

## BALANCED TRUNCATION FOR MODEL ORDER REDUCTION OF LINEAR DYNAMICAL SYSTEMS WITH QUADRATIC OUTPUTS\*

ROLAND PULCH<sup>†</sup> AND AKIL NARAYAN<sup>‡</sup>

**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.

**Key words.** linear dynamical system, quadratic-bilinear system, model order reduction, balanced truncation, Lyapunov equation, Hankel singular values

**AMS subject classifications.** 65L05, 34C20, 93B40

**DOI.** 10.1137/17M1148797

**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, 11, 12, 30]). 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. Differential equations with quadratic outputs arise in mechanical applications like mass-spring-damper systems (see [9], for example). Maxwell's equations with quadratic outputs were examined using finite element methods in [16]. In stochastic models, the variance of a quantity of interest represents a quadratic function (cf. [14]). 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 and co-authors [7, 8] 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.

---

\*Submitted to the journal's Methods and Algorithms for Scientific Computing section September 22, 2017; accepted for publication (in revised form) April 22, 2019; published electronically July 23, 2019.

<https://doi.org/10.1137/17M1148797>

<sup>†</sup>Institute of Mathematics and Computer Science, University of Greifswald, D-17489 Greifswald, Germany (roland.pulch@uni-greifswald.de).

<sup>‡</sup>Department of Mathematics, and Scientific Computing and Imaging (SCI) Institute, University of Utah, Salt Lake City, UT 84112 (akil@sci.utah.edu).

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.

The theory of differential balancing was designed for general nonlinear dynamical systems (see [18, 29]). This concept also applies to our case of linear dynamical systems with quadratic outputs. Concerning the two involved Gramian matrices, one is constant and the other is state-dependent. Thus the numerical solution of the defining matrix-valued equations becomes too costly for an efficient MOR.

The strategy we employ is to 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, whose solutions are constant Gramian matrices. We perform a stabilization of our quadratic-bilinear system to ensure the solvability of the Lyapunov equations. When analyzing the structure of the resulting quadratic-bilinear system, MOR can be accomplished by solving just two linear Lyapunov equations. In particular, approximate methods like the alternating direction implicit (ADI) scheme can be used to solve the linear Lyapunov equations numerically [21, 23, 25, 26, 27].

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 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

$$(1) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= x(t)^\top Mx(t). \end{aligned}$$

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 small ( $n_{\text{in}} \ll n$ ). 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

$$(2) \quad x(0) = x_0$$

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 symmetric matrix  $M \in \mathbb{R}^{n \times n}$ . The system (1) is multiple-input-single-output (MISO).

*Remark 2.1.* Note that, if  $T \in \mathbb{R}^{n \times n}$  is any matrix, then

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

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).

We assume a situation with large dimension  $n$ . (See section 4 for examples of such systems.) It is in this regime when 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  $\bar{y}$ , such that the difference  $y - \bar{y}$  should be sufficiently small for all relevant times.

**2.1. Transformation to linear dynamical systems with multiple outputs.** For comparison, we adopt the approach in [8], where a linear dynamical system with a single quadratic output is transformed into a linear dynamical system with multiple (linear) outputs. We can consider again the system (1) but with a new output of interest  $z$ :

$$(3) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ z(t) &= L^\top x(t). \end{aligned}$$

The matrix  $L \in \mathbb{R}^{n \times m}$  is any matrix satisfying  $M = LL^\top$  with  $m = \text{rank}(M) \leq n$  and  $z : [0, t_{\text{end}}] \rightarrow \mathbb{R}^m$ . The initial values are as in (2). The system (3) is multiple-input-multiple-output (MIMO). The relationship between  $L$  and  $M$  immediately yields a relationship between the output  $z$  in (3) and the output  $y$  in (1):

$$(4) \quad y(t) = x(t)^\top M x(t) = (L^\top x(t))^\top (L^\top x(t)) = z(t)^\top z(t) = \|z(t)\|_2^2,$$

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

The matrix  $L$  can be identified via an arbitrary symmetric decomposition  $M = LL^\top$ . For example, if  $M$  is positive semidefinite, then the pivoted Cholesky factorization [17] provides one method of symmetric decomposition. Negative semidefinite matrices  $M$  can be replaced by  $-M$  and the negative Euclidean norm from (4) is used. In the case of indefinite symmetric matrices, an eigen-decomposition yields

$$(5) \quad M = S(D_+ - D_-)S^\top = SD_+S^\top - SD_-S^\top =: M_+ - M_-$$

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 semidefinite so that the decomposition (5) can be used to construct matrices  $L_\pm$  satisfying

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

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

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ \begin{pmatrix} z_+(t) \\ z_-(t) \end{pmatrix} &= \begin{pmatrix} L_+^\top \\ L_-^\top \end{pmatrix} x(t), \end{aligned}$$

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 (3) 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 given by a matrix  $M$  of full rank. Some MOR methods achieve accuracy only when  $\text{rank}(M) \ll n$  (see [8, p. 231], for example) 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 directly applying MOR to (3).

**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  $\dot{y} = 2\dot{x}^\top Mx$ , and using (1) results in

$$\dot{y} = 2(Ax + Bu)^\top Mx = x^\top (2A^\top M)x + u^\top (2B^\top M)x.$$

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 Remark 2.1 shows that we can replace this matrix by

$$(6) \quad S := A^\top M + M^\top A = A^\top M + MA.$$

Furthermore, let  $b_1, b_2, \dots, b_{n_{\text{in}}}$  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

$$(7) \quad \begin{aligned} \dot{\tilde{x}}(t) &= \tilde{A}\tilde{x}(t) + \tilde{B}u(t) + \sum_{j=1}^{n_{\text{in}}} u_j(t) \tilde{N}_j \tilde{x}(t) + \tilde{H}(\tilde{x}(t) \otimes \tilde{x}(t)), \\ \tilde{y}(t) &= \tilde{c}^\top \tilde{x}(t) \end{aligned}$$

with

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

for  $j = 1, \dots, n_{\text{in}}$ . The quadratic term involves the Kronecker product  $\tilde{x} \otimes \tilde{x} \in \mathbb{R}^{(n+1)^2}$ . The matrix  $\tilde{H} \in \mathbb{R}^{(n+1) \times (n+1)^2}$  can be defined in terms of the columns  $s_1, \dots, s_n \in \mathbb{R}^n$  of the symmetric matrix  $S$  from (6):

$$(8) \quad \tilde{H} = \begin{pmatrix} 0_{n \times n} & 0_n & 0_{n \times n} & 0_n & \cdots & 0_{n \times n} & 0_n & 0_{n \times n} & 0_n \\ s_1^\top & 0 & s_2^\top & 0 & \cdots & s_n^\top & 0 & 0_{1 \times n} & 0 \end{pmatrix},$$

where only the last row is occupied. Thus the quadratic-bilinear system (7) 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 (7) with

$$\tilde{x}(0) = \begin{pmatrix} x_0 \\ x_0^\top Mx_0 \end{pmatrix}.$$

The quadratic output  $y$  of (1) and the linear output  $\tilde{y}$  of (7) coincide.

Our approach is to apply MOR methods for quadratic-bilinear systems, where an advantage is that the system (7) is MISO with a low number of inputs by assumption. We will show that the specific structure of (7) allows for an efficient MOR computation. Note that our system (7) is well-defined for arbitrary output matrices  $M$ , i.e., including indefinite matrices.

**2.2.1. Relationship to tensors.** The system (7) can be related to more standard or classical definitions of quadratic dynamical systems. A quadratic dynamical system is defined by a three-dimensional tensor  $\mathcal{H} \in \mathbb{R}^{k \times k \times k}$ . In the system (7), the matrix  $\tilde{H} = \mathcal{H}^{(1)}$  represents the 1-matricization of this tensor. Since the matrix  $S$

in (6) is symmetric, its rows and columns coincide. Consequently, the 2-matricization  $\mathcal{H}^{(2)}$  and the 3-matricization  $\mathcal{H}^{(3)}$  are identical, i.e.,

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

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 (7) 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  $\mathcal{I}_{b_j} \subset \{1, \dots, n\}$  be the support indices of  $b_j$ , and let  $\mathcal{I}_M \subset \{1, \dots, n\}$  denote the subset of state variables, which are involved in the quadratic output. It holds that

$$(9) \quad \mathcal{I}_{b_j} \cap \mathcal{I}_M = \emptyset \quad \Rightarrow \quad b_j^\top M = 0.$$

If the premise of (9) is satisfied for all  $j = 1, \dots, n_{\text{in}}$ , then the bilinear term in (7) vanishes.

**2.2.3. Stabilization.** The linear term  $\tilde{A}\tilde{x}$  of the system (7) 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. Hence we stabilize via a tunable parameter  $\varepsilon > 0$ , resulting in the system

$$(10) \quad \begin{aligned} \dot{\tilde{x}}(t) &= \tilde{A}(\varepsilon)\tilde{x}(t) + \tilde{B}u(t) + \sum_{j=1}^{n_{\text{in}}} u_j(t)\tilde{N}_j\tilde{x}(t) + \tilde{H}(\tilde{x}(t) \otimes \tilde{x}(t)), \\ \tilde{y}(t) &= \tilde{C}^\top \tilde{x}(t) \end{aligned}$$

with the matrix

$$(11) \quad \tilde{A}(\varepsilon) = \begin{pmatrix} A & 0 \\ 0 & -\varepsilon \end{pmatrix}.$$

The asymptotic stability of the system (10), and thus the  $\varepsilon > 0$  assumption, is crucial in our balanced truncation method, which will be demonstrated in section 3.2.

**3. Balanced truncation.** Our goal is now to compute an MOR of the stabilized quadratic-bilinear system (10). To achieve this we apply a method of balanced truncation. The ultimate accomplishment of balanced truncation is determination of transformation matrices  $\tilde{T}_r, \tilde{T}_l \in \mathbb{R}^{(n+1) \times (n+1)}$ . Truncations of these matrices to size  $\mathbb{R}^{(n+1) \times r}$  with  $r \ll n$  are subsequently used to transform the system (10) of size  $n+1$  into an approximating system of size  $r$ . Since  $r \ll n$ , this therefore accomplishes MOR. Our ultimate identification of these transformation matrices is given by (28) in section 3.3, and an appropriate truncation is prescribed in section 3.4. An overview of the entire algorithm we propose is demonstrated in section 3.6.

The first step in a balanced truncation approach is the computation of two Gramian matrices, and our approach to compute these matrices is given in section 3.2. However, we first take a short detour in section 3.1 to illustrate why a related but alternative MOR approach, namely, the approach differential balancing, includes substantial challenges and is not attractive for our class of problems.

**3.1. Differential balancing.** For comparison against balanced truncation, we examine the concept of differential balancing from [18, 29] applied to our class of nonlinear problems. The purpose of this section is to illustrate that differential balancing presents significant challenges. In section 3.2 we will see that such challenges do not arise using our approach of balanced truncation.

**3.1.1. Gramian matrices.** We consider a general nonlinear dynamical system

$$(12) \quad \begin{aligned} \dot{x}(t) &= f(t, x(t)) + g(t, x(t))u(t), \\ y(t) &= h(t, x(t)). \end{aligned}$$

The Gramian matrices of the system (12) may depend on time and/or the state space. For a matrix-valued function  $A(t, x) = (a_{ij}) \in \mathbb{R}^{n \times n}$  and a vector-valued function  $f(t, x) \in \mathbb{R}^n$ , we use the notation  $\delta_f(A) = (\frac{\partial a_{ij}}{\partial t} + \frac{\partial a_{ij}}{\partial x} f)$ . The reachability Gramian  $P \in \mathbb{R}^{n \times n}$  associated to (12) satisfies the equations

$$(13) \quad \begin{aligned} -\delta_f(P(t, x)) + \frac{\partial f(t, x)}{\partial x} P(t, x) + P(t, x) \frac{\partial f(t, x)}{\partial x}^\top + g(t, x)g(t, x)^\top &= 0 \\ -\delta_{g_j}(P(t, x)) + \frac{\partial g_j(t, x)}{\partial x} P(t, x) + P(t, x) \frac{\partial g_j(t, x)}{\partial x}^\top &= 0 \end{aligned}$$

for  $j = 1, \dots, n_{\text{in}}$ . The observability Gramian  $Q$  is the solution of the equations

$$(14) \quad -\delta_f(Q(t, x)) + \frac{\partial f(t, x)}{\partial x}^\top Q(t, x) + Q(t, x) \frac{\partial f(t, x)}{\partial x} + \frac{\partial h(t, x)}{\partial x}^\top \frac{\partial h(t, x)}{\partial x} = 0.$$

The existence and uniqueness of solutions to the above equations is not guaranteed a priori.

**3.1.2. Linear dynamical system with quadratic output.** Only the output is nonlinear in the dynamical system (1). The following definitions of the functions in (12) yield the special case (1),

$$f(t, x) = Ax, \quad g(t, x) = B, \quad h(t, x) = x^\top Mx$$

without an explicit time-dependence. On the one hand, let  $P$  be the constant matrix solving the linear Lyapunov equation

$$(15) \quad AP + PA^\top + BB^\top = 0$$

of the linear case. It follows that the matrix  $P$  satisfies all equations (13). Hence the differential balancing coincides with the linear concept. On the other hand, the equations (14) become

$$-\delta_f(Q(x)) + A^\top Q(x) + Q(x)A + 4Mxx^\top M = 0$$

assuming a time-invariant solution. This problem is much more complicated than the linear case, since the solution still depends on the state space and partial differential equations emerge.

In [18, p. 3302], the technique of generalized differential balancing (gDB) was introduced to achieve an MOR with a reasonable computation work. This approach requires an input term  $B(t)u(t)$  and an output  $y(t) = C(t)x(t)$ , where the matrices do not depend on the state space. Hence gDB cannot be directly applied to the dynamical system (1) due to the nonlinear output.

**3.1.3. Quadratic-bilinear dynamical system.** The function definitions that cast the general system (12) into the special quadratic-bilinear system (7) are

$$f(t, \tilde{x}) = \begin{pmatrix} Ax \\ x^\top Sx \end{pmatrix}, \quad g(t, \tilde{x}) = \begin{pmatrix} B \\ 2x^\top MB \end{pmatrix}, \quad h(t, \tilde{x}) = \tilde{c}^\top \tilde{x}.$$

Again there is no explicit time-dependence. The Jacobian matrices with respect to the state space read as

$$\frac{\partial f}{\partial \tilde{x}} = \begin{pmatrix} A & 0 \\ 2x^\top S & 0 \end{pmatrix} \quad \text{and} \quad \frac{\partial g_j}{\partial \tilde{x}} = \begin{pmatrix} 0 & 0 \\ 2b_j^\top M & 0 \end{pmatrix}$$

for  $j = 1, \dots, n_{\text{in}}$ . The first part of the equations (13) becomes

$$(16) \quad -\delta_f(\tilde{P}(x)) + \begin{pmatrix} A & 0 \\ 2x^\top S & 0 \end{pmatrix} \tilde{P}(x) + \tilde{P}(x) \begin{pmatrix} A^\top & 2Sx \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} BB^\top & 2BB^\top Mx \\ 2x^\top MBB^\top & 4x^\top MBB^\top Mx \end{pmatrix} = 0$$

with a state-dependent solution  $\tilde{P}(x)$ . Let

$$(17) \quad \tilde{P}(x) = \begin{pmatrix} P & v(x) \\ v(x)^\top & w(x) \end{pmatrix}$$

with  $P \in \mathbb{R}^{n \times n}$  satisfying the Lyapunov equation (15) and  $v(x) \in \mathbb{R}^n$ ,  $w(x) \in \mathbb{R}$ . The ansatz (17) solves the left upper minor of the matrix-valued system (16). However, the other minors remain state-dependent. The other parts of the equations (13) are simpler, because the Jacobian matrices  $\frac{\partial g_j}{\partial \tilde{x}}$  are constant, but still represent differential equations. The equations (14) simplify to

$$-\delta_f(\tilde{Q}(x)) + \begin{pmatrix} A^\top & 2Sx \\ 0 & 0 \end{pmatrix} \tilde{Q}(x) + \tilde{Q}(x) \begin{pmatrix} A & 0 \\ 2x^\top S & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = 0$$

with a state-dependent solution  $\tilde{Q}(x)$ . The lower right entry shows  $(\delta_f(\tilde{Q}))_{\tilde{n}, \tilde{n}} = 1$  with  $\tilde{n} = n + 1$ , which implies that the matrix  $\tilde{Q}$  is not constant.

In conclusion, the reachability Gramian  $P$  as well as observability Gramian  $Q$  require the solution of matrix-valued differential equations, which is computationally expensive. Our approach in the following section derives constant (time- and state-independent) Gramian matrices.

If it holds that  $N_j \neq 0$  for some  $j$ , then the gDB technique cannot be applied to the quadratic-bilinear system (7), because the input part does not exhibit the simple form  $B(t)u(t)$ . If it holds that  $N_j = 0$  for all  $j$ , then gDB is feasible. The generalized differential Gramians are nonunique solutions of matrix inequalities (see [18, p. 3302]), which become constant matrices in this special case. Yet the matrix inequalities depend on the state variables and have to be satisfied for all states. Thus the complexity is still high in comparison to linear or quadratic Lyapunov equations with constant coefficients.

**3.2. Gramians of quadratic-bilinear system.** The reachability and observability Gramian matrices are defined for general quadratic-bilinear systems with multiple inputs and multiple outputs (MIMO) in [5]. The theorems on Gramian matrices

require a quadratic-bilinear system with an asymptotically stable linear part. Moreover, in [5] the existence of the Gramian matrices is assumed, i.e., the solvability of quadratic Lyapunov equations alone does not imply that a solution represents a Gramian (cf. Theorems 3.1 and 3.2 of [5]). However, the existence of these Gramians can be used to give bounds on observability and controllability energy functionals, similar to the linear case (cf. Theorems 4.1 and 4.2 of [5]). Furthermore, truncations of these Gramian matrices also provide observability and controllability estimates for reduced systems. Therefore, computation of these matrices, and subsequent truncations of them, are of fundamental importance for our MOR strategy.

The reachability Gramian  $\tilde{P} \in \mathbb{R}^{(n+1) \times (n+1)}$  associated to the stabilized system (10) is the solution of the quadratic Lyapunov equation

$$(18) \quad \tilde{A}(\varepsilon)\tilde{P} + \tilde{P}\tilde{A}(\varepsilon)^\top + \tilde{B}\tilde{B}^\top + \tilde{H}(\tilde{P} \otimes \tilde{P})\tilde{H}^\top + \sum_{j=1}^{n_{\text{in}}} \tilde{N}_j \tilde{P} \tilde{N}_j^\top = 0.$$

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

$$(19) \quad \tilde{A}(\varepsilon)^\top \tilde{Q} + \tilde{Q}\tilde{A}(\varepsilon) + \tilde{c}\tilde{c}^\top + \tilde{H}^{(2)}(\tilde{P} \otimes \tilde{Q})(\tilde{H}^{(2)})^\top + \sum_{j=1}^{n_{\text{in}}} \tilde{N}_j^\top \tilde{Q} \tilde{N}_j = 0.$$

If the reachability Gramian  $\tilde{P}$  is given, then the Lyapunov equations (19) represent a linear system for the unknown entries of the observability Gramian  $\tilde{Q}$ . Both Gramian matrices are symmetric and positive semidefinite since  $\varepsilon > 0$ .

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

LEMMA 3.1. 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)  $\tilde{N}_j \tilde{P} \tilde{N}_j^\top = 4 \begin{pmatrix} 0 & 0 \\ 0 & b_j^\top M P M b_j \end{pmatrix}$  for each  $j$ ,
- (ii)  $\sum_{j=1}^{n_{\text{in}}} \tilde{N}_j^\top \tilde{Q} \tilde{N}_j = 4q' \begin{pmatrix} M B B^\top M & 0 \\ 0 & 0 \end{pmatrix}$ ,
- (iii)  $\tilde{H}(\tilde{P} \otimes \tilde{P})\tilde{H}^\top = \text{tr}((PS)^2) \tilde{c}\tilde{c}^\top$ ,
- (iv)  $\tilde{H}^{(2)}(\tilde{P} \otimes \tilde{Q})(\tilde{H}^{(2)})^\top = q' \begin{pmatrix} S P S & 0 \\ 0 & 0 \end{pmatrix}$ .

*Proof.*

(i) We calculate directly

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

(ii) Likewise, it holds that

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

for  $j = 1, \dots, n_{\text{in}}$ . The sum over  $j$  yields the formula.



(iii) The structure (8) of  $\tilde{H}$  implies that the matrix  $\tilde{H}(\tilde{P} \otimes \tilde{P})\tilde{H}^\top$  has only one nonzero entry in the position  $(n+1, n+1)$ . The vector  $\tilde{c}$  is the  $(n+1)$ th unit vector. Hence we obtain  $\tilde{H}(\tilde{P} \otimes \tilde{P})\tilde{H}^\top = \gamma \tilde{c} \tilde{c}^\top$  with a scalar  $\gamma$  to be determined.

Let  $t_{ij}$  for  $i, j = 1, \dots, n$  be the entries of the nonsymmetric matrix  $PS$ . It holds that

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

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

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

The definition of the matrix  $\tilde{S}$  implies

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

which shows the statement.  $\square$

The left-hand sides of (i)–(iv) are terms in the Lyapunov equations (18) and (19). Direct evaluation of terms in the Lyapunov equations can be expensive, but Lemma 3.1 indicates that this effort can be significantly reduced because of the special structure of the system (10) 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 3.2.** *A reachability Gramian  $\tilde{P}$  satisfying (18) is given by*

$$(20) \quad \tilde{P} = \begin{pmatrix} P & 0 \\ 0 & p' \end{pmatrix},$$

where  $P$  solves the linear Lyapunov equation (15), i.e.,  $AP + PA^\top + BB^\top = 0$ , and

$$(21) \quad p' = \frac{1}{2\varepsilon} p'' \quad \text{with} \quad p'' = \text{tr}((PS)^2) + 4 \sum_{j=1}^{n_{\text{in}}} b_j^\top M P M b_j \geq 0.$$

An observability Gramian  $\tilde{Q}$  satisfying (19) is given by

$$(22) \quad \tilde{Q} = \frac{1}{2\varepsilon} \begin{pmatrix} Q & 0 \\ 0 & 1 \end{pmatrix},$$

where  $Q$  satisfies the linear Lyapunov equations

$$(23) \quad A^\top Q + QA + SPS + 4MBB^\top M = 0.$$

*Proof.* Inserting the ansatz (20) into the quadratic Lyapunov equations (18) yields the linear Lyapunov equations (15) for the first part. Due to Lemma 3.1(i) and (iii), the second part becomes

$$-2\varepsilon p' + \operatorname{tr}((PS)^2) + 4 \sum_{j=1}^{n_{\text{in}}} b_j^\top M P M b_j = 0,$$

which uniquely defines the scalars  $p'$  and  $p''$ , respectively.

Now the ansatz (22) is inserted into the Lyapunov equations (19). The second part is fulfilled immediately due to  $-2\varepsilon \cdot \frac{1}{2\varepsilon} + 1 = 0$ . Lemma 3.1(ii) and (iv) yield the linear Lyapunov equations

$$A^\top \frac{1}{2\varepsilon} Q + A \frac{1}{2\varepsilon} Q + \frac{1}{2\varepsilon} S P S + 4 \frac{1}{2\varepsilon} M B B^\top M = 0$$

as the first part. A multiplication by the factor  $2\varepsilon$  results in the Lyapunov equations (23).

Finally, we show the lower bound  $p'' \geq 0$  in (21). The solution  $P$  of the Lyapunov equation (15) is always symmetric and positive semidefinite. The matrix  $S$  is just symmetric. It follows that  $b_j^\top M P M b_j \geq 0$  for all  $j = 1, \dots, n_{\text{in}}$ . Using the matrix square root  $P = P^{\frac{1}{2}} P^{\frac{1}{2}}$ , we obtain

$$\operatorname{tr}((PS)^2) = \operatorname{tr}(P^{\frac{1}{2}} P^{\frac{1}{2}} S P S) = \operatorname{tr}(P^{\frac{1}{2}} S P S P^{\frac{1}{2}}) = \operatorname{tr}((S P^{\frac{1}{2}})^\top P (S P^{\frac{1}{2}}))$$

due to a property of the trace. Obviously, the matrix  $(S P^{\frac{1}{2}})^\top P (S P^{\frac{1}{2}})$  is symmetric and positive semidefinite again. Hence its trace is nonnegative.  $\square$

We have shown the existence of symmetric positive semidefinite solutions satisfying the quadratic Lyapunov equations (18) and (19). The proof of Theorem 3.2 demonstrates that the matrices (20) and (22) are the unique solutions of the Lyapunov equations in the set of block-diagonal matrices of the used form. We just assume that the solutions are also unique in the set of all matrices. All these properties have to be assumed in the case of general quadratic-bilinear systems (cf. [6, p. 13]).

Theorem 3.2 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), (23) are independent of the stabilization parameter  $\varepsilon$ . The observability Gramian (22) is directly proportional to  $\frac{1}{\varepsilon}$ . This  $1/\varepsilon$  dependence suggests that the Lyapunov equations cannot be solved if we set  $\varepsilon = 0$ . We codify this fact below.

**COROLLARY 3.3.** *If  $\varepsilon = 0$  and  $P^{\frac{1}{2}} S P^{\frac{1}{2}} \neq 0$ , then the Lyapunov equation (18) does not have a solution. If  $\varepsilon = 0$ , then the Lyapunov equation (19) does not have a solution.*

*Proof.* Let  $\varepsilon = 0$  in the matrix (11). We investigate the component  $(n+1, n+1)$  in each quadratic Lyapunov equation. In the Lyapunov equation (18), this component yields  $p'' = 0$  with  $p''$  defined in (21). We showed that  $p'' \geq 0$  in the proof of Theorem 3.2. A necessary condition for  $p'' = 0$  is  $P^{\frac{1}{2}} S P S P^{\frac{1}{2}} = 0$ , which is excluded due to  $(P^{\frac{1}{2}} S P^{\frac{1}{2}})^2 \neq 0$ . In the Lyapunov equation (19), the left-hand side of this component becomes 1 due to Lemma 3.1(ii) and (iv). This contradicts the fact that this component on the right-hand side of (19) must be 0.  $\square$

**3.3. 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 of (15) and (23), respectively. From Theorem 3.2, 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

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

We remark that one need not actually form the Gramian matrices in order to compute low-rank Cholesky factors [15, 20]. In the method of balanced truncation, we require the singular value decomposition (SVD)

$$(24) \quad \tilde{L}_Q^\top \tilde{L}_P = \tilde{U} \tilde{\Sigma} \tilde{V}^\top.$$

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

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

We use the SVD

$$(25) \quad L_Q^\top L_P = U \Sigma V^\top,$$

where the diagonal matrix  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$  includes the singular values in *ascending* order  $\sigma_1 \leq \sigma_2 \leq \dots \leq \sigma_n$ . The SVD (25) is independent of the stabilization parameter  $\varepsilon$ . Thus the SVD component matrices in (24) are given by

$$(26) \quad \tilde{U} = \begin{pmatrix} U & 0 \\ 0 & 1 \end{pmatrix}, \quad \tilde{V}^\top = \begin{pmatrix} V^\top & 0 \\ 0 & 1 \end{pmatrix}, \quad \tilde{\Sigma} = \frac{1}{\sqrt{2\varepsilon}} \begin{pmatrix} \Sigma & 0 \\ 0 & \sqrt{\frac{p''}{2\varepsilon}} \end{pmatrix}.$$

If the stabilization parameter  $\varepsilon$  is sufficiently small, then the maximum singular value is  $\sqrt{p''}/(2\varepsilon)$ . However, the associated singular vector is independent of  $p''$  and  $\varepsilon$ . The singular values of the quadratic-bilinear system (7) are

$$(27) \quad \frac{1}{\sqrt{2\varepsilon}} \left( \sigma_1, \sigma_2, \dots, \sigma_n, \sqrt{\frac{p''}{2\varepsilon}} \right).$$

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 in

$$(28) \quad \begin{aligned} \tilde{T}_l &= \tilde{L}_Q \tilde{U} \tilde{\Sigma}^{-\frac{1}{2}} = \frac{1}{\sqrt[4]{2\varepsilon}} \begin{pmatrix} L_Q U \Sigma^{-\frac{1}{2}} & 0 \\ 0 & \sqrt[4]{\frac{2\varepsilon}{p''}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt[4]{2\varepsilon}} L_Q U \Sigma^{-\frac{1}{2}} & 0 \\ 0 & p''^{-\frac{1}{4}} \end{pmatrix}, \\ \tilde{T}_r &= \tilde{L}_P \tilde{V} \tilde{\Sigma}^{-\frac{1}{2}} = \sqrt[4]{2\varepsilon} \begin{pmatrix} L_P V \Sigma^{-\frac{1}{2}} & 0 \\ 0 & \sqrt[4]{\frac{p''}{2\varepsilon}} \end{pmatrix} = \begin{pmatrix} \sqrt[4]{2\varepsilon} L_P V \Sigma^{-\frac{1}{2}} & 0 \\ 0 & p''^{\frac{1}{4}} \end{pmatrix} \end{aligned}$$

with  $p''$  from (21). It holds that  $\tilde{T}_l^\top \tilde{T}_r = I$  with the identity matrix  $I$ . For  $\varepsilon \rightarrow 0$  some parts of the matrices converge to zero and the other parts tend to infinity, and hence the limits do not exist. However, the balanced system can be written in a form which allows for further interpretations.

LEMMA 3.4. *The balanced system of dimension  $n + 1$  reads as*

$$(29) \quad \begin{aligned} \dot{\bar{x}}(t) &= \bar{A}(\varepsilon)\bar{x}(t) + \bar{B}(\varepsilon)u(t) + \sum_{j=1}^{n_{\text{in}}} u_j(t)\bar{N}_j(\varepsilon)\bar{x}(t) + \bar{H}(\varepsilon)(\bar{x}(t) \otimes \bar{x}(t)), \\ \bar{y}(t) &= \bar{c}^\top \bar{x}(t) \end{aligned}$$

with

$$\bar{A}(\varepsilon) = \begin{pmatrix} \bar{A}' & 0 \\ 0 & -\varepsilon \end{pmatrix}, \quad \bar{B}(\varepsilon) = \frac{1}{\sqrt[4]{2\varepsilon}} \begin{pmatrix} \bar{B}' \\ 0 \end{pmatrix}, \quad \bar{N}_j(\varepsilon) = \sqrt[4]{2\varepsilon} \begin{pmatrix} 0 & 0 \\ \bar{N}'_j & 0 \end{pmatrix}, \quad \bar{H}(\varepsilon) = \sqrt{2\varepsilon} \bar{H}'$$

and  $\bar{A}', \bar{B}', \bar{N}'_j, \bar{H}', \bar{c}$  independent of  $\varepsilon$ .

*Proof.* It holds that  $\bar{c}^\top = \tilde{c}^\top \tilde{T}_r = (0, \dots, 0, \sqrt[4]{p''})^\top$ . We obtain  $\bar{A}(\varepsilon) = \tilde{T}_1^\top \tilde{A}(\varepsilon) \tilde{T}_r$  and  $\bar{B}(\varepsilon) = \tilde{T}_1^\top \tilde{B}$ . It follows that

$$\bar{A}' = \left(\Sigma^{-\frac{1}{2}}\right)^\top U^\top L_Q^\top A L_P V \Sigma^{-\frac{1}{2}} \quad \text{and} \quad \bar{B}' = \left(\Sigma^{-\frac{1}{2}}\right)^\top U^\top L_Q^\top B.$$

The matrices of the bilinear part become

$$\bar{N}_j(\varepsilon) = \tilde{T}_1^\top \tilde{N}_j(\varepsilon) \tilde{T}_r = \sqrt[4]{2\varepsilon} \begin{pmatrix} 0 & 0 \\ 2p''^{-\frac{1}{4}} b_j^\top M L_P V \Sigma^{-\frac{1}{2}} & 0 \end{pmatrix} \quad \text{for } j = 1, \dots, n_{\text{in}}.$$

The quadratic part exhibits the structure (8). It holds that  $\bar{H}(\varepsilon) = \tilde{T}_1^\top \tilde{H}(\tilde{T}_r \otimes \tilde{T}_r)$ . We obtain  $\tilde{T}_1^\top \tilde{H} = p''^{-\frac{1}{4}} \tilde{H}$  due to the structure of  $\tilde{T}_1$ . Let

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

It follows that  $\bar{H}(\varepsilon) = \sqrt{2\varepsilon} p''^{-\frac{1}{4}} \tilde{H}(\tilde{T}'_r \otimes \tilde{T}'_r)$ , where a factor  $\sqrt[4]{2\varepsilon}$  comes from each  $\tilde{T}_r$  and thus  $\sqrt[4]{2\varepsilon} \cdot \sqrt[4]{2\varepsilon} = \sqrt{2\varepsilon}$ .  $\square$

Lemma 3.4 shows that the differential equations of the balanced system can be decoupled into the parts

$$(30) \quad \begin{aligned} \dot{\bar{x}}^*(t) &= \bar{A}' \bar{x}^*(t) + \bar{B}' \left( \frac{1}{\sqrt[4]{2\varepsilon}} u(t) \right) \\ \dot{\bar{x}}_{n+1}(t) &= -\varepsilon \bar{x}_{n+1}(t) + \sum_{j=1}^{n_{\text{in}}} u_j(t) \bar{N}'_j(\sqrt[4]{2\varepsilon} \bar{x}^*(t)) + \bar{H}''((\sqrt[4]{2\varepsilon} \bar{x}^*(t)) \otimes (\sqrt[4]{2\varepsilon} \bar{x}^*(t))) \end{aligned}$$

with a row vector  $\bar{H}''$ . Initial values are transformed via

$$\bar{x}(0) = \tilde{T}_r^{-1} \tilde{x}(0) = \begin{pmatrix} \frac{1}{\sqrt[4]{2\varepsilon}} \Sigma^{\frac{1}{2}} V^\top L_P^{-1} x_0 \\ p''^{-\frac{1}{4}} x_0^\top M x_0 \end{pmatrix}.$$

It follows that the solution  $\bar{x}^*$  is directly proportional to  $\frac{1}{\sqrt[4]{2\varepsilon}}$  (change in input signal as well as change in initial values). For  $0 < \varepsilon < \frac{1}{2}$ , this amplification is canceled by

multiplication of  $\bar{x}^*$  with the factor  $\sqrt[4]{2\varepsilon}$  in the last equation of (30). It follows that the system

$$(31) \quad \begin{aligned} \dot{\bar{x}}^*(t) &= \bar{A}'\bar{x}^*(t) + \bar{B}'u(t), \\ \dot{\bar{x}}_{n+1}(t) &= -\varepsilon\bar{x}_{n+1}(t) + \sum_{j=1}^{n_{\text{in}}} u_j(t)\bar{N}'_j\bar{x}^*(t) + \bar{H}''(\bar{x}^*(t) \otimes \bar{x}^*(t)) \end{aligned}$$

with initial values  $\bar{x}^*(0) = \Sigma^{\frac{1}{2}}V^\top L_P^{-1}x_0$  exhibits the same solution  $\bar{x}_{n+1}$  as in the system (30). Now (31) includes the parameter  $\varepsilon$  only in the scalar term of the last equation. We may arrange  $\varepsilon \rightarrow 0$  to eliminate the dependence on the stabilization parameter completely.

**3.4. 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. In contrast to the linear case, error estimates are not available in the case of quadratic-bilinear systems yet.

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

$$(32) \quad \sigma_{n+1} := \sqrt{\frac{p''}{2\varepsilon}} > \sigma_{n-r}.$$

Hence  $\sigma_{n+1}$  belongs to the set of the  $r$  dominant singular values. We partition the SVD (24) into

$$(33) \quad \tilde{L}_Q^\top \tilde{L}_P = (\tilde{U}_1 \quad \tilde{U}_2) \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & \tilde{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \tilde{V}_1^\top \\ \tilde{V}_2^\top \end{pmatrix}$$

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  $\tilde{T}_{1/r} \in \mathbb{R}^{(n+1) \times r}$  read as

$$(34) \quad \tilde{T}_1 = \tilde{L}_Q \tilde{U}_2 \tilde{\Sigma}_2^{-\frac{1}{2}} \quad \text{and} \quad \tilde{T}_r = \tilde{L}_P \tilde{V}_2 \tilde{\Sigma}_2^{-\frac{1}{2}}.$$

Due to the ascending order of the singular values and the condition (32), 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 (10) becomes

$$(35) \quad \begin{aligned} \dot{\bar{x}}(t) &= \bar{A}\bar{x}(t) + \bar{B}u(t) + \sum_{j=1}^{n_{\text{in}}} u_j(t)\bar{N}_j\bar{x}(t) + \bar{H}(\bar{x}(t) \otimes \bar{x}(t)), \\ \bar{y}(t) &= \bar{c}^\top \bar{x}(t) \end{aligned}$$

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

$$(36) \quad \begin{aligned} \bar{A} &= \tilde{T}_1^\top \tilde{A} \tilde{T}_r, & \bar{B} &= \tilde{T}_1^\top \tilde{B}, & \bar{c}^\top &= \tilde{c}^\top \tilde{T}_r, \\ \bar{N}_j &= \tilde{T}_1^\top \tilde{N}_j \tilde{T}_r, & \bar{H} &= \tilde{T}_1^\top \tilde{H} (\tilde{T}_r \otimes \tilde{T}_r). \end{aligned}$$

The output vector is

$$\bar{c}^\top = (0, \dots, 0, \sqrt[4]{p''}).$$

Thus the output is just a multiple of the final state variable as in the quadratic-bilinear system (10). Initial values  $\bar{x}(0)$  have to be determined from (2). The balanced truncation method preserves the local asymptotic stability of the equilibrium  $\tilde{x}_{\text{eq}} = 0$  in autonomous systems (7) with  $u \equiv 0$  (see [5]).

The computational effort for the matrices  $\tilde{N}_j$  in (36) is negligible because only the last row of  $\tilde{N}_j$  is nonzero. The computation of the matrix  $\tilde{H} \in \mathbb{R}^{r \times r^2}$  represents the most expensive part in the projections (36). In [4, p. 245], an algorithm is outlined to construct  $\tilde{H}$  without calculating the Kronecker product  $\tilde{T}_r \otimes \tilde{T}_r$  explicitly. In our case, the effort becomes even lower, since the matrix  $\tilde{H}$  exhibits the structure (8) of  $\tilde{H}$ . The entries of the matrix

$$(37) \quad \bar{S} = \tilde{T}_r^\top S \tilde{T}_r \in \mathbb{R}^{r \times r},$$

with the symmetric matrix  $S$  from (6), yield the last row of the intermediate matrix  $\tilde{H}(\tilde{T}_r \otimes \tilde{T}_r)$ , whereas the other rows are zero. Thus the effort mainly consists in the computation of matrix-matrix products in (37).

*Remark 3.5.* We obtain

$$(38) \quad \bar{A} = \begin{pmatrix} \bar{A}^* & 0 \\ 0 & -\varepsilon \end{pmatrix}$$

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  $\varepsilon \rightarrow 0$  in the matrix (38).

The ROM (35) exhibits the same structure as the quadratic-bilinear system (10) in the nonlinear terms.

**THEOREM 3.6.** *Let  $\bar{x} = (\bar{x}_1, \dots, \bar{x}_r)^\top$  and  $\bar{x}^* = (\bar{x}_1, \dots, \bar{x}_{r-1})^\top$ . The reduced system (35) has the equivalent form*

$$(39) \quad \begin{aligned} \dot{\bar{x}}^*(t) &= \bar{A}^* \bar{x}^*(t) + \bar{B}^* u(t), \\ \dot{\bar{x}}_r(t) &= -\varepsilon \bar{x}_r(t) + \sum_{j=1}^{n_{\text{in}}} u_j(t) \bar{N}_j^* \bar{x}^*(t) + \bar{H}^*(\bar{x}^*(t) \otimes \bar{x}^*(t)), \\ \bar{y}(t) &= \sqrt[p'']{\bar{p}''} \bar{x}_r(t) \end{aligned}$$

with  $p''$  from (21), the matrix  $\bar{A}^*$  from (38), and a modified matrix  $\bar{B}^*$  and modified row vectors  $\bar{N}_j^*, \bar{H}^*$ .

The proof is straightforward. The structure of the system (39) 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.

Moreover, the matrix  $\bar{H}^*$  in (39) has the structure (8). We collect its nonzero entries in a symmetric matrix  $\bar{S}^* \in \mathbb{R}^{(r-1) \times (r-1)}$ . Now let  $\varepsilon = 0$ . In view of (6), we consider the symmetric solution  $\bar{M}^*$  of the Lyapunov equation

$$\bar{A}^{*\top} \bar{M}^* + \bar{M}^* \bar{A}^* = \bar{S}^*.$$

If the linear dynamical system with quadratic output

$$(40) \quad \begin{aligned} \dot{\bar{x}}^*(t) &= \bar{A}^* \bar{x}^*(t) + \bar{B}^* u(t), \\ \bar{y}(t) &= \bar{x}^*(t)^\top \bar{M}^* \bar{x}^*(t) \end{aligned}$$

is transformed into a quadratic-bilinear system as in section 2.2, then the quadratic term of system (39) exactly appears. However, the bilinear part becomes different, even if all matrices  $\tilde{N}_j$  (and thus  $\tilde{N}_j^*$ ) were zero. Hence the structure of the original system (1) cannot be retrieved by this MOR. Only in the autonomous case ( $u \equiv 0$ ), the dynamical systems (39) and (40) are equivalent ( $\tilde{y} = \sqrt[4]{p''}\tilde{y}$ ).

Lemma 3.4 implies that the stabilization parameter  $\varepsilon$  influences only the scalar term of the last differential equation in the ROM (39). If we choose  $\varepsilon = 0$  in (38) or, equivalently, in the scalar term of (39), then the output becomes independent of the parameter. The value  $\varepsilon$  just has to be sufficiently small such that the last singular value in (27) belongs to the dominant singular values used to determine the ROM.

**3.5. Low-rank approximations.** An MOR for the linear dynamical system (1) with quadratic output can be performed by using the linear dynamical system (3) with multiple outputs or the quadratic-bilinear system (7) with single output. Two criteria determine the efficiency of the approaches in balanced truncation:

1. The decay of the 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 (23). A general linear Lyapunov equation reads as

$$(41) \quad AG + GA^\top + F = 0$$

for the unknown matrix  $G \in \mathbb{R}^{n \times n}$  with a given symmetric positive semidefinite matrix  $F \in \mathbb{R}^{n \times n}$ . If we apply direct methods of linear algebra to solve (41), then the computational complexity is  $O(n^3)$  and nearly independent of the rank of  $F$ . Consequently, we could solve the linear dynamical system (3) 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 (41). Efficient algorithms are achieved by iterations based on the alternating direction implicit (ADI) technique (see [21, 23]). A low-rank approximation reads as  $G \approx Z_G Z_G^\top$  with  $Z_G \in \mathbb{R}^{n \times k}$  for some  $k \ll n$ . An ADI technique requires a symmetric factorization  $F \approx Z_F Z_F^\top$  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 [26, p. 10], the property  $k_F \ll n$  is assumed in the ADI method.

Concerning both systems (3) and (10), an iterative computation of the reachability Gramian can be easily devised because  $F = BB^\top$  and thus  $k_F = n_{\text{in}}$  due to our assumption of a low number of inputs. We obtain a low-rank approximation  $P \approx Z_P Z_P^\top$  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 (21).

However, the observability Gramian requires a factorization  $F \approx Z_F Z_F^\top$  in the case of the linear dynamical system (3), where the rank may be large (possibly close to  $n$ ). In the case of the quadratic-bilinear system (10), the input matrix for (23) becomes

$$(42) \quad \hat{F} = SPS + 4MBB^\top M \approx (SZ_P)(SZ_P)^\top + 4(MB)(MB)^\top.$$

Now we obtain directly an approximate factorization of (42) by

$$(43) \quad Z_{\hat{F}} \approx ( (SZ_P), (2MB) ) \in \mathbb{R}^{n \times (k_P + n_{\text{in}})},$$

where the number of columns is low since  $k_P, n_{\text{in}} \ll n$ . The rank of the factor (43) may be smaller than  $k_P + n_{\text{in}}$ , which allows for a simplification to a full-rank factor  $Z_{\hat{F}} \in \mathbb{R}^{n \times k_{\hat{F}}}$  for some  $k_{\hat{F}} < k_P + n_{\text{in}}$ . Furthermore, a reduced factor  $Z_{\hat{F}}$  with  $k_{\hat{F}}$  columns can be obtained by using just the first  $k'_P < k_P$  columns of the factor  $Z_P$ . An iterative scheme solving (23) yields the factorization  $Q \approx Z_Q Z_Q^\top$  with  $Z_Q \in \mathbb{R}^{n \times k_Q}$  for the observability Gramian. For a general Lyapunov equation (41) with  $F = Z_F Z_F^\top$  and  $Z_F \in \mathbb{R}^{n \times k_F}$ ,  $j$  iterations of the ADI method generate a factor with  $jk_F$  columns (see [26]).

The balanced truncation approach using approximate low-rank factors represents a well-known strategy (see [13]). The detailed formulas can be found in [31], for example. We reduce to a dimension  $r < n$ , assuming the condition (32) 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), (23). Symmetric decompositions  $\tilde{P} \approx \tilde{Z}_P \tilde{Z}_P^\top$  and  $\tilde{Q} \approx \tilde{Z}_Q \tilde{Z}_Q^\top$  for the matrices from (20), (22) read as

$$\tilde{Z}_P = \begin{pmatrix} Z_P & 0 \\ 0 & \sqrt{\frac{p''}{2\varepsilon}} \end{pmatrix} \quad \text{and} \quad \tilde{Z}_Q = \frac{1}{\sqrt{2\varepsilon}} \begin{pmatrix} Z_Q & 0 \\ 0 & 1 \end{pmatrix}.$$

Now we compute an SVD of the small matrix  $\tilde{Z}_Q^\top \tilde{Z}_P \in \mathbb{R}^{(k_Q+1) \times (k_P+1)}$ . A partition (33) of this SVD is used again assuming an ascending order of the singular values. We suppose that the condition (32) is also satisfied in this approximation. The projection matrices for the reduction (36) are

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

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.

**3.6. Algorithm overview.** We summarize here the main steps of our method. We recall that we have rewritten the original MISO system (1) into an equivalent quadratic-bilinear system with a single output (10). The following steps provide an algorithm for the construction of an ROM in form of a small quadratic-bilinear system:

1. Solve the linear Lyapunov equation (15) for  $P$ , and subsequently the linear Lyapunov equation (23) for  $Q$ , where  $S$  is defined in (6). This is the most costly portion of the MOR procedure, and in our numerical experiments we will solve these equations using two approaches: (i) direct linear algebraic methods and (ii) approximate iterative methods, namely, ADI iteration.
2. Compute factorizations  $P = L_P L_P^\top$  and  $Q = L_Q L_Q^\top$ , if not already obtained within the solution procedure of step 1.
3. Choose a small stabilization parameter  $\varepsilon > 0$  and compute  $p''$  using (21).
4. Compute the SVD (25) of  $L_Q^\top L_P$ .
5. Assemble the matrices from (26) used to form the SVD in (24).
6. Given some rank  $r \ll n$ , perform the partition (33), and subsequently form  $\tilde{T}_l$  and  $\tilde{T}_r$  in (34).
7. Construct the size- $r$  reduced system (35) via the matrices in (36).

In our overview above, we omit the technical details of many straightforward algebraic manipulations, which can be employed to substantially reduce the cost of direct



computation in some of the steps. For example, one needs not explicitly assemble the full matrices in (26) and just the dominant part of the SVD (25) is required.

**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 (3) with multiple outputs by direct algorithms of linear algebra,
- (ii) the quadratic-bilinear system (10) with single output by direct algorithms of linear algebra,
- (iii) the quadratic-bilinear system (10) with single output using ADI iteration.

The numerical computations were performed by the software package MATLAB [24] (version R2016a), where the machine precision is around  $\varepsilon_0 = 2 \cdot 10^{-16}$ . We used the ADI algorithm from the Matrix Equation Sparse Solver (M.E.S.S.) toolbox in MATLAB (see [25]). We note that alternative iterative approaches can effectively solve the Lyapunov equations encountered in this paper (for example, low-rank rational Krylov methods [10, 19]). We focus on ADI methods in this paper for simplicity, but acknowledge that alternative and perhaps better choices exist that could improve the reported performance of our approach in this section. 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.,

$$(44) \quad E_{\text{abs}} = \max_{t \in [0, T]} |\bar{y}(t) - y(t)| \quad \text{and} \quad E_{\text{rel}} = \frac{1}{T} \int_0^T \frac{|\bar{y}(t) - y(t)|}{|y(t)|} dt$$

with  $y$  from a full-order model (FOM) and  $\bar{y}$  from an ROM. On the one hand, the absolute error measures the maximum pointwise discrepancy between the FOM trajectory and the ROM trajectory. On the other hand, the relative error measures the discrepancy between these two averaged over the trajectory. Since  $y$  can have values close to zero, we expect that the relative error can be large compared to the absolute error. Our numerical results will observe this.

**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 pseudorandom 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  $B = (1, 1, \dots, 1)^\top$ . 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

$$(45) \quad u(t) = \sin(k(t)t) \quad \text{with} \quad k(t) = k_0 t$$

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

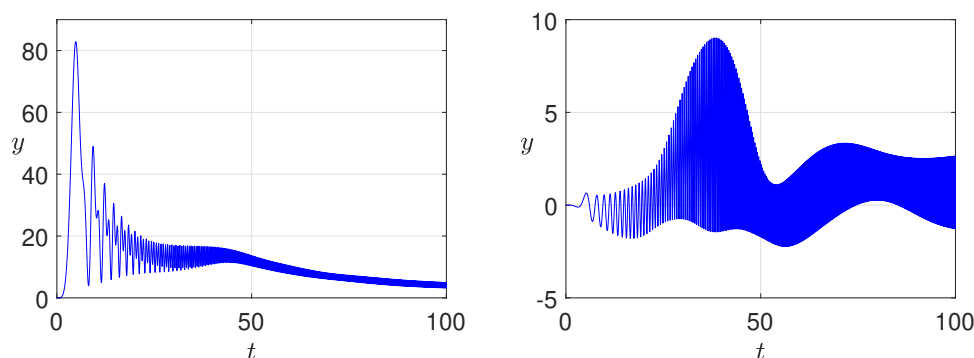


FIG. 1. Quadratic outputs of the linear dynamical systems for the identity matrix (left) and indefinite matrix (right).

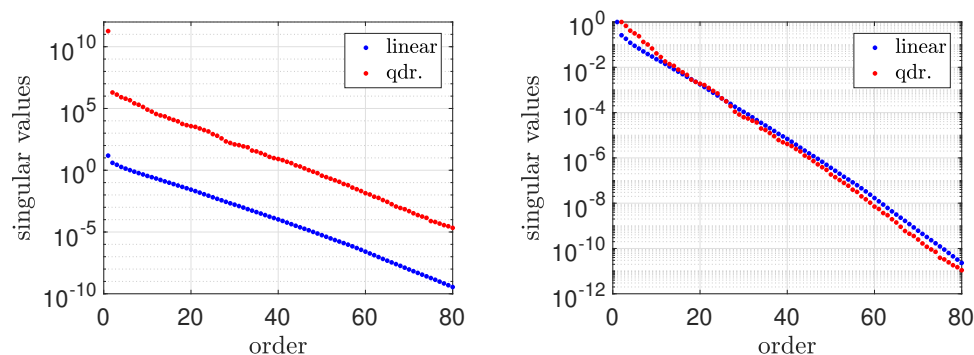


FIG. 2. Singular values (left) and their normalized values (right) for the two dynamical systems associated to the identity output matrix.

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 (3), where it holds that  $L^\top = I$ . Hence the number of outputs is equal to  $n$  in (3). The balanced truncation technique yields the Hankel singular values in the linear case. On the other hand, we derive the quadratic-bilinear system (10) including the stabilization parameter  $\varepsilon = 10^{-8}$ . The balanced truncation scheme from section 3.3 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 (10) has a special role (see (27)). Thus we normalize the first singular value of (3) and the second singular value of (10) 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 (28) 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 (3) and from the quadratic-bilinear system (10) for  $r = 5, 6, \dots, 80$ . In the quadratic-bilinear case, we choose  $\varepsilon = 0$  only

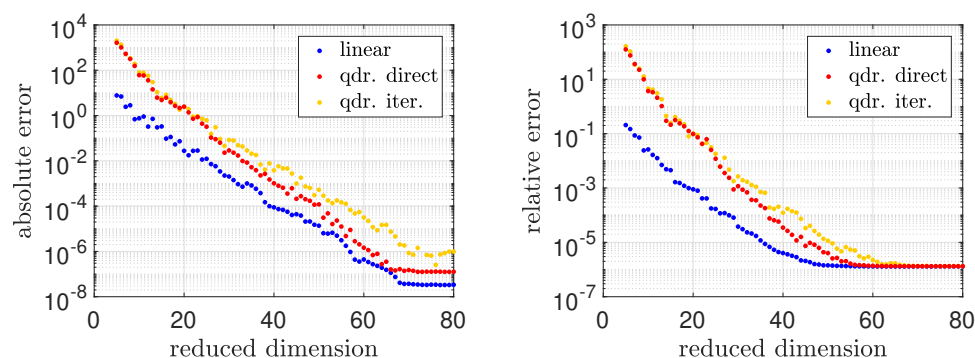


FIG. 3. Maximum absolute errors (left) and mean relative errors (right) for ROMs of different dimensions in the case of the identity output matrix.

in the reduction of the matrix (38), because numerical results show that the errors become smaller as  $\varepsilon \rightarrow 0$  in this matrix.

Furthermore, we solve the linear Lyapunov equations (15) and (23) iteratively by an ADI method for each  $r$  separately as described in section 3.5. 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 (23) 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 (43). Each pair of iterative solutions implies an associated ROM.

The transient simulation of the FOM (1) 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 (44) 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 (3) in comparison to the quadratic-bilinear system (10), 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 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 seconds. 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.

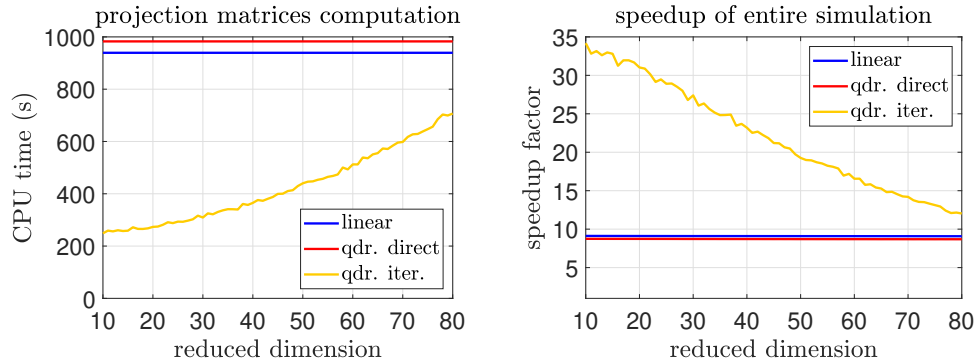


FIG. 4. CPU times for computation of projection matrices (left) and total speedups (right) in the case of the identity output matrix.

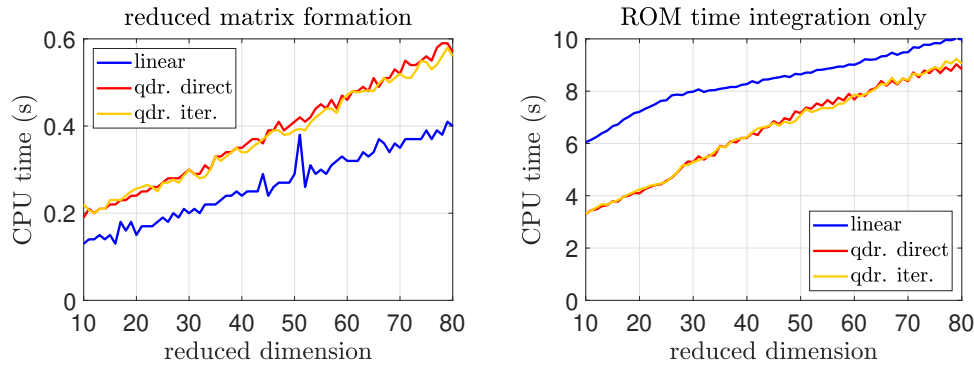


FIG. 5. CPU times for computation of reduced matrices (left) and transient simulation of ROMs (right) in the case of the identity output matrix.

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 speedup 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.

We conclude that the quadratic-bilinear approach using iterative solvers can achieve error comparable to the linear approach with substantially increased computational efficiency. For a fixed reduced dimension, choosing the quadratic approach decreases accuracy when compared to the linear approach; this accuracy drop is slightly exacerbated by use of an iterative solver for computing the Gramian matrices (cf. Figure 3). However, the quadratic bilinear approach using an iterative solver is significantly faster than the linear approach. It is so efficient that one can compute a quadratic-bilinear reduced-order model of significantly increased rank (and hence accuracy) for a fixed computational budget. For example, to achieve a relative error of approximately  $10^{-5}$ , we require a reduced dimension of approximately  $r \approx 50$  for the quadratic-bilinear case (see Figure 3 (right)). However, even at reduced rank 50, the

TABLE 1

Maximum difference between outputs from ROMs of dimension  $r = 20$  for different stabilization parameters  $\varepsilon$  to reference value  $\varepsilon = 10^{-8}$  (differences are rounded to one digit).

Value $\varepsilon$	$10^{-1}$	$10^{-2}$	$10^{-3}$	$10^{-4}$	$10^{-5}$	$10^{-6}$	$10^{-7}$
Difference	$4 \cdot 10^{-8}$	$3 \cdot 10^{-8}$	$8 \cdot 10^{-9}$	$5 \cdot 10^{-9}$	$2 \cdot 10^{-8}$	$2 \cdot 10^{-8}$	$2 \cdot 10^{-8}$

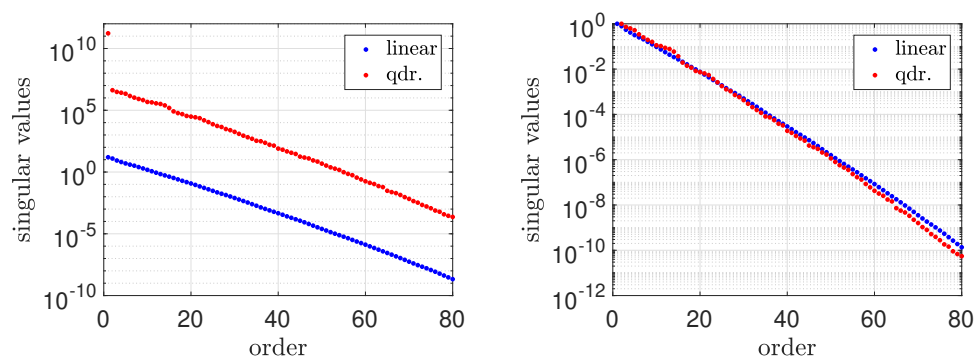


FIG. 6. Singular values (left) and their normalized values (right) for the two dynamical systems with respect to the indefinite output matrix.

quadratic-bilinear iterative approach is still approximately twice as fast as the linear or quadratic-bilinear direct approaches (see Figure 4 (right)).

Finally, we investigate the choice of different stabilization parameters  $\varepsilon$ . The focus is on ROMs of dimension  $r = 20$ , where the transient outputs are computed by the time integration described above. The value  $\varepsilon = 0$  is used in the reduced matrices (38) again. We compute the maximum difference in time between the outputs for different  $\varepsilon$  to the reference value  $\varepsilon = 10^{-8}$ . Table 1 depicts the numerical results. The differences are tiny and exhibit the same order or magnitude for all  $\varepsilon$ . This property confirms the theoretical results in section 3.3, which imply that the ROMs are independent of  $\varepsilon$  except for the scalar entry in (38).

**4.2. Indefinite output matrix.** We arrange a linear dynamical system (1) 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 pseudorandom numbers). We fill a matrix  $M' \in \mathbb{R}^{n \times n}$  by pseudorandom numbers associated to a uniform distribution in  $[-1, 1]$ . Now the output matrix  $M := \frac{1}{2}(M' + M'^T)$  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. Thus the construction of the output matrix in the linear dynamical system (3) requires the complete eigen-decomposition of  $M$ . We apply the chirp signal (45) 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 (3) with  $n$  outputs and the quadratic-bilinear system (10) 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

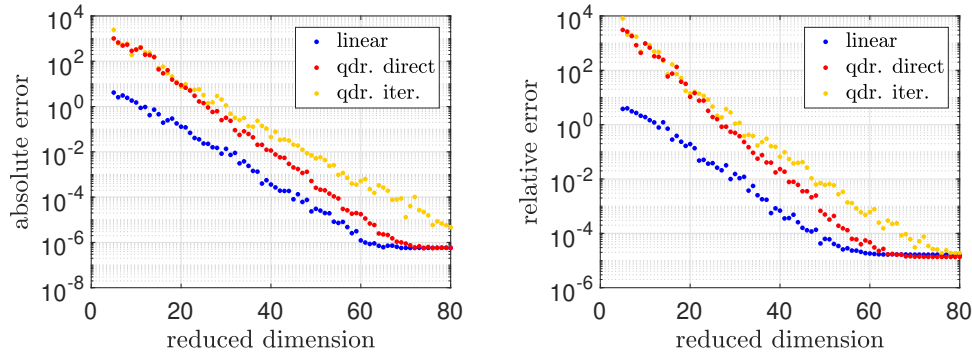


FIG. 7. Maximum absolute errors (left) and mean relative errors (right) for ROMs of different dimensions in the case of an indefinite output matrix.

values are zero. Figure 7 shows the resulting discrete approximations of the error measures (44). 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 [22]. The associated linear dynamical system consists of 8 ordinary differential equations including 14 physical parameters and is single-input-single-output. In [28], this test example was investigated in the context of stochastic modeling, 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 [32]).

The stochastic Galerkin approach yields a larger linear dynamical system (single-input-multiple-output)

$$(46) \quad \begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ w(t) &= Cx(t), \end{aligned}$$

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

$$(47) \quad \text{Var}(t) = \sum_{i=2}^m w_i(t)^2.$$

The details of the above modeling can be found in [28].

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 (46) 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 (1) from the stochastic Galerkin system (46), whose quadratic output is the variance (47). Define  $L^\top \in \mathbb{R}^{(m-1) \times n}$  as the matrix  $C \in \mathbb{R}^{m \times n}$  in (46) with its first row omitted. It follows that  $M = LL^\top$

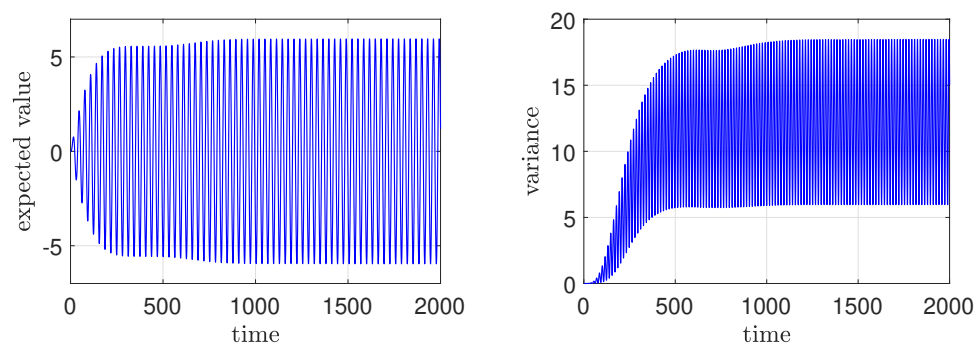


FIG. 8. Expected value (left) and variance (right) of random output in the mass-spring-damper configuration.

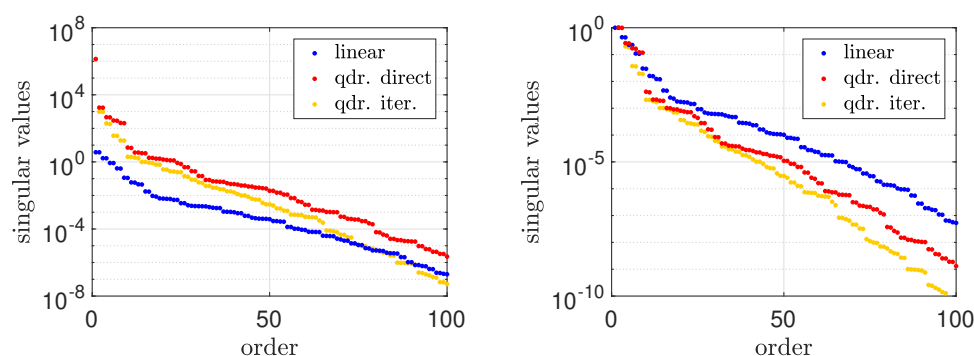


FIG. 9. Singular values (left) and their normalized values (right) for the two dynamical systems in the mass-spring-damper example.

in (1) is symmetric and positive semidefinite of rank  $m - 1$ . Thus the equivalent system (3) with  $m - 1$  linear outputs defined by  $L^T$  is already available in this application.

The associated quadratic-bilinear system (7) is without bilinearity, because the property (9) is satisfied. We apply the stabilized system (10) with a parameter  $\varepsilon = 10^{-8}$  again.

We examine the MOR by balanced truncation for this problem comparing the reduction of the linear system (3) and the quadratic system (10). 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  $r_{\max} = 100$  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 in each of the three reductions, which are the Hankel singular values in the linear case. Figure 9 illustrates the dominant singular values in descending order. We recognize a faster decay of the singular values in the quadratic system (7). 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.



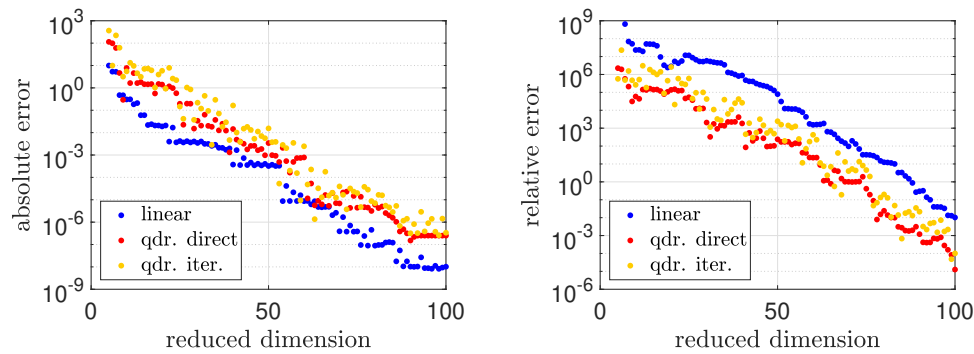


FIG. 10. Maximum absolute errors (left) and mean relative errors (right) for ROMs of different dimensions in the mass-spring-damper example.

Given the projection matrices with  $r_{\max}$  columns, we choose the dominant part to obtain ROMs of dimension  $r = 5, 6, \dots, 100$ . We arrange  $\varepsilon = 0$  in the matrix (38) 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  $5 \cdot 10^5$  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 3.6, nonlinear systems of algebraic equations are omitted in the quadratic ROMs (39). 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. The errors decrease exponentially in each approach. The absolute errors decay exponentially until reduced dimension  $r \approx 90$  and stagnate thereafter. Our tests suggest that this stagnation is due to the accuracy of the time integration method. Including more time steps in the integration routine would remove this stagnation. Furthermore, the iteration technique produces approximations of the same quality as the direct approach. 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 MOR 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 MOR approaches implement the method of balanced truncation for each of these two recast systems. Balanced truncation requires solutions to Lyapunov equations. We find that the manipulation of large matrices necessary to solve the Lyapunov equations motivates 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 MOR 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 one 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, AND I. JAIMOUKHA, *Krylov subspace methods for model reduction of quadratic bilinear systems*, IET Control Theory Appl., 10 (2016), pp. 2010–2018.
- [2] A. ANTOULAS, *Approximation of Large-Scale Dynamical Systems*, SIAM, Philadelphia, 2005.
- [3] P. BENNER, M. HINZE, AND E. J. W. TER MATEN, EDS., *Model Reduction for Circuit Simulation*, Lect. Notes Electr. Eng. 74, Springer, New York, 2011.
- [4] P. BENNER AND T. BREITEN, *Two-sided projection methods for nonlinear model order reduction*, SIAM J. Sci. Comput., 37 (2015), pp. B239–B260.
- [5] P. BENNER AND P. GOYAL, *Balanced Truncation Model Order Reduction for Quadratic-Bilinear Control Systems*, arXiv:1705.00160v1, 2017.
- [6] P. BENNER, P. GOYAL, AND S. GUGERCIN,  *$H_2$ -quasi-optimal model order reduction for quadratic-bilinear control systems*, SIAM J. Matrix Anal. Appl., 39 (2018), pp. 983–1032.
- [7] R. VAN BEEUMEN AND K. MEERBERGEN, *Model reduction by balanced truncation of linear systems with a quadratic output*, in Proceedings of the International Conference on Numerical Analysis and Applied Mathematics, T. E. Simons, G. Psihoyios, and Ch. Tsitouras, eds., 2010, pp. 2033–2036.
- [8] R. VAN BEEUMEN, K. VAN NIMMEN, G. LOMBAERT, AND K. MEERBERGEN, *Model reduction for dynamical systems with quadratic output*, Internat. J. Numer. Methods Engrg., 91 (2012), pp. 229–248.
- [9] C.A. DEPKEN, *The Observability of Systems with Linear Dynamics and Quadratic Output*, Ph.D. thesis, Georgia Institute of Technology, 1971.
- [10] V. DRUSKIN, L. KNIZHNERMAN, AND V. SIMONCINI, *Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation*, SIAM J. Numer. Anal., 49 (2011), pp. 1875–1898.
- [11] R. FREUND, *Model reduction methods based on Krylov subspaces*, Acta Numer., 12 (2003), pp. 267–319.
- [12] S. GUGERCIN, A.C. ANTOULAS, AND C. BEATTIE,  *$H_2$  model reduction for large-scale linear dynamical systems*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 609–638.
- [13] S. GUGERCIN AND J. R. LI, *Smith-type methods for balanced truncation of large sparse systems*, in Dimension Reduction of Large-Scale Systems, P. Benner, D.C. Sorensen, and V. Mehrmann, eds., Springer, New York, 2003, pp. 49–82.
- [14] B. HAASDONK, K. URBAN, AND B. WIELAND, *Reduced basis methods for parameterized partial differential equations with stochastic influences using the Karhunen–Loève expansion*, SIAM/ASA J. Uncertain. Quantif., 1 (2013), pp. 79–105.
- [15] S. J. HAMMARLING, *Numerical solution of stable non-negative definite Lyapunov equation*, IMA J. Numer. Anal., 2 (1982), pp. 303–323.
- [16] M. HAMMERSCHMIDT, S. HERRMANN, J. POMPLUN, L. ZSCHIEDRICH, S. BURGER, AND F. SCHMIDT, *Reduced basis method for Maxwell’s equations with resonance phenomena*, in Proceedings of SPIE 9630, 2015.
- [17] N. J. HIGHAM, *Analysis of the Cholesky decomposition of a semi-definite matrix*, in Reliable Numerical Computation, M. G. Cox and S. H. Hammarling, eds., Oxford University Press, Oxford, 1990, pp. 161–185.
- [18] Y. KAWANO AND J. M. A. SCHERPEN, *Model reduction by differential balancing based on non-linear Hankel operators*, IEEE Trans. Automat. Control, 62 (2017), pp. 3293–3308.
- [19] D. KOLESNIKOV AND I. OSELEDETS, *From low-rank approximation to a rational Krylov subspace method for the Lyapunov equation*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1622–1637.
- [20] P. KÜRSCHNER, *Efficient Low-Rank Solution of Large-Scale Matrix Equations*, Ph.D. thesis, Otto-von-Guericke Universität Magdeburg, Shaker Verlag, Herzogenrath, Germany, 2016.
- [21] J.-R. LI AND J. WHITE, *Low rank solution of Lyapunov equations*, SIAM J. Matrix Anal. Appl., 24 (2002), pp. 260–280.
- [22] B. LOHMANN AND R. EID, *Efficient order reduction of parametric and nonlinear models by superposition of locally reduced models*, in Methoden und Anwendungen der Regelungstechnik, G. Roppenecker and B. Lohmann, eds., Shaker Verlag, Herzogenrath, 2009.
- [23] A. LU AND E. L. WACHSPRESS, *Solution of Lyapunov equations by alternating direction implicit iteration*, Comput. Math. Appl., 21 (1991), pp. 43–58.
- [24] *MATLAB, version 9.0.0.341360 (R2016a)*, MathWorks, Natick, MA, 2016.

- [25] J. SAAK, M. KÖHLER, AND P. BENNER, *M-M.E.S.S.-1.0.1—The Matrix Equations Sparse Solvers Library*, doi:10.5281/zenodo.50575, 2016.
- [26] 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.
- [27] T. PENZL, *A cyclic low-rank Smith method for large sparse Lyapunov equations*, SIAM J. Sci. Comput., 21 (2000), pp. 1401–1418.
- [28] R. PULCH, *Model order reduction and low-dimensional representations for random linear dynamical systems*, Math. Comput. Simulation, 144 (2018), pp. 1–20.
- [29] J. M. A. SCHERPEN, *Balancing for nonlinear systems*, Systems Control Lett., 21 (1993), pp. 143–153.
- [30] W. H. A. SCHILDERS, M. A. VAN DER VORST, AND J. ROMMES, EDS., *Model Order Reduction: Theory, Research Aspects and Applications*, Math. Ind. 13, Springer, New York, 2008.
- [31] T. WOLF, H. PANZER, AND B. LOHMANN, *Model order reduction by approximate balanced truncation: A unifying framework*, Automatisierungstechnik, 61 (2013), pp. 545–556.
- [32] D. XIU, *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press, Princeton, NJ, 2010.