Balanced truncation for model order reduction of linear dynamical systems with quadratic outputs

Roland Pulch$^1$ and Akil Narayan$^2$

$^1$Institute of Mathematics and Computer Science, University of Greifswald, Walther-Rathenau-Str. 47, D-17489 Greifswald, Germany. Email: roland.pulch@uni-greifswald.de

$^2$Scientific Computing and Imaging Institute, Department of Mathematics, University of Utah, 72 Central Campus Dr., Salt Lake City, UT 84112, United States. Email: akil@sci.utah.edu

Abstract

We investigate model order reduction (MOR) for linear dynamical systems, where a quadratic output is defined as a quantity of interest. The system can be transformed into a linear dynamical system with many linear outputs. MOR is feasible by the method of balanced truncation, but suffers from the large number of outputs in approximate methods. To ameliorate this shortcoming we derive an equivalent quadratic-bilinear system with a single linear output and analyze the properties of this system. We examine MOR for this system via the technique of balanced truncation, which requires a stabilization of the system. Therein, the solution of two quadratic Lyapunov equations is traced back to the solution of just two linear Lyapunov equations. We present numerical results for several test examples comparing the two MOR approaches.

Keywords: linear dynamical system, quadratic-bilinear system, model order reduction, balanced truncation, Lyapunov equation, Hankel singular values.

*corresponding author
1 Introduction

A mathematical modeling of physical or industrial applications often yields dynamical systems. Automatic model generation typically generates dynamical systems of enormous dimension, and thus repeated transient simulations may become too costly. In this situation methods of model order reduction (MOR) are required to decrease the computational complexity of the system. Efficient MOR techniques already exist for linear dynamical systems, see [2, 3, 8, 9, 18]. MOR by balanced truncation or moment matching, for example, is based on an approximation of the input-output mapping, which can be described by a transfer function in the frequency domain. In contrast, the design of efficient and accurate algorithms for MOR of nonlinear dynamical systems is still a challenging task.

In this paper we investigate linear dynamical systems in the form of ordinary differential equations. However, the output quantity of interest for our system is the time-dependent trajectory of a quadratic function of the state. This nonlinear (quadratic) dependence of outputs on inputs precludes the direct applicability of transfer function methods for linear time-invariant systems.

There is some existing literature that addresses MOR for linear dynamical systems with quadratic outputs: Van Beeumen et al. [6, 7] consider a linear dynamical system with a single quadratic output in the frequency domain. Those authors transform this system into an equivalent form with multiple linear outputs. This approach is also applicable to our case of a linear dynamical system with a single quadratic output in the time domain, but often suffers from a very large number of outputs and hence is computationally expensive.

Alternatively, we derive a quadratic-bilinear dynamical system, whose single linear output coincides with the quadratic output of the original linear dynamical system. This allows us to leverage MOR methods for quadratic-bilinear systems, e.g. [1, 4]. We find the balanced truncation technique introduced by Benner and Goyal [5] particularly effective. Therein, quadratic Lyapunov equations must be solved, and we perform a stabilization of our quadratic-bilinear system to obtain the solvability of the Lyapunov equations. When analyzing the structure of the resulting quadratic-bilinear system, the MOR can be accomplished by solving two linear Lyapunov equations only. In particular, approximate methods like the alternating direction implicit (ADI) scheme can be used to solve the linear Lyapunov equations numerically [12, 14].

The paper is organized as follows. We derive both the linear dynamical system with multiple outputs and the quadratic-bilinear system with single output in Section 2. The balanced truncation technique is applied to the quadratic-bilinear system in Section 3. Therein, we analyze the structure and deduce an efficient
solution of the Lyapunov equations. Section 4 presents results of numerical computations for three test examples, and both the MOR approaches are compared with respect to accuracy and computation work.

2 Problem definition: Linear dynamical systems with quadratic outputs

Let a linear time-invariant system be given with a quadratic output in the form

\[ \dot{x}(t) = Ax(t) + Bu(t) \]
\[ y(t) = x(t)^\top M x(t). \]  

(1)

The state variables \( x : [0, t_{\text{end}}] \rightarrow \mathbb{R}^n \) are determined by the matrix \( A \in \mathbb{R}^{n \times n} \), the matrix \( B \in \mathbb{R}^{n \times n_{\text{in}}} \) and given inputs \( u : [0, t_{\text{end}}] \rightarrow \mathbb{R}^{n_{\text{in}}} \). We suppose that the number of inputs is low (\( n_{\text{in}} \ll n \)). We assume that the system is asymptotically stable, i.e., all eigenvalues of \( A \) exhibit a (strictly) negative real part. An initial value problem is defined by

\[ x(0) = x_0 \]  

(2)

with a predetermined \( x_0 \in \mathbb{R}^n \). The quantity of interest \( y : [0, t_{\text{end}}] \rightarrow \mathbb{R} \) represents a quadratic output defined by the matrix \( M \in \mathbb{R}^{n \times n} \). Note that, if \( T \in \mathbb{R}^{n \times n} \) is any matrix, then

\[ x^\top T x = x^\top \left( \frac{1}{2} (T + T^\top) \right) x \quad \text{for all } x \in \mathbb{R}^n. \]  

(3)

Thus we can assume that the matrix \( M \) in (1) is symmetric (since the relation above implies it can be replaced by its symmetric part). The system (1) is multiple-input-single-output (MISO).

We assume a situation with large dimension \( n \). (See Section 4 for examples of such systems.) It is in this regime when model order reduction (MOR) algorithms are advantageous. The aim of MOR is to construct a (linear or quadratic) dynamical system of a much lower dimension \( r \ll n \), and an associated output \( \tilde{y} \), such that the difference \( y - \tilde{y} \) should be sufficiently small for all relevant times.

2.1 Transformation to linear dynamical systems with multiple outputs

We adopt the strategy of [7], where the approach is transformation of a linear dynamical system with a single quadratic output into a linear dynamical system
with multiple (linear) outputs. We can consider again the system (1) but with a new output of interest $z$:

$$\dot{x}(t) = Ax(t) + Bu(t),$$
$$z(t) = L^\top x(t).$$

(4)

The matrix $L \in \mathbb{R}^{n \times m}$ is any matrix satisfying $M = LL^\top$ with $m = \text{rank}(M) \leq n$ and $z : [0, t_{\text{end}}] \to \mathbb{R}^m$. The initial values are as in (2). The system (4) is multiple-input-multiple-output (MIMO). The relationship between $L$ and $M$ immediately yields a relationship between the output $z$ in (4) and the output $y$ in (1):

$$y(t) = x(t)^\top M x(t) = (L^\top x(t))^\top (L^\top x(t)) = z(t)^\top z(t) = \|z(t)\|_2^2,$$

(5)

where $\| \cdot \|_2$ is the Euclidean norm.

We can apply an arbitrary symmetric decomposition $M = LL^\top$. For example, if $M$ is positive semi-definite, then the pivoted Cholesky factorization \[11\] provides one method of symmetric decomposition. Negative semi-definite matrices $M$ can be replaced by $-M$ and the negative Euclidean norm from (5) is used. In the case of indefinite symmetric matrices, an eigen-decomposition yields $M = S(D_+ - D_-)S^\top = SD_+S^\top - SD_-S^\top =: M_+ - M_-$

(6)

for an orthogonal matrix $S \in \mathbb{R}^{n \times n}$ and diagonal matrices $D_+, D_-$ including the positive eigenvalues and the modulus of the negative eigenvalues, respectively. The matrices $M_+$ and $M_-$ are both positive semi-definite so that the decomposition (6) can be used to construct matrices $L_\pm$ satisfying:

$$M_\pm = L_\pm^\top L_\pm, \quad L_\pm \in \mathbb{R}^{n \times m_\pm}, \quad \text{rank}(M) = m_+ + m_-.$$

Now it holds that $M = L_+L_+^\top - L_-L_-^\top$. Thus we arrange the linear dynamical system in the form

$$\begin{pmatrix} z_+(t) \\ z_-(t) \end{pmatrix} = \begin{pmatrix} L_+^\top \\ L_-^\top \end{pmatrix} x(t),$$

so that $y(t) = \|z_+(t)\|^2_2 - \|z_-(t)\|^2_2$ relates the quadratic output of (1) to the linear outputs $z_\pm$.

If the rank of the matrix $M$ is low, then just a few outputs arise in the system (4) and efficient MOR methods are available to reduce the linear dynamical system. However the situation becomes more difficult when the matrix $M$ has a large rank since MOR approaches often suffer limitations from the presence of a large number of outputs, with the extreme case is given by a matrix $M$ of full rank. Some MOR methods achieve accuracy only when the reduced dimension $r$ satisfies $r \approx \text{rank}(M)$, so that little to no computational benefit is gained in this case.

We will see for one of our capstone examples in Section 4 a realistic example where $M = I$, which presents a significant disadvantage to the methodology of this section.
2.2 Transformation to quadratic-bilinear systems

This section presents a strategy to overcome the limitations described at the end of the previous section. The idea is to construct a dynamical system that includes the quadratic quantity of interest \( y \) as an additional state variable. Differentiation of the quadratic output from (1) yields

\[
\dot{y} = 2 \dot{x}^\top M x = x^\top (2A^\top M) x + u^\top (2B^\top M) x.
\]

This reveals that \( \dot{y} \) as a function of the state \( x \) can be expressed as the sum of a quadratic term and a bilinear term. The matrix \( 2A^\top M \) is not symmetric in general, but the observation (3) shows that we can replace this matrix by

\[
S := A^\top M + M^\top A.
\]

Furthermore, let \( b_1, b_2, \ldots, b_{n_{in}} \) be the columns of the matrix \( B \).

By constructing a new vector \( \tilde{x} := (x^\top, y)^\top \in \mathbb{R}^{n+1} \), we obtain

\[
\dot{\tilde{x}}(t) = \tilde{A}\tilde{x}(t) + \tilde{B}u(t) + \sum_{j=1}^{n_{in}} u_j(t) \tilde{N}_j(t) \tilde{x}(t) + \tilde{H}(\tilde{x}(t) \otimes \tilde{x}(t))
\]

and

\[
\tilde{y}(t) = \tilde{c}^\top \tilde{x}(t)
\]

and

\[
\tilde{A} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{B} = \begin{pmatrix} B \\ 0 \end{pmatrix}, \quad \tilde{c}^\top = (0, \ldots, 0, 1), \quad \tilde{N}_j = \begin{pmatrix} 0 & 0 \\ 2b_j^\top M & 0 \end{pmatrix}
\]

for \( j = 1, \ldots, n_{in} \). The quadratic term involves the Kronecker product \( \tilde{x} \otimes \tilde{x} \in \mathbb{R}^{(n+1)^2} \). The matrix \( \tilde{H} \in \mathbb{R}^{(n+1)x(n+1)^2} \) can be defined in terms of the the columns \( s_1, \ldots, s_n \in \mathbb{R}^n \) of the symmetric matrix \( S \) from (7):

\[
\tilde{H} = \begin{pmatrix} 0_{n \times n} & 0_n & 0_n & \cdots & 0_n_{n \times n} & 0_n & 0_n \times n \times n & 0_n \\ s_1^\top & 0 & s_2^\top & \cdots & s_n^\top & 0 & 0_{1 \times n} & 0 \end{pmatrix},
\]

where only the last row is occupied. Thus the quadratic-bilinear system (8) produces just a single linear output, which is identical to the \((n + 1)\)th state variable. The initial values (2) imply that we must augment (8) with

\[
\tilde{x}(0) = \begin{pmatrix} x_0 \\ x_0^\top M x_0 \end{pmatrix}
\]

The quadratic output \( y \) of (1) and the linear output \( \tilde{y} \) of (8) coincide.

Our approach is to apply MOR methods for quadratic-bilinear systems, where an advantage is that the system (8) is MISO with a low number of inputs by assumption. We will show that the specific structure of (8) allows for an efficient MOR computation. Note that our system (8) is well-defined for arbitrary output matrices \( M \), i.e., including indefinite matrices.
2.2.1 Relationship to tensors

The system (8) can be related to more standard or classical definitions of quadratic dynamical systems. A quadratic dynamical system is defined by a three-dimensional tensor $H \in \mathbb{R}^{n \times n \times n}$. In the system (8), the matrix $\tilde{H}^{(1)}$ represents the 1-matricization of this tensor. Since the matrix $S$ in (7) is symmetric, its rows and columns coincide. Consequently, the 2-matricization $\tilde{H}^{(2)}$ and the 3-matricization $\tilde{H}^{(3)}$ are identical, i.e.,

$$\tilde{H}^{(2)} = \tilde{H}^{(3)} = \begin{pmatrix}
0_{n \times n} & s_1 & 0_{n \times n} & s_2 & \cdots & 0_{n \times n} & s_n & 0_{n \times n} & 0_n \\
0_{1 \times n} & 0 & 0_{1 \times n} & 0 & \cdots & 0_{1 \times n} & 0 & 0_{1 \times n} & 0
\end{pmatrix}.$$ 

It follows that the underlying tensor is symmetric, which allows for advantageous algebraic manipulations. However, an unsymmetric tensor can always be symmetrized. The definition of the tensor matricizations and symmetric tensors can be found in [4], for example.

2.2.2 Simplification of the bilinear term

The matrices of the bilinear part in (8) become $\tilde{N}_j = 0$ when $b_j^\top M = 0$ for each $j$. This simplification can happen quite often in practice, in particular when output-relevant state variables have equations that do not include the input. Let $I_{b_j} \subset \{1, \ldots, n\}$ be the support index of $b_j$, and let $I_M \subset \{1, \ldots, n\}$ denote the subset of state variables which are involved in the quadratic output. It holds that

$$I_{b_j} \cap I_M = \emptyset \quad \Rightarrow \quad b_j^\top M = 0. \quad (10)$$

If the premise of (10) is satisfied for all $j = 1, \ldots, n_{in}$ then the bilinear term in (8) vanishes.

2.2.3 Stabilization

The linear term $\tilde{A} \tilde{x}$ of the system (8) is just Lyapunov stable and not asymptotically stable, because the matrix $\tilde{A}$ has a zero eigenvalue that is simple. However, asymptotic stability is mandatory in some MOR methods, e.g., balanced truncation. Hence we stabilize via a tunable parameter $\varepsilon > 0$, resulting in the system

$$\begin{align*}
\dot{\tilde{x}}(t) &= \tilde{A}(\varepsilon)\tilde{x}(t) + \tilde{B}u(t) + \sum_{j=1}^{n_{in}} u_j(t)\tilde{N}_j\tilde{x}(t) + \tilde{H}(\tilde{x}(t) \otimes \tilde{x}(t)) \\
\tilde{y}(t) &= \tilde{c}^\top \tilde{x}(t)
\end{align*} \quad (11)$$

with the matrix

$$\tilde{A}(\varepsilon) = \begin{pmatrix} A & 0 \\
0 & -\varepsilon \end{pmatrix}.$$
3 Balanced Truncation

Our goal is now to compute an MOR of the stabilized quadratic-bilinear system \((\text{III})\). To accomplish this we apply the method of balanced truncation.

3.1 Reachability Gramian and observability Gramian

The first step in balanced truncation is computation of two Gramian matrices. For general quadratic-bilinear systems with multiple inputs and multiple outputs (MIMO), the reachability and observability Gramian matrices are defined in [5, Sect. 3]. The linear part of the quadratic-bilinear system must be asymptotically stable to obtain existence, uniqueness and positive (semi-)definiteness of the Gramians. The reachability Gramian \(\tilde{P} \in \mathbb{R}^{(n+1) \times (n+1)}\) associated to the stabilized system \((\text{III})\) is the solution of the quadratic Lyapunov equation

\[
\tilde{A}(\varepsilon)\tilde{P} + \tilde{P}\tilde{A}(\varepsilon)^\top + \tilde{B}\tilde{B}^\top + \tilde{H}(\tilde{P} \otimes \tilde{P})\tilde{H}^\top + \sum_{j=1}^{n_{\text{in}}} \tilde{N}_j \tilde{P} \tilde{N}_j^\top = 0. \tag{12}
\]

The observability Gramian \(\tilde{Q} \in \mathbb{R}^{(n+1) \times (n+1)}\) satisfies the quadratic Lyapunov equation

\[
\tilde{A}(\varepsilon)^\top \tilde{Q} + \tilde{Q}\tilde{A}(\varepsilon) + \tilde{c}\tilde{c}^\top + \tilde{H}^{(2)}(\tilde{P} \otimes \tilde{Q})(\tilde{H}^{(2)})^\top + \sum_{j=1}^{n_{\text{in}}} \tilde{N}_j^\top \tilde{Q} \tilde{N}_j = 0. \tag{13}
\]

If the reachability Gramian \(\tilde{P}\) is given, then the Lyapunov equations (13) represent a linear system for the unknown entries of the observability Gramian \(\tilde{Q}\). Both Gramian matrices are symmetric and positive semi-definite since \(\varepsilon > 0\).

The following lemma compiles relations that are used to evaluate the terms in the Lyapunov equations.

Lemma 1. Let \(\tilde{P}, \tilde{Q} \in \mathbb{R}^{(n+1) \times (n+1)}\) be partitioned into

\[
\tilde{P} = \begin{pmatrix} P & 0 \\ 0 & p' \end{pmatrix} \quad \text{and} \quad \tilde{Q} = \begin{pmatrix} Q & 0 \\ 0 & q' \end{pmatrix}
\]

with symmetric matrices \(P, Q \in \mathbb{R}^{n \times n}\) and \(p', q' \in \mathbb{R}\). It follows that

i) \(\tilde{N}_j \tilde{P} \tilde{N}_j^\top = 4 \begin{pmatrix} 0 & 0 \\ 0 & b_j^\top MPMb_j \end{pmatrix}\) for each \(j\),

ii) \(\sum_{j=1}^{n_{\text{in}}} \tilde{N}_j^\top \tilde{Q} \tilde{N}_j = 4q' \begin{pmatrix} MBB^\top M & 0 \\ 0 & 0 \end{pmatrix}\).
iii) \( \hat{H}(\tilde{P} \otimes \tilde{P})\hat{H}^\top = \text{tr}((PS)^2) \tilde{c}\tilde{c}^\top, \)

iv) \( \hat{H}^{(2)}(\tilde{P} \otimes \tilde{Q})(\hat{H}^{(2)})^\top = q' \begin{pmatrix} SP & 0 \\ 0 & 0 \end{pmatrix}. \)

**Proof.**

i) We calculate directly

\[
\tilde{N}_j \tilde{P} \tilde{N}_j^\top = \begin{pmatrix} 0 & 0 \\ 2b_j^\top M & 0 \end{pmatrix} \begin{pmatrix} P & 0 \\ 0 & p' \end{pmatrix} \begin{pmatrix} 0 & 2Mb_j \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 4b_j^\top MPMb_j \end{pmatrix}.
\]

ii) Likewise, it holds that

\[
\tilde{N}_j^\top \tilde{Q} \tilde{N}_j = \begin{pmatrix} 0 & 2Mb_j \\ 0 & 0 \end{pmatrix} \begin{pmatrix} Q & 0 \\ 0 & q' \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 2b^\top M & 0 \end{pmatrix} = \begin{pmatrix} 4q'(Mb_j)(b_j^\top M) & 0 \\ 0 & 0 \end{pmatrix}
\]

for \( j = 1, \ldots, n_{\text{in}} \). The sum over \( j \) yields the formula.

iii) The structure of \( \tilde{H} \) implies that the matrix \( \tilde{H}(\tilde{P} \otimes \tilde{P})\tilde{H}^\top \) has only one non-zero entry in the position \( (n+1, n+1) \). The vector \( \tilde{c} \) is the \( (n+1) \)th unit vector. Hence we obtain

\[
\hat{H}(\tilde{P} \otimes \tilde{P})\hat{H}^\top = \gamma \tilde{c}\tilde{c}^\top \]

with a scalar \( \gamma \) to be determined.

Let \( t_{ij} \) for \( i, j = 1, \ldots, n \) be the entries of the non-symmetric matrix \( PS \). It holds that

\[
\gamma = \sum_{i,j=1}^n p_{ji}s_j^\top Ps_i = \sum_{i,j,k,\ell=1}^n p_{ij}s_{jk}p_{\ell k}s_{i\ell} = \sum_{i,k=1}^n t_{ik}t_{ki}.
\]

Furthermore, we obtain the entries

\[
((PS)^2)_{ij} = \sum_{k=1}^n t_{ik}t_{kj} \quad \text{and thus} \quad ((PS)^2)_{ii} = \sum_{k=1}^n t_{ik}t_{ki}
\]

for \( i, j = 1, \ldots, n \). The sum over \( i \) yields the trace.

iv) We define the symmetric matrix \( \tilde{S} \in \mathbb{R}^{(n+1)\times(n+1)} \) by

\[
\tilde{S} = \begin{pmatrix} S & 0 \\ 0 & 0 \end{pmatrix}.
\]

It holds that \( \hat{H}^{(2)} = \tilde{S} \otimes \tilde{c}^\top \in \mathbb{R}^{(n+1)\times(n+1)^2} \), since \( \tilde{c} \) is the \( (n+1) \)th unit vector. The rule for matrix multiplications with the Kronecker product yields

\[
\hat{H}^{(2)}(\tilde{P} \otimes \tilde{Q})(\hat{H}^{(2)})^\top = (\tilde{S} \otimes \tilde{c}^\top)(\tilde{P} \otimes \tilde{Q})(\tilde{S} \otimes \tilde{c}) = (\tilde{S}\tilde{P}\tilde{S}) \otimes (\tilde{c}^\top \tilde{Q}\tilde{c}) = q'\tilde{S}\tilde{P}\tilde{S}.
\]

The definition of the matrix \( \tilde{S} \) implies

\[
\tilde{S}\tilde{P}\tilde{S} = \begin{pmatrix} SP & 0 \\ 0 & 0 \end{pmatrix},
\]

which shows the statement. \( \square \)
The left-hand sides of (i)-(iv) are terms in the Lyapunov equations \((12)\) and \((13)\). Direct evaluation of terms in the Lyapunov equations can be expensive, but Lemma \(\text{II}\) indicates that this effort can be significantly reduced because of the special structure of the system \((11)\) under consideration. The terms (i) and (ii) together require mainly two matrix-vector multiplications with the columns of \(B\). The term (iv) is obtained by two matrix-matrix multiplications for \(SPS\). One additional matrix-matrix multiplication yields \((PS)^2\) and thus the term (iii).

The next result characterizes the solutions of the quadratic Lyapunov equations.

**Theorem 1.** The reachability Gramian \(\tilde{P}\) satisfying (12) is given by

\[
\tilde{P} = \begin{pmatrix} P & 0 \\ 0 & p' \end{pmatrix},
\]

where \(P\) solves the linear Lyapunov equations

\[
AP + PA^\top + BB^\top = 0
\]

and

\[
p' = \frac{1}{2\varepsilon} p'' \quad \text{with} \quad p'' = \text{tr}((PS)^2) + 4 \sum_{j=1}^{n_{in}} b_j^\top MPMb_j.
\]

The observability Gramian \(\tilde{Q}\) defined by (13) is given by

\[
\tilde{Q} = \frac{1}{2\varepsilon} \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix},
\]

where \(Q\) satisfies the linear Lyapunov equations

\[
A^\top Q + QA + SPS + 4MBB^\top M = 0.
\]

**Proof.**

Inserting the ansatz (14) into the quadratic Lyapunov equations (12) yields the linear Lyapunov equations (15) for the first part. Due to Lemma \(\text{II}\) (i) and (iii), the second part becomes

\[-2\varepsilon p' + \text{tr}((PS)^2) + 4 \sum_{j=1}^{n_{in}} b_j^\top MPMb_j = 0,
\]

which defines the scalars \(p'\) and \(p''\), respectively.

Now the ansatz (17) is inserted into the Lyapunov equations (13). The second part is fulfilled immediately due to \(-2\varepsilon \cdot \frac{1}{2\varepsilon} + 1 = 0\). Lemma \(\text{II}\) (ii) and (iv) yield the linear Lyapunov equations

\[
A^\top \frac{1}{2\varepsilon} Q + A \frac{1}{2\varepsilon} Q + \frac{1}{2\varepsilon} SPS + 4 \frac{1}{2\varepsilon} MBB^\top M = 0
\]
as the first part. A multiplication by the factor $2\varepsilon$ results in the Lyapunov equations (18).

Theorem 1 reveals the explicit dependence of the reachability and observability Gramians $\tilde{P}$ and $\tilde{Q}$, respectively, on the stabilization parameter $\varepsilon$. In particular, the matrices $P, Q$ satisfying the linear Lyapunov equations (15), (18) are independent of the stabilization parameter $\varepsilon$. The observability Gramian (17) is directly proportional to $\frac{1}{\varepsilon}$.

### 3.2 Balancing the system

In order to perform balanced truncation, we require symmetric decompositions of the two Gramian matrices. Let $P = L_P L_P^\top$ and $Q = L_Q L_Q^\top$ be the Cholesky decompositions of the solutions of the linear Lyapunov equations (15) and (18), respectively. We obtain factorizations of the reachability Gramian $\tilde{P} = \tilde{L}_P \tilde{L}_P^\top$ and the observability Gramian $\tilde{Q} = \tilde{L}_Q \tilde{L}_Q^\top$ by

$$\tilde{L}_P = \begin{pmatrix} L_P & 0 \\ 0 & \sqrt{\frac{\nu''}{2\varepsilon}} \end{pmatrix} \quad \text{and} \quad \tilde{L}_Q = \frac{1}{\sqrt{2\varepsilon}} \begin{pmatrix} L_Q & 0 \\ 0 & 1 \end{pmatrix}.$$

In the method of balanced truncation, we require the singular value decomposition (SVD)

$$\tilde{L}_Q \tilde{L}_P = \tilde{U} \tilde{\Sigma} \tilde{V}^\top. \quad (19)$$

In our case, the matrix on the left-hand side reads as

$$\tilde{L}_Q^\top \tilde{L}_P = \frac{1}{\sqrt{2\varepsilon}} \begin{pmatrix} L_Q^\top L_P & 0 \\ 0 & \sqrt{\frac{\nu''}{2\varepsilon}} \end{pmatrix}.$$

We use the SVD

$$L_Q^\top L_P = U \Sigma V^\top, \quad (20)$$

where the diagonal matrix $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$ includes the singular values in ascending order $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$. The SVD (20) is independent of the stabilization parameter $\varepsilon$. Thus an SVD (19) is given by

$$\tilde{U} = \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix}, \quad \tilde{V}^\top = \begin{pmatrix} V^\top & 0 \\ 0 & 1 \end{pmatrix}, \quad \tilde{\Sigma} = \frac{1}{\sqrt{2\varepsilon}} \begin{pmatrix} \Sigma & 0 \\ 0 & \sqrt{\frac{\nu''}{2\varepsilon}} \end{pmatrix}.$$

If the stabilization parameter is sufficiently small, then the maximum singular value is $\sqrt{\nu''/(2\varepsilon)}$. However, the associated singular vector is independent of $\nu''$ and $\varepsilon$. The singular values of the quadratic-bilinear system (8) are

$$\frac{1}{\sqrt{2\varepsilon}} \begin{pmatrix} \sigma_1, \sigma_2, \ldots, \sigma_n; \sqrt{\frac{\nu''}{2\varepsilon}} \end{pmatrix}. \quad (21)$$
These real numbers represent the analogue of the Hankel singular values in the case of linear dynamical systems. The two transformation matrices, which achieve a balanced system, result to

$$
\tilde{T}_l = \tilde{L}_Q \tilde{U} \tilde{\Sigma}^{-\frac{1}{2}} = \frac{1}{\sqrt{2\varepsilon}} \begin{pmatrix} L_Q U \Sigma^{-\frac{1}{2}} & 0 \\ 0 & \sqrt{\frac{2\varepsilon}{p''}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2\varepsilon}} L_Q U \Sigma^{-\frac{1}{2}} & 0 \\ 0 & p''^{-\frac{1}{4}} \end{pmatrix},
$$

$$
\tilde{T}_r = \tilde{L}_P \tilde{V} \tilde{\Sigma}^{-\frac{1}{2}} = \sqrt{2\varepsilon} \begin{pmatrix} L_P V \Sigma^{-\frac{1}{2}} & 0 \\ 0 & \sqrt{\frac{2\varepsilon}{p''}} \end{pmatrix} = \begin{pmatrix} \sqrt{2\varepsilon} L_P V \Sigma^{-\frac{1}{2}} & 0 \\ 0 & p''^{-\frac{1}{4}} \end{pmatrix},
$$

with $p''$ from $[16]$. It holds that $\tilde{T}_l^\top \tilde{T}_r = I$ with the identity matrix $I$. For $\varepsilon \to 0$ some parts of the matrices converge to zero and the other parts tend to infinity, hence the limits do not exist.

### 3.3 Reduced-order model

The concepts of reachability and observability allow one to devise an MOR strategy: State variables components that require a large energy to achieve (reach) or generate a low energy in the output (observation) should be truncated. In the balanced systems, a state variable is hard to reach if and only if it produces a low output energy.

Given a reduced order dimension $r \in \mathbb{N}$, we assume that the stabilization parameter $\varepsilon > 0$ is chosen sufficiently small such that

$$
\sigma_{n+1} := \sqrt{\frac{p''}{2\varepsilon}} > \sigma_{n-r}.
$$

Hence $\sigma_{n+1}$ belongs to the set of the $r$ dominant singular values. We partition the SVD $[19]$ into

$$
\tilde{L}_Q^\top \tilde{L}_P = (\tilde{U}_1 \quad \tilde{U}_2) \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & \tilde{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \tilde{V}_1^\top \\ \tilde{V}_2^\top \end{pmatrix}
$$

with $\tilde{\Sigma}_2 \in \mathbb{R}^{r \times r}$, $\tilde{U}_2, \tilde{V}_2 \in \mathbb{R}^{(n+1) \times r}$. The associated projection matrices $\tilde{T}_{l/r} \in \mathbb{R}^{(n+1) \times r}$ read as

$$
\tilde{T}_l = \tilde{L}_Q \tilde{U}_2 \tilde{\Sigma}_2^{-\frac{1}{2}} \quad \text{and} \quad \tilde{T}_r = \tilde{L}_P \tilde{V}_2 \tilde{\Sigma}_2^{-\frac{1}{2}}.
$$

Due to the ascending order of the singular values and the condition $[23]$, the MOR truncates state variables, which are both hard to reach and difficult to observe.
The reduced-order model (ROM) of the quadratic-bilinear system (11) becomes

\[
\dot{\bar{x}}(t) = \bar{A}\bar{x}(t) + \bar{B}u(t) + \sum_{j=1}^{n_{\text{in}}} u_j(t) N_j \bar{x}(t) + \bar{H}(\bar{x}(t) \otimes \bar{x}(t))
\]
\[
\bar{y}(t) = \bar{c}^\top \bar{x}(t)
\]

with the solution \(\bar{x} : [0, t_{\text{end}}] \to \mathbb{R}^r\) and the downsized matrices

\[
\bar{A} = \bar{T}_r^\top \tilde{A} \bar{T}_r, \quad \bar{B} = \bar{T}_r^\top \tilde{B}, \quad \bar{c}^\top = \bar{T}_r^\top \tilde{c},
\]
\[
\bar{N}_j = \bar{T}_r^\top \tilde{N}_j \bar{T}_r, \quad \bar{H} = \bar{T}_r^\top \tilde{H}(\bar{T}_r \otimes \bar{T}_r).
\]

The output vector is

\[
\bar{c}^\top = \left(0, \ldots, 0, \sqrt{p''} \right).
\]

Thus the output is just a multiple of the final state variable as in the quadratic-bilinear system (11).

The computational effort for the matrices \(\tilde{N}_j\) in (26) is negligible because only the last row of \(\tilde{N}_j\) is non-zero. The computation of the matrix \(\bar{H} \in \mathbb{R}^{r \times r_2}\) represents the most expensive part in the projections (26). In [4, Sect. 3.2], an algorithm is outlined to construct \(\bar{H}\) without calculating the Kronecker product \(\bar{T}_r \otimes \bar{T}_r\) explicitly. In our case, the effort becomes even lower since the matrix \(\bar{H}\) exhibits the structure (9) of \(H\) again. The entries of the matrix

\[
\bar{S} = \bar{T}_r^\top S \bar{T}_r \in \mathbb{R}^{r \times r}
\]

with the symmetric matrix \(S\) from (7) yields the last row of the intermediate matrix \(\bar{H}(\bar{T}_r \otimes \bar{T}_r)\), whereas the other rows are zero. Thus the effort mainly consists in the computation of matrix-matrix products in (27).

On the one hand, we have

\[
\bar{A} = \begin{pmatrix} \bar{A}^* & 0 \\ 0 & -\varepsilon \end{pmatrix}
\]

with a matrix \(\bar{A}^* \in \mathbb{R}^{(r-1) \times (r-1)}\) independent of \(\varepsilon\). This allows us in principle to consider the limit \(\varepsilon \to 0\) in the matrix (28). On the other hand, taking the limits \(\varepsilon \to 0\) for other matrices and vectors is dubious, since these limits either do not exist or become zero.

The ROM (25) exhibits the same structure as the quadratic-bilinear system (11) in the nonlinear terms.
Theorem 2. Let $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_r)^\top$ and $\bar{x}^* = (\bar{x}_1, \ldots, \bar{x}_{r-1})^\top$. The reduced system (25) features the equivalent form

\begin{align*}
\dot{\bar{x}}^*(t) &= \bar{A}^*\bar{x}^*(t) + \bar{B}^*u(t) \\
\dot{\bar{x}}_r(t) &= -\varepsilon\bar{x}_r(t) + \sum_{j=1}^{n_m} u_j(t)\bar{N}_j^*\bar{x}^*(t) + \bar{H}^*(\bar{x}^*(t) \otimes \bar{x}^*(t)) \\
\bar{y}(t) &= \sqrt[4]{p''(t)} \bar{x}_r(t)
\end{align*}

with $p''$ from (16), the matrix $\bar{A}^*$ from (28), a modified matrix $\bar{B}^*$ and row vectors $\bar{N}_j^*$, $\bar{H}^*$.

The proof is straightforward. The structure of the system (29) implies two benefits for solving initial value problems in comparison to general quadratic-bilinear systems:

1. If an implicit time integration scheme is used, then nonlinear systems of algebraic equations can be avoided and only linear systems have to be solved.
2. Since the output is just a constant multiple of a single state variable, an adaptive time step size selection can be performed by a local error control of this state variable only.

3.4 Low-rank approximations

An MOR for the linear dynamical system (1) with quadratic output can be performed by using the linear dynamical system (4) with multiple outputs or the quadratic-bilinear system (8) with single output. Two criteria determine the efficiency of the approaches in balanced truncation:

1. The decay of the (Hankel) singular values. A faster decay typically allows for a sufficiently accurate ROM of a lower dimension.
2. The computational effort to construct an ROM.

Numerical computations indicate that the rate of decay is similar for the singular values in test examples. Thus an advantage in the quadratic-bilinear system formulation can be achieved only by decreasing the computational effort.

The main part of the computational work for the balanced truncation technique consists in the solution of the linear Lyapunov equations (15) and (18). A general linear Lyapunov equation reads as

\[ AG + GA^\top + F = 0 \]
for the unknown matrix $G \in \mathbb{R}^{n \times n}$ with a given symmetric positive semi-definite matrix $F \in \mathbb{R}^{n \times n}$. If we apply direct methods of linear algebra to solve (30), then the computational complexity is $O(n^3)$ and nearly independent of the rank of $F$. Consequently, we could solve the linear dynamical system (11) including many outputs as well, where $F$ exhibits a high rank.

Alternatively, approximate methods yield low-rank factorizations of the solution of the Lyapunov equation (30). Efficient algorithms are achieved by iterations based on the alternating direction implicit (ADI) technique, see [12, 14]. A low-rank approximation reads as $G \approx Z_GZ_G^\top$ with $Z_G \in \mathbb{R}^{n \times k}$ for some $k \ll n$. An ADI technique requires a symmetric factorization $F \approx Z_FZ_F^\top$ with $Z_F \in \mathbb{R}^{n \times k_F}$ and $k_F \leq n$ as input. However, the convergence properties as well as the computational efficiency suffer from a large rank $k_F$. In [16, p. 10], the property $k_F \ll n$ is assumed in the ADI method.

Concerning both systems (11) and (11), an iterative computation of the reachability Gramian can be easily devised because $F = BB^\top$ and thus $k_F = n_{in}$ due to our assumption of a low number of inputs. We obtain a low-rank approximation $P \approx Z_PZ_P^\top$ solving (15) with $Z_P \in \mathbb{R}^{n \times k_P}$. A low-rank factorization of the reachability Gramian also allows for a fast computation of the value (16).

However, the observability Gramian requires a factorization $F \approx Z_FZ_F^\top$ in the case of the linear dynamical system (11), where the rank may be large (possibly close to $n$). In the case of the quadratic-bilinear system (11), the input matrix for (18) becomes

$$\hat{F} = SPS + 4MBB^\top M \approx (SZ_P)(SZ_P)^\top + 4(MB)(MB)^\top.$$  

(31)

Now we obtain directly an approximate factorization of (31) by

$$Z_{\hat{F}} \approx (SZ_P, (2MB)) \in \mathbb{R}^{n \times (k_P + n_{in})},$$  

(32)

where the number of columns is low due to $k_P, n_{in} \ll n$. The rank of the factor (32) may be smaller than $k_P + n_{in}$, which allows for a simplification to a full-rank factor $Z_{\hat{F}} \in \mathbb{R}^{n \times k_{\hat{F}}}$ for some $k_{\hat{F}} < k_P + n_{in}$. Furthermore, a reduced factor $Z_{\hat{F}}$ with $k_{\hat{F}}$ columns can be obtained by using just the first $k_{\hat{F}} < k_P$ columns of the factor $Z_P$. An iterative scheme solving (15) yields the factorization $Q \approx Z_QZ_Q^\top$ with $Z_Q \in \mathbb{R}^{n \times k_Q}$ for the observability Gramian. For a general Lyapunov equation (30) with $F = Z_FZ_F^\top$ and $Z_F \in \mathbb{R}^{n \times k_F}$, $j$ iterations of the ADI method generate a factor with $jk_F$ columns, see [16].

Our balanced truncation approach follows the strategy described in [19]. We reduce to a dimension $r < n$, assuming the condition (23) is satisfied. Thus approximate factors $Z_P, Z_Q$ with ranks $k_P, k_Q \geq r - 1$ are required for the linear Lyapunov equations (15), (18). Symmetric decompositions $\hat{P} \approx \hat{Z}_P\hat{Z}_P^\top$ and $\hat{Q} \approx$
\( \tilde{Z}_Q \tilde{Z}_Q^T \) for the matrices from (14), (17) read as

\[
\tilde{Z}_P = \left( \begin{array}{cc} Z_P & 0 \\ 0 & \sqrt{\frac{y''}{2\varepsilon}} \end{array} \right) \quad \text{and} \quad \tilde{Z}_Q = \frac{1}{\sqrt{2\varepsilon}} \left( \begin{array}{cc} Z_Q & 0 \\ 0 & 1 \end{array} \right).
\]

Now we compute an SVD of the small matrix \( \tilde{Z}_Q^T \tilde{Z}_P \in \mathbb{R}^{(k_Q+1) \times (k_P+1)} \). A partition (24) of this SVD is used again assuming an ascending order of the singular values. We suppose that the condition (23) is also satisfied in this approximation. The projection matrices for the reduction (26) are

\[
\tilde{T}_1 = \tilde{Z}_Q \tilde{U}_2 \tilde{\Sigma}_2^{-\frac{1}{2}} \quad \text{and} \quad \tilde{T}_r = \tilde{Z}_P \tilde{V}_2 \tilde{\Sigma}_2^{-\frac{1}{2}}
\]

with \( \tilde{U}_2 \in \mathbb{R}^{(k_Q+1) \times r} \) and \( \tilde{V}_2 \in \mathbb{R}^{(k_P+1) \times r} \). We require just the \( r \) largest singular values and their singular vectors for the computation of the ROM.

4 Numerical results

We apply the reduction approaches from the previous sections to three test examples. In each case, three types of MOR using balanced truncation are examined:

i) the linear dynamical system (4) with multiple outputs by direct algorithms of linear algebra,

ii) the quadratic-bilinear system (11) with single output by direct algorithms of linear algebra,

iii) the quadratic-bilinear system (11) with single output using ADI iteration.

The numerical computations were performed by the software package MATLAB \([15]\) (version R2016a), where the machine precision is around \( \varepsilon_0 = 2 \cdot 10^{-16} \). We used the ADI algorithm from the package LYAPACK, see [16]. The CPU times were measured on an iMac with 3.4 GHz Inter Core i7 processor and operation system OS X El Capitan.

In each test example, we compute discrete approximations of the maximum absolute error and the integral mean value of the relative error on a time interval [0, T], i.e.,

\[
E_{\text{abs}} = \max_{t \in [0, T]} |\tilde{y}(t) - y(t)| \quad \text{and} \quad E_{\text{rel}} = \frac{1}{T} \int_0^T \frac{|\tilde{y}(t) - y(t)|}{|y(t)|} \, dt
\]

with \( y \) from a full-order model (FOM) and \( \tilde{y} \) from an ROM.
4.1 Positive definite output matrix

We construct a linear dynamical system (1) of dimension $n = 5000$. A matrix $A' \in \mathbb{R}^{n \times n}$ is arranged using pseudo random numbers with respect to a standard Gaussian distribution. Let $\gamma$ be the largest real part of the eigenvalues of $A'$. We define the dense matrix $A = A' - \lceil \gamma \rceil I$, which implies an asymptotically stable system. Furthermore, a single input is introduced using the vector $B = (1, 1, \ldots, 1)^\top$. We use the identity matrix $M = I$ in the definition of the quadratic output. This matrix is obviously symmetric and positive definite. Even though this choice is simple, the identity matrix cannot be well-approximated by a low-rank matrix.

As time-dependent input, we supply a chirp signal

$$u(t) = \sin(k(t)t) \quad \text{with} \quad k(t) = k_0 t$$

and the constant $k_0 = 0.1$. All initial values are zero. The total time interval $[0, 100]$ is considered in the transient simulations. We use an explicit embedded Runge-Kutta method of convergence order 4(5) for computing numerical solutions of our initial value problems (ode45 in MATLAB). This procedure uses step size selection by a local error control with relative tolerance $\varepsilon_{\text{rel}} = 10^{-6}$ and absolute tolerance $\varepsilon_{\text{abs}} = 10^{-8}$ in all state variables. Thus high accuracy requirements are imposed. Figure 1 (left) shows the quadratic output resulting from the numerical solution of (1).

On the one hand, we arrange the linear dynamical system (4), where it holds that $L^\top = I$. Hence the number of outputs is equal to $n$ in (4). The balanced truncation technique yields the Hankel singular values. On the other hand, we derive the quadratic-bilinear system (11) including the stabilization parameter.
Figure 2: Hankel singular values (left) and their normalized values (right) for the two dynamical systems associated to identity output matrix.

$\varepsilon = 10^{-8}$. The balanced truncation scheme from Section 3.2 produces other singular values. The dominant singular values up to order 80 are illustrated in descending magnitudes by Figure 2 (left). The largest singular value of the quadratic-bilinear system (11) has a special role, see (21). Thus we normalize the first singular value of (11) and the second singular value of (11) to one. Figure 2 (right) shows the normalized singular values. We observe the same rate of decay in both sets of singular values, which indicates a similar potential for MOR by balanced truncation.

Since we apply direct linear algebra methods, each balanced truncation approach yields square transformation matrices, see (22) for the quadratic-bilinear case. The projection matrices of the MOR result from the dominant columns of these square matrices. Consequently, we obtain an ROM for an arbitrary dimension $r < n$. We compute ROMs from the linear dynamical systems (4) and from the quadratic-bilinear system (11) for $r = 5, 6, \ldots, 80$. In the quadratic-bilinear case, we choose $\varepsilon = 0$ only in the reduction of the matrix (28), because numerical results show that the errors become smaller as $\varepsilon \to 0$ in this matrix.

Furthermore, we solve the linear Lyapunov equations (15) and (18) iteratively by an ADI method for each $r$ separately as described in Section 3.4. On the one hand, the low-rank factor $Z_P$ of the reachability Gramian is computed with rank $k_P = r + 10$ by $j = k_P$ iteration steps. On the other hand, only the first $r$ columns of $Z_P$ are inserted in the Lyapunov equation (18) and $j = 10$ iteration steps are performed. It follows that the low-rank factor $Z_Q$ of the observability Gramian has rank $k_Q = j(r + 1)$ due to (32). Each pair of iterative solutions implies an associated ROM.

The transient simulation of the FOM (11) yields approximations at many time points with variable step sizes. We integrate the ROMs by the same Runge-
Kutta method with local error control including the tolerances from above. The integrator produces approximations of identical convergence order at the predetermined time points. We obtain discrete approximations of the errors (33) by the differences in each time point, which are depicted in Figure 3. Both the maximum absolute error and the mean relative error decrease exponentially for increasing dimensions of the ROMs. The errors start to stagnate at higher dimensions, since the errors of the time integration become dominant. Furthermore, tiny values of the exact solution appear close to the initial time, which locally causes large relative errors. The errors are lower for the linear dynamical system (4) in comparison to the quadratic-bilinear system (11), although the associated singular values exhibit a similar rate of decay. We suspect that the balanced truncation strategy works in general better for linear dynamical systems. This suspicion can be corroborated by other facts: For example, ROMs for linear dynamical systems have error bounds that depend on the Hankel singular values, but error bounds depending on the singular values of a quadratic-bilinear case are not known. The iterative solution of the Lyapunov equations associated to the quadratic-bilinear system results in larger errors than the direct solution, since several approximations are made in this procedure.

Now we analyze the CPU times of the three MOR techniques. The CPU time for the time integration of the FOM (1) is 8602.9. Figure 4 (left) illustrates the effort for the calculation of the projection matrices in the balanced truncation, which includes the solution of Lyapunov equations and thus represents the main part of the computation work. In the direct approaches, the complete transformation matrices have to be computed, which is indicated by constant CPU times. In the iterative approach, the effort grows just slowly with increasing reduced dimension. Figure 5 depicts the CPU time for the computation of the reduced matrices (left) and the time integration of the ROMs including the calculation of
the quantity of interest (right). We observe that the computation work for the matrices is negligible. The total speed up is shown in Figure 4 (right), where both the construction and the transient simulation of the ROMs is compared to the solution of the FOM. The speedup is nearly constant for different reduced dimensions in the direct approaches because the balanced truncation part dominates. The iterative method exhibits a significantly higher speedup for small dimensions, whereas the speedup decreases for larger dimensions.

4.2 Indefinite output matrix

We arrange a linear dynamical system \( \mathbf{A} \) of dimension \( n = 5000 \) with a system matrix \( A \) and a vector \( B \) as in Section 4.1 (But using a different realization.
of the pseudo random numbers.) We fill a matrix $M' \in \mathbb{R}^{n \times n}$ by pseudo random numbers associated to a uniform distribution in $[-1, 1]$. Now the output matrix $M := \frac{1}{2}(M' + M'^\top)$ is dense, full-rank and symmetric. The matrix is indefinite, and for our realization it has exactly $\frac{n}{2}$ positive eigenvalues and $\frac{n}{2}$ negative eigenvalues. We apply the chirp signal (34) with $k_0 = 0.1$ as input again. Figure 1 (right) illustrates the quadratic output of the system (1), which exhibits both positive and negative values.

We use balanced truncation by direct linear algebra methods for the linear dynamical system (4) with $n$ outputs and the quadratic-bilinear system (11) with single output, where the stabilization parameter is $\varepsilon = 10^{-8}$. The resulting dominant singular values are depicted in Figure 6. Again the rate of decay is nearly identical in both cases.

The ROMs are computed as in Section 4.1. The same Runge-Kutta method with identical tolerances is used for solving the initial value problems, where all initial values are zero. Figure 7 shows the resulting discrete approximations of the error measures (33). The behavior of the errors is similar to the previous example in Section 4.1. Concerning the relative errors, note that the exact quadratic output features many zeros.

### 4.3 Stochastic Galerkin method and variance

We consider a mass-spring-damper configuration from Lohmann and Eid [13]. The associated linear dynamical system consists of 8 ordinary differential equations including 14 physical parameters and is single-input-single-output (SISO). In [17], this test example was investigated in the context of stochastic modeling,
Figure 7: Maximum absolute errors (left) and mean relative errors (right) for ROMs of different dimensions in the case of indefinite output matrix.

where the physical parameters are replaced by independent random variables. The state variables as well as the output are expressed as a polynomial chaos expansion with \( m = 680 \) basis polynomials, see [20].

The stochastic Galerkin approach yields a larger linear dynamical system (SIMO)

\[
\begin{align*}
\dot{x}(t) &= Ax(t) + Bu(t) \\
w(t) &= Cx(t),
\end{align*}
\]

\( x \in \mathbb{R}^n \), with dimension \( n = 5440 \) and outputs \( w = (w_1, \ldots, w_m)^\top \). The first output \( w_1 \) represents an approximation of the expected value for the original single output. The other outputs \( w_2, \ldots, w_m \) produce an approximation of its variance by

\[
\text{Var}(t) = \sum_{i=2}^{m} w_i^2.
\]  

The details of the above modeling can be found in [17].

As single input, we choose the harmonic oscillation \( u(t) = \sin(\omega t) \) with frequency \( \omega = 0.2 \). Initial value problems are solved with starting values zero in the time interval \([0, T] \) with \( T = 2000 \). Since the stochastic Galerkin system (35) is mildly stiff, we apply the implicit trapezoidal rule. Figure 8 shows the approximations of the expected value as well as the variance obtained from the transient simulation. Driven by the input signal, the solutions become nearly periodic functions after an initial phase.

We construct a linear dynamical system \( \ddot{x} = A\dot{x} + Bu + Cw \) from the stochastic Galerkin system (35), whose quadratic output is the variance (36). Define \( L^\top \in \mathbb{R}^{(m-1) \times n} \) as the matrix \( C \in \mathbb{R}^{m \times n} \) in (35) with its first row omitted. It follows that \( M = LL^\top \).
in (11) is symmetric and positive semi-definite of rank $m - 1$. Thus the equivalent system (11) with $m - 1$ linear outputs defined by $L^\top$ is already available in this application.

The associated quadratic-bilinear system (8) is without bilinearity, because the property (10) is satisfied. We apply the stabilized system (11) with a parameter $\varepsilon = 10^{-8}$ again.

We examine the MOR by balanced truncation for this problem comparing the reduction of the linear system (4) and the quadratic system (11). In the linear system, projection matrices are obtained directly by linear algebra algorithms. In the quadratic system, both a direct method and an iterative method using the ADI technique are employed. We compute projection matrices for the reduced dimension $r_{\text{max}} = 100$ in each approach. Within the ADI iteration, an approximate factor for the reachability Gramian is computed with rank $k_P = 200$. Just the first $k_P' = 20$ columns are applied with $j = 20$ iterations for the calculation of an approximate factor of the observability Gramian with rank $k_Q = 400$.

The balanced truncation techniques yield singular values of Hankel type in each of the three reductions. Figure 9 illustrates the dominant singular values in descending order. We recognize a faster decay of the singular values in the quadratic system (8). However, the faster decrease of the singular values in the iterative method represents an error by the approximation, because the direct approach produces much more accurate values.

Given the projection matrices with $r_{\text{max}}$ columns, we choose the dominant part to obtain ROMs of dimension $r = 5, 6, \ldots, 100$. We arrange $\varepsilon = 0$ in the matrix (28) again. To investigate the errors of the MOR, we solve initial value problems highly accurate in the time interval $[0, T]$ by the trapezoidal rule with constant
time step size using $10^5$ time steps. The constant step size allows for reusing
$LU$-decompositions in all systems. The original system (1) yields the reference
solution. Due to Theorem 2, nonlinear systems of algebraic equations are omitted
in the quadratic-bilinear ROMs (29). We solve the ROMs for each dimension $r$.
The maximum absolute errors and the integral mean values of the relative errors
are depicted in Figure 10. For low dimensions (say $r < 50$), the errors decrease
exponentially in each approach. For higher dimensions, the errors stagnate in
the quadratic system, i.e., the approximations do not improve any more. The
iteration technique produces approximations of the same quality as the direct
approach for low dimensions. However, two outliers appear for high dimensions
with very large errors (partly outside the shown range in the figure). The relative
error is very large for low dimensions in all approaches, because the exact values
of the output are close to zero for small times.
5 Conclusions

We have investigated two approaches for model order reduction of linear dynamical systems with an output of interest that is quadratic in the state variables. This problem can be recast, equivalently, as a linear dynamical system with multiple outputs, or as a quadratic-bilinear system with a single output. Our model order reduction approaches implement the method of balanced truncation for each of these two recast systems. Balanced truncation requires solutions to Lyapunov equations. We find that manipulation of large matrices necessary to solve the Lyapunov equations motivate the use of approximate or iterative approaches, notably the alternating direction implicit method, in both linear and quadratic systems.

Our numerical examples demonstrate that both model order reduction approaches can achieve significant accuracy with a much smaller dynamical system. For computing the output quantities of interest, the quadratic-bilinear systems are advantageous because they require computation of only a single scalar output. In contrast, computation of the output quantity of interest from linear dynamical systems requires to compute multiple outputs. The alternating direction implicit method is not possible in the case of a large number of outputs. Alternatively, our numerical computations show that the iterative solution is both feasible and faster than a direct solution in the case of the quadratic-bilinear system. We suppose that a further tuning of the iteration settings may still improve the efficiency of the used technique, which takes further investigations.
References

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

[2] A. Antoulas, Approximation of Large-Scale Dynamical Systems, SIAM Publications, 2005.

[3] P. Benner, M. Hinze, E.J.W. ter Maten (eds.), Model Reduction for Circuit Simulation, Lect. Notes in Electr. Engng. Vol. 74, Springer, 2011.

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

[5] P. Benner, P. Goyal, Balanced truncation model order reduction for quadratic-bilinear control systems. arXiv:1705.00160v1, 29 Apr 2017.

[6] R. Van Beeumen, K. Meerbergen, Model reduction by balanced truncation of linear systems with a quadratic output, in: T.E. Simons, G. Psihoyios, Ch. Tsitouras (eds.), International Conference on Numerical Analysis and Applied Mathematics (ICNAAM), 2010, pp. 2033–2036.

[7] R. Van Beeumen, K. Van Nimmen, G. Lombaert, K. Meerbergen, Model reduction for dynamical systems with quadratic output, Int. J. Numer. Meth. Engng. 91 (2012) 229-248.

[8] R. Freund, Model reduction methods based on Krylov subspaces, Acta Numerica 12 (2003) 267–319.

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

[10] S.J. Hammarling, Numerical solution of stable non-negative definite Lyapunov equation, IMA J. Numer. Anal. 2 (1982) 303–323.

[11] N.J. Higham, Analysis of the Cholesky decomposition of a semi-definite matrix, in: M.G. Cox, S.H. Hammarling (eds.), Reliable Numerical Computation, Oxford University Press, 1990, pp. 161–185.

[12] J.-R. Li, J. White, Low rank solution of Lyapunov equations, SIAM J. Matrix Anal. & Appl. 24 (2002) 260–280.

[13] B. Lohmann, R. Eid, Efficient order reduction of parametric and nonlinear models by superposition of locally reduced models, in: G. Roppenecker, B. Lohmann (eds.), Methoden und Anwendungen der Regelungstechnik. Shaker, 2009.
[14] A. Lu, E.L. Wachspress, Solution of Lyapunov equations by alternating direction implicit iteration, Computers Math. Appl. 21 (1991), 43–58.

[15] MATLAB, version 9.0.0.341360 (R2016a), The Mathworks Inc., Natick, Massachusetts, 2016.

[16] T. Penzl, LYAPACK: A MATLAB Toolbox for Large Lyapunov and Riccati Equations, Model Reduction Problems, and Linear Quadratic Optimal Control Problems, Users’ Guide (Version 1.0), 1999.

[17] R. Pulch, Model order reduction and low-dimensional representations for random linear dynamical systems. Math. Comput. Simulat. (2017) DOI 10.1016/j.matcom.2017.05.007

[18] W.H.A. Schilders, M.A. van der Vorst, J. Rommes (eds.), Model Order Reduction: Theory, Research Aspects and Applications, Mathematics in Industry Vol. 13, Springer, 2008.

[19] T. Wolf, H. Panzer, B. Lohmann, Model order reduction by approximate balanced truncation: a unifying framework, Automatisierungstechnik 61 (2013) 545–556.

[20] D. Xiu, Numerical methods for stochastic computations: a spectral method approach, Princeton University Press, 2010.