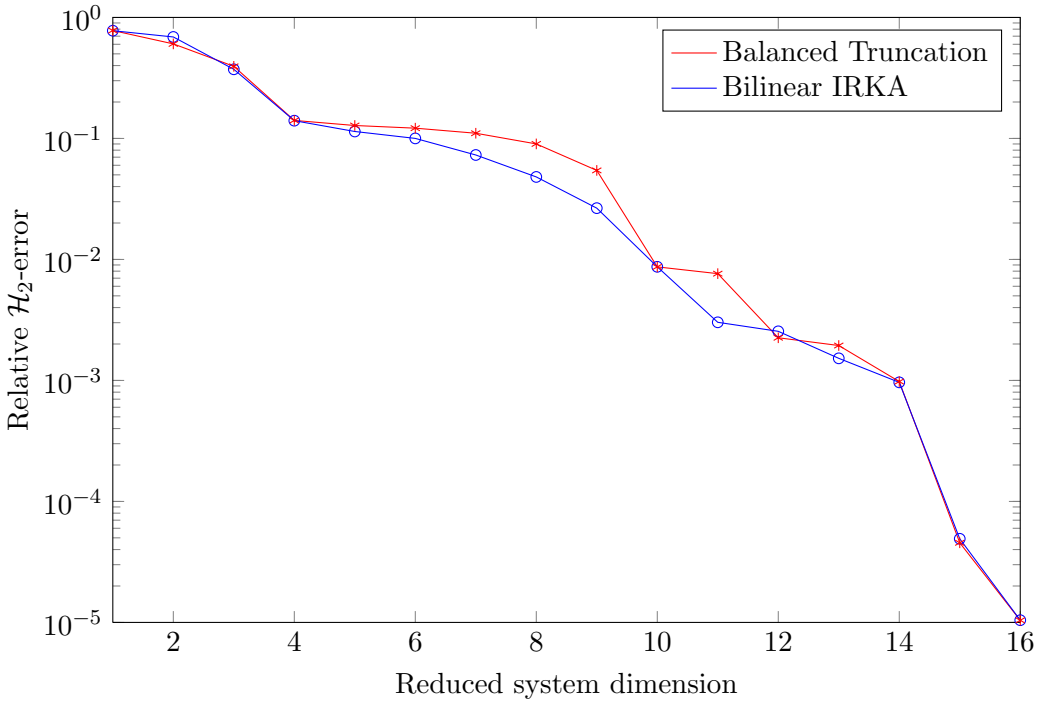

Figure 1: **Power System**. Comparison of relative $\mathcal{H}_2$-error between balanced truncation and B-IRKA.

with $\sigma = \frac{2}{3}$ and $V(x,u) = W(x,t) + \Phi(x,u_t) = (x^2-1)^2 - xu - x$. As mentioned in [19], we might alternatively consider the underlying probability distribution function

$$\rho(x,t)dx = \mathbf{P}\left[X_t \in [x, x+dx)\right]$$

which is described by the Fokker-Planck equation

$$\begin{aligned} \frac{\partial \rho}{\partial t} &= \sigma \Delta \rho + \nabla \cdot (\rho \nabla V), && (x,t) \in (a,b) \times (0,T], \\ 0 &= \sigma \nabla \rho + \rho \nabla B, && (x,t) \in \{a,b\} \times [0,T], \\ \rho_0 &= \rho, && (x,t) \in (a,b) \times 0. \end{aligned}$$

After a semi-discretization resulting from a finite difference scheme consisting of 500 nodes in the interval $[-2,2]$, we obtain a single-input single-output bilinear control system, where we choose the output matrix C to be the discrete characteristic function of the interval $[0.95, 1.05]$. Since we only pointed out the most important parameters of the model, we once more refer to [19] for gaining a more detailed insight into this topic. In Figure 3, we again compare the relative $\mathcal{H}_2$-errors between balanced truncation and B-IRKA for varying system dimensions. Despite the fact that we do not observe convergence for $\hat{n} = 2$, our new method clearly outperforms balanced truncation.

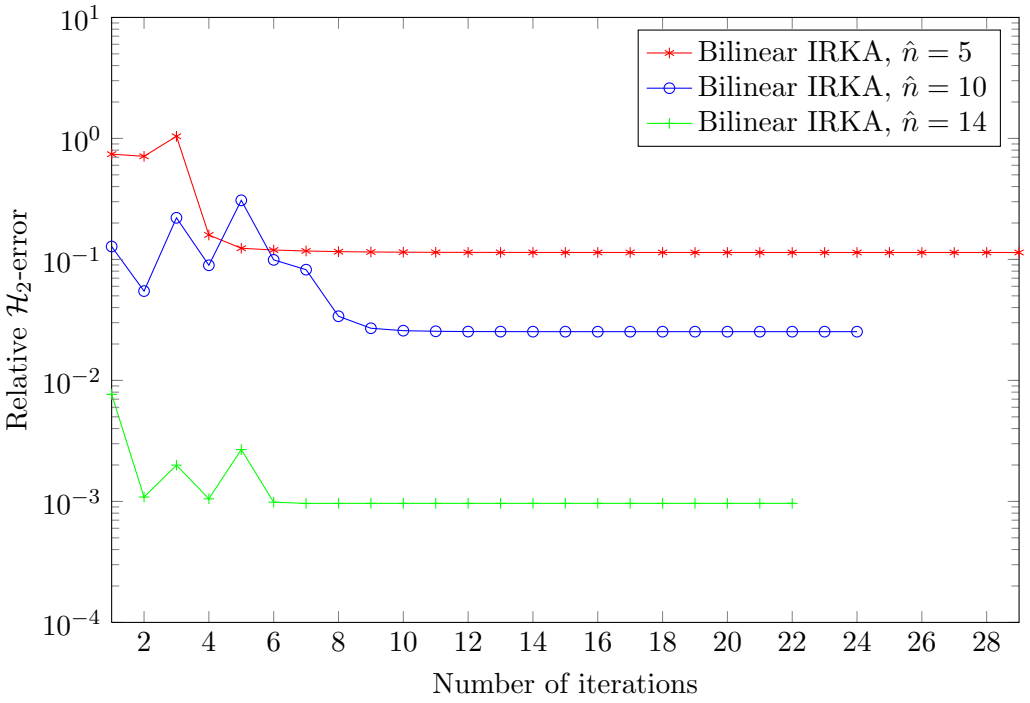

Figure 2: **Power System**. Convergence history of the relative $\mathcal{H}_2$-error.

6.3 Viscous Burgers equation

Next, let us consider the viscous Burgers equation

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = \nu\frac{\partial^2 v}{\partial x^2}, \quad (x,t) \in (0,1) \times (0,T),$$

subject to initial and boundary conditions

$$v(x,0) = 0, \quad x \in [0,1], \qquad v(0,t) = u(t), \quad v(1,t) = 0, \quad t \geq 0.$$

Introduced in [8], after a spatial semi-discretization of this nonlinear partial differential equation using k nodes in a finite difference scheme, we end up with an ordinary differential equation including a quadratic nonlinearity. As is well-known, the Carleman linearization technique, see e.g. [27], allows to approximate this system by a bilinearized system of dimension $n = k + k^2$. The simulations are generated with $\nu = 0.1$ and $k = 30$. The measurement vector C is chosen to yield the spatial average value for the quantity v. As shown in Figure 4, in all cases the relative $\mathcal{H}_2$-error for the systems constructed by B-IRKA is smaller than that resulting from balanced truncation. Moreover, except for $\hat{n} = 11$, there are no convergence problems at all although we again use random data for the initialization.

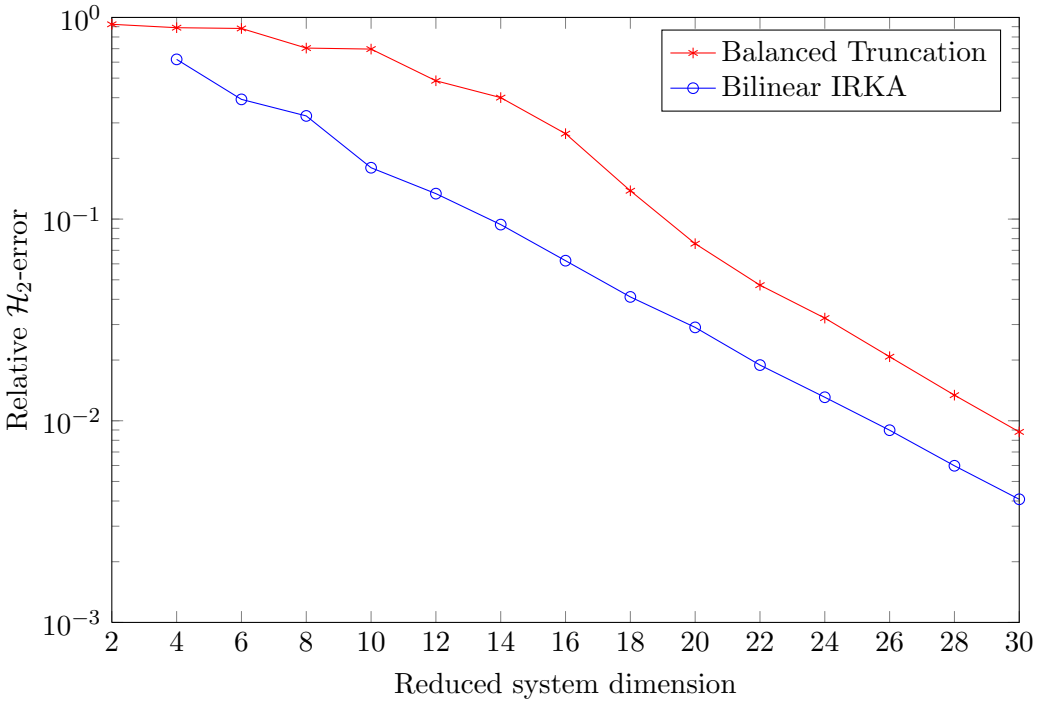

Figure 3: **Fokker-Planck Equation**. Comparison of relative $\mathcal{H}_2$-error between balanced truncation and B-IRKA.

6.4 A heat transfer model

Finally, we want to study another standard bilinear test example resulting from a boundary controlled heat transfer system. Formally, the dynamics are described by the heat equation subject to Dirichlet and Robin boundary conditions, i.e.

$$\begin{aligned} x_t &= \Delta x && \text{in } (0,1)\times(0,1), \\ n\cdot\nabla x &= 0.75\cdot u_{1,2,3}(x-1) && \text{on } \Gamma_1,\Gamma_2,\Gamma_3, \\ x &= 0.75\cdot u_4 && \text{on } \Gamma_4, \end{aligned}$$

where $\Gamma_1, \Gamma_2, \Gamma_3$ and Γ_4 denote the boundaries of Ω. Hence, a spatial discretization using k^2 grid points now yields a bilinear system of dimension $n = k^2$, with 4 inputs and 1 output, chosen to be the average temperature on the grid. In order to show that our algorithm also works in large-scale settings, we implement the above system with 10 000 grid points. The results for reduced system dimensions $\hat{n} = 2, \ldots, 30$, are given in Figure 5 and demonstrate that we can improve the approximation quality with regard to the $\mathcal{H}_2$-norm with a numerically efficient interpolation-based framework. Moreover, in order to show the superiority of the new approach we further plot the results for the reduced systems obtained by IRKA as well as those generated by the new interpolation framework together with some clever, but non-optimal interpolation points. This means,

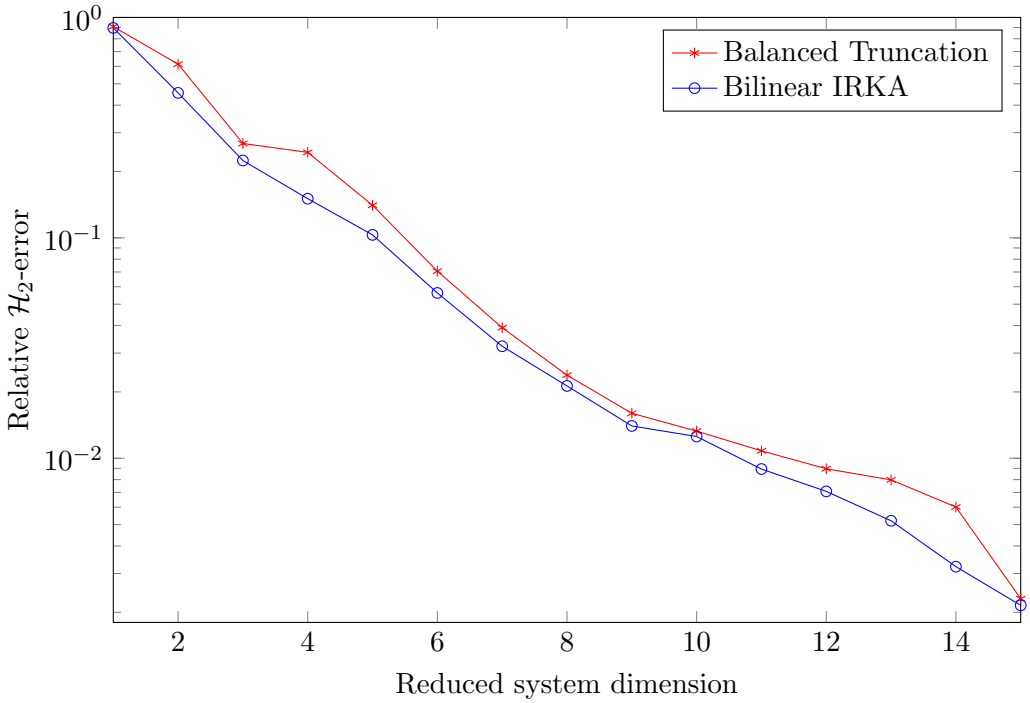

Figure 4: **Burgers equation**. Comparison of relative $\mathcal{H}_2$-error between balanced truncation and B-IRKA.

we use real equi-distributed and Chebyshev interpolations points between the smallest and largest real part of the mirror images of the eigenvalues of the system matrix A and stop Algorithm 2 after the first iteration step. However, the relative $\mathcal{H}_2$-error is only computed when the corresponding reduced systems are stable, leading to positive definite solutions of the Gramians of the error systems. Moreover, as can be seen in Figure 5, the linear iterative rational Krylov algorithm only converges for reduced system dimensions up to $\hat{n} = 18$ at all.

Since so far most bilinear reduction methods have been evaluated by means of comparing the relative error for outputs corresponding to typical system inputs, we compute the time response to an input of the form $u_k(t) = \cos(k\pi t),\ k = 1, 2, 3, 4$. The results are plotted in Figure 6, where we test the performance for an original bilinear system of order $n = 2500$ and different scaling values γ. This means, the matrices N_k and B, respectively are multiplied with γ, while the input signal $u(t)$ is replaced with $\frac{1}{\gamma}u(t)$. Similar experiments are studied in [6]. Interestingly enough, while the convergence results for B-IRKA do not change significantly, the relative error is smaller for smaller values of γ. However, all tested values γ can certainly compete with the approximation quality obtained from balanced truncation.

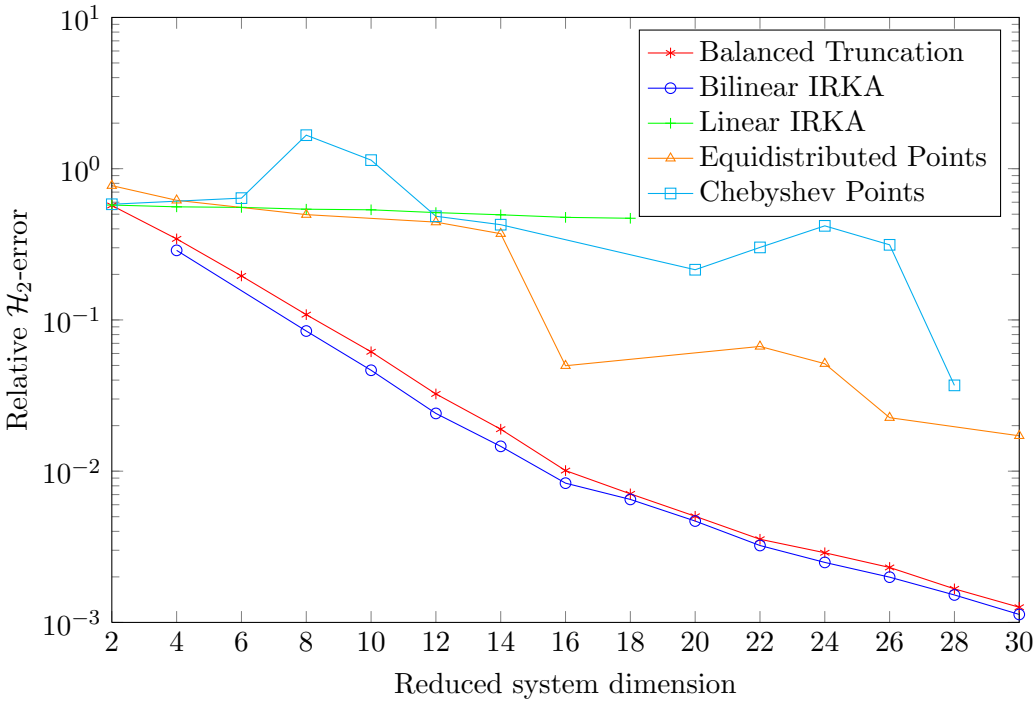

Figure 5: **Heat Transfer Model**. Comparison of relative $\mathcal{H}_2$-error between balanced truncation and B-IRKA.

7 Conclusions

In this paper, we have studied the problem of $\mathcal{H}_2$-model reduction for bilinear systems. Based on an existing generalization of the linear $\mathcal{H}_2$-norm, we have derived first-order necessary conditions for optimality. As has been shown, these can be interpreted as an extension of those obtained for the linear case and lead to a generalization of the iterative rational Krylov algorithm (IRKA). We have further proposed an equivalent iterative procedure that requires solving certain generalized Sylvester equations. The efficiency of our approaches has been evaluated by several bilinear test examples for which they yield better results than the popular method of balanced truncation. Finally, it was shown that the new method can additionally compete when the approximation quality is measured in terms of the transient response in time domain. However, so far we did not investigate the effect of choosing reasonable initial data in order to improve convergence rates of the algorithms as well as efficient solution techniques for the special generalized Sylvester equations one has to solve in each iteration step.

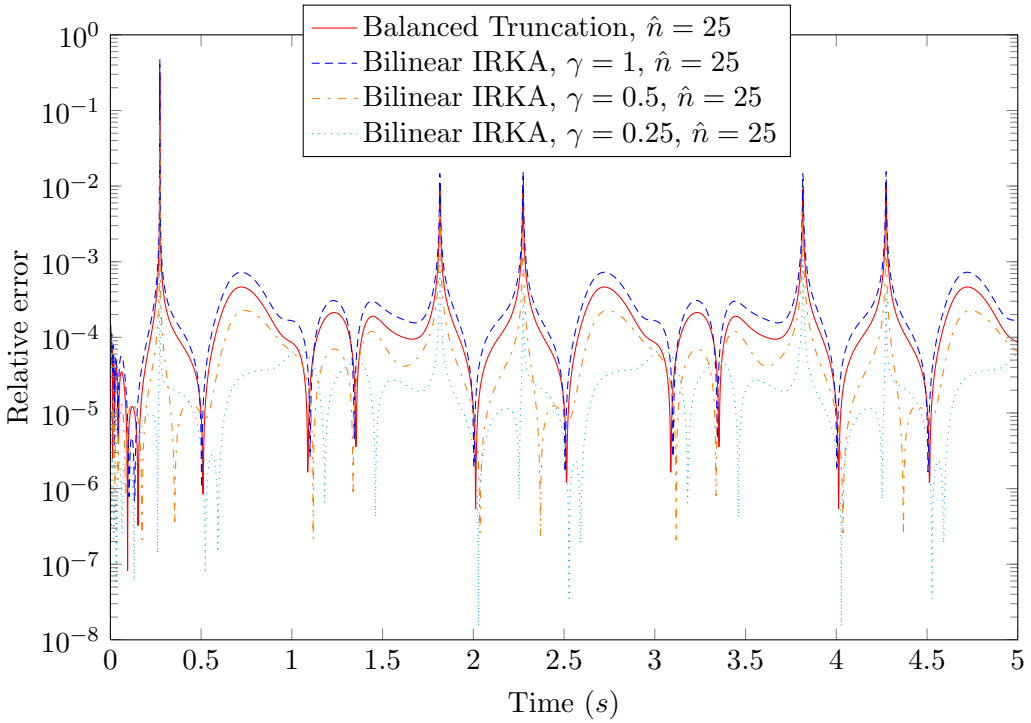

Figure 6: **Heat Transfer Model**. Comparison of relative error to an input of the form $u_k(t) = \cos(k\pi t)$ for a bilinear system of order $n = 2500$ between balanced truncation and B-IRKA for varying scaling factors γ.

Acknowledgements

We thank Carsten Hartmann from FU Berlin for kindly providing us with the data used in the numerical example in Section 6.2.

References

[1] S. Al-Baiyat. Model reduction of bilinear systems described by input-output difference equation. *Internat. J. Syst. Sci.*, 35(9):503–510, 2004.

[2] S. Al-Baiyat, AS Farag, and M. Bettayeb. Transient approximation of a bilinear two-area interconnected power system. *Electric Power Systems Research*, 26(1):11–19, 1993.

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

[4] Z. Bai. Krylov subspace techniques for reduced-order modeling of nonlinear dynamical systems. *Appl. Numer. Math.*, 43:9–44, 2002.

[5] Z. Bai and D. Skoogh. A projection method for model reduction of bilinear dynamical systems. *Linear Algebra Appl.*, 415(2–3):406–425, 2006.

[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 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`.

[8] T. Breiten and T. Damm. Krylov subspace methods for model order reduction of bilinear control systems. *Sys. Control Lett.*, 59(8):443–450, 2010.

[9] C. Bruni, G. DiPillo, and G. Koch. On the mathematical models of bilinear systems. *Automatica*, 2:11–26, 1971.

[10] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, and D. Wilczek. Necessary optimality conditions for $\mathcal{H}_2$-norm optimal model reduction. Preprint, 2007.

[11] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, and D. Wilczek. h2-norm optimal model reduction for large scale discrete dynamical MIMO systems. *J. Comput. Appl. Math.*, 233(5):1202–1216, 2010.

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

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

[14] A. Dunoyer, L. Balmer, K.J. Burnham, and D.J.G. James. On the discretization of single-input single-output bilinear systems. *Internat. J. Control*, 68(2):361–372, 1997.

[15] L. Feng and P. Benner. A note on projection techniques for model order reduction of bilinear systems. In *Numerical Analysis and Applied Mathematics, AIP Conference Proceedings*, volume 936, pages 208–211, 2007.

[16] D. Galbally, K. Fidkowski, K. Willcox, and O. Ghattas. Non-linear model reduction for uncertainty quantification in large-scale inverse problems. *Internat. J. Numer. Methods Engrg.*, 81(12):1581–1608, 2010.

[17] K Gallivan, A. Vandendorpe, and P. Van Dooren. Model reduction of MIMO systems via tangential interpolation. *SIAM J. Matrix Anal. Appl.*, 26(2):328–349, 2004.

[18] 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.

[19] 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.

[20] T. Hinamoto and S. Maekawa. Approximation of polynomial state-affine discrete-time systems. *IEEE Trans. Circuits and Systems*, 31:713–721, 1984.

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

[22] R.R. Mohler. *Bilinear Control Processes*. New York Academic Press, 1973.

[23] R.R. Mohler. *Nonlinear Systems (vol. 2): Applications to Bilinear Control*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1991.

[24] R.R. Mohler. Natural bilinear control processes. *IEEE Transactions on Systems Science and Cybernetics*, 6(3):192–197, 2007.

[25] J.R. Phillips. Projection frameworks for model reduction of weakly nonlinear systems. In *Proceedings of DAC 2000*, pages 184–189, 2000.

[26] J.R. Phillips. Projection-based approaches for model reduction of weakly nonlinear, time-varying systems. *IEEE Trans. Circuits and Systems*, 22(2):171–187, 2003.

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

[28] H. Schwarz. Stability of discrete-time equivalent homogeneous bilinear systems. *Contr. Theor. Adv. Tech.*, 3:263–269, 1987.

[29] T. Siu and M. Schetzen. Convergence of Volterra series representation and BIBO stability of bilinear systems. *Int. J. Syst. Sci.*, 22(12):2679–2684, 1991.

[30] P. Van Dooren, K.A. Gallivan, and P.A. Absil. $\mathcal{H}_2$-optimal model reduction of MIMO systems. *Appl. Math. Lett.*, 21(12):1267–1273, 2008.

[31] D.A. Wilson. Optimum solution of model-reduction problem. *Proc. IEE-D*, 117(6):1161–1165, 1970.

[32] L. Zhang and J. Lam. On $\mathcal{H}_2$ model reduction of bilinear systems. *Automatica*, 38(2):205–216, 2002.

[33] L. Zhang, J. Lam, B. Huang, and G. Yang. On gramians and balanced truncation of discrete-time bilinear systems. *Internat. J. Control*, 76(4):414–427, 2003.