Max Planck Institute Magdeburg
Preprints

Mian Ilyas Ahmad, Peter Benner, Pawan Goyal, Jan Heiland

Moment-Matching Based Model Reduction for Stokes-Type Quadratic-Bilinear Descriptor Systems

MPIMD/15-18 **October 22, 2015**

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

We discuss a Krylov subspace projection method for model order reduction of a special class of quadratic-bilinear descriptor systems. The goal is to extend the two-sided moment-matching method for quadratic-bilinear ODEs to descriptor systems in an efficient and reliable way. Recent results have shown that the direct application of interpolation based model reduction techniques to linear descriptor systems, without any modifications, may lead to poor reduced-order systems. Therefore, for the analysis, we transform the quadratic-bilinear descriptor system into an equivalent quadratic-bilinear ODE system for which the moment-matching is performed. In view of implementation, we provide algorithms that identify the required Krylov subspaces without explicitly computing the projectors used in the analysis. The benefits of our approach are illustrated for the quadratic-bilinear descriptor system of semi-discretized Navier-Stokes equations.

Keywords: Model-order reduction, moment-matching, descriptor systems, tensor matricization.

Corresponding author's addresses:

Pawan Goyal
Computational Methods in Systems and Control Theory
Max Planck Institute for Dynamics of Complex Technical Systems
Sandtorstr. 1
39106 Magdeburg, Germany
email address: goyalp@mpi-magdeburg.mpg.de

1 Introduction

We discuss model order reduction for a single-input single-output (SISO) quadratic-bilinear descriptor system of the form

$$E_{11}\dot{v}(t) = A_{11}v(t) + A_{12}p(t) + H_1 v(t) \otimes v(t) + N_1 v(t)u(t) + b_1 u(t), \tag{1a}$$
$$v(0) = \alpha, \tag{1b}$$
$$0 = A_{21}v(t) + b_2 u(t), \tag{1c}$$
$$y(t) = c_1 v(t) + c_2 p(t) + Du(t), \tag{1d}$$

for time $t > 0$ and an initial value $\alpha \in \mathbb{R}^{n_1}$, where $E_{11}, A_{11}, N_1 \in \mathbb{R}^{n_1 \times n_1}$, $A_{21}, A_{12}^T \in \mathbb{R}^{n_2 \times n_1}$, $H_1 \in \mathbb{R}^{n_1 \times n_1^2}$, $b_1, c_1^T \in \mathbb{R}^{n_1}$, $b_2, c_2^T \in \mathbb{R}^{n_2}$ are the coefficients, $v(t) \in \mathbb{R}^{n_1}$ and $p(t) \in \mathbb{R}^{n_2}$ are state vectors, where D is a scalar, and where $u(t)$ is an input to the system. It is assumed that E_{11} is invertible as is $A_{21}E_{11}^{-1}A_{12}$. This means that the system in (1) reduces to an index-2 linear system, cf. [17], if $N_1 = 0$ and $H_1 = 0$. The somewhat particular structure of (2) arises in semi-discretizations of Navier-Stokes equations and reflects the divergence free constraint and the quadratic nonlinearity in the velocity.

It is an appealing task in the field of numerical analysis to identify efficient numerical methods that can be used to analyze and study engineering problems for complex dynamical processes. These dynamical processes are often described by ordinary differential equations (ODEs) or partial differential equations (PDEs). Spatial discretization of such governing equations leads to large scale systems of ODEs or the more general differential algebraic equations (DAEs). Simulation or design of these large scale systems often is computationally cumbersome, and it is hardly possible to get fast and accurate solutions. A remedy to this problem is model order reduction (MOR) which can play an important role in improving the simulation time. For linear systems, well-established model reduction techniques have been proposed in the literature [1, 8, 24]. Some of these techniques have already been extended to nonlinear systems, however, there are many open questions that need further research.

Model reduction techniques for nonlinear systems with the quadratic-bilinear nonlinearities can be classified broadly into two classes, trajectory-based methods and moment-matching methods. The proper orthogonal decomposition method [2, 9, 10, 11, 21] is a well-known trajectory-based method, where a set of snapshots of the state trajectory are used to compute a Galerkin projection of the nonlinear system. Another approach is the so-called trajectory piecewise linear (TPWL) method [22], in which the nonlinear state equation is written as a weighted combination of the linear systems which allows us to employ linear reduction techniques. We refer to [1, 15], for details on these trajectory based methods, which can identify highly accurate reduced-order systems with the main drawback of depending strongly on the training input. This means that the reduced-order system obtained from these methods may not be suitable for applications in control or in optimization, where input variation is the main goal.

On the other hand, moment-matching methods tend to approximate the input-output behavior of the system well and therefore, unlike trajectory based methods,

these methods are not bound to a specific input. Extending well-known results for linear ODEs, the moment-matching problem has been considered in [4, 14] for quadratic-bilinear ODEs for one-sided moments. The latest extension was to two-sided moment-matching for the SISO quadratic-bilinear ODEs [6].

In this paper, we study two-sided moment-matching technique for model reduction of the Stokes-type quadratic-bilinear descriptor systems (1). This class of quadratic bilinear descriptor systems is different from the one considered in [13]. Here, we propose a structured approach is required, since the direct implementation of [4, 6, 14] to the descriptor system might lead to an unbounded error in some norm, cf., in particular, [16], where it was shown that the direct extension of moment-matching techniques for linear ODEs to linear DAEs may lead to unbounded $\mathcal{H}_2$ or $\mathcal{H}_\infty$ error. An extension of the ideas presented in [16] to a special class of bilinear descriptor systems is presented in [12].

The first contribution of this paper is to transform the system in (1) into an equivalent quadratic-bilinear ODE system. This is done by introducing projectors similar to those used in [16, 19] for linear systems. The second contribution is to reduce the equivalent ODE system by constructing basis matrices for Krylov subspaces without explicitly computing the projectors.

The paper is organized as follows: Section 2 contains the background theory on the two-sided moment-matching technique for model reduction of quadratic-bilinear ODEs; Section 3 presents the transformation of the system (1) to an equivalent ODE system and shows how two-sided moment matching can be used to obtain a reduced-order system and discusses the implementation issues of the two-sided moment-matching technique for the equivalent system; Section 4 shows how the general case with $b_2 \neq 0$ can be treated with the same technique. Finally, in Section 5, we provide results of numerical tests for semi-discretized Navier-Stokes equations with quadratic-bilinear nonlinearities.

2 Quadratic-Bilinear DAEs and Background Work

In this section, we briefly review some properties of the general quadratic-bilinear differential algebraic equations (QBDAEs),

$$\Sigma_{QB}: \quad \begin{cases} E\dot{x}(t) = Ax(t) + Hx(t) \otimes x(t) + Nx(t)u(t) + bu(t), \\ \quad y(t) = cx(t), \quad x(0) = 0, \end{cases} \tag{2}$$

with $E, A, N \in \mathbb{R}^{n\times n}$, $H \in \mathbb{R}^{n\times n^2}$, $b, c^T \in \mathbb{R}^n$, and where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, and $y(t) \in \mathbb{R}$ are the state, input, and output of the system, respectively. It is assumed that E is nonsingular. Notice that the system (2) can be seen as combination of a purely quadratic system and a bilinear control system and it is shown in [14] that a large class of nonlinear systems can be transformed to this kind of representation allowing it to be used for many applications.

In frequency domain, the system (2) can be represented by a nonlinear input-output map that involves an infinite set of multivariate functions [23], often called the gener-

alized transfer functions of the subsystems associated with Σ_{QB}. These multivariate transfer functions can be identified explicitly, for example, via the growing exponential approach [6, 23]. In its symmetric form, the first two generalized transfer functions are

$$G^{(1)}(s_1) = c\underbrace{(s_1E - A)^{-1}b}_{L(s_1)}$$

and

$$G^{(2)}(s_1, s_2) = \frac{1}{2}c\left((s_1 + s_2)E - A\right)^{-1}\big[N\left(L(s_1) + L(s_2)\right) + H\left(L(s_1) \otimes L(s_2) + L(s_2) \otimes L(s_1)\right)\big].$$

A goal of interpolation based model reduction for Σ_{QB} is to find a reduced quadratic-bilinear DAE of similar form whose leading k generalized transfer functions interpolate the original one. That is

$$G^{(i)}(\sigma_1, \ldots, \sigma_i) = G_r^{(i)}(\sigma_1, \ldots, \sigma_i), \quad i = 1, \ldots, k,$$

where the σ_is are the interpolation points for the generalized transfer function corresponding to s_i. In the literature, cf., e.g., [6, 14], often k is set to 2, so that the interpolation concept is analyzed for the first two transfer functions. We will also restrict ourself to this setting. Notice that for $i = 1$, the problem reduces to the well known interpolation concept for linear systems [1]. Thus, a series expansion of $G^{(1)}(s_1)$ at an interpolation point σ can identify the so-called moments of $G^{(1)}$ and if we construct a reduced-order system $G_r^{(1)}(s_1)$ whose first q moments coincide with the original system moments, then $G_r^{(1)}(s_1)$ should be locally equal to $G^{(1)}(s_1)$. The issue is, however, to identify a reduced-order system which in addition also achieves moment-matching for $G^{(2)}(s_1, s_2)$ so that $G_r^{(2)}(s_1, s_2)$ is also locally equal to $G^{(2)}(s_1, s_2)$. To achieve this, a one-sided projection technique has been introduced in [14], which is then extended in [6] to a two-sided projection framework.

Similar to the linear case, the projection scheme involves identifying suitable basis matrices $\mathcal{V} \in \mathbb{R}^{n\times r}$ and $\mathcal{W} \in \mathbb{R}^{n\times r}$, approximating the state vector as $x(t) \approx x_r(t) = \mathcal{V}x(t)$ and ensuring the Petrov-Galerkin condition:

$$\begin{aligned}&\mathcal{W}^T\left(E\mathcal{V}\dot{x}_r(t) - A\mathcal{V}x_r(t) - H\mathcal{V}x_r(t) \otimes \mathcal{V}x_r(t) - N\mathcal{V}x_r(t)u(t) - bu(t)\right) = 0,\\ &y_r(t) = c\mathcal{V}x_r(t), \quad x_r(0) = 0.\end{aligned}$$

This means that the reduced-order system matrices are of the form

$$\begin{aligned}&E_r = \mathcal{W}^TE\mathcal{V}, \quad A_r = \mathcal{W}^TA\mathcal{V}, \quad H_r = \mathcal{W}^TH\mathcal{V} \otimes \mathcal{V},\\ &N_r = \mathcal{W}^TN\mathcal{V}, \quad b_r = \mathcal{W}^Tb, \qquad c_r = c\mathcal{V}.\end{aligned} \tag{3}$$

Recall that, in case of one-sided projection, $\mathcal{W} = \mathcal{V}$ and therefore moment-matching is related only to the choice of $\mathcal{V}$. However in case of two-sided projection, both $\mathcal{V}$ and $\mathcal{W}$ play an important role for matching more moments of $G^1(s_1)$ and $G^2(s_1, s_2)$. Next, we outline a theorem, preceded by some definitions, which suggests a choice of $\mathcal{V}$ and $\mathcal{W}$ that ensures the required moment-matching criteria.

Definition 2.1 (Matricization, cf., e.g., [6, 20]). *By $X^{(k)}$, we denote the matrix that is obtained by unfolding the K-dimensional tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$ along the kth dimension, $k \in 1, 2, \ldots, K$. This* k-matricization *is formally obtained via the mapping of the tensor indices $(i_1, i_2, \ldots, i_K)$ onto the matrix indices (i_k, j) via*

$$j = 1 + \sum_{l=1, l \neq k}^{K} (i_l - 1) J_l, \quad \text{where} \quad J_l := \prod_{m=1, m \neq l}^{l-1} I_m.$$

We observe that the matrix $H \in \mathbb{R}^{n \times n^2}$ in (2) can be interpreted as a 1-matricization of a 3-dimensional tensor $\mathcal{H} \in \mathbb{R}^{n \times n \times n}$ and the remaining two matricizations $H^{(2)}$ and $H^{(3)}$ are related to H as follows. Without loss of generality [6] we can assume that $H(u \otimes v) = H(v \otimes u)$ which implies that

$$w^T H(u \otimes v) = u^T H^{(2)}(v \otimes w) = u^T H^{(3)}(v \otimes w),$$

where $u, v, w \in \mathbb{R}^n$ are arbitrary vectors and H is such that it holds $H^{(2)} = H^{(3)}$. The above relation is used in Theorem 2.2 which explains the appearance of $H^{(2)}$ in the provided algorithm. For later reference we state the main results and the basic algorithm on two-sided moment-matching for general DAEs as developed in [6].

Theorem 2.2 ([6], Thm. 4.1). *Let E, A, H, N, b, c be the coefficient matrices of a quadratic-bilinear DAE as in (2). Suppose $F(s) := sE - A$ and $\sigma_i \in \mathbb{C}$ be the interpolation points such that $F(\tilde{\sigma}_i)$ is invertible for $\tilde{\sigma}_i = \{\sigma_i, 2\sigma_i\}$, $i = 1, \ldots, r$. Assume that a reduced quadratic-bilinear DAE is constructed by using a Petrov-Galerkin projection resulting in E_r, A_r, H_r, N_r, b_r and c_r as defined in (3). Then, if the basis matrices $\mathcal{V}$ and $\mathcal{W}$ are such that*

$$\operatorname{span}\{\mathcal{V}\} = \bigcup_{i=1,\ldots,r} \operatorname{span}\{\mathcal{V}^{(i)}\} \qquad \text{and} \qquad \operatorname{span}\{\mathcal{W}\} = \bigcup_{i=1,\ldots,r} \operatorname{span}\{\mathcal{W}^{(i)}\},$$

where $\mathcal{V}^i$ and $\mathcal{W}^{(i)}$ are

$$\begin{aligned}
\operatorname{span}\{\mathcal{V}^{(i)}\} &= \operatorname{span}\left\{F(\sigma_i)^{-1} b, F(2\sigma_i)^{-1}\left[H(F(\sigma_i)^{-1} b \otimes F(\sigma_i)^{-1} b) + N F(\sigma_i)^{-1} b\right]\right\}, \\
\operatorname{span}\{\mathcal{W}^{(i)}\} &= \operatorname{span}\Big\{F(2\sigma_i)^{-T} c^T, F(\sigma_i)^{-T}\Big[H^{(2)}(F(\sigma_i)^{-1} b \otimes F(2\sigma_i)^{-T} c^T) \\
&\qquad + \tfrac{1}{2} N^T F(2\sigma_i)^{-T} c^T\Big]\Big\}.
\end{aligned}$$

Then, the following moments are matched:

$$\begin{aligned}
G^{(1)}(\sigma_i) &= G_r^{(1)}(\sigma_i), & G^{(1)}(2\sigma_i) &= G_r^{(1)}(2\sigma_i), \\
G^{(2)}(\sigma_i, \sigma_i) &= G_r^{(2)}(\sigma_i, \sigma_i), & \frac{\partial}{\partial s_1 s_2} G^{(2)}(\sigma_i, \sigma_i) &= \frac{\partial}{\partial s_1 s_2} G_r^{(2)}(\sigma_i, \sigma_i), \quad j = 1, 2,
\end{aligned}$$

for $i = 1, \ldots, r$. (4)

This theoretical result gives rise to the following algorithm:

Algorithm 1 Model Reduction for general QBDAEs

1: Input: $E, A, H, N, b, c, \sigma_1, \ldots, \sigma_r$.
2: Construct the projection matrices $\mathcal{V}$ and $\mathcal{W}$ using Theorem 2.2.
3: Determine the reduced-order system as:

$$E_r = \mathcal{W}^T E \mathcal{V}, \qquad A_r = \mathcal{W}^T A \mathcal{V}, \qquad H_r = \mathcal{W}^T H \mathcal{V} \otimes \mathcal{V},$$
$$N_r = \mathcal{W}^T N \mathcal{V}, \qquad b_r = \mathcal{W}^T b, \qquad c_r = c\mathcal{V}.$$

Note that, in general, $\mathcal{V}$ is a dense matrix and the computation of $\mathcal{V} \otimes \mathcal{V}$ can be numerically infeasible. However, one can rely on concepts used in tensor theory to compute $\mathcal{W}^T H \mathcal{V} \otimes \mathcal{V}$ efficiently [6].

Remark 2.1. *Theorem 2.2 does not differentiate between E singular and E non-singular. As long as the inverse $(sE - A)^{-1}$ evaluated at $s = \sigma_i$ and $s = 2\sigma_i$ exists, the reduced quadratic-bilinear system will satisfy (4). However, if E is singular, the application of Theorem 2.2 may lead to a reduced-order system with an unbounded approximation error, i.e., $\|G^{(1)} - G_r^{(1)}\|_{\mathcal{H}_2} = \infty$ or $\|G^{(2)} - G_r^{(2)}\|_{\mathcal{H}_2} = \infty$, cf. the discussion of model reduction for linear descriptor systems in [16]. Turning the arguments around, in the ODE case, i.e., if E is not singular, we can expect a good performance of Algorithm 1.*

3 Model Reduction of Stokes-Type Quadratic-Bilinear DAEs

In view of the issues raised in Remark 2.1, in this section, we discuss the transformation of the Stokes-type quadratic-bilinear descriptor system (1) to an equivalent ODE system. In addition, we address some computation aspects related to the intended model reduction procedure.

3.1 Transformation of the Stokes-type QBDAE

We consider the Stokes-type quadratic-bilinear descriptor systems (1) which can be transformed into an equivalent quadratic-bilinear ODE system. We consider the system equations (1) with $B_2 = 0$ and $\alpha = 0$:

$$E_{11}\dot{v}(t) = A_{11}v(t) + A_{12}p(t) + H_1 v(t) \otimes v(t) + N_1 v(t)u(t) + b_1 u(t), \tag{5a}$$
$$v(0) = 0, \tag{5b}$$
$$0 = A_{21}v(t), \tag{5c}$$
$$y(t) = c_1 v(t) + c_2 p(t) + Du(t). \tag{5d}$$

For the analysis, we consider a decoupling, cf. [18, Thm. 8.6], of (5) into an algebraic and a differential part as

$$p(t) = -(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}\big(A_{11}v(t) + H_1 v(t) \otimes v(t) + N_1 v(t)u(t) + b_1 u(t)\big) \tag{6}$$

and

$$\begin{aligned} E_{11}\dot{v}(t) &= \Pi A_{11}v(t) + \Pi H_1 v(t)\otimes v(t) + \Pi N_1 v(t)u(t) + \Pi b_1 u(t), \quad v(0)=0,\\ y(t) &= \mathcal{C}v(t) + \mathcal{C}_H v(t)\otimes v(t) + \mathcal{C}_N v(t)u(t) + \mathcal{D}u(t), \end{aligned} \tag{7}$$

where

$$\begin{aligned} \mathcal{C} &= c_1 - c_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}A_{11}, \quad \mathcal{C}_H = -c_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}H_1,\\ \mathcal{C}_N &= -c_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}N_1, \qquad \mathcal{D} = D - c_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}b_1, \end{aligned}$$

and

$$\Pi = I - A_{12}(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}. \tag{8}$$

In what follows, we assume that $A_{21} = A_{12}^T$. The arguments for the unsymmetrical case are laid out in [16] and can be readily applied in the current nonlinear setting.

Note that Π is the discrete *Helmholtz* projector that is commonly used [16, 18, 19] to transform Stokes type DAEs into ODEs and that has the following properties:

$$\Pi^2 = \Pi,\ E_{11}\Pi = \Pi^T E_{11},\ \ker(\Pi) = \text{range}\,(A_{12}),\ \text{and}\ \text{range}\,(\Pi) = \ker(A_{12}^T E_{11}^{-1}).$$

Using these properties of Π, one can derive that

$$A_{12}^T z = 0 \quad \text{if, and only if,} \quad \Pi^T z = z. \tag{9}$$

By construction, a solution $v(t)$ of (7) fulfills $A_{12}^T v(t) = 0$, so that in (7), we can replace $v(t)$ by $\Pi^T v(t)$ and, using $\Pi = \Pi^2$ and $E_{11}\Pi = \Pi^T E_{11}$, we obtain the following equivalent system

$$\Pi E_{11}\Pi^T \dot{v}(t) = \Pi A_{11}\Pi^T v(t) + \Pi H_1\big(\Pi^T v(t)\otimes \Pi^T v(t)\big) + \Pi N_1 \Pi^T v(t)u(t) + \Pi b_1 u(t), \tag{10a}$$

$$y(t) = \mathcal{C}\Pi^T v(t) + \mathcal{C}_H\big(\Pi^T v(t)\otimes \Pi^T v(t)\big) + \mathcal{C}_N\Pi^T v(t)u(t) + \mathcal{D}u(t), \tag{10b}$$

with $v(0) = 0$. The above dynamical system (10) lies in the $n_1 - n_2$ dimensional null space of Π. Therefore, as in [19], we can decompose the projector Π as

$$\Pi = \phi_1\phi_2^T, \tag{11}$$

with $\phi_1, \phi_2 \in \mathbb{R}^{n_1\times n_1-n_2}$ satisfying

$$\phi_1^T\phi_2 = I.$$

This decomposition allows us to write (10) in the following form

$$\phi_2^T E_{11}\phi_2\dot{\tilde{v}}(t) = \phi_2^T A_{11}\phi_2\tilde{v}(t) + \phi_2^T H_1\big(\phi_2\tilde{v}(t)\otimes\phi_2\tilde{v}(t)\big) + \phi_2^T N_1\phi_2\tilde{v}(t)u(t) + \phi_2^T b_1 u(t), \tag{12a}$$

$$y(t) = \mathcal{C}\phi_2\tilde{v}(t) + \mathcal{C}_H\big(\phi_2\tilde{v}(t)\otimes\phi_2\tilde{v}(t)\big) + \mathcal{C}_N\phi_2\tilde{v}(t)u(t) + \mathcal{D}u(t), \tag{12b}$$

with $\tilde{v}(t) = \phi_1^T v(t)$ and $\tilde{v}(0) = 0$. Thus, model reduction of the original quadratic-bilinear DAE is equivalent to the reduction of (10) and (12). However, the equivalent system in (12) has the advantage that the matrix $\phi_2^T E_{11}\phi_2$ is nonsingular (ϕ_2 has full column rank). Therefore, standard Krylov subspace technique for model reduction of a quadratic-bilinear system, discussed in Section 2, can be employed to the system (12) in order to obtain a reduced-order system.

Note that the output equation of the system (12) involves nonlinear terms in the state and input. It is still an open problem to consider the nonlinear terms in the output equations in order to compute the projection matrices and need further research in this direction. In this paper we restrict ourself to the linear relation between the state and the output by neglecting the nonlinear terms in the output as far as the computation of the projection matrices $\mathcal{V}$ and $\mathcal{W}$ are concerned. Having neglected these terms in (12), we consider the following equation

$$\begin{aligned} \phi_2^T E_{11}\phi_2\dot{\tilde{v}}(t) &= \phi_2^T A_{11}\phi_2\tilde{v}(t) + \phi_2^T H_1\big(\phi_2\tilde{v}(t)\otimes\phi_2\tilde{v}(t)\big) + \phi_2^T N_1\phi_2\tilde{v}(t)u(t) + \phi_2^T b_1 u(t), \\ \tilde{y}(t) &= \mathcal{C}\phi_2\tilde{v}(t), \end{aligned} \tag{13}$$

for the definition of the projection matrices by applying Theorem 2.2 in order to identify the reduced quadratic-bilinear system.

Remark 3.1. *Algorithm 1 may safely be applied (cf. Remark 2.1) to the system* (7)*, which is an equivalent quadratic-bilinear ODE system too. To avoid the projector* Π *and to stay in line with [16, 19], we rather consider the* condensed *projected formulation* (13)*.*

The following subsection shows how the computation of ϕ_2 or Π can be avoided in an implementation.

3.2 Computational Issues

We will use the ODE system (13) to compute the projection matrices $\mathcal{V}$ and $\mathcal{W}$ which we will apply to reduce the original coefficients E_{11}, A_{11}, H_1, N_1, b_1, $\mathcal{C}$, $\mathcal{C}_H$, and $\mathcal{C}_N$ to give

$$\begin{aligned} E_r &= \mathcal{W}^T E_{11}\mathcal{V}, & A_r &= \mathcal{W}^T A_{11}\mathcal{V}, & H_{1r} &= \mathcal{W}^T H_1\mathcal{V}\otimes\mathcal{V}, & N_{1r} &= \mathcal{W}^T N_1\mathcal{V}, \\ b_r &= \mathcal{W}^T b_1, & \mathcal{C}_r &= \mathcal{C}\mathcal{V}, & \mathcal{C}_{Hr} &= \mathcal{C}_H\mathcal{V}\otimes\mathcal{V}, & \mathcal{C}_{Nr} &= \mathcal{C}_N\mathcal{V}. \end{aligned}$$

Thus the reduced system will be independent of ϕ_2. However, the definition of $\mathcal{V}$ and $\mathcal{W}$ will involve ϕ_2 as defined in (9) which might not be easily accessible.

Theorem 3.1. *Let* $\mathcal{F}(s)$ *be defined as*

$$\mathcal{F}(s) := \phi_2(s\phi_2^T E_{11}\phi_2 - \phi_2^T A_{11}\phi_2)^{-1}\phi_2^T.$$

Also, let $\mathcal{V}$ *and* $\mathcal{W}$ *be such that*

$$\operatorname{span}\{\mathcal{V}\} = \bigcup_{i=1,\ldots,r}\operatorname{span}\{\mathcal{V}^{(i)}\} \qquad \textit{and} \qquad \operatorname{span}\{\mathcal{W}\} = \bigcup_{i=1,\ldots,r}\operatorname{span}\{\mathcal{W}^{(i)}\}, \tag{14}$$

in which for $i = 1 : r$

$$\operatorname{span}\{\mathcal{V}^{(i)}\} = \operatorname{span}\left\{\mathcal{F}(\sigma_i)b_1, \mathcal{F}(2\sigma_i)\left[H_1(\mathcal{F}(\sigma_i)b_1 \otimes \mathcal{F}(\sigma_i)b_1) + N_1\mathcal{F}(\sigma_i)b_1\right]\right\}, \quad (15)$$

$$\operatorname{span}\{\mathcal{W}^{(i)}\} = \operatorname{span}\Big\{\mathcal{F}(2\sigma_i)\mathcal{C}^T, \mathcal{F}(\sigma_i)\Big[\mathcal{H}_1^{(2)}(\mathcal{F}(\sigma_i)b_1 \otimes \mathcal{F}(2\sigma_i)^T\mathcal{C}^T) + \tfrac{1}{2}N_1^T\mathcal{F}(2\sigma_i)^T\mathcal{C}^T\Big]\Big\}, \quad (16)$$

Then, the following moments are matched:

$$\begin{aligned} \mathcal{G}^{(1)}(\sigma_i) &= \mathcal{G}_r^{(1)}(\sigma_i), & \mathcal{G}^{(1)}(2\sigma_i) &= \mathcal{G}_r^{(1)}(2\sigma_i), \\ \mathcal{G}^{(2)}(\sigma_i,\sigma_i) &= \mathcal{G}_r^{(2)}(\sigma_i,\sigma_i), & \frac{\partial}{\partial s_1 s_2}\mathcal{G}^{(2)}(\sigma_i,\sigma_i) &= \frac{\partial}{\partial s_1 s_2}\mathcal{G}_r^{(2)}(\sigma_i,\sigma_i), \quad j = 1, 2. \end{aligned} \quad (17)$$

Here $\mathcal{G}^{(1)}(s_1)$ *and* $\mathcal{G}^{(2)}(s_1, s_2)$ *are the first two generalized transfer functions of the equivalent system* (13)*. Similarly,* $\mathcal{G}_r^{(1)}(s_1)$ *and* $\mathcal{G}_r^{(2)}(s_1, s_2)$ *are the first two generalized transfer functions of the reduced-order system.*

Proof. We first define the state matrices in (13) as

$$\begin{aligned} \bar{E}_{11} &= \phi_2^T E_{11}\phi_2, & \bar{A}_{11} &= \phi_2^T A_{11}\phi_2, & \bar{H}_1 &= \phi_2^T H_1(\phi_2 \otimes \phi_2) \\ \bar{N}_1 &= \phi_2^T H_1\phi_2, & \bar{b}_1 &= \phi_2^T b_1, & \bar{\mathcal{C}} &= \mathcal{C}\phi_2 \end{aligned} \quad (18)$$

For these matrices, if the projection matrices $\bar{V}$ and $\bar{W}$ are computed according to Theorem 2.2, then we can relate them to $\mathcal{V}$ and $\mathcal{W}$ as

$$\mathcal{V} = \phi_2\bar{V}, \quad \text{and} \quad \mathcal{W} = \phi_2\bar{W}, \quad (19)$$

respectively. Together with the results required for the proof of Theorem 2.2, we can easily prove (17) by using the above relations. In the following, we only prove the first equation in (17) and the remaining equations follow analogously. Since

$$\bar{V}(\sigma_i\bar{W}^T\bar{E}_{11}\bar{V} - \bar{W}^T\bar{A}_{11}\bar{V})^{-1}\bar{W}^T\bar{b}_1 = (\sigma_i\bar{E}_{11} - \bar{A}_{11})^{-1}\bar{b}_1, \quad (20)$$

we use (18) and (19) to have

$$\bar{V}(\sigma_i\mathcal{W}^T E_{11}\mathcal{V} - \mathcal{W}^T A_{11}\mathcal{V})^{-1}\mathcal{W}^T b_1 = (\sigma_i\phi_2^T E_{11}\phi_2 - \phi_2^T A_{11}\phi_2)^{-1}\phi_2^T b_1, \quad (21)$$

Pre-multiplying $\mathcal{C}\phi_2$ on both sides of the above equation one can find that $\mathcal{G}^{(1)}(\sigma_i) = \mathcal{G}_r^{(1)}(\sigma_i)$ holds. Similarly, we can prove the other three equalities in (17). □

Theorem 3.1 is based on a moment matching framework that matches the generalized transfer functions and their first partial derivatives. However, it is also possible to use an interpolation scheme, as suggested in [5], that also matches higher derivatives of the first two generalized transfer functions. The following corollary shows the results for (13) with higher moments matched.

Corollary 3.1. *Let $\mathcal{F}(s)$ be as defined in Theorem 3.1 and let $\underline{\mathcal{V}}$ and $\underline{\mathcal{W}}$ be such that*

$$\operatorname{span}\{\underline{\mathcal{V}}\} = \bigcup_{i=1,\dots,r} \operatorname{span}\{\underline{\mathcal{V}}_1^{(i)},\dots,\underline{\mathcal{V}}_q^{(i)}\}, \quad \operatorname{span}\{\underline{\mathcal{W}}\} = \bigcup_{i=1,\dots,r} \operatorname{span}\{\underline{\mathcal{W}}_1^{(i)},\dots,\underline{\mathcal{W}}_q^{(i)}\},$$

in which $\underline{\mathcal{V}}_1^{(i)}$ and $\underline{\mathcal{W}}_1^{(i)}$ are the same as $\mathcal{V}^{(i)}$ and $\mathcal{W}^{(i)}$ in Theorem 3.1 and $\underline{\mathcal{V}}_2^{(i)},\dots,\underline{\mathcal{V}}_q^{(i)}$ are of the form

$$\begin{aligned}\operatorname{span}\{\underline{\mathcal{V}}_2^{(i)}\} =& \operatorname{span}\Big\{\mathcal{F}(\sigma_i)E_{11}\mathcal{F}(\sigma_i)b_1, \mathcal{F}(2\sigma_i)E_{11}\mathcal{F}(2\sigma_i)\\ &[H_1(\mathcal{F}(\sigma_i)b_1\otimes\mathcal{F}(\sigma_i)b_1)+N_1\mathcal{F}(\sigma_i)b_1]\Big\}, \qquad (22)\\ \operatorname{span}\{\underline{\mathcal{V}}_q^{(i)}\} =& \operatorname{span}\Big\{(\mathcal{F}(\sigma_i)E_{11})^{q-1}\mathcal{F}(\sigma_i)b_1, (\mathcal{F}(2\sigma_i)E_{11})^{q-1}\mathcal{F}(2\sigma_i)\\ &[H_1(\mathcal{F}(\sigma_i)b_1\otimes\mathcal{F}(\sigma_i)b_1)+N_1\mathcal{F}(\sigma_i)b_1]\Big\} \qquad (23)\end{aligned}$$

Similarly we can define $\underline{\mathcal{W}}_2^{(i)},\dots,\underline{\mathcal{W}}_q^{(i)}$. With such projection matrices, the reduced system does not only ensure (17) *but also the matching of higher moments, i.e., the following equations also hold:*

$$\begin{aligned}\frac{\partial^p}{\partial s_1^p}\mathcal{G}^{(1)}(\sigma_i) = \frac{\partial^p}{\partial s_1^p}\mathcal{G}_r^{(1)}(\sigma_i), \quad \frac{\partial^p}{\partial s_1^p}\mathcal{G}^{(1)}(2\sigma_i) = \frac{\partial^p}{\partial s_1^p}\mathcal{G}_r^{(1)}(2\sigma_i),\\ \frac{\partial^{i+j}}{\partial s_1^i s_2^j}\mathcal{G}^{(2)}(\sigma_i,\sigma_i) = \frac{\partial^{i+j}}{\partial s_1^i s_2^j}\mathcal{G}_r^{(2)}(\sigma_i,\sigma_i)\end{aligned}$$

for $p = 1,\dots,q-1$ and $i+j < 2q-1$.

Note that the projection matrices $\mathcal{V}$ and $\mathcal{W}$ in Theorem 3.1 (and also $\underline{\mathcal{V}}$ and $\underline{\mathcal{W}}$ in Corollary 3.1) are projections for the original subsystems $G^{(1)}(s_1) = \mathcal{C}(sE_{11}-A_{11})^{-1}b_1$ and $G^{(2)}(s_1,s_2)$, that are independent of ϕ_2. Now, as a second step, we need to show that the matrices $\mathcal{V}^{(i)}$and $\mathcal{W}^{(i)}$ can also be constructed such that they do not require the computation of ϕ_2.

We start with the subspace associated with $\mathcal{V}^{(i)}$. As shown in (15), the column vectors of $\mathcal{V}^{(i)}$, for $i = 1,\dots,r$ can be written as

$$\mathcal{V}^{(i)} = \{\mathcal{F}(\sigma_i)b_1, \mathcal{F}(2\sigma_i)\,[H_1(\mathcal{F}(\sigma_i)b_1\otimes\mathcal{F}(\sigma_i)b_1)+N_1\mathcal{F}(\sigma_i)b_1]\}\,. \qquad (24)$$

The first column of $\mathcal{V}^{(i)}$, which is $\mathcal{F}(\sigma_i)b_1$, can be identified by solving a linear system that does not resort to ϕ_2 [16, 19].

Lemma 3.1 ([16], Lem. 6.3)**.** *Let $(\tilde\sigma\phi_2^T E_{11}\phi_2 - \phi_2^T A_{11}\phi_2)^{-1}$ exist for $\tilde\sigma = \sigma$ or $\tilde\sigma = 2\sigma$. Then $v = \mathcal{F}(\tilde\sigma)f$ solves*

$$\begin{bmatrix}\tilde\sigma E_{11}-A_{11} & A_{12}\\ A_{12}^T & 0\end{bmatrix}\begin{bmatrix}v\\ z\end{bmatrix} = \begin{bmatrix}f\\ 0\end{bmatrix}, \qquad (25)$$

and $w = \mathcal{F}(\tilde\sigma)^T g$ solves

$$\begin{bmatrix}\tilde\sigma E_{11}^T-A_{11}^T & A_{12}\\ A_{12}^T & 0\end{bmatrix}\begin{bmatrix}w\\ q\end{bmatrix} = \begin{bmatrix}g\\ 0\end{bmatrix}, \qquad (26)$$

where f and g are arbitrary vectors or matrices of appropriate sizes.

This means that for $\tilde{E} := \begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix}$ and $\tilde{A} := \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0 \end{bmatrix}$, relation (25) implies that the first column of $\mathcal{V}^{(i)}$ can be given by $\begin{bmatrix} I_{n_1} & 0 \end{bmatrix} (\sigma\tilde{E} - \tilde{A})^{-1}\tilde{b}_1$, where I_{n_1} is the identity matrix of size $n_1 \times n_1$ and $\tilde{b}_1^T = \begin{bmatrix} b_1^T & 0 \end{bmatrix}$.

Lemma 3.1 can also be used to construct the second column of $\mathcal{V}^{(i)}$ as well via setting $f = H_1(\mathcal{V}_1^i(:,1) \otimes \mathcal{V}^i(:,1)) + N_1\mathcal{V}^{(i)}(:,1)$, where $\mathcal{V}^{(i)}(:,1)$ is the first column of $\mathcal{V}^{(i)}$. Thus we can identify $\mathcal{V}$ without explicitly computing ϕ_2. Similarly, $\mathcal{W}$ can also be identified from such settings. The framework can also be extended to $\underline{\mathcal{V}}$ and $\underline{\mathcal{W}}$, such that they are also independent of ϕ_2. The following algorithm summarizes all the steps for the reduction of the system in (13).

Algorithm 2 Model Reduction for Stokes-type QBDAEs

1: **Input:** $E_{11}, A_{11}, H_1, N_1, b_1, \mathcal{C}, \mathcal{C}_H, \mathcal{C}_N, \sigma_1, \ldots, \sigma_r$.
2: **Output:** $E_r, A_r, H_{1r}, N_{1r}, b_r, \mathcal{C}_r, \mathcal{C}_{Hr}, \mathcal{C}_{Nr}$.
3: Construct $\mathcal{V}^{(i)}$ and $\mathcal{W}^{(i)}$

$$\operatorname{span}\{\mathcal{V}^{(i)}\} = \operatorname{span}\left\{ I_0\tilde{F}(\sigma_i)^{-1}\tilde{B}, I_0\tilde{F}(2\sigma_i)^{-1}\mathcal{P} \right\},$$

$$\text{where } \tilde{B} = \begin{bmatrix} b_1 \\ 0 \end{bmatrix}, \ \mathcal{P} = \begin{bmatrix} H_1\big(\mathcal{V}^{(i)}(:,1) \otimes \mathcal{V}^{(i)}(:,1)\big) + N_1\mathcal{V}^{(i)}(:,1) \\ 0 \end{bmatrix},$$

$$\tilde{F}(s) = s\tilde{E} - \tilde{A} \text{ and } I_0 = [I_{n_1}, 0].$$

$$\operatorname{span}\{\mathcal{W}^{(i)}\} = \operatorname{span}\left\{ I_0\tilde{F}(2\sigma_i)^{-T}\tilde{\mathcal{C}}^T, I_0\tilde{F}(\sigma_i)^{-1}\mathcal{Q} \right\},$$

$$\text{where } \tilde{\mathcal{C}}^T = \begin{bmatrix} \mathcal{C}^T \\ 0 \end{bmatrix}, \ \mathcal{Q} = \begin{bmatrix} \mathcal{H}_1^2\big(\mathcal{V}^{(i)}(:,1) \otimes \mathcal{W}^{(i)}(:,1)\big) + (1/2)N_1^T\mathcal{W}^{(i)}(:,1) \\ 0 \end{bmatrix}.$$

4: $\operatorname{span}\{\mathcal{V}\} = \underset{i=1,\ldots,r}{\cup} \operatorname{span}\{\mathcal{V}^{(i)}\}$, and $\operatorname{span}\{\mathcal{W}\} = \underset{i=1,\ldots,r}{\cup} \operatorname{span}\{\mathcal{W}^{(i)}\}$.
5: Construct the reduced-order system:

$$E_r = \mathcal{W}^T E_{11}\mathcal{V}, \quad A_r = \mathcal{W}^T A_{11}\mathcal{V}, \quad H_{1r} = \mathcal{W}^T H_1\mathcal{V} \otimes \mathcal{V}, \quad N_{1r} = \mathcal{W}^T N_1\mathcal{V},$$
$$b_r = \mathcal{W}^T b_1, \quad \mathcal{C}_r = \mathcal{C}\mathcal{V}, \quad \mathcal{C}_{Hr} = \mathcal{C}_H\mathcal{V} \otimes \mathcal{V}, \quad \mathcal{C}_{Nr} = \mathcal{C}_N\mathcal{V}.$$

4 The General Case $b_2 \neq 0$

In the preceding section, we have derived the equivalent representations of the descriptor system (1) under the assumption that $b_2 = 0$. Now, we consider the general case with $b_2 \neq 0$ and show how this can be brought back to the $b_2 = 0$ case. The approach is similar to the approach for linear DAEs with $b_2 \neq 0$ taken in [16, 19]. In the nonlinear setting, some extra terms need to be considered. Analogously to the linear case, we decompose $v(t)$ as

$$v(t) = v_0(t) + v_g(t),$$

with

$$v_g(t) = -\underbrace{E_{11}^{-1}A_{12}(A_{12}^T E_{11}^{-1}A_{12})^{-1}b_2}_{\Upsilon} u(t) \tag{27}$$

so that and $v_0(t)$ satisfies $0 = A_{12}^T v_0(t)$. In the state equations (1), we substitute $v(t)$ by its decomposition which leads to the following system

$$\begin{aligned} E_{11}\dot{v}_0(t) &= A_{11}v_0(t) + A_{12}p(t) + H_1 v_0(t)\otimes v_0(t) + \mathcal{N}_1 v_0(t)u(t) + \mathcal{B}_1 u(t) + \mathcal{B}_u u^2(t) \\ &\quad + A_{12}(A_{12}^T E_{11}^{-1} A_{12})^{-1} b_2 \dot{u}(t), \qquad (28a) \\ v_0(0) &= \alpha - v_g(0) \qquad (28b) \\ 0 &= A_{12}^T v_0(t), \qquad (28c) \\ y(t) &= c_1 v_0(t) + c_2 p(t) + (D - c_1 E_{11}^{-1} A_{12}(A_{12}^T E_{11}^{-1} A_{12})^{-1} b_2) u(t), \qquad (28d) \end{aligned}$$

where

$$\mathcal{N}_1 = N_1 - H_1(\Upsilon \otimes I + I \otimes \Upsilon), \quad \mathcal{B}_1 = b_1 - A_{11}\Upsilon, \quad \mathcal{B}_u = H_1\,(\Upsilon \otimes \Upsilon) - N_1\Upsilon.$$

Using (28c) and (28a), we can explicitly compute $p(t)$ as

$$\begin{aligned} p(t) = -(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{12}^T E_{11}^{-1} \Big(& A_{11}v(t) + H_1 v_0(t)\otimes v_0(t) + \mathcal{N}_1 v_0(t)u(t) \\ & + \mathcal{B}_1 u(t) + \mathcal{B}_u u^2(t) + A_{12}(A_{12}^T E_{11}^{-1} A_{12})^{-1} b_2 \dot{u}(t) \Big). \end{aligned} \qquad (29)$$

Remark 4.1. *With p given through* (29)*, the system* (28) *is of the form* (5) *but with terms containing $u(t), u^2(t)$ and $\dot{u}(t)$. Although these terms are functions of $u(t)$, in a forward simulation, we can consider them as three different inputs. Accordingly, we can use the transformation steps discussed in Section 3 to obtain an associated ODE system in $v_0(t)$.*

In fact, substituting $p(t)$ into (28a), using that $A_{12}^T v_0(t) = 0$ implies $\Pi^T v_0(t) = v_0(t)$, and premultiplying the resulting system by Π, we arrive at an ODE for v_0:

$$\begin{aligned} \Pi E_{11}\dot{v}_0(t) &= \Pi A_{11} v(t) + \Pi H_1(v_0(t)\otimes v_0(t)) + \Pi\mathcal{N}_1 v_0(t)u(t) + \Pi\mathcal{B}\tilde{u}(t), \qquad (30a) \\ v_0(0) &= \Pi^T(\alpha - v_g(0)), \qquad (30b) \\ y(t) &= \mathcal{C}\Pi^T v_0(t) + \mathcal{C}_H(\Pi\otimes\Pi)^T(v_0(t)\otimes v_0(t)) + \mathcal{C}_N\Pi^T v_0(t) \\ &\quad + \mathcal{D}\tilde{u}(t) - c_2(A_{12}^T E_{11}^{-1} A_{12})^T b_2 \dot{u}(t), \qquad (30c) \end{aligned}$$

where $\mathcal{B} = [\mathcal{B}_1, \mathcal{B}_u]$, with the new input $\tilde{u}(t) := [u(t), u^2(t)]^T$, and where

$$\begin{aligned} \mathcal{C} &= c_1 - c_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{12}^T E_{11}^{-1} A_{11}, \\ \mathcal{C}_H &= -c_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{12}^T E_{11}^{-1} H_1, \\ \mathcal{C}_N &= -c_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{12}^T E_{11}^{-1}\mathcal{N}_1, \end{aligned}$$

and $\mathcal{D} = [\mathcal{D}_1, \mathcal{D}_2]$, with

$$\begin{aligned} &\mathcal{D}_1 = D - c_1 E_{11}^{-1} A_{12} A_{12}^T E_{11}^{-1} A_{12})^{-1} b_2 - c_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{12}^T E_{11}^{-1}\mathcal{B}_1, \quad \text{and} \\ &\mathcal{D}_2 = c_2(A_{12}^T E_{11}^{-1} A_{12})^{-1} A_{12}^T E_{11}^{-1}\mathcal{B}_u. \end{aligned}$$

Note that, because of $\Pi A_{12}(A_{12}^T E_{11}^{-1} A_{12})^{-1} b_2 = 0$, the term associated with $\dot{u}(t)$ in (28a) vanishes.

In Section 3, we have referred to *equivalent systems* as systems that have the same solution set. In the current setting, where $v(0) \neq 0$ and $b_2 \neq 0$, this equivalence is bound to the *consistency* of the initial value. Bluntly put, if the initial value is consistent then the decoupled and the original system have the same solution set. If the initial value is not consistent, then the decoupled system can have a solution although the original system is not solvable, cf. the following Theorem 4.1.

Theorem 4.1. *Consider the state equations* (2) *and a given input function* $u \in \mathcal{C}([0,T])$*. A solution* $(v,p) \in \mathcal{C}^1((0,T];\mathbb{R}^{n_1}) \times \mathcal{C}([0,T];\mathbb{R}^{n_2})$ *can only exist if the initial value* $\alpha \in \mathbb{R}^{n_1}$ *fulfills*

$$\alpha = \alpha_0 + E_{11}^{-1} A_{12}(A_{12}^T E_{11}^{-1} A_{12})^{-1} b_2 u(0) \tag{31}$$

for a $\alpha_0 \in \ker A_{12}^T$*. If this is the case, then* (v,p) *is defined through* $v = v_0 + v_g$ *and* p*, where* v_0*,* v_g*, and* p *solve the decoupled system (30a-c),* (27)*, and* (29)*.*

Proof. The claim is a direct consequence of [18, Thm. 8.6] considering also [18, Rem. 8.7]. □

Remark 4.2. *In general, the consistency condition* (31) *can never hold for all inputs* $u \in \mathcal{C}([0,T])$*, which is not seen by the ODE* (30a) *for* v_0*. In our numerical experiments, for the time being, we will ensure consistency and, thus, equivalence of the reformulation, by fixing the initial value of the inputs and adjusting the initial value* $v(0)$ *accordingly.*

Considering the issues raised in Remark 4.1 and Remark 4.2, we can employ System (30) to determine a reduced-order system. With $u(t)$ and $u^2(t)$ being considered as two different inputs, the equivalent system (30) is a multi-input system and the algorithms for the SISO case of Section 3 do not readily apply. In our numerical examples, for the case with $b_2 \neq 0$, we will consider constant inputs for which the SISO case approach still works. The discussion of moment matching for multi-input multi-output (MIMO) Stokes-type quadratic-bilinear system we leave to a forthcoming paper.

5 Numerical Results

In this section, we examine the performance of the proposed approach for Stokes-type quadratic-bilinear descriptor systems by comparing it to the general version of the algorithm. To the newly proposed specification of the moment matching algorithm for Stokes-type systems, we refer as ind2QBmm and to the general moment matching implementation on the basis of Algorithm 1, we refer as genQBmm.

5.1 Lid Driven Cavity

As a first test case, we consider a lid driven cavity on the unit square $\Omega = [0,1]^2$, with the boundary Γ, for time $t \in (0,2]$, modelled by the Navier-Stokes equations for the velocity $\mathfrak{v}$ and the pressure $\mathfrak{p}$

$$\dot{\mathfrak{v}} + (\mathfrak{v} \cdot \nabla)\mathfrak{v} - \frac{1}{\mathrm{Re}}\Delta\mathfrak{v} + \nabla\mathfrak{p} = 0, \tag{32a}$$

$$\nabla \cdot \mathfrak{v} = 0, \tag{32b}$$

$$\mathfrak{v}\big|_{\Gamma} = g, \tag{32c}$$

$$\mathfrak{v}\big|_{t=0} = \mathfrak{v}_0, \tag{32d}$$

where the Reynolds number Re is a parameter depending on the geometry, a characteristic velocity, and the kinematic viscosity, where g is the Dirichlet condition that models the driven lid, i.e. $\mathfrak{v} = \begin{bmatrix} 1 & 0 \end{bmatrix}$ at the upper boundary and $\mathfrak{v} = 0$ elsewhere at the boundary, and where $\mathfrak{v}_0$ is an initial condition. We apply a finite element discretization to (32) using the *Taylor-Hood* scheme on a uniform mesh which leads to a system for the discretized velocity v and pressure p of the form

$$E_{11}\dot{v} = A_{11}v + Hv \otimes v + A_{12}p + f, \qquad v(0) = v_s \tag{33a}$$

$$0 = A_{12}^T v, \tag{33b}$$

where E_{11} is the mass matrix and, thus, positive definite, A_{11} models the discrete diffusion, H models the discretized convection, and A_{12} is the discrete gradient operator with its transpose modelling the divergence. The source term f contains the Dirichlet boundary conditions. For the initial value we choose associated steady-state solution. The chosen a spatial discretization resulted in a state space dimension of $n_1 + n_2 = 1681 + 255 = 1936$.

The system (33) is extended to a descriptor system as follows. To model the input, we add $b_1 u := fu$ to (33a). As the output, we take the average pressure in the subdomain $\Omega_o = [0.45, 0.55] \times [0.7, 0.8]$, i.e.

$$y(t) = c_2 p(t) := \frac{1}{|\Omega_o|} \int_{\Omega_o} p(t,x)dx \tag{34}$$

evaluated in the corresponding finite element space. See, Figure 1 for an illustration of the problem and its discretization. We consider the input to output behavior for an actuation $b_1 u(t)$. A model-order reduction for lid driven cavity was also considered in [13] where the quadratic-bilinear DAE was approximated as a bilinear DAE via *Carleman* bilinearization to determine a reduced-order system employing an $\mathcal{H}_2$ optimal model reduction strategy. However, unlike in this paper, in [13], the driven cavity was modelled using the Navier-Stokes equations in vorticity-stream function formulations.

As it is common practice in optimal control and model reduction, we consider the deviation from a reference state instead of the actual state. Therefore, we decompose $v = v_s + v_\delta$ and $p = p_s + p_\delta$, where (v_s, p_s) is the associated steady state solution of (33), and obtain the system for the deviations (v_δ, p_δ) as

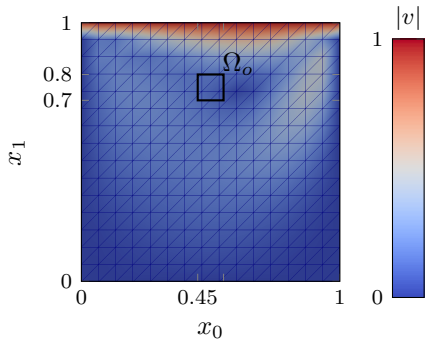


Figure 1: Illustration of the driven cavity for Re = 10, the grid used for the *Taylor-Hood* discretization, and the domain of observation Ω_o.

$$E_{11}\dot{v}_\delta = (A_{11} + L)v_\delta + Hv_\delta \otimes v_\delta + A_{12}p_\delta + b_1 u(t), \qquad v_\delta(0) = 0, \tag{35a}$$

$$0 = A_{12}^T v_\delta, \tag{35b}$$

$$y(t) = c_2 p_\delta, \tag{35c}$$

where $L := H(v_\delta \otimes I + I \otimes v_\delta)$. Note that no parts of the equations have been discarded as it is typically done in stabilization setups [3, 7]. We use ind2QBmm to approximate the system described through (35) and compare the results to the results obtained via genQBmm for two Reynolds numbers Re = 10 and Re = 75.

Reynolds number Re= 10

We first consider the low Reynolds number problem and seek to determine the reduced-order systems via both approaches. The interpolation points σ are identified by using *IRKA* on the linear part of the quadratic bilinear DAE, cf. [16]. The number of interpolation points is set to 10, which leads to a reduced-order system of order 20. The time domain simulations for the original and the reduced systems via one-sided projection (by setting $\mathcal{W} = \mathcal{V}$) for $u(t) = e^{-t}(2 + \sin(2\pi t))$ and the absolute error are shown in Figure 2. In order to match more moments, we set $\mathcal{W} \neq \mathcal{V}$ and identify them by using Algorithm 2. In case of two-sided moment-matching, we compare results for the full state vectors which are shown in Figure 3 for the same $u(t)$ and absolute error in the velocity on the full grid are shown in 4 . These figures clearly show that the specific approach ind2QBmm captures the dynamics of the original system way much better than the general approach.

Reynolds number Re= 75

As for the previous example, also for the higher Reynolds number Re = 75 setup, we determine the interpolation points by applying the *IRKA* on the linear part. We

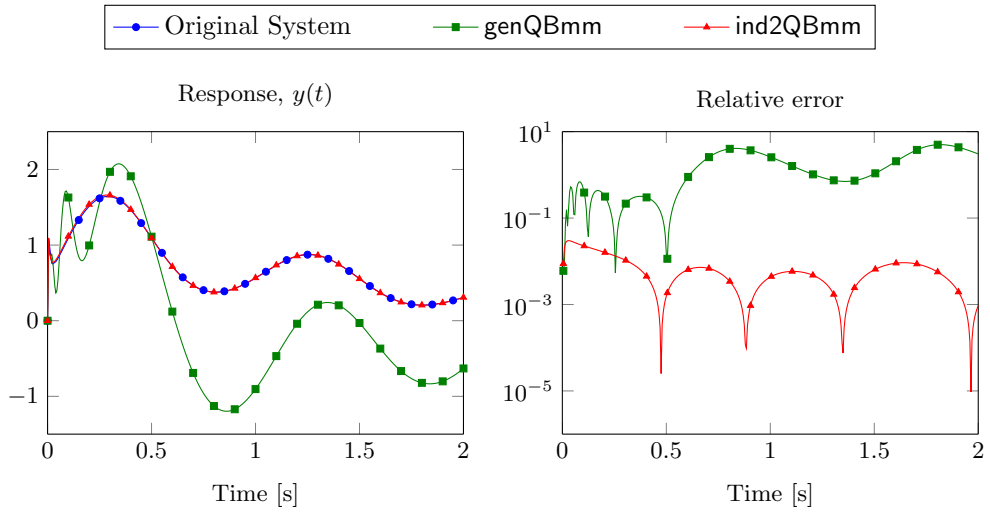


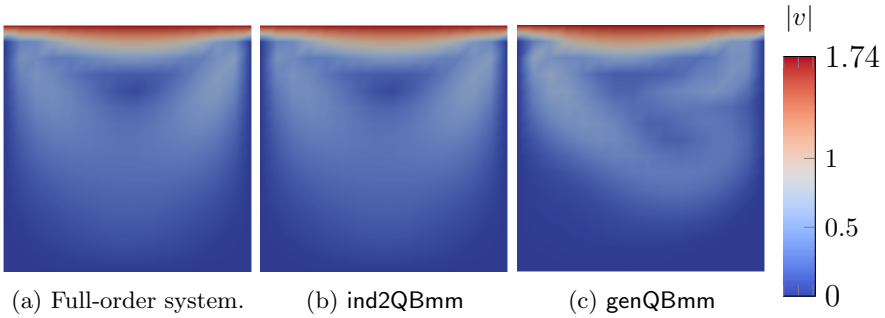
Figure 2: Comparison of the reduced-order systems obtained from the genQBmm and ind2QBmm implementation for one-sided moment-matching for the input $u(t) = e^{-t}(2 + \sin(2\pi t))$.

determined reduced-order systems of order 20 using the newly developed approach ind2QBmm and the generally valid approach genQBmm using one-sided projection. The results of the time domain simulations for the original and the reduced systems for $u(t) = 0.5e^{-t}(2+\sin(2\pi t))$ as well as the absolute error between the approximations are shown in Figure 5. We observe that unlike the general implementation genQBmm, the specific approach ind2QBmm well captures the input-output behavior of the original system, see the plot of the relative error in Figure 5. In the case of two-sided projections, we were unable to determine a stable reduced-order system using either approach.

5.2 Cylinder Wake

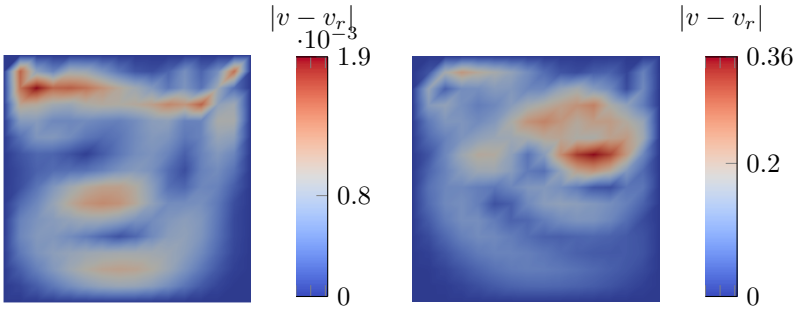
As a second test case, we consider the cylinder wake. The continuous model equations are the same as (32) but with a spatial domain Ω as illustrated in Figure 6 and different boundary conditions. Namely, we prescribe a parabolic inflow profile at the left boundary, *do nothing* boundary conditions for the outflow at the right boundary, and *no-slip* conditions, i.e., $\mathfrak{v} = 0$, elsewhere. An application of the *Taylor-Hood* finite element scheme gave a system of type (33) of state space dimensions $n_1 + n_2 = 5812 + 805 = 6617$.

Again, as the actuation, we add another instance of the source terms, accounting for the Dirichlet boundary conditions, scaled by the constant scalar value β to the system. The output was defined as in (34) with the domain of observation $\Omega_o = [0.6, 0.64] \times [0.18, 0.22]$.



(a) Full-order system. (b) ind2QBmm (c) genQBmm

Figure 3: Comparison of $|v|$ obtained from full and reduced-order for two-sided moment-matching at $t = 2s$ for the input $u(t) = e^{-t}(2 + \sin(2\pi t))$.



(a) ind2QBmm (b) genQBmm

Figure 4: Absolute error of velocity $|v - v_r|$ obtained from full and reduced-order systems for two-sided moment-matching at $t = 2s$.

Here $b_2 \neq 0$, because of the normal component of the parabolic inflow profile at the left boundary. So, the discretized system looks as follows:

$$E_{11}\dot{v}(t) = A_{11}v(t) + A_{12}p(t) + H_1 v(t) \otimes v(t) + f_1 + b_1\beta, \qquad v_0 = v_s - \mathcal{X}, \tag{36a}$$
$$0 = A_{21}v(t) + f_2 + b_2\beta, \tag{36b}$$
$$y(t) = C_1 v(t), \tag{36c}$$

where v_s is the steady-state solution and $\mathcal{X}$ accounts for $b_2\beta$ as defined below. Analogously to the previous example, here we also consider the system for the deviation (v_δ, p_δ) from the steady-state solution (v_s, p_s) with $\beta = 0$ which is given as follows:

$$E_{11}\dot{v}_\delta(t) = \mathcal{A}_{11}v_\delta(t) + A_{12}p_\delta(t) + H_1 v_\delta(t) \otimes v_\delta(t) + b_1\beta, \qquad v_\delta(0) = -\mathcal{X}, \tag{37a}$$
$$0 = A_{21}v_\delta(t) + b_2\beta, \tag{37b}$$
$$y(t) = c_2 p_\delta(t), \tag{37c}$$

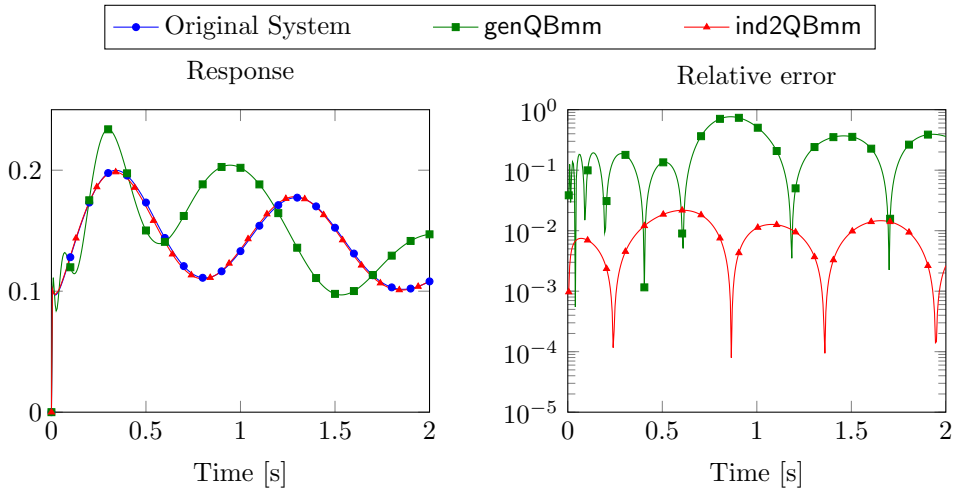


Figure 5: Comparison of reduced-order systems obtained from the two implementations for one-sided moment-matching for the input $u(t) = 0.5e^{-t}(2 + \sin(2\pi t))$.

where $\mathcal{A}_{11} = A_{11} + H(v_s \otimes I + I \otimes v_s)$. As discussed in Section 4, the $b_2 \neq 0$ case can be appropriately transformed into $b_2 = 0$ by substituting $v_\delta(t) = v_0(t) - \mathcal{X}$ where $\mathcal{X} = (E_{11}^{-1}A_{12})(A_{12}^T E_{11}^T A_{12})^{-1} b_2 \beta$. This results in the following equivalent system

$$E_{11}\dot{v}_0(t) = \tilde{A}_{11}v_0(t) + A_{12}p(t) + H_1 v_0(t) \otimes v_0(t) + \tilde{B}, \qquad v_0(0) = 0, \tag{38a}$$

$$0 = A_{21}v_0(t), \tag{38b}$$

$$y(t) = c_2 p_\delta(t), \tag{38c}$$

where $\tilde{A}_{11} = \mathcal{A}_{11} - H(I \otimes \mathcal{X} + \mathcal{X} \otimes I)$ and $\tilde{B} = b_1\beta + H(\mathcal{X} \otimes \mathcal{X}) - \hat{A}_{11}\mathcal{X}$. Note that $v_0(0) = v_\delta(0) + \mathcal{X} = 0$. We set the number of interpolation points to 15 and identify their location by employing *IRKA* on the linear part of the quadratic-bilinear DAE, cf. [16]. This gives us a reduced-order systems of order $r = 30$ using both methods. For these settings, we compare the time-domain simulations for two different Reynolds numbers.

Reynolds number Re=10

For small Reynolds number (Re = 10), we determine the reduced-order systems using both approaches for one-sided projection (by setting $\mathcal{W} = \mathcal{V}$). We plot the time-domain simulations obtained with both reduced-order systems in Figure 7 for $\beta = 0.5$. We observe that the reduced-order system, determined by the Stokes-type specific approach ind2QBmmwith one-sided projections, replicates the input-output dynamics of the original system very well, whereas the general implementation genQBmm fails to do so. For this example, we were not able to get stable reduced-order systems in case of two-sided projections for any of both approaches.

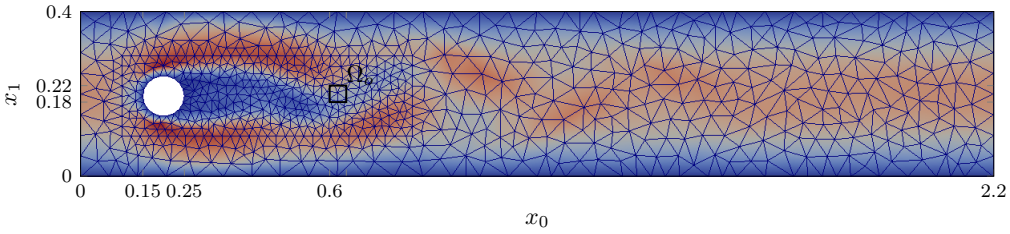


Figure 6: Illustration of the cylinder wake for $\nu = 10^{-1}$, the grid used for the *Taylor-Hood* discretization, and the domain of observation Ω_o.

For Reynolds number Re= 100

Analogously, we determine reduced-order systems for higher Reynolds number (Re= 100) and compare the time-domain simulations of the reduced-order system obtained via both approaches in Figure 8 for $\beta = 0.1$. Again, we observe that ind2QBmm outperforms the general approach genQBmm. However, we also observed that as β increases, for one-sided moment-matching, also the specific approach ind2QBmm fails to replicate the input-output behavior of the original system. A visual comparison is performed through plotting the velocity approximations computed via the original system and the reduced-order system obtained by the specific approach ind2QBmm (Figure 9) as well as the absolute error (Figure 10) on the full grid. One can see, that the reduced-order system also captures the dynamics on the full grid quite accurately.

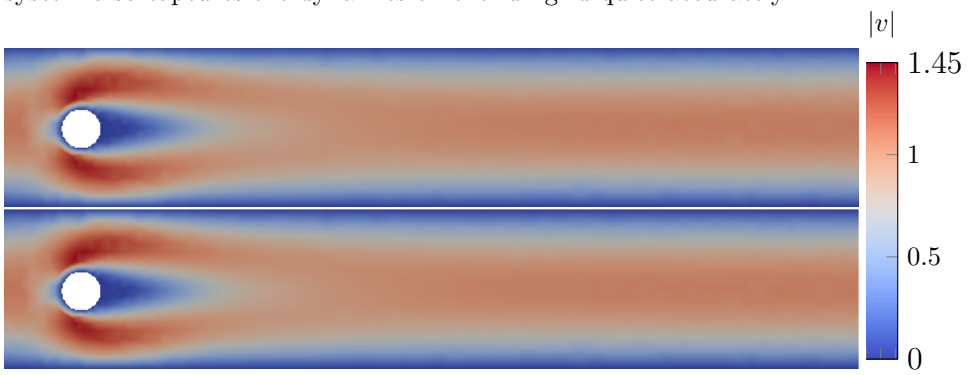


Figure 9: Comparison of $|v|$ obtained from full-order (top) and reduced-order (below) models on the full grid.

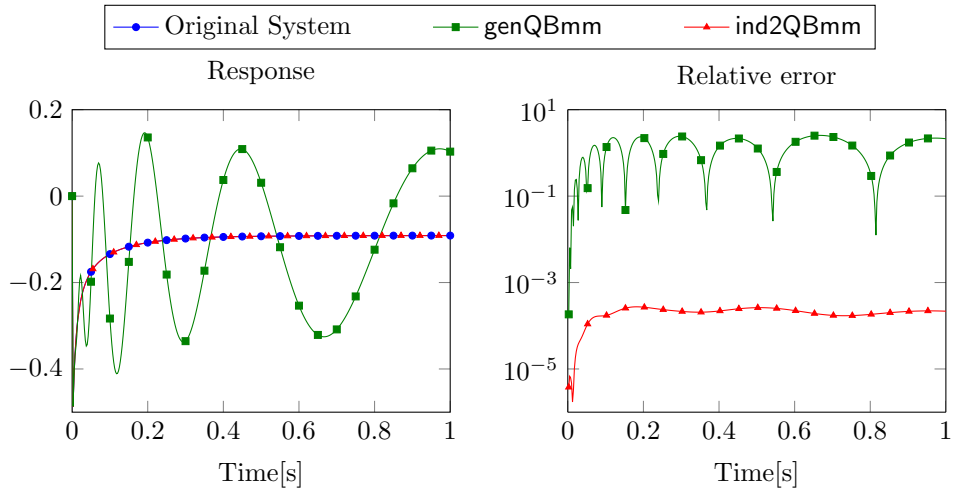


Figure 7: Comparison of the reduced-order systems obtained through the general approach genQBmm and through the specific ind2QBmm implementation for one-sided moment-matching.

6 Conclusions

We have proposed a two-sided moment-matching method for a special class of SISO quadratic-bilinear descriptor systems. We have applied two-sided moment-matching to an equivalent quadratic-bilinear ODE, so that a growing unbounded error due to the systems differential-algebraic nature will not occur. In view of efficient implementation, we have provided an algorithm that avoids the explicit computation of the projectors used for decoupling the DAEs into ODEs and purely algebraic equations. For the example of semi-discretized Navier-Stokes equation, we have shown the efficiency of the proposed method by comparing it to the approach to reduce quadratic-bilinear systems that was taken in [6]. For both two-sided and one-sided moment-matching, we could report significant improvements in the approximations by our proposed approach.

As we have observed in the numerical results, the two-sided moment-matching does not guarantee the stability of the reduced-order systems. Therefore, as a future avenue it is very important to address the stability of reduced-order systems obtained via moment-matching. Moreover, in our tests, we have chosen the interpolation points as they are obtained by applying *IRKA* [16] to the linear part that fulfill the optimality properties for linear systems. It is still an open question how to choose the optimal interpolation points which minimize the error system in some measure for quadratic-bilinear systems. Also, an interesting follow-up is the extension of the two-sided moment-matching to the MIMO case.

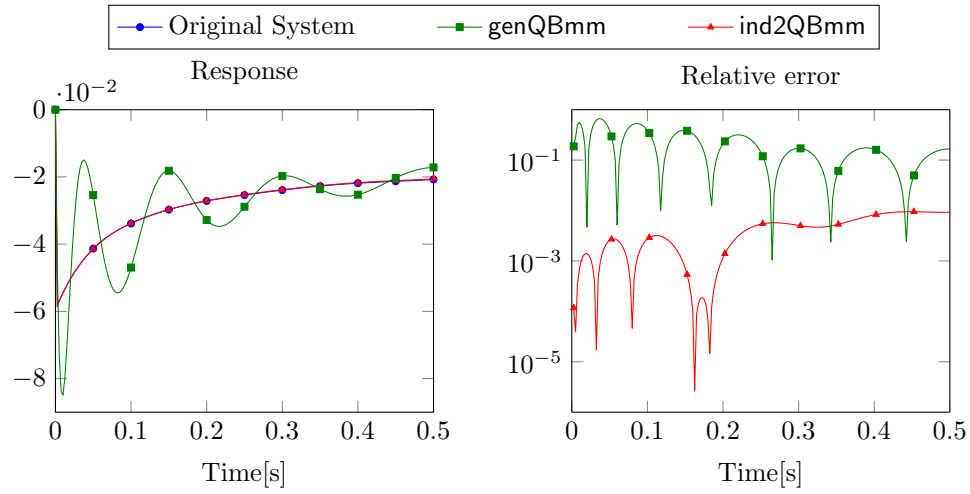


Figure 8: Comparison of reduced-order systems obtained from both implementations for one-sided moment-matching.

Figure 10: Absolute error in $|v|$ obtained from the full-order and reduced-order systems on the full grid.

References

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

[2] P. Astrid, S. Weiland, K. Willcox, and T. Backx. Missing point estimation in models described by proper orthogonal decomposition. *IEEE Trans. Automat. Control*, 53(10):2237–2251, 2008.

[3] E. Bänsch, P. Benner, J. Saak, and H. K. Weichelt. Riccati-based boundary feedback stabilization of incompressible Navier–Stokes flows. *SIAM J. Sci. Comput.*, 37(2):A832–A858, 2015.

[4] P. Benner and T. Breiten. Krylov-subspace based model reduction of nonlinear circuit models using bilinear and quadratic-linear approximations. In M. Günther, A. Bartel, M. Brunk, S. Schöps, and M. Striebel, editors, *Progress in Industrial*

Mathematics at ECMI 2010, Mathematics in Industry, volume 17, pages 153–159. Springer-Verlag, Berlin, 2012.

[5] P. Benner and T. Breiten. Two-sided projection methods for nonlinear model order reduction. Preprint MPIMD/12-12, Max Planck Institute Magdeburg, 2012. Available from `http://www.mpi-magdeburg.mpg.de/preprints/`.

[6] P. Benner and T. Breiten. Two-sided projection methods for nonlinear model order reduction. *SIAM J. Sci. Comput.*, 37(2):B239–B260, 2015.

[7] P. Benner and J. Heiland. LQG-balanced truncation low-order controller for stabilization of laminar flows. In R. King, editor, *Active Flow and Combustion Control 2014*, volume 127 of *Notes on Numerical Fluid Mechanics and Multidisciplinary Design*, pages 365–379. Springer International Publishing, 2015.

[8] P. Benner, V. Mehrmann, and D. C. Sorensen. *Dimension Reduction of Large-Scale Systems*, volume 45 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin/Heidelberg, Germany, 2005.

[9] T. Bui-Thanh, M. Damodaran, and K. Willcox. Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition. *AIAA J.*, 42(8):1505–1516, 2004.

[10] W. Cazemier, R.W.C.P. Verstappen, and A.E.P. Veldman. Proper orthogonal decomposition and low-dimensional models for driven cavity flows. *Phy. of Fluids*, 10(7):1685–1699, 1998.

[11] S. Chaturantabut and D. C. Sorensen. Nonlinear model reduction via discrete empirical interpolation. *SIAM J. Sci. Comput.*, 32(5):2737–2764, 2010.

[12] P. Goyal, M. I. Ahmad, and P. Benner. Krylov subspace-based model reduction for a class of bilinear descriptor systems. Preprint MPIMD/15-07, Max Planck Institute Magdeburg, 2015. Available from `http://www.mpi-magdeburg.mpg.de/preprints/`.

[13] P. Goyal, M. I. Ahmad, and P. Benner. Model reduction of quadratic-bilinear descriptor systems via Carleman bilinearization. In *Proceeding in European Control Conference*, pages 1171–1176, 2015.

[14] C. Gu. QLMOR: A projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. *IEEE Trans. Computer-Aided Design Integrated Circuits Systems*, 30(9):1307–1320, 2011.

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

[16] S. Gugercin, T. Stykel, and S. Wyatt. Model reduction of descriptor systems by interpolatory projection methods. *SIAM J. Sci. Comput.*, 35(5):B1010–B1033, 2013.

[17] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II - Stiff and Differential-Algebraic Problems*. Springer Series in Computational Mathematics. Springer, second edition, 2002.

[18] J. Heiland. *Decoupling and Optimization of Differential-Algebraic Equations with Application in Flow Control*. PhD thesis, TU Berlin, 2014.

[19] M. Heinkenschloss, D. C. Sorensen, and K. Sun. Balanced truncation model reduction for a class of descriptor systems with applications to the Oseen equations. *SIAM J. Sci. Comput.*, 30(2):1038–1063, 2008.

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

[21] K. Kunisch and S. Volkwein. Proper orthogonal decomposition for optimality systems. *ESAIM Math. Model. Numer. Anal.*, 42(1):1–23, 2008.

[22] M. J. Rewieński. *A Trajectory Piecewise-Linear Approach to Model Order Reduction of Nonlinear Dynamical Systems*. PhD Thesis, Massachusetts Institute of Technology, 2003.

[23] W. J. Rugh. *Nonlinear System Theory*. Johns Hopkins University Press, Baltimore, 1981.

[24] W. H. A. Schilders, H. A. Van der Vorst, and J. Rommes. *Model Order Reduction: Theory, Research Aspects and Applications*. Springer-Verlag, Berlin, Heidelberg, 2008.

Max Planck Institute Magdeburg Preprints