Multivariate moment matching for model order reduction of quadratic-bilinear systems using error bounds

Muhammad Altaf Khattak[a,1], Mian Ilyas Ahmad[a,1,*], Lihong Feng[b], Peter Benner[b]

[a] *Research Center for Modelling and Simulation, NUST H-12, Islamabad 44000 Pakistan*
[b] *Computational Methods for Systems and Control, Max Planck Institute Magdeburg, Sandtorstrasse 1, 39106 Germany*

Abstract

We propose an adaptive moment-matching framework for model order reduction of quadratic-bilinear descriptor systems. In this framework, an important issue is the selection of those shift frequencies where moment-matching is to be achieved. Often, the choice is random or linked to the linear part of the non-linear system. In this paper, we extend the use of an existing a posteriori error bound for general linear time invariant systems to quadratic-bilinear systems and develop a greedy-type framework to select a good choice of interpolation points for the construction of the projection matrices. The results are compared with standard quadratic-bilinear projection methods and we observe that the approximations obtained by the proposed method yield high accuracy.

Keywords: quadratic-bilinear systems, model order reduction, projection/moment matching, error bounds

[*] Corresponding author
Email address: m.ilyas@rcms.nust.edu.pk (Mian Ilyas Ahmad)
[1] This author is supported by HEC Pakistan under NRPU Project ID 10176.

1. Introduction

There are different applications where the dynamics of the system can be represented by quadratic-bilinear differential algebraic equations (QBDAEs). These include simulation of distribution networks [1], fluid flow problems [2] and nonlinear VLSI circuits [3, 4]. In addition, a large class of nonlinear systems can be written in quadratic-bilinear form by using exact transformations [4]. Most of these applications involve large number of equations, which make simulation, control and optimization computationally inefficient. A remedy to this issue is the use of model order reduction (MOR).

We consider the problem of MOR for a single-input single-output quadratic-bilinear descriptor system of the form:

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Nx(t)u(t) + Qx(t) \otimes x(t) + Bu(t), \\ y(t) &= Cx(t), \end{aligned} \tag{1}$$

where E, A, $N \in \mathbb{R}^{n\times n}$, $Q \in \mathbb{R}^{n\times n^2}$, B, $C^T \in \mathbb{R}^n$ are the coefficient matrices and vectors. $x(t) \in \mathbb{R}^n$ is the state vector and $u(t)$, $y(t) \in \mathbb{R}$ are the input and output of the system. The matrix E may or may not be singular but the pencil is assumed to be regular, i.e., $\lambda E - A$ is singular only for finitely many values of $\lambda \in \mathbb{C}$ [5].

The goal of MOR is to construct a reduced system of dimension $r \ll n$:

$$\begin{aligned} E_r\dot{x}_r(t) &= A_r x_r(t) + N_r x_r(t)u(t) + Q_r x_r(t) \otimes x_r(t) + B_r u(t), \\ y_r(t) &= C_r x_r(t), \end{aligned} \tag{2}$$

with the output response $y_r(t)$ approximately equal to $y(t)$. In case of linear systems (where Q and N are null matrices), there are various techniques in the literature to compute reduced-order models (ROMs), cf., [6, 7]. Among these methods, projection-based moment-matching methods [8, 9] are well used and are recently extended to quadratic-bilinear systems [4, 10, 11]. Projection involves approximating the state vector $x(t)$ in an r-dimensional subspace spanned by the column vectors of $V \in \mathbb{R}^{n\times r}$, so that the residual in the state equation is orthogonal to another r-dimensional subspace spanned by the column vectors of $W \in \mathbb{R}^{n\times r}$. That is, we approximate $x(t) \approx Vx_r(t)$ such that the Petrov-Galerkin orthogonality condition holds:

$$W^T\Big(EV\dot{x}_r(t) - \big(AVx_r(t) + NVx_r(t)u(t) + QVx_r(t)\otimes Vx_r(t) + Bu(t)\big)\Big) = 0,$$
$$\hat{y}(t) = CVx_r(t). \tag{3}$$

If $W = V$, the projection is orthogonal and is often called one-sided projection, otherwise it is oblique and is called two-sided projection. The oblique projection framework leads to a set of reduced system matrices of the form:

$$\begin{gathered} E_r = W^TEV, \quad A_r = W^TAV, \quad Q_r = W^TQ(V\otimes V), \quad N_r = W^TNV, \\ B_r = W^TB, \quad C_r = CV. \end{gathered} \tag{4}$$

In case of linear systems, a suitable choice of the basis matrices V and W, implicitly ensure moment-matching, where moments are the coefficients of the series expansion of the transfer function at some predefined shift frequencies. Thus for projection-based moment-matching, the choice of V and W is related to the transfer function of the system. However, nonlinear systems have no universal input-output representation though for some classes of nonlinear systems, including the QBDAE system, it is possible to generalise the transfer function concept by utilising the Volterra theory [12], where the input-output relationship is represented by a set of high-order transfer functions. This makes the concept of moment-matching slightly complex in the nonlinear case, since the structure of the basis matrices V and W in (4) now depends on multiple high-order transfer functions. To achieve moment-matching, some simplifications are made in the literature [4, 10] for computing the ROMs. For example, [10] constructs V and W such that the reduced system matches the moments of the first- and second-order transfer functions. In [11], simplified forms of high-order transfer functions are derived, which also enable the projection based techniques to match moments of high-order transfer functions. In addition, all the existing moment-matching/interpolation approaches [4, 10, 11] are based on the simplification that the interpolation points for each frequency variable is the same. We discuss these results further in Section 2.

Recently a new framework [13] for quadratic-bilinear systems has been proposed that is based on generalized Sylvester-type matrix equations. The approach involves truncated solution of two complex matrix equations to identify a good choice for the basis matrices V and W. Another approach is the extension of the Loewner framework from linear/bilinear systems [14, 15] to quadratic bilinear systems [16] . Also an indirect approach for MOR of the QBDAE system is proposed in [17], where the basis matrices are constructed from the bilinear part of the quadratic-bilinear system. In [18], the linear-bilinear part of the system is viewed as a linear parametric system and a posteriori error bound is used to select the interpolation points and construct the basis matrices adaptively. All these techniques are using the first two or three high-order transfer functions and their structure is different from the one identified in [10]. Since our target is moment-matching for QBDAEs, we will mainly focus on the two-sided moment-matching technique of Benner and Breiten [10].

In this paper, we identify a good choice of interpolation points for the quadratic-bilinear system by utilizing a greedy type framework based on error bounds for quadratic-bilinear systems motivated by the recently proposed error bound for linear parametric systems in [19]. Here we relax the restriction of using the same interpolation points for different frequency variables. The approach starts from some initial interpolation points that are iteratively updated to identify a set of interpolation points corresponding to the maximal values of certain error bounds. For each choice of interpolation points, we interpolate, not only, the original transfer function and its first derivative but also higher derivatives, so that the quadratic-bilinear system is well approximated. The iterations stops when the approximation error is less than the prescribed tolerance level. Each iteration contributes to constructing a better set of basis matrices V and W, until a given error tolerance is achieved. The main difference from the work in [18] is that the quadratic part of the system is also involved in basis construction in the proposed framework based on a posteriori error bound for quadratic-bilinear systems, whereas only the bilinear part is considered for the basis matrix computation in [18]. The error estimator used in [18] only estimates the error of the linear-bilinear part.

The remaining part of the paper is organized as follows. Section 2 reviews the existing projection based moment-matching techniques for quadratic-bilinear systems. Section 3 presents the error bound expressions for quadratic bilinear systems and Section 4 utilises these error bounds in a greedy-type algorithmto select interpolation points. Finally in Section 5, numerical results are shown for some benchmark examples.

2. Background

In this section, we briefly review the concept of moment-matching discussed in [10, 11] for quadratic-bilinear systems. Before going into the details of nonlinear moment-matching, we begin with the structure of high-order transfer functions.

2.1. Multivariate Transfer Functions

The input-output representation for single input quadratic-bilinear systems can be expressed by the Volterra series expansion of the output $y(t)$ with quantities analogous to the standard convolution operator. That is,

$$y(t) = \sum_{k=1}^{\infty} \int_0^t \int_0^{t_1} \cdots \int_0^{t_{k-1}} h_k(t_1, \ldots, t_k) u(t - t_1) \cdots u(t - t_k) dt_k \cdots dt_1, \tag{5}$$

where it is assumed that the input signal is one-sided, i.e., $u(t) = 0$ for $t < 0$. In addition, each of the generalized impulse responses, $h_k(t_1, \ldots, t_k)$, also called the k-dimensional kernel of the subsystem, is assumed to be one-sided. In terms of the multivariate Laplace transform, the k-dimensional subsystem can be represented as,

$$Y_k(s_1, \ldots, s_k) = H_k(s_1, \ldots, s_k) U(s_1) \cdots U(s_k), \tag{6}$$

where $H_k(s_1, \ldots, s_k)$ is the multivariate transfer function of the k-dimensional subsystem. The generalized transfer functions in the output expression (6) are in the so-called triangular form [12]. We denote the k-dimensional triangular form by $H_{tri}^{[k]}(s_1, \dots, s_k)$. There are some other useful forms such as the symmetric form and the regular form of the multivariate transfer functions as discussed in [12]. The triangular form is related to the symmetric form by the following expression

$$H_{sym}^{[k]}(s_1, \dots, s_k) = \frac{1}{n!} \sum_{\pi(\cdot)} H_{tri}^{[k]}(s_{\pi(1)}, \dots, s_{\pi(k)}), \tag{7}$$

where the summation includes all $k!$ permutations of $s_1, \dots, s_k$. Also, the triangular form can be connected to the regular form of the transfer function by using

$$H_{tri}^{[k]}(s_1, \dots, s_k) = H_{reg}^{[k]}(s_1, s_1 + s_2, \dots, s_1 + s_2 + \dots + s_k). \tag{8}$$

According to [12], the structure of the generalized symmetric transfer functions can be identified by the growing exponential approach. The structure of these symmetric transfer functions for the first two subsystems of the quadratic-bilinear system (1) can be written as

$$\begin{aligned} H_1(s_1) &= C(s_1 E - A)^{-1} B, \\ H_2(s_1, s_2) &= C((s_1 + s_2)E - A)^{-1} B(s_1, s_2), \end{aligned} \tag{9}$$

here

$$B(s_1, s_2) =: Q(x_1(s_1) \otimes x_1(s_2)) + \frac{1}{2!} N(x_1(s_1) + x_1(s_2)), \tag{10}$$

in which $x_1(s) := (sE - A)^{-1} B$ and Q satisfies $Q(u \otimes v) = Q(v \otimes u)$ for all $u, v \in \mathbb{R}^n$. Defining $x_2(s_1, s_2) := ((s_1 + s_2)E - A)^{-1} B(s_1, s_2)$, the first two (first- and second-order) symmetric transfer functions can be written as

$$\begin{aligned} H_1(s_1) &= C x_1(s_1), \\ H_2(s_1, s_2) &= C x_2(s_1, s_2). \end{aligned} \tag{11}$$

Before going into the partial differentiation of these multivariate transfer functions, we introduce the concept of matricization. The process of reshaping a tensor into a matrix is called matricization. In [10], the matrix $Q \in \mathbb{R}^{n \times n^2}$ is considered as the mode-1 matricization of a 3 dimensional tensor $\mathcal{Q} \in \mathbb{R}^{n \times n \times n}$.

The $n \times n$ components of Q are the frontal slices $\mathcal{Q}_i \in \mathbb{R}^{n\times n}$ of the tensor $\mathcal{Q}$, i.e. $Q = \begin{bmatrix} \mathcal{Q}_1 & \cdots & \mathcal{Q}_n \end{bmatrix}$. The mode-2 and mode-3 matricization can be defined as

$$\begin{aligned} Q^{(2)} &= \begin{bmatrix} \mathcal{Q}_1^T & \cdots & \mathcal{Q}_n^T \end{bmatrix}, \\ Q^{(3)} &= \begin{bmatrix} vec(\mathcal{Q}_1) & \cdots & vec(\mathcal{Q}_n) \end{bmatrix}^T. \end{aligned}$$

It is observed that the following property holds

$$w^T Q(u \otimes v) = u^T Q^{(2)}(v \otimes w), \tag{12}$$

where $w, u, v \in \mathbb{R}^n$ are arbitrary and Q is symmetric in the sense that $Q(u\otimes v) = Q(v \otimes u)$, see [20]. Let $G(s) := sE - A$, then by using

$$\frac{\partial G(s)^{-1}}{\partial s} = -G(s)^{-1}\frac{\partial G(s)}{\partial s}G(s)^{-1},$$

and (12), we have

$$\begin{aligned} \frac{\partial H_2(s_1,s_2)}{\partial s_1} &= -y_1(s_1+s_2)^T E x_2(s_1,s_2) \\ &\quad - x_1(s_1)^T E^T y_2(s_1,s_2) \end{aligned} \tag{13}$$

where $y_1(s) := (sE-A)^{-T}C^T$ and $y_2(s_1,s_2) := (s_1E-A)^{-T}C(s_1,s_2)^T$ in which

$$C(s_1,s_2) = Q^{(2)}\big(x_1(s_2) \otimes y_1(s_1+s_2)\big) + \frac{1}{2!}N^T y_1(s_1+s_2)$$

Similarly

$$\begin{aligned} \frac{\partial H_2(s_1,s_2)}{\partial s_2} &= -y_1(s_1+s_2)^T E x_2(s_1,s_2) \\ &\quad - x_1(s_2)^T E^T y_2(s_2,s_1) \end{aligned} \tag{14}$$

Notice that when $s_1 = s_2 = \sigma$, the two partial differentiations are the same. This condition on interpolation points is assumed in [10] to show the moment-matching properties of the ROM. In the following, we show moment-matching in the multivariate settings where $s_1 \neq s_2$ ($s_1 = \sigma_{1i}$ and $s_2 = \sigma_{2i}$).

2.2. Moment-Matching for QBDAE

The goal of a moment-matching based reduction approach is to ensure that the high-order transfer functions are well approximated. In case of symmetric transfer functions, we can represent it as

$$H_k(s_1,\ldots,s_k) \approx \hat{H}_k(s_1,\ldots,s_k), \quad \text{for } k = 1,\ldots,K, \tag{15}$$

with $\hat{H}_k(s_1, \ldots, s_k)$ being the k-th order multivariate transfer function of the reduced system (2). With the task in (15) achieved for some K, we can expect that the output $y(t)$ is well approximated by $\hat{y}(t)$. To get recursive relations between vectors for approximation subspaces, it is assumed in [10] that $s_1 = s_2 = \sigma$. With these settings, the second-order transfer function becomes

$$H_2(\sigma, \sigma) = y(2\sigma)^T \Big(Q \left(x_1(\sigma) \otimes x_1(\sigma) \right) + N x_1(\sigma) \Big).$$

The following Lemma summarizes the result introduced in [10].

Lemma 1. *Let $\sigma_i \in \mathbb{C}$ be the interpolation points and $\sigma_i \notin \{\Lambda(A, E), \Lambda(A_r, E_r)\}$, where $\Lambda(A, E)$ represents the generalized eigenvalues of the matrix pencil $\lambda E - A$. Assume that $\hat{E} = W^T E V$ is nonsingular and $\hat{A}$, $\hat{Q}$, $\hat{N}$, $\hat{B}$, $\hat{C}$ are as in* (4) *with full rank matrices $V, W \in \mathbb{R}^{n \times r}$ such that*

$$\begin{aligned} \operatorname{span}(V) &= \operatorname*{span}_{i=1,\ldots,k} \{x_1(\sigma_i),\ x_2(\sigma_i, \sigma_i)\}, \\ \operatorname{span}(W) &= \operatorname*{span}_{i=1,\ldots,k} \{y_1(2\sigma_i),\ y_2(\sigma_i, \sigma_i)\}, \end{aligned}$$

then the reduced QBDAE satisfies the following (Hermite) interpolation conditions:

$$\begin{aligned} H_1(\sigma_i) &= \hat{H}_1(\sigma_i), & H_1(2\sigma_i) &= \hat{H}_1(2\sigma_i), \\ H_2(\sigma_i, \sigma_i) &= \hat{H}_2(\sigma_i, \sigma_i), & \frac{\partial}{\partial s_j} H_2(\sigma_i, \sigma_i) &= \frac{\partial}{\partial s_j} \hat{H}_2(\sigma_i, \sigma_i), \quad j = 1, 2. \end{aligned}$$

See [10] for a proof. Next, we present moment-matching properties in the multivariable settings, where $s_1 \neq s_2$.

Lemma 2. *Let $\sigma_{1i}, \sigma_{2i} \in \mathbb{C}$ with $\sigma_{1i}, \sigma_{2i} \notin \{\Lambda(A, E), \Lambda(A_r, E_r)\}$. Assume that $\hat{E} = W^T E V$ is nonsingular and $\hat{A}$, $\hat{Q}$, $\hat{N}$, $\hat{B}$, $\hat{C}$ are as in* (4) *with full rank matrices $V, W \in \mathbb{R}^{n \times r}$ such that*

$$\begin{aligned} \operatorname{span}(V) &= \operatorname*{span}_{i=1,\ldots,k} \{x_1(\sigma_{1i}),\ x_1(\sigma_{2i}),\ x_2(\sigma_{1i}, \sigma_{2i})\} \\ \operatorname{span}(W) &= \operatorname*{span}_{i=1,\ldots,k} \{y_1(\sigma_{1i} + \sigma_{2i}),\ y_2(\sigma_{1i}, \sigma_{2i}),\ y_2(\sigma_{2i}, \sigma_{1i})\}. \end{aligned}$$

Then the reduced QBDAE satisfies the following (Hermite) interpolation conditions:

$$H_1(\sigma_{1i}) = \hat{H}_1(\sigma_{1i}), \quad H_1(\sigma_{2i}) = \hat{H}_1(\sigma_{2i}), \quad H_1(\sigma_{1i}+\sigma_{2i}) = \hat{H}_1(\sigma_{1i}+\sigma_{2i}),$$
$$H_2(\sigma_{1i},\sigma_{2i}) = \hat{H}_2(\sigma_{1i},\sigma_{2i}), \quad \frac{\partial}{\partial s_1}H_2(\sigma_{1i},\sigma_{2i}) = \frac{\partial}{\partial s_1}\hat{H}_2(\sigma_{1i},\sigma_{2i}),$$
$$\frac{\partial}{\partial s_2}H_2(\sigma_{2i},\sigma_{1i}) = \frac{\partial}{\partial s_2}\hat{H}_2(\sigma_{2i},\sigma_{1i}).$$

The proof of the statement is similar to Lemma 1 and therefore omitted. Note that the statement in Lemma 2 reduces to Lemma 1, if $\sigma_{1i} = \sigma_{2i}$. In the remaining part of the paper, our goal is to identify a good choice of the interpolation points σ_{1i} and σ_{2i}.

3. Error Bound for QBDAE's

In this section, we show how the error bound expression, derived initially in [19] for parametric linear time invariant systems, can be extended to the quadratic-bilinear DAEs. We begin with a brief overview of the error bound for the first subsystem, as in [19] and then discuss the extension to the second subsystem of QBDAE.

3.1. Error bound for $H_1(s_1)$

Here the error bound provides an estimate for the error between $H_1(s_1)$ and $\hat{H}_1(s_1)$. To this end, we define the primal and the dual systems as:

$$(s_1E - A)x_1(s_1) = B, \tag{16}$$
$$(s_1E - A)^T x_1^{du}(s_1) = -C^T, \tag{17}$$

respectively, where T denotes transpose of the matrix. The error bound is constructed so that it is based on two residuals, which result from MOR of the primal and the dual system, respectively. The primal system is reduced using the matrix pair V_1 and W_1, where

$$\operatorname{span}(V_1) = \operatorname*{span}_{i=1,\dots,k}\{x_1(\sigma_{1i})\}, \quad \operatorname{span}(W_1) = \operatorname*{span}_{i=1,\dots,k}\{x_1^{du}(\sigma_{1i})\}. \tag{18}$$

As a result, the reduced primal system is,

$$(s_1\hat{E}_1 - \hat{A}_1)z_1(s_1) = \hat{B},$$

where $\hat{E}_1 = W_1^T E V_1$, $\hat{A}_1 = W_1^T A V_1$, $\hat{B}_1 = W_1^T B$ and $\hat{C}_1 = CV_1$. Here $\hat{x}_1(s_1) := V_1 z_1(s_1)$ is the approximation of $x_1(s_1)$. Due to the dual relation between (16) and (17), the dual system can be reduced by using $V_1^{du} = W_1$ and $W_1^{du} = V_1$. The reduced dual system is

$$(s_1\tilde{E}_1 - \tilde{A}_1)^T z_1^{du}(s_1) = -\tilde{C}_1^T,$$

where $\tilde{E}_1 = V_1^T E W_1$, $\tilde{A}_1 = V_1^T A W_1$, $\tilde{C}_1 = W_1^T C^T$. Also $\tilde{x}_1^{du}(s_1) := W_1 z_1^{du}(s_1)$ is the approximation of $x_1^{du}(s_1)$. The residuals associated with the reduction of the primal and the dual systems can be written as

$$\begin{aligned} r_1^{pr}(s_1) &= B - (s_1E - A)V_1 z_1(s_1), \\ r_1^{du}(s_1) &= -C^T - (s_1E - A)^T W_1 z_1^{du}(s_1). \end{aligned} \tag{19}$$

With these quantities, the following result provides an a posteriori upper bound on the approximation error, $|H_1(s_1) - \hat{H}_1(s_1)|$:

Theorem 1. *[19] The upper bound on the approximation of the transfer function $H_1(s_1) = C(s_1E - A)^{-1}B$ can be written as $|H_1(s_1) - \hat{H}_1(s_1)| = \Delta_1(s_1)$, where*

$$\Delta_1(s_1) := \frac{\|r_1^{du}(s_1)\|_2 \|r_1^{pr}(s_1)\|_2}{\beta_1(s_1)}, \tag{20}$$

in which $\beta_1(s_1) = \sigma_{\min}(G(s_1))$, where $\sigma_{\min}$ indicates the smallest singular value of $G(s_1)$.

3.2. Error Bound for $H_2(s_1, s_2)$

Analogous to $H_1(s_1)$, we define the primal and dual systems as:

$$G(s_1 + s_2)x_2(s_1, s_2) = B(s_1, s_2), \tag{21}$$

$$G^T(s_1 + s_2)x_2^{du}(s_1, s_2) = -C^T, \tag{22}$$

respectively. The interpolation points for $H_1(s_1)$ can be identified through the error bound $\Delta_1(s_1)$ by using a greedy framework as presented in [19]. This means that we can select σ_{1i} for $i = 1, \dots, r$ as the interpolation points corresponding to the maximal values of the error bound at subsequent iterations of the greedy algorithm in [19].With these interpolation points fixed for s_1, we can also express error bound for the second subsystem. The error bound is constructed based on two residuals, which result from MOR of the primal and the dual systems in (21) (22), respectively. The primal system is reduced using the matrix pair V_2 and W_2, where

$$\operatorname{span}(V_2) = \operatorname*{span}_{j=1,\dots,k} \{x_2(\sigma_{1i}, \sigma_{2j})\}, \quad \operatorname{span}(W_2) = \operatorname*{span}_{j=1,\dots,k} \{x_2^{du}(\sigma_{1i}, \sigma_{2j})\}. \tag{23}$$

As a result, the reduced primal system is

$$((s_1 + s_2)\hat{E}_2 - \hat{A}_2)z_2(s_1, s_2) = \hat{B}(s_1, s_2),$$

where $\hat{E}_2 = W_2^T E V_2$, $\hat{A}_2 = W_2^T A V_2$, $\hat{B}(s_1, s_2) = W_2^T B(s_1, s_2)$ and $\hat{C}_2 = C V_2$. Similarly, the dual system is reduced using the matrix pair V_2^{du} and W_2^{du},

$$\operatorname{span}(V_2^{du}) = \operatorname*{span}_{i=1,\dots,k} \{x_2^{du}(\sigma_{1i}, \sigma_{2i})\}, \quad \operatorname{span}(W_2^{du}) = \operatorname*{span}_{i=1,\dots,k} \{x_2(\sigma_{1i}, \sigma_{2i})\}. \tag{24}$$

The reduced dual system is

$$((s_1 + s_2)\tilde{E}_2 - \tilde{A}_2)^T z_2^{du}(s_1, s_2) = -\tilde{C}_2^T,$$

where $\tilde{E}_2 = (W_2^{du})^T E V_2^{du}$, $\tilde{A}_2 = (W_2^{du})^T A V_2^{du}$, $\tilde{C}_2^T = (V_2^{du})^T C^T$. The residuals associated with the reduction of the primal and dual systems can be written as

$$\begin{aligned} r_2^{pr}(s_1, s_2) &= B(s_1, s_2) - ((s_1 + s_2)E - A)V_2 z_2(s_1, s_2), \\ r_2^{du}(s_1, s_2) &= -C^T - ((s_1 + s_2)E - A)^T V_2^{du} z_2^{du}(s_1, s_2). \end{aligned} \tag{25}$$

With these quantities, the following result provides an a posteriori upper bound on the approximation error, $|H_2(s_1, s_2) - \hat{H}_2(s_1, s_2)|$:

Theorem 2. *The upper bound on the approximation of $H_2(s_1, s_2) = C((s_1 + s_2)E - A)^{-1}B(s_1, s_2)$ can be written as $|H_2(s_1, s_2) - \hat{H}_2(s_1, s_2)| = \Delta_2(s_1, s_2)$, where*

$$\Delta_2(s_1, s_2) := \frac{\|r_2^{du}(s_1, s_2)\|_2 \|r_2^{pr}(s_1, s_2)\|_2}{\beta_2(s_1, s_2)}, \tag{26}$$

in which $\beta_2(s_1, s_2) = \sigma_{\min}(G(s_1{+}s_2))$*, where* $\sigma_{\min}$ *indicates the smallest singular value of* $G(s_1 + s_2) = (s_1 + s_2)E - A$*.*

The proof is similar to Theorem 1 and therefore is omitted.

4. Interpolation Points using Error Bounds

As discussed in Section 2, the projection matrices V and W defined in Lemma 2 require a good choice of interpolation points σ_{1i} and σ_{2i} which also serve as interpolation points for MOR of the primal and dual systems in (16)-(17) and (21)-(22). In this section, we show the use of the error bound expressions derived previously to select the interpolation points.

The idea is to identify interpolation points corresponding to the maximal bound $\Delta_1(s_1)$. Assuming that σ_{1i} are the selected interpolation points for s_1, the remaining interpolation points for s_2 correspond to the maximal bound $\Delta_2(\sigma_{1i}, s_2)$ for each value of σ_{1i}. In this way, the error bound can be used iteratively to select a good choice of interpolation points in a predefined sample space, starting from an initial choice of sigma's. The selected interpolation points are then used to construct and update the required basis matrices V and W, by using the multimoment-matching technique described before. It is interesting to see that although we need to construct the ROMs for the primal and the dual systems in (16)-(17) and (21)-(22), the projection matrices for those ROMs are obtained without extra computations, since V_1, W_1 and V_2, W_2 are part of V, W by definition. Therefore, V, W can be obtained by orthogonalizing V_1 with V_2 and W_1 with W_2 as indicated in Step 9 of Algorithm 1, where a greedy framework for selecting interpolation points is presented. For an initial pair of interpolation points, the ROMs of the primal and the dual systems in (16)-(17) and (21)-(22) are constructed and the error bounds Δ_1, Δ_2 are computed. A new pair is selected such that the corresponding error bounds Δ_1 and Δ_2 are maximized at these points. With the selected interpolation points, we enrich the projection matrices V, W for MOR of the original quadratic-bilinear system iteratively during the greedy algorithm. Finally, the reduced quadratic bilinear system is constructed using V, W that are derived upon convergence of Algorithm 1. Algorithm 1 stops when $\Delta := \Delta_1 + \Delta_2$ is below the tolerance ϵ_{tol}, where Δ includes the errors introduced by approximating the first and second transfer functions. Since the interpolation points are selected according to the error bounds Δ_1 and Δ_2, it is important that the error bounds dynamically reflect the decay of the true error with the iteration of the greedy algorithm. Ideally, the error bounds should be very close to the true error. Numerical tests in the next section show that the error bounds really control the true error robustly.

Algorithm 1 An adaptive framework for selection of interpolation points

Inputs: σ_{10}, σ_{20}, E, A, N, H, B, C and S_{sample}: a set of the samples of $\mu := (s_1, s_2)$, which covers the domain of the two frequency variables.

Outputs: μ, V and W

Initialization: $V = [\ ]$; $W = [\ ]$; $V_1 = [\ ]$; $W_1 = [\ ]$; $V_2 = [\ ]$; $W_2 = [\ ]$; $\epsilon = 1$; $i = -1$; $j = 0$; $\epsilon_{tol} < 1$, $\mu^0 = (\sigma_{10}, \sigma_{20})$.

```
WHILE $\epsilon > \epsilon_{tol}$
    $i = i + 1$; $j = j + 1$;
    compute $\bar{V}_i(\sigma_{1i})$ and $\bar{W}_i(\sigma_{1i})$ using (18)
    $V_1 = orth[V_1, \bar{V}_i]$;  $W_1 = orth[W_1, \bar{W}_i]$;
    $\sigma_{1j} = \arg \max\limits_{\sigma_1 \in S_{1\ \text{sample}}} \Delta_1(\sigma_1)$;
    compute $V_i(\sigma_{1i}, \sigma_{2i})$ and $W_i(\sigma_{1i}, \sigma_{2i})$ using (23)
    $V_2 = orth[V_2, V_i]$;  $W_2 = orth[W_2, W_i]$;
    $\sigma_{2j} = \arg \max\limits_{\sigma_2 \in S_{2\ \text{sample}}} \Delta_2(\sigma_{1i}, \sigma_2)$;
    $\mu^j = [\sigma_{1j}, \sigma_{2j}]$;
    $V = orth[V_1, V_2]$;  $W = orth[W_1, W_2]$;
    $\Delta(\mu^j) := \Delta_1(\mu^j) + \Delta_2(\mu^j)$;   $\epsilon = \Delta(\mu^j)$;
END WHILE.
```

5. Numerical results

We consider three benchmark examples for our results on MOR of QBDAE systems. The results are compared with the one-sided and two-sided projection methods, where the interpolation points are computed by IRKA, implemented on the linear part of the system. We represented the proposed method by 1s/2s-greedy(One-sided/two-sided projection with greedy based interpolation points) and the method from literature by 1s/2s-IRKA (One-sided/two- sided projection with IRKA interpolation points). The Max. True error in the following tables is defined as $\max_{s_1,s_2 \in S_{\text{sample}}} |H_1(s_1) - \hat{H}_1(s_1)| + |H_2(s_1,s_2) - \hat{H}_2(s_1,s_2)|$ and the Max. error bound is $\max_{s_1,s_2 \in S_{\text{sample}}} \Delta(s_1,s_2)$.

5.1. Nonlinear RC circuit

The nonlinear RC circuit was first considered in [21] and since then it has been used in many papers for nonlinear MOR [5]. Consider v be the voltage and $g(v)$ be the current function then I-V characteristics can be represented as: $g(v) = e^{40v} + v - 1$. The nonlinearity in the current function results in nonlinear model. All the capacitances are fixed to $C = 1$. Figure 1 shows the complete circuit.

It is shown in [4] that the nonlinearity in the RC circuit can be written in the quadratic-bilinear form as in (1) by introducing some auxiliary variables. The transformation is exact, but the dimension of the system increases to $n = 2 \cdot l$, where l represents the number of nodes in Figure 1, and it is also the dimension of the original nonlinear system.

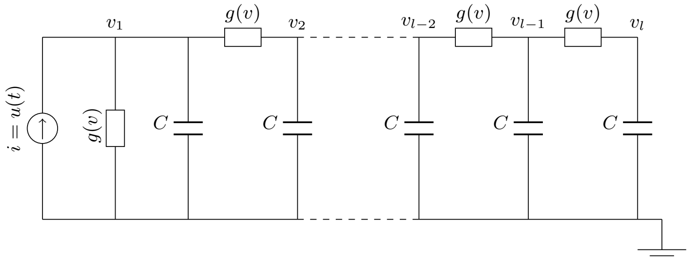


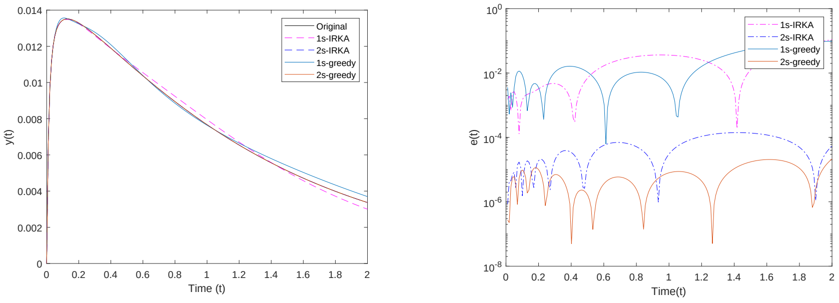
Figure 1: Nonlinear RC circuit

For our results, we set $l = 50$, so $n = 100$ and use two-sided projection method to reduce the system. Table 1 shows the results with tolerance $\epsilon_{tol} = 1e^{-5}$ and an initial choice of interpolation points as $\sigma_1 = \sigma_{20} = 119.5642$.

S.No.	Interpolation points	Max. True Error	Max. Est. Error
	$\{\sigma_{1i}, \sigma_{2i}\}$		
1	$119.5642, 119.5642$	1.8616×10^{-2}	0.109183
2	$0.9875, 0.9875$	1.3683×10^{-3}	8.4421×10^{-3}
3	$4.9567, 0.9875$	1.6127×10^{-4}	4.0341×10^{-4}
4	$18.1107, 5.5319$	4.2956×10^{-5}	7.22×10^{-5}
5	$2.0292, 4.4445$	8.239×10^{-6}	9.6404×10^{-6}

Table 1: Error estimation results for RC circuit

The second column of Table 1 shows interpolation points that are identified by the greedy framework and are based on the error bound. It is clear that the error bound tightly catches the true error and can be used as a surrogate of the true error to select the interpolation points. The size of the ROM obtained from both approaches has been kept the same i.e. $r_1 = r_2 = 12$. For the input $u(t) = e^{-t}$, the output of the original model and ROMs along with corresponding relative errors are shown in Figure 2.



(a) Comparison of transient response

(b) Comparison of relative error

Figure 2: Non-linear RC circuit

Figure 2a shows the comparison of transient response of the two approaches, while Figure 2b plots relative errors of the two approaches. It is clearly seen that 1s-greedy and 2s-greedy outperform 1s-IRKA and 2s-IRKA respectively in terms of accuracy.

5.2. Burgers' Equation

In nonlinear MOR, 1D burgers' equation is commonly used [2],[10]. Mathematical model of 1D burger's equation with $\Gamma = (0,1) \times (1,T)$ is:

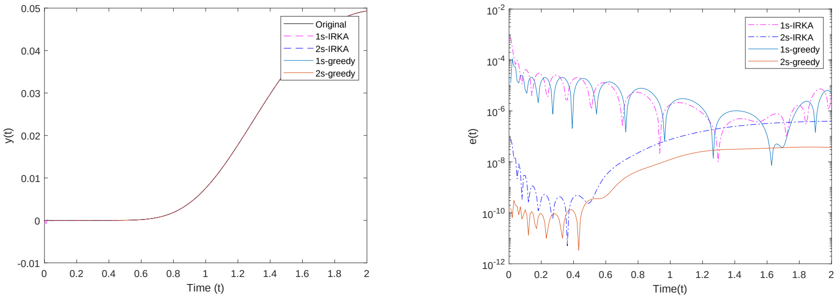
$$\begin{aligned} \upsilon_t + \upsilon\upsilon_x &= \nu \cdot \upsilon\upsilon_{xx}, && in\ \Gamma, \\ \alpha\upsilon(0,t) + \beta x(0,t) = u(t), &\quad \upsilon_x(1,t) = 0, && t \in (0,T), \\ \upsilon(x,0) = \upsilon_0(x), &\quad \upsilon_0(x) = 0, && x \in (0,1), \end{aligned} \tag{27}$$

we use it as an example to test our proposed method. We keep the size of the original model as n = 1000. Table 2 shows our results with tolerance $\epsilon_{tol} = 1e^{-4}$ and an initial choice of interpolation points as $\sigma_{10} = \sigma_{20} = 5.4124$.

S.No.	Interpolation points $\{\sigma_{1i}, \sigma_{2i}\}$	Max. True Error	Max. Est. Error
1	$5.4124, 5.4124$	1.1299×10^{-3}	32.4786
2	$31.6141, 1.383$	1.0259×10^{-3}	3.2407
3	$2.9603, 1.0818$	1.0746×10^{-3}	4.2125×10^{-1}
4	$9.2633 - 11.3351\iota, 24.9534$	1.416×10^{-4}	4.3411×10^{-4}
5	$7.4119 - 3.622\iota, 1.0818$	1.785×10^{-5}	1.7869×10^{-5}

Table 2: Error estimation results for burgers equation

The second column of the table shows interpolation points that are based on the error bound and identified by the greedy framework. Similarly, the error bound again tightly bounds the true error and therefore is reliable for choosing the interpolation points in the greedy algorithm. The sizes of the ROMs obtained from both approaches are kept same i.e. $r_1 = r_2 = 16$. The ROMs constructed from IRKA interpolation points and the proposed framework are shown in Figure 3 for input $u(t) = cos(\pi t)$.



(a) Comparison of transient response

(b) Comparison of absolute error

Figure 3: Burger's equation

Figure 3a shows the transient responses of the burgers equation computed from simulating the original model and the two different MOR approaches, while Figure 3b compares the absolute response errors of the ROMs derived using two approaches. The absolute error of ROM constructed using the proposed methodology of choosing interpolation points is less than that of the ROM constructed using IRKA interpolation points, especially for the two-sided projection.

5.3. FitzHugh - Nagumo System

We use the FitzHugh - Nagumo system as our third example to check our results. The FitzHugh - Nagumo system can be represented as[13]:

$$\begin{aligned} \epsilon \upsilon_t(x,t) &= \epsilon^2 \upsilon_{xx}(x,t) + f(\upsilon(x,t)) - w(x,t) + g, \\ w_t(x,t) &= h\upsilon(x,t) - \gamma w(x,t) + g, \end{aligned} \tag{28}$$

with $f(\upsilon) = \upsilon(\upsilon - 0.1)(1-\upsilon)$ and boundary conditions:

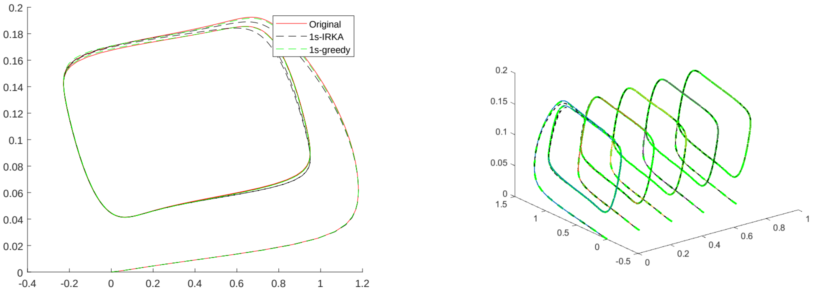
$$\begin{aligned} \upsilon(x,0) &= 0, & w(x,0) &= 0, \\ \upsilon_x(0,t) &= -i_0(t), & \upsilon_x(1,t) &= 0. \end{aligned} \tag{29}$$

Here, we choose $\epsilon = 0.015$, $h = 0.5$, $\gamma = 0.05$ and $i_0(t) = 5 \times 10^4 t^3 e^{-15t}$. When standard finite difference method is applied to numerically discretize the PDEs in (28), a system of ODEs with cubic non-linearities is obtained. We can get a quadratic bilinear system by introducing new variables. For an original discretized system with size $\bar{n}$, a quadratic bilinear system has the size of $n = 3\bar{n}$. we set $\bar{n} = 100$, which gives rise to quadratic bilinear system of $n = 300$. We choose interpolation points using the proposed greedy framework to construct the ROM of size $r = 26$ and then compare it with the ROM of the same size, which is constructed from the interpolation points using IRKA. Table 3 shows our results with tolerance $\epsilon_{tol} = 1e^{-6}$ and the interpolation points $\sigma_{10} = \sigma_{20} = 534.69$.

S.No.	Interpolation points	Max. True Error	Max. Est. Error
	$\{\sigma_{1i}, \sigma_{2i}\}$		
1	$534.69, 534.69$	0.282519	1152.4511
2	$1.38, 1.08$	4.7413×10^{-1}	8.4587
3	$3.91 - 5.45\iota, 1.38$	1.2373×10^{-4}	4.3284×10^{-3}
4	$39.38, 1.08$	2.5379×10^{-6}	5.9555×10^{-5}
5	$110.46, 1.08$	8.2393×10^{-6}	2.1293×10^{-5}
6	$3.96, 1.08$	4.3429×10^{-5}	7.1251×10^{-4}
7	$17.63, 1.08$	7.6047×10^{-6}	4.6707×10^{-5}
8	$4.83 - 4.72\iota, 1.08$	9.7775×10^{-8}	1.932×10^{-7}

Table 3: Error estimation results for the FitzHugh - Nagumo model

The table 3 shows interpolation points that are selected by the error bound and the decay of the true error and the error bound at each iteration of the greedy algorithm. The error bound once more, estimates the true error accurately, implicating that the selected interpolation points indeed nearly corresponds to the largest error. The sizes of ROMs obtained from both approaches have been kept the same i.e. $r_1 = r_2 = 26$. Figure 4 shows the transient responses of the FitzHugh - Nagumo system computed from simulating the original model and two approaches.



(a) Comparison of transient response

(b) Comparison of transient response (3D)

Figure 4: FitzHugh - Nagumo equation

The input signal is $u(t) = 50000t^3e^{-15t}$. It is seen that the 1s-greedy performs better than the 1s-IRKA when the outputs in both cases are compared with that of the original model; however, 2s-greedy and 2s-IRKA produce unstable responses.

6. Conclusions

In this paper, the proposed methodology of choosing interpolation points for construction of ROM of the first- and second-order transfer functions of quadratic-bilinear systems has been tested for three different models. The results have also been compared with ROMs of the same size constructed using the interpolation points chosen by linear IRKA. In each case, the ROMs constructed using interpolation points from the greedy framework yield better approximation of the output than the ROMs constructed from IRKA.

References

[1] S. Grundel, N. Hornung, B. Klaassen, P. Benner, T. Clees, Computing Surrogates for Gas Network Simulation using Model Order Reduction, Springer New York, 2013, pp. 189–212. doi:10.1007/978-1-4614-7551-4_9.

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

[3] J. R. Phillips, Projection-based approaces for model reduction of weakly nonlinear, time-varying systems, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 22 (2) (2003) 171–187.

[4] C. Gu, QLMOR: a projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems, IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst. 30 (9) (2011) 1307–1320.

[5] R. W. Freund, The SPRIM algorithm for structure-preserving order reduction of general RCL circuits, in: Model Reduction for Circuit Simulation, Springer, 2011, pp. 25–52.

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

[7] U. Baur, P. Benner, L. Feng, Model order reduction for linear and nonlinear systems: a system-theoretic perspective, Archives of Computational Methods in Engineering 21 (4) (2014) 331–358.

[8] E. J. Grimme, Krylov projection methods for model reduction, Phd thesis, Univ. of Illinois at Urbana-Champaign, USA (1997).

[9] A. C. Antoulas, D. C. Sorensen, S. Gugercin, A survey of model reduction methods for large-scale systems, Contemp. Math. 280 (2001) 193–219.

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

[11] M. Ahmad, P. Benner, I. Jaimoukha, Krylov subspace methods for model reduction of quadratic-bilinear systems, IET Control Theory Appl. 10 (2016) 2010–2018(8).

[12] R. J. Rugh, Nonlinear System Theory, Johns Hopkins University Press Baltimore, MD, 1981.

[13] P. Benner, P. Goyal, S. Gugercin, $\mathcal{H}_2$-quasi-optimal model order reduction for quadratic-bilinear control systems, arXiv preprint arXiv:1610.03279 (2016).

[14] A. J. Mayo, A. C. Antoulas, A framework for the solution of the generalized realization problem, Linear Algebra Appl. 425 (2-3) (2007) 634–662, special Issue in honor of P.A. Fuhrmann, Edited by A.C. Antoulas, U. Helmke, J. Rosenthal, V. Vinnikov, and E. Zerz.

[15] A. C. Ionita, A. C. Antoulas, Data-driven parametrized model reduction in the loewner framework, SIAMSciComp 36 (3) (2014) A984–A1007. doi:10.1137/130914619.

[16] I. V. Gosea, A. C. Antoulas, Model reduction of linear and nonlinear systems in the loewner framework: A summary, in: European Control Conference (ECC), IEEE, 2015, pp. 345–349.

[17] M. Ahmad, L. Feng, P. Benner, A new interpolatory model reduction for quadratic bilinear descriptor systems, Proc. Appl. Math. Mech. 15 (1) (2015) 589 – 590.

[18] M. I. Ahmad, P. Benner, L. Feng, Interpolatory model reduction for quadratic-bilinear systems using error estimators, Engineering Computations (2018).

[19] L. Feng, A. C. Antoulas, P. Benner, Some a posteriori error bounds for reduced-order modelling of (non-) parametrized linear systems, ESAIM: Mathematical Modelling and Numerical Analysis 51 (6) (2017) 2127–2158.

[20] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM Rev. 51 (3) (2009) 455–475.

[21] Y. Chen, Model reduction for nonlinear systems, Master's thesis, Massachusetts Institute of Technology (1999).