Advances in Computational Mathematics (2024) 50:88
https://doi.org/10.1007/s10444-024-10186-9

Balanced truncation for quadratic-bilinear control systems

Peter Benner[1,2] · **Pawan Goyal**[1]

Received: 3 May 2023 / Accepted: 17 July 2024

Abstract
We discuss model order reduction (MOR) for large-scale quadratic-bilinear (QB) systems based on balanced truncation. The method for linear systems mainly involves the computation of the Gramians of the system, namely reachability and observability Gramians. These Gramians are extended to a general nonlinear setting in Scherpen (Systems Control Lett. **21**, 143-153 1993). These formulations of Gramians are not only challenging to compute for large-scale systems but hard to utilize also in the MOR framework. This work proposes algebraic Gramians for QB systems based on the underlying Volterra series representation of QB systems and their Hilbert adjoint systems. We then show their relation to a certain type of generalized quadratic Lyapunov equation. Furthermore, we quantify the reachability and observability subspaces based on the proposed Gramians. Consequently, we propose a balancing algorithm, allowing us to find those states that are simultaneously hard to reach and hard to observe. Truncating such states yields reduced-order systems. We also study sufficient conditions for the existence of Gramians, and a local stability of reduced-order models obtained using the proposed balanced truncation scheme. Finally, we demonstrate the proposed balancing-type MOR for QB systems using various numerical examples.

Keywords Model order reduction · Balanced truncation · Reachability and observability · Hilbert adjoint operator and Lyapunov stability

Mathematics Subject Classification (2010) 93A15 · 93C10 · 93C15 · 37E99 · 93B05 · 93B07

Communicated by: Tobias Breiten

✉ Pawan Goyal
goyalp@mpi-magdeburg.mpg.de

Peter Benner
benner@mpi-magdeburg.mpg.de

[1] Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, Magdeburg 39106, Germany

[2] Faculty of Mathematics, Otto von Guericke University Magdeburg, Universitätsplatz 2, Magdeburg 39106, Germany

Published online: 08 August 2024

1 Introduction

Numerical simulation is a primary tool to study dynamical systems, e.g., for prediction and design studies. High-fidelity modeling is an essential step to gain deep insight into the behavior of complex dynamical systems. Even though computational resources have been developing extensively over the last few decades, fast numerical simulation of such high-fidelity systems, whose number of state variables can easily be of order $\mathcal{O}(10^5 - 10^6)$, is still a huge computational burden. This makes the usage of these large-scale systems very difficult and inefficient, for instance, in optimization and control design. One approach to mitigate this problem is *model order reduction* (MOR). MOR seeks to substitute large-scale dynamical systems with low-dimensional (reduced-order) systems such that the input-output behaviors of both original and reduced-order systems are close enough, and the reduced-order systems preserve some important properties, for example, stability and passivity of the original system, allowing to use the reduced-order models as a surrogate in the control design process.

MOR techniques for linear systems are well-established and are widely applied in various application areas, see, e.g., [1–5]. In many applications where the dynamics are governed by nonlinear partial differential equations (PDEs), such as Navier-Stokes equations, a linear system can also be obtained via linearization of the system around a suitable expansion point, e.g., the steady-state solution so that the linearized system captures the dynamics very well locally. However, as it moves away from the expansion point, the linearized system might not be able to capture the system dynamics accurately. Therefore, there is often the need to take nonlinear terms into consideration, thus resulting in a more accurate system. Consider a nonlinear system of the form

$$\begin{aligned} \dot{x}(t) &= f(x(t)) + g(x(t))u(t), \qquad x(0) = x_0, \\ y(t) &= h(x(t)), \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^m$ and $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ are nonlinear smooth functions, and $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ denote the state, input, and output vectors of the system at time t, respectively. In this work, we assume homogeneous initial conditions, i.e., $x_0 = 0$. The main goal of MOR is to construct a low-dimensional system, having a similar form as system (1). That is as follows:

$$\begin{aligned} \dot{\widehat{x}}(t) &= \widehat{f}(\widehat{x}(t)) + \widehat{g}(\widehat{x}(t))u(t), \qquad \widehat{x}(0) = 0, \\ \widehat{y}(t) &= \widehat{h}(\widehat{x}(t)), \end{aligned} \tag{2}$$

in which $\widehat{f} : \mathbb{R}^{\widehat{n}} \to \mathbb{R}^{\widehat{n}}$, $\widehat{g} : \mathbb{R}^{\widehat{n}} \to \mathbb{R}^{\widehat{n}} \times \mathbb{R}^m$ and $\widehat{h} : \mathbb{R}^{\widehat{n}} \times \mathbb{R}^m \to \mathbb{R}^p$ with $\widehat{n} \ll n$ so that the output of both systems are close to each other if both systems are excited by the same input signal, for any feasible input function.

MOR techniques for general nonlinear systems, namely trajectory-based MOR techniques, have been widely applied in the literature to determine reduced-order systems for nonlinear systems; see, e.g., [6–8]. The proper orthogonal decomposition (POD) method is a very powerful trajectory-based MOR technique, which depends on a Galerkin projection $\mathcal{P} = VV^T$, where V is a projection matrix such that $x(t) \approx V\widehat{x}(t)$. The nonlinear functions $\widehat{f}(\widehat{x})$ can be given as $\widehat{f}(\widehat{x}(t)) = V^T f(V\widehat{x}(t))$, and similar

expressions can also be derived for $\widehat{g}(\widehat{x}(t))$ and $\widehat{h}(\widehat{x}(t))$. This method preserves the structure of the original system in the reduced-order system, but the reduced-order system still requires the computation of the nonlinear functions on the full grid. This may obstruct the success of MOR; however, there are many advanced hyperreduction methodologies, see, e.g., [9] and therein references. In recent years, reduced basis methods have been successfully applied to nonlinear systems to obtain reduced-order systems [10, 11].

In this article, we consider a certain class of nonlinear control systems, namely quadratic-bilinear (QB) control systems. The advantage of this class of nonlinear systems is that systems containing smooth mono-variate nonlinearities such as exponentials and polynomial functions can also be rewritten in a QB form by introducing some new variables in the state vector [12]. Note that this transformation is exact, i.e., it requires no approximation and does not introduce any approximation error, but this transformation may not be unique.

Related to MOR for QB systems, the idea of one-sided moment-matching has been extended from linear or bilinear systems to QB systems; see, e.g., [12–16], where a reduced system is determined by capturing the input-output behavior of the original system, given by generalized transfer functions. More recently, there have been extensions to two-sided moment-matching in [17–20], ensuring more moments to be matched for a given order of the reduced system. Even though these methods have evolved as effective MOR techniques for nonlinear systems, shortcomings of these methods, however, are choosing the appropriate order of the reduced system and selecting good interpolation points. Moreover, the applicability of the two-sided moment-matching method [20] is limited to single-input single-output QB systems, and the stability of the obtained reduced-order system is a major issue in this method. Furthermore, the construction of $\mathcal{H}_2$-optimal approach to choose the optimal interpolation points and the corresponding tangential directions has been discussed in [21], but this requires to have the order of reduced models a priori.

Here, our focus instead lies on balancing-type MOR techniques for QB systems. This technique mainly depends on the reachability and observability energy functionals, or in other words, the Gramians of the system. This method is presented for linear systems, e.g., in [1, 22], and later on, a theory of balancing for general nonlinear systems is developed in a sequence of papers [23–27]. In the general nonlinear case, the balancing requires the solutions of the state-dependent nonlinear Hamilton-Jacobi equation, which are, firstly, very expensive to solve for large-scale dynamical systems; secondly, it is not straightforward to use them in the MOR context. Along with these, it may happen that the reduced-order systems obtained from nonlinear balancing do not preserve the structure of the original nonlinearities. However, for some weakly nonlinear systems, the so-called bilinear systems, reachability, and observability Gramians have been studied in [28–32], which are solutions to generalized algebraic Lyapunov equations. Moreover, these Gramians, when used to define appropriate quadratic forms, approximate energy functionals of bilinear systems (in the neighborhood of the origin), see [29, 30].

In this direction of balanced truncation for quadratic systems, our first goal is to develop reachability and observability Gramians for these systems, which are state-independent matrices and suitable for the MOR purpose. Additionally, we show how

these Gramians can describe the reachable and observable subspaces, providing motivation for MOR. In Section 2, we propose the reachability Gramian for QB systems based on the underlying Volterra series of the system. Additionally, we determine the observability Gramian based on the dual system associated with the QB system. Furthermore, we establish relations between the solutions of a certain type of quadratic Lyapunov equations and these Gramians. We also discuss a truncated version of Gramians using the leading integral-kernels of the Volterra series. In Section 3, we discuss a connection between the proposed Gramians and reachable/observable subspaces. Consequently, we utilize these Gramians for balancing QB systems, allowing us to determine those states that are hard to reach as well as hard to observe. Truncation of such states leads to reduced systems. In Section 4, we theoretically analyze sufficient conditions for the existence of these Gramians, and the stability of these reduced systems obtained using truncated Gramians. In Section 5, we test the efficiency of the proposed balanced truncation MOR technique for various semi-discretized nonlinear PDEs and compare it with moment-matching techniques for QB systems.

We note that this manuscript is based on our preprint [33] and mostly reflects our findings by 2017. Later on, some more works on balanced truncation of QB systems has appeared that cite our preprint, see, e.g., [34–36]. Here, we do not discuss these works further, but would like to mention that they have advanced the theory and applicability of balanced truncation of QB systems further.

2 Quadratic-Bilinear systems and their Gramians

In this work, we consider quadratic-bilinear systems of the following form:

$$\dot{x}(t) = Ax(t) + H\,(x(t) \otimes x(t)) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \quad x(0) = 0, \tag{3a}$$

$$y(t) = Cx(t), \tag{3b}$$

where $A, N_k \in \mathbb{R}^{n\times n}$, $H \in \mathbb{R}^{n\times n^2}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$. Furthermore, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^p$ denote the state, input, and output vectors of the system, respectively, and '$\otimes$' denotes the Kronecker product. In this section, we aim at determining algebraic Gramians for QB systems, which can also be related to the reachable and observable subspaces of the QB systems. We begin by deriving the reachability Gramian of the QB system and by showing its connection with a certain type of quadratic Lyapunov equation.

2.1 Reachability Gramian for QB systems

To derive the reachability Gramian, we first formulate the Volterra series for the QB system (3). Before we proceed further, for ease, we define the following short-hand notation:

$$u^{(k)}_{(t_1,\ldots,t_l)}(t) := u_k(t - t_1 \cdots - t_l) \quad \text{and} \quad x_{(t_1,\ldots,t_l)}(t) := x(t - t_1 \cdots - t_l),$$

where u_k denotes the k-th element of the input vector u. Adopting the analysis for bilinear systems from [37] to QB systems, we can write the solution $x(t)$ as follows:

$$\begin{aligned} x(t) = &\int_0^t e^{At_1} B u_{t_1}(t) dt_1 + \sum_{k=1}^{m} \int_0^t e^{At_1} N_k x_{t_1}(t) u_{t_1}^{(k)}(t) dt_1 \\ &+ \int_0^t e^{At_1} H \left(x_{t_1}(t) \otimes x_{t_1}(t) \right) dt_1. \end{aligned} \tag{4}$$

Based on the above equation, we can obtain an expression for $x_{t_1}(t)$ as follows:

$$\begin{aligned} x_{t_1}(t) = &\int_0^{t-t_1} e^{At_2} B u_{(t_1,t)2}(t) dt_2 + \sum_{k=1}^{m} \int_0^{t-t_1} e^{At_2} N_k x_{(t_1,t_2)}(t) u_{(t_1,t_2)}^{(k)}(t) dt_2 \\ &+ \int_0^{t-t_1} e^{At_2} H \left(x_{(t_1,t_2)}(t) \otimes x_{(t_1,t_2)}(t) \right) dt_2, \end{aligned}$$

and substitute it in (4) to get

$$\begin{aligned} x(t) = &\int_0^t e^{At_1} B u_{t_1}(t) dt_1 + \sum_{k=1}^{m} \int_0^t \int_0^{t-t_1} e^{At_1} N_k e^{At_2} B u_{t_1}^{(k)}(t) u_{(t_1,t_2)}(t) dt_1 dt_2 \\ &+ \int_0^t \int_0^{t-t_1} \int_0^{t-t_1} e^{At_1} H(e^{At_2} B \otimes e^{At_3} B) \left(u_{(t_1,t_2)}(t) \otimes u_{(t_1,t_3)}(t) \right) dt_1 dt_2 dt_3 \\ &+ \cdots. \end{aligned} \tag{5}$$

Repeating this process of substituting for the state yields the Volterra series for the QB system [38]. Having analyzed the *integral-kernels* of the Volterra series for the QB system, we define the reachability mapping $\bar{P}$ as follows:

$$\bar{P} = [\bar{P}_1, \ \bar{P}_2, \ \bar{P}_3, \dots], \tag{6}$$

where the $\bar{P}_i$'s are:

$$\bar{P}_1(t_1) = e^{At_1} B, \tag{7a}$$

$$\bar{P}_2(t_1, t_2) = e^{At_2} \left[N_1, \dots, N_m \right] \left(I_m \otimes \bar{P}_1(t_1) \right), \tag{7b}$$

$$\vdots \qquad \vdots$$

$$\begin{aligned} \bar{P}_i(t_1, \dots, t_i) = e^{At_i} \Big[&H\big[\bar{P}_1(t_1) \otimes \bar{P}_{i-2}(t_2, \dots, t_{i-1}), \ \bar{P}_2(t_1, t_2) \otimes \bar{P}_{i-3}(t_3, \dots, t_{i-1}), \\ &\dots, \bar{P}_{i-2}(t_1, \dots, t_{i-2}) \otimes \bar{P}_1(t_{i-1}) \big], \\ &\left[N_1, \dots, N_m \right] \left(I_m \otimes \bar{P}_{i-1}(t_1, \dots, t_{i-1}) \right) \Big], \ \forall \ i \geq 3. \end{aligned} \tag{7c}$$

Using the mapping $\bar{P}$ from (6), we define the reachability Gramian P as follows:

$$P = \sum_{i=1}^{\infty} P_i \quad \text{with} \quad P_i = \int_0^{\infty} \cdots \int_0^{\infty} \bar{P}_i(t_1, \ldots, t_i) \bar{P}_i^T(t_1, \ldots, t_i) \mathrm{d}t_1 \cdots \mathrm{d}t_i, \tag{8}$$

assuming that the series defining P converges and all improper integrals exit.

In what follows, we show the equivalence between the above-proposed reachability Gramian and a solution of a certain type of quadratic Lyapunov equation.

Theorem 1 *Consider the QB system (3) with a stable matrix A, i.e., all the eigenvalues of the matrix A strictly lies in the negative half plane. If the reachability Gramian P of the system defined as in (8) exists, meaning the sum of the infinite series in (8) converges, then it satisfies the generalized quadratic Lyapunov equation, given by*

$$AP + PA^T + H(P \otimes P)H^T + \sum_{k=1}^{m} N_k P N_k^T + BB^T = 0. \tag{9}$$

Proof We begin by considering the first term in the summation (8). This is,

$$P_1 = \int_0^{\infty} \bar{P}_1(t_1) \bar{P}_1^T(t_1) \mathrm{d}t_1 = \int_0^{\infty} e^{At_1} BB^T e^{A^T t_1} \mathrm{d}t_1.$$

As shown, e.g., in [1], P_1 satisfies the following Lyapunov equation, provided A is stable:

$$AP_1 + P_1 A^T + BB^T = 0. \tag{10}$$

Next, we consider the second term in the summation (8):

$$\begin{aligned}
P_2 &= \int_0^{\infty} \int_0^{\infty} \bar{P}_2(t_1, t_2) \bar{P}_2^T(t_1, t_2) \mathrm{d}t_1 \mathrm{d}t_2 \\
&= \int_0^{\infty} \int_0^{\infty} e^{At_2} \left[N_1, \ldots, N_m\right] \left(I_m \otimes \left(e^{At_1} BB^T e^{A^T t_1}\right)\right) \left[N_1, \ldots N_m\right]^T e^{A^T t_2} \mathrm{d}t_1 \mathrm{d}t_2 \\
&= \sum_{k=1}^{m} \int_0^{\infty} e^{At_2} N_k \Big(\int_0^{\infty} e^{At_1} BB^T e^{A^T t_1} \mathrm{d}t_1 \Big) N_k^T e^{A^T t_2} \mathrm{d}t_1 \mathrm{d}t_2 \\
&= \sum_{k=1}^{m} \int_0^{\infty} e^{At_2} N_k P_1 N_k^T e^{A^T t_2} \mathrm{d}t_2.
\end{aligned}$$

Again using the integral representation of the solution to Lyapunov equations [1], we see that P_2 is the solution of the following Lyapunov equation:

$$AP_2 + P_2 A^T + \sum_{k=1}^{m} N_k P_1 N_k^T = 0. \tag{11}$$

Finally, we consider the ith term, for $i \geq 3$, which is

$$\begin{aligned} P_i &= \int_0^\infty \cdots \int_0^\infty \bar{P}_i(t_1, \ldots, t_i) \bar{P}_i^T(t_1, \ldots, t_i) \mathrm{d}t_1 \cdots \mathrm{d}t_i \\ &= \int_0^\infty e^{At_i} \Bigg[H \Bigg[\int_0^\infty \mathcal{F}\left(\bar{P}_1(t_1)\right) \mathrm{d}t_1 \otimes \int_0^\infty \cdots \int_0^\infty \mathcal{F}\left(\bar{P}_{i-2}(t_2, \ldots, t_{i-1})\right) \mathrm{d}t_2 \cdots \mathrm{d}t_{i-1} \\ &\quad + \cdots + \int_0^\infty \cdots \int_0^\infty \mathcal{F}\left(\bar{P}_{i-2}(t_1, \ldots, t_{i-2})\right) \mathrm{d}t_1 \cdots \mathrm{d}t_{i-2} \otimes \int_0^\infty \mathcal{F}\left(\bar{P}_1(t_{i-1})\right) \mathrm{d}t_{i-1} \Bigg] H^T \\ &\quad + \sum_{k=1}^m N_k \left(\int_0^\infty \cdots \int_0^\infty \mathcal{F}\left(\bar{P}_{i-1}(t_1, \ldots, t_{i-1})\right) \right) N_k^T \Bigg] e^{A^T t_i} \mathrm{d}t_i, \end{aligned}$$

where we use the shorthand $\mathcal{F}(\mathcal{A}) := \mathcal{A}\mathcal{A}^T$. Thus, we have

$$P_i = \int_0^\infty e^{At_i} \Big[H(P_1 \otimes P_{i-2} + \cdots + P_{i-2} \otimes P_1) H^T + \sum_{k=1}^m N_k P_{i-1} N_k^T \Big] e^{A^T t_i} \mathrm{d}t_i.$$

Similar to P_1 and P_2, we can show that P_i satisfies the following Lyapunov equation, given in terms of the preceding P_k, for $k = 1, \ldots, i-1$:

$$A P_i + P_i A^T + H(P_1 \otimes P_{i-2} + \cdots + P_{i-2} \otimes P_1) H^T + \sum_{k=1}^m N_k P_{i-1} N_k^T = 0. \quad (12)$$

Furthermore, let us define

$$P^{(L)} := \sum_{i=1}^L P_i,$$

which satisfies

$$A P^{(L)} + P^{(L)} A^T + H \mathcal{X}_P^{(L)} H^T + \sum_{k=1}^m N_k P^{(L-1)} N_k^T + B B^T = 0, \quad (13)$$

where

$$\mathcal{X}_P^{(L)} := \sum_{j=1}^{L-2} \sum_{i=1}^{L-2} P_i \otimes P_j \quad \text{with} \quad i + j \leq L - 1.$$

We know that the Gramian $P = \sum_{i=1}^\infty P_i = \lim_{L \to \infty} P^{(L)}$. Thus, with noting $\lim_{L \to \infty} \mathcal{X}_P^{(L)} = P \otimes P$, we obtain

$$A P + P A^T + H (P \otimes P) + \sum_{k=1}^m N_k P N_k^T + B B^T = 0.$$

Hence, it satisfies the generalized quadratic Lyapunov equation stated in (9). □

Remark 1 We note that it might be the case that there exist multiple $P \succeq 0$ satisfying the quadratic Lyapunov equation (9). In this case, we should consider the solution which also satisfy the definition (8). Later, in Subsection 4.1, we discuss a fixed point scheme, which, under certain conditions, can yield a solution for the quadratic Lyapunov equation that also satisfy the definition (8).

2.2 Dual system and observability Gramian for QB systems

We next derive the dual system for the QB system since it plays an important role in determining the observability Gramian for the QB system (3). Based on it, we seek to determine the observability Gramian in a similar fashion as done for the reachability Gramian in the preceding subsection. From linear and bilinear systems, we know that the observability Gramian is the reachability Gramian of the dual system. Here also, we consider the same analogy.

Using [39, Corollary 1], we first write down the state-space realization of the adjoint operator of the QB system as follows:

$$\begin{aligned}
\dot{x}(t) &= Ax(t) + H(x(t) \otimes x(t)) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \quad x(0) = 0, \\
\dot{z}(t) &= -A^T z(t) - (x(t)^T \otimes I) H^T z(t) - \sum_{k=1}^{m} N_k^T z(t) u_k(t) - C^T u^{(d)}(t), \; z(\infty) = 0, \\
y^{(d)}(t) &= B^T z(t),
\end{aligned} \tag{14}$$

where $z(t) \in \mathbb{R}^n$, $u^{(d)}(t) \in \mathbb{R}$ and $y^{(d)} \in \mathbb{R}$ can be interpreted as the dual state, dual input, and dual output vectors of the system, respectively.

Moreover, using techniques from tensor algebra, we have the following relation:

$$(x(t) \otimes I) H^T z(t) = \mathcal{H}^{(2)} (x(t) \otimes z(t)),$$

where $\mathcal{H}^{(2)}$ is the mode-2 matricization of a tensor $\mathcal{H} \in \mathbb{R}^{n \times n \times n}$ with $\mathcal{H}$ being such that its mode-1 matricization is H. For more details on tensor algebra, we refer to [20, 40]. Hence, we can rewrite the system (14) as:

$$\dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \qquad x(0) = 0, \tag{15a}$$

$$\dot{z}(t) = -A^T z(t) - \mathcal{H}^{(2)} (x(t) \otimes z(t)) - \sum_{k=1}^{m} N_k^T u_k(t) z(t) - C^T u^{(d)}(t), \; z(\infty) = 0, \tag{15b}$$

$$y^{(d)}(t) = B^T z(t). \tag{15c}$$

Now, we focus on determining the observability Gramian for the QB system by utilizing the state-space realization of the Hilbert adjoint operator (dual system). For

this, we follow the same steps used to determine the reachability Gramian. Using the dual system (15), one can write the dual state $z(t)$ of the dual system at time t as follows:

$$z(t) = \int_{\infty}^{t} e^{-A^T(t-t_1)} C^T u^{(d)}(t_1) dt_1 + \sum_{k=1}^{m} \int_{\infty}^{t} e^{-A^T(t-t_1)} N_k^T z(t_1) u_k(t_1) dt_1 + \int_{\infty}^{t} e^{-A^T(t-t_1)} \mathcal{H}^{(2)} \left(x(t_1) \otimes z(t_1)\right) dt_1,$$

which, after an appropriate change of variables, leads to

$$\begin{aligned} z(t) = & \int_{\infty}^{0} e^{A^T t_1} C^T u^{(d)}(t + t_1) dt_1 \\ & + \sum_{k=1}^{m} \int_{\infty}^{0} e^{A^T t_1} N_k^T z(t + t_1) u_k(t + t_1) dt_1 \\ & + \int_{\infty}^{0} e^{A^T t_1} \mathcal{H}^{(2)} \big(x(t + t_1) \otimes z(t + t_1)\big) dt_1. \end{aligned} \tag{16}$$

Eqn. (15a) gives the expression for $x(t + t_1)$. This is

$$\begin{aligned} x(t + t_1) = & \int_{0}^{t+t_1} e^{At_2} B u(t + t_1 - t_2) dt_2 + \sum_{k=1}^{m} \int_{0}^{t+t_1} \Big(e^{At_2} N_k x(t + t_1 - t_2) \\ & \times u_k(t + t_1 - t_2) \Big) dt_2 + \int_{0}^{t+t_1} e^{At_2} H(x(t + t_1 - t_2) \otimes x(t + t_1 - t_2)) dt_2. \end{aligned}$$

We substitute for $x(t+t_1)$ using the above equation, and $z(t+t_1)$ using (16), which gives rise to the following expression:

$$\begin{aligned} z(t) = & \int_{\infty}^{0} e^{A^T t_1} C^T u^{(d)}(t + t_1) dt_1 + \sum_{k=1}^{m} \int_{\infty}^{0} \int_{\infty}^{0} e^{A^T t_1} N_k^T \\ & \times e^{A^T t_2} C^T u^{(d)}(t + t_1 + t_2) u_k(t + t_1) dt_1 dt_2 + \int_{\infty}^{0} \int_{0}^{t+t_1} \int_{\infty}^{0} e^{A^T t_1} \\ & \times \mathcal{H}^{(2)} \Big(e^{At_2} B \otimes e^{A^T t_3} C^T \Big) u(t + t_1 - t_2) u^{(d)}(t + t_1 + t_3) dt_1 dt_2 dt_3 + \cdots. \end{aligned} \tag{17}$$

By repeatedly substituting for the state x and the dual state z, we derive the Volterra series for the dual system, although the notation becomes much more complicated. After inspecting the integral-kernels of the Volterra series of the dual system, we define the observability mapping $\bar{Q}$, similar to the reachability mapping, as follows:

$$\bar{Q} = [\bar{Q}_1, \ \bar{Q}_2, \ \bar{Q}_3, \ldots], \tag{18}$$

in which

$$\begin{aligned}
\bar{Q}_1(t_1) &= e^{A^T t_1} C^T, \\
\bar{Q}_2(t_1, t_2) &= e^{A^T t_2} \left[N_1^T, \ldots, N_m^T \right] \left(I_m \otimes \bar{Q}_1(t_1) \right), \\
&\vdots \qquad\qquad \vdots \\
\bar{Q}_i(t_1, \ldots, t_i) &= e^{A^T t_i} \Big[\mathcal{H}^{(2)} \big[\bar{P}_1(t_1) \otimes \bar{Q}_{i-2}(t_2, \ldots, t_{i-1}), \\
&\qquad \ldots, \bar{P}_{i-2}(t_1, \ldots, t_{i-2}) \otimes \bar{Q}_1(t_{i-1}) \big], \\
&\qquad \left[N_1^T, \ldots, N_m^T \right] \left(I_m \otimes \bar{Q}_{i-1}(t_1, \ldots, t_{i-1}) \right) \Big], \ \forall\, i \geq 3,
\end{aligned}$$

where $\bar{P}_i(t_1, \ldots, t_i)$ are defined in (7). Based on the above observability mapping, we define the observability Gramian Q of the QB system as follows:

$$Q = \sum_{i=1}^{\infty} Q_i \quad \text{with} \quad Q_i = \int_0^{\infty} \cdots \int_0^{\infty} \bar{Q}_i \bar{Q}_i^T \, \mathrm{d}t_1 \cdots \mathrm{d}t_i, \tag{19}$$

assuming that the series defining Q converges and all improper integrals exit. Analogous to the reachability Gramian, we next show a relation between the observability Gramian and the solution of a generalized Lyapunov equation.

Theorem 2 *Consider the QB system (3) with a stable matrix A, and let Q, defined in (19), be the observability Gramian of the system and assume it exists. Then, the Gramian Q satisfies the following Lyapunov equation:*

$$A^T Q + QA + \mathcal{H}^{(2)}(P \otimes Q)(\mathcal{H}^{(2)})^T + \sum_{k=1}^{m} N_k^T Q N_k + C^T C = 0, \tag{20}$$

where P is the reachability Gramian of the system, i.e., the solution of the generalized quadratic Lyapunov equation, given in (8).

Proof The proof of the above theorem is analogous to the proof of Theorem 1; therefore, we skip it for the brevity of the paper. □

Remark 2 As one would expect, the Gramians for QB systems boil down to the Gramians for bilinear systems [29] if the quadratic term is zero, i.e., $H = 0$. We shall discuss sufficient conditions for the existence of solutions of (9) and (20) in Subsection 4.1.

2.3 Truncated Gramians based on leading integral-kernels

It would also be interesting to look at a truncated version of the Gramians of the QB system based on the leading integral-kernels of the Volterra series, like done for bilinear systems in [30]. We refer to them as *truncated Gramians* of QB systems. For

this, let us consider approximate reachability and observability mappings using the leading integral-kernels as follows:

$$\widetilde{P}_{\mathcal{T}} = \left[\widetilde{P}_1, \widetilde{P}_2, \widetilde{P}_3\right], \qquad \widetilde{Q}_{\mathcal{T}} = \left[\widetilde{Q}_1, \widetilde{Q}_2, \widetilde{Q}_3\right],$$

where

$$\begin{aligned}
\widetilde{P}_1(t_1) &= e^{At_1}B,\\
\widetilde{Q}_1(t_1) &= e^{A^T t_1}C^T,\\
\widetilde{P}_2(t_1,t_2) &= e^{At_2}\left[N_1,\ldots,N_m\right]\left(I_m \otimes \widetilde{P}_1(t_1)\right),\\
\widetilde{Q}_2(t_1,t_2) &= e^{A^T t_2}\left[N_1^T,\ldots,N_m^T\right]\left(I_m \otimes \widetilde{Q}_1(t_1)\right),\\
\widetilde{P}_3(t_1,t_2,t_3) &= e^{At_3}H(\widetilde{P}_1(t_1)\otimes\widetilde{P}_1(t_2)),\\
\widetilde{Q}_3(t_1,t_2,t_3) &= e^{A^T t_3}\mathcal{H}^{(2)}(\widetilde{P}_1(t_1)\otimes\widetilde{Q}_1(t_2)).
\end{aligned}$$

Then, one can define the truncated reachability and observability Gramians in a similar fashion as the Gramians of the system:

$$P_{\mathcal{T}} = \sum_{i=1}^{3}\widehat{P}_i, \quad \text{where} \quad \widehat{P}_i = \int_0^\infty \widetilde{P}_i(t_1,\ldots,t_i)\widetilde{P}_i^T(t_1,\ldots,t_i)\mathrm{d}t_1\cdots\mathrm{d}t_i, \tag{21a}$$

$$Q_{\mathcal{T}} = \sum_{i=1}^{3}\widehat{Q}_i, \quad \text{where} \quad \widehat{Q}_i = \int_0^\infty \widetilde{Q}_i(t_1,\ldots,t_i)\widetilde{Q}_i^T(t_1,\ldots,t_i)\mathrm{d}t_1\cdots\mathrm{d}t_i, \tag{21b}$$

respectively. Similar to the Gramians P and Q, in the following, we derive the relation between these truncated Gramians and the solutions of the Lyapunov equations.

Corollary 2.1 *Let $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ be the truncated Gramians of the QB system as defined in (21) with a stable matrix A. Then, $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ satisfy the following Lyapunov equations:*

$$AP_{\mathcal{T}} + P_{\mathcal{T}}A^T + H(\widehat{P}_1\otimes\widehat{P}_1)H^T + \sum_{k=1}^{m}N_k\widehat{P}_1N_k^T + BB^T = 0, \quad \textit{and} \tag{22a}$$

$$A^TQ_{\mathcal{T}} + Q_{\mathcal{T}}A + \mathcal{H}^{(2)}(\widehat{P}_1\otimes\widehat{Q}_1)(\mathcal{H}^{(2)})^T + \sum_{k=1}^{m}N_k^T\widehat{Q}_1N_k + C^TC = 0, \tag{22b}$$

respectively, where $\widehat{P}_1$ and $\widehat{Q}_1$ are solutions to the following Lyapunov equations:

$$A\widehat{P}_1 + \widehat{P}_1A^T + BB^T = 0, \quad \textit{and} \tag{23}$$

$$A^T\widehat{Q}_1 + \widehat{Q}_1A + C^TC = 0, \quad \textit{respectively.} \tag{24}$$

Proof We begin by showing the relation between the truncated reachability Gramian $P_{\mathcal{T}}$ and the solution of the Lyapunov equation. First, note that the first two terms of the reachability Gramian P in (21a) and the truncated reachability Gramian $P_{\mathcal{T}}$ in (8) are the same, i.e., $\widehat{P}_1 = P_1$ and $\widehat{P}_2 = P_2$, and $\widehat{P}_1$ and $\widehat{P}_2$ are the unique solutions of the following Lyapunov equations for a stable matrix A:

$$A\widehat{P}_1 + \widehat{P}_1 A^T + BB^T = 0, \quad \text{and} \tag{25}$$

$$A\widehat{P}_2 + \widehat{P}_2 A^T + \sum_{k=1}^{m} N_k \widehat{P}_1 N_k^T = 0. \tag{26}$$

Now, we consider the third term in the summation (21a). That is

$$\begin{aligned}
\widehat{P}_3 &= \int_0^\infty \int_0^\infty \int_0^\infty \widetilde{P}_3(t_1,t_2,t_3)\widetilde{P}_3^T(t_1,t_2,t_3)\mathrm{d}t_1\mathrm{d}t_2\mathrm{d}t_3 \\
&= \int_0^\infty \int_0^\infty \int_0^\infty e^{At_3} H(\widetilde{P}_1(t_1)\widetilde{P}^T(t_1) \otimes \widetilde{P}_1(t_2)\widetilde{P}^T(t_2))H^T e^{A^T t_3}\mathrm{d}t_1\mathrm{d}t_2\mathrm{d}t_3 \\
&= \int_0^\infty e^{At_3} H\left(\left(\int_0^\infty \widetilde{P}_1(t_1)\widetilde{P}^T(t_1)\mathrm{d}t_1\right)\otimes\left(\int_0^\infty \widetilde{P}_1(t_2)\widetilde{P}^T(t_2)\mathrm{d}t_2\right)\right)H^T e^{A^T t_3}\mathrm{d}t_3 \\
&= \int_0^\infty e^{At_3} H\left(\widehat{P}_1 \otimes \widehat{P}_1\right)H^T e^{A^T t_3}\mathrm{d}t_3.
\end{aligned}$$

Here, we have used that the infinite integrals exist due to the stability of A. Furthermore, we use the relation between the above integral representation and the solution of the Lyapunov equation to show that $\widehat{P}_3$ solves:

$$A\widehat{P}_3 + \widehat{P}_3 A^T + H(\widehat{P}_1 \otimes \widehat{P}_1)H^T = 0. \tag{27}$$

Summing (25), (26) and (27) yields

$$AP_{\mathcal{T}} + \mathcal{P}_{\mathcal{T}} A^T + H(\widehat{P}_1 \otimes \widehat{P}_1) + \sum_{k=1}^{m} N_k \widehat{P}_1 N_k + BB^T = 0. \tag{28}$$

Analogously, we can show that $Q_{\mathcal{T}}$ solves (22b), thus concluding the proof. □

Next, we study how these Gramians characterize reachable and observable subspaces of QB systems.

3 Characterization of reachable and observable subspaces using Gramians

In this section, our objective is first to provide an interpretation of the proposed Gramians—that is, the connection of Gramians with the reachability and observability

of the system. For the observability energy functional, we consider the output y of the following *homogeneous* QB system:

$$\begin{aligned} \dot{x}(t) &= Ax + H\left(x(t) \otimes x(t)\right) + \sum_{k=1}^{m} N_k x(t) u_k(t), \\ y(t) &= Cx(t), \qquad x(0) = x_0, \end{aligned} \tag{29}$$

inspired from the bilinear systems [29, 32]. However, it might also be possible to consider an *inhomogeneous* system by setting the control input u completely zero, as shown in [25]. We first investigate how the proposed Gramians are related to the reachability and observability of the QB systems, analogues to derivation for bilinear systems in [29].

Theorem 3

(a) Consider the QB system (3), and assume the reachability Gramian P exists and satisfies (9). If the system is steered from 0 to $x_0 \notin \mathsf{Im}P$, then the controllability energy functional $L_c(x_0) = \infty$ for all input functions u.

(b) Furthermore, consider the homogeneous *QB system (29) and assume $P \succ 0$ and Q to be the reachability and observability Gramians of the QB system, which are solutions of (9) and (20), respectively. If the initial state satisfies $x_0 \in \mathsf{Null}Q$, then the observability energy functional $L_o(x_0) = 0$.*

Proof

(a) By assumption, P satisfies

$$AP + PA^T + H(P \otimes P)H^T + \sum_{k=1}^{m} N_k P N_k^T + BB^T = 0. \tag{30}$$

Next, we consider a vector $v \in \mathsf{Null}P$ and multiply the above equation from the left and right with v^T and v, respectively, to obtain

$$\begin{aligned} 0 &= v^T APv + v^T PA^T v + v^T H(P \otimes P)H^T v + \sum_{k=1}^{m} v^T N_k P N_k^T v + v^T BB^T v \\ &= v^T H(P \otimes P)H^T v + \sum_{k=1}^{m} v^T N_k P N_k^T v + v^T BB^T v. \end{aligned}$$

This implies $B^T v = 0$, $PN_k^T v = 0$ and $(P \otimes P)H^T v = 0$. From (30), we thus obtain $PA^T v = 0$. Now, we consider an arbitrary state vector $x(t)$, which is the solution of (3) at time t for any given input function u. If $x(t) \in \mathsf{Im}P$ for some t, then we have

$$\dot{x}(t)^T v = x(t)^T A^T v + \left(x(t) \otimes x(t)\right)^T H^T v + \sum_{k=1}^{m} u_k(t) x(t)^T N_k^T v + u(t) B^T v = 0.$$

The above relation indicates that $\dot{x}(t) \perp v$ if $v \in \mathsf{Null}P$ and $x(t) \in \mathsf{Im}P$. It shows that $\mathsf{Im}P$ is invariant under the dynamics of the system. Since the initial condition 0 lies in $\mathsf{Im}P$, $x(t) \in \mathsf{Im}P$ for all $t \geq 0$. This reveals that if the final state $x_0 \notin \mathsf{Im}P$, then it cannot be reached from 0; hence, $L_c(x_0) = \infty$.

(b) Following the above discussion, we can show that $(P \otimes Q)\left(\mathcal{H}^{(2)}\right)^T \mathsf{Null}Q = 0$, $QN_k\mathsf{Null}Q = 0$, $QA\mathsf{Null}Q = 0$, and $C\mathsf{Null}Q = 0$. Moreover, if $P > 0$, then $(I \otimes Q)\left(\mathcal{H}^{(2)}\right)^T \mathsf{Null}Q = 0$. Let $x(t)$ denote the solution of the homogeneous system at time t. If $x(t) \in \mathsf{Null}Q$ and a vector $\widetilde{v} \in \mathsf{Im}Q$, then we have

$$\begin{aligned}\widetilde{v}^T\dot{x}(t) &= \underbrace{\widetilde{v}^T Ax(t)}_{=0} + \widetilde{v}^T H(x(t) \otimes x(t))) + \sum_{k=1}^{m} \underbrace{\widetilde{v}^T N_k x(t) u_k(t)}_{=0} \\ &= x(t)^T \mathcal{H}^{(2)}(x(t) \otimes \widetilde{v}) = \underbrace{x(t)^T \mathcal{H}^{(2)}(I \otimes \widetilde{v})}_{=0} x(t) = 0.\end{aligned}$$

This implies that if $x(t) \in \mathsf{Null}Q$, then $\dot{x}(t) \in \mathsf{Null}Q$. Therefore, if the initial condition $x_0 \in \mathsf{Null}Q$, then $x(t) \in \mathsf{Null}Q$ for all $t \geq 0$, resulting in $y(t) = C \underbrace{x(t)}_{\in \mathsf{Null}Q} = 0$; hence, $L_o(x_0) = 0$. □

The above theorem suggests that the state components belonging to $\mathsf{Null}P$ or $\mathsf{Null}Q$, do not play a major role as far as the system dynamics are concerned. This shows that the states which belong to $\mathsf{Null}P$, are unreachable, and similarly, the states, lying in $\mathsf{Null}Q$ are unobservable once the unreachable states are removed. In addition to this, note that the reachability Gramian P is defined by inspecting the integral-kernels of the underlying Volterra series, see Subsection 2.1. The Gramian P encodes the information about all infinite integral-kernels, thus also has the information about the subspace of the state x. Consequently, the dominant reachable subspaces of the infinite integral-kernels can be determined by the dominant singular vectors of Gramian P. Likewise also holds for the dominant observable subspaces, which can be determined by the dominant singular values of Q. These align with the concept of balanced truncation, which aims to identify the dominant reachable and observable subspaces and remove the less dominant reachable and observable subspaces. To find the states, which are simultaneously dominant reachable and observable, we utilize the balancing tools similar to the linear case; see, e.g., [1, 28]. For this, one needs to determine the Cholesky factors of the Gramians as $P =: S^T S$ and $Q =: R^T R$, and compute the SVD of $SR^T =: U\Sigma V^T$, resulting in a transformation matrix $T = S^T U \Sigma^{-\frac{1}{2}}$. Using the matrix T, we obtain an equivalent QB system

$$\begin{aligned}\dot{\widetilde{x}}(t) &= \widetilde{A}\widetilde{x}(t) + \widetilde{H}\widetilde{x}(t) \otimes \widetilde{x}(t) + \sum_{k=1}^{m} \widetilde{N}_k \widetilde{x}(t) u_k(t) + \widetilde{B}u(t), \\ y(t) &= \widetilde{C}\widetilde{x}(t), \quad \widetilde{x}(0) = 0,\end{aligned} \tag{31}$$

with

$$\widetilde{A} = T^{-1}AT, \quad \widetilde{H} = T^{-1}H(T \otimes T), \quad \widetilde{N}_k = T^{-1}N_k T, \quad \widetilde{B} = T^{-1}B, \quad \widetilde{C} = CT.$$

Then, the above transformed system (31) is balanced, as the Gramians $\widetilde{P}$ and $\widetilde{Q}$ of the system (31) are equal and diagonal, i.e., $\widetilde{P} = \widetilde{Q} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$. The attractiveness of the balanced system is that it allows finding state components corresponding to small singular values of both $\widetilde{P}$ and $\widetilde{Q}$. If $\sigma_1 \gg \sigma_{\widehat{n}+1}$, for some $\widehat{n} \in \mathbb{N}$, then in analogy to the linear case, we interpret those as hard to reach and hard to observe simultaneously; hence, they can be eliminated. In order to determine a reduced system of order $\widehat{n}$, we partition $T = \begin{bmatrix} T_1 & T_2 \end{bmatrix}$ and $T^{-1} = \begin{bmatrix} S_1^T & S_2^T \end{bmatrix}^T$, where $T_1, S_1^T \in \mathbb{R}^{n \times \widehat{n}}$, and define the reduced-order system's realization as follows:

$$\widehat{A} = S_1 A T_1, \quad \widehat{H} = S_1 H (T_1 \otimes T_1), \quad \widehat{N}_k = S_1 N_k T_1, \quad \widehat{B} = S_1 B, \quad \widehat{C} = C T_1, \tag{32}$$

which is generally a locally good approximation of the original system. It is not a straightforward task to estimate the error occurring due to the truncation of the QB system, unlike in the case of linear systems. Still, we can consider the sum of truncated singular values as an error indicator.

Remark 3 Similar to linear and bilinear systems, it would be interesting to establish a connection between energy functional and a quadratic form of the Gramians. In this direction, an attempt has been made in [33, 41], but it requires further investigation to make these connections and arguments concrete.

4 Further analysis and remarks

In this section, we provide some analysis related to the existence of Gramians, and the computational advantages of considering the truncated Gramians in the MOR framework. Towards the end, we analyze a local Lyapunov stability of the reduced-order systems obtained by using the truncated Gramians.

4.1 Existence of Gramians

We have shown that the reachability and observability Gramians satisfy (9) and (20), respectively, provided the sum of the infinite series defining Gramians converge. Beginning with the reachability Gramian, we discuss a iterative type of scheme to obtain a solution of (9) and satisfy also the sum of the underlying infinite sum. A simple scheme can be designed based on fixed-point iteration method, as follows:

$$\begin{aligned} A\widetilde{P}_1 + \widetilde{P}_1 A^T + BB^T &= 0, \\ A\widetilde{P}_g + \widetilde{P}_g A^T + H\left(\widetilde{P}_{g-1} \otimes \widetilde{P}_{g-1}\right) H^T & \\ + \sum_{k=1}^{m} N_k \widetilde{P}_{g-1} N_k^T + BB^T &= 0, \quad \text{for } g \geq 2. \end{aligned} \tag{33}$$

If the above fixed-point iteration converges, then it leads to a solution of (9). Moreover, with analytic calculations, we can also show that $\lim_{g\to\infty} \widetilde{P}_g$ can be cast as an infinite sum which will match with (8).

Similarly, if we define the following fixed-point iteration:

$$\begin{aligned} A^T \widetilde{Q}_1 + \widetilde{Q}_1 A + C^T C &= 0, \\ A^T \widetilde{Q}_g + \widetilde{Q}_g A + \mathcal{H}^{(2)} \left(\widetilde{P}_\infty \otimes \widetilde{Q}_{g-1} \right) & \\ + \sum_{k=1}^{m} N_k^T \widetilde{Q}_{g-1} N_k + C^T C &= 0, \quad \text{for } g \geq 2, \end{aligned} \tag{34}$$

then it gives the solution of (20). Next, we discuss sufficient conditions under which the fixed point iterations that are given in (33) and (34) converge. This would also imply the existence of Gramians.

Theorem 4 *Consider a QB system as defined in* (3) *and let P and Q be its reachability and observability Gramians, respectively, and assume their existence. Further, assume that the Gramians P and Q are determined using fixed point iterations as shown in* (33) *and* (34)*, respectively. These iterations converge if*

(i) *A is stable, i.e., there exist $0 < \alpha \leq -\max(\mathrm{Re}\,(\lambda_i(A)))$ and $\beta > 0$ such that $\|e^{At}\| \leq \beta e^{-\alpha t}$.*

(ii) $\dfrac{\beta^2 \Gamma_N}{2\alpha} < 1$, *where* $\Gamma_N := \sum_{k=1}^{m} \|N_k\|^2$.

(iii) $1 > \mathcal{D}^2 - \dfrac{\beta^2 \Gamma_H}{\alpha} \dfrac{\beta^2 \Gamma_B}{\alpha} > 0$, *where* $\mathcal{D} := 1 - \dfrac{\beta^2 \Gamma_N}{2\alpha}$, *and* $\Gamma_B := \|BB^T\|$, $\Gamma_H := \|H\|^2$.

Moreover, the norm $\lim_{g \to \infty} \|\widetilde{P}_g\| =: \widetilde{P}_\infty$ *is bounded by*

$$\|\widetilde{P}_\infty\| \leq \frac{2\alpha}{\beta^2 \Gamma_H} \left(\mathcal{D} - \sqrt{\mathcal{D}^2 - 4 \frac{\beta^2 \Gamma_H}{2\alpha} \frac{\beta^2 \Gamma_B}{2\alpha}} \right) =: \widetilde{\mathcal{P}}_\infty. \tag{35}$$

Furthermore, the iteration for $\widetilde{Q}_g$ also converges to a positive semidefinite solution Q of the linear matrix equation (20) *if in addition to the above conditions i–iii, the following condition is satisfied:*

$$\frac{\beta^2}{2\alpha} \left(\Gamma_N + \widetilde{\Gamma}_H \mathcal{P}_\infty \right) < 1, \tag{36}$$

where $\widetilde{\Gamma}_H := \|\mathcal{H}^{(2)}\|^2$. *Moreover,* $\lim_{g \to \infty} \|Q_g\| =: \widetilde{Q}_\infty$ *is bounded by*

$$\|\widetilde{Q}_\infty\| \leq \frac{\beta^2}{2\alpha} \Gamma_C \left(1 - \frac{\beta^2}{2\alpha} \left(\Gamma_N + \widetilde{\Gamma}_H \mathcal{P}_\infty \right) \right)^{-1}, \tag{37}$$

where $\Gamma_C := \|C^T C\|$.

Proof Let us first consider the equation corresponding to $\widetilde{P}_1$:

$$A\widetilde{P}_1 + \widetilde{P}_1 A^T + BB^T = 0. \tag{38}$$

Alternatively, if A is stable, we can write $\widetilde{P}_1$ in the integral form as

$$\widetilde{P}_1 = \int_0^\infty e^{At} BB^T e^{A^T t} \mathrm{d}t, \tag{39}$$

implying

$$\|\widetilde{P}_1\| \leq \beta^2 \|BB^T\| \int_0^\infty e^{-2\alpha t} \mathrm{d}t = \frac{\beta^2 \Gamma_B}{2\alpha}, \tag{40}$$

where $\Gamma_B := \|BB^T\|$. Next, we look at the equation corresponding to $\widetilde{P}_g$, which is given in terms of $\widetilde{P}_{g-1}$:

$$A\widetilde{P}_g + \widetilde{P}_g A^T + H(\widetilde{P}_{g-1} \otimes \widetilde{P}_{g-1})H^T + \sum_{k=1}^m N_k \widetilde{P}_{g-1} N_k + BB^T = 0. \tag{41}$$

We can also write $\widetilde{P}_g$ in an integral form, provided A is stable:

$$\begin{aligned}
\widetilde{P}_g &= \int_0^\infty e^{At} \left(H(\widetilde{P}_{g-1} \otimes \widetilde{P}_{g-1})H^T + \sum_{k=1}^m N_k \widetilde{P}_{g-1} N_k + BB^T \right) e^{A^T t} \mathrm{d}t \\
&\leq \beta^2 \left(\Gamma_H \|\widetilde{P}_{g-1}\|^2 + \Gamma_N \|\widetilde{P}_{g-1}\| + \Gamma_B \right) \int_0^\infty e^{-2\alpha t} \mathrm{d}t \\
&\leq \beta^2 \frac{\left(\Gamma_H \|\widetilde{P}_{g-1}\|^2 + \Gamma_N \|\widetilde{P}_{g-1}\| + \Gamma_B \right)}{2\alpha},
\end{aligned}$$

where $\Gamma_H := \|H\|^2$ and $\Gamma_N := \sum_{k=1}^m \|N_k\|^2$. If we consider an upper bound for the norm of $\widetilde{P}_{g-1}$ in order to provide an upper bound for $\widetilde{P}_g$ and apply Appendix A, then we know that $\lim_{g\to\infty} \|\widetilde{P}_g\|$ is bounded if

$$1 > \mathcal{D}^2 - 4\frac{\beta^2\Gamma_H}{2\alpha}\frac{\beta^2\Gamma_B}{2\alpha} \geq 0, \quad \textit{where } \mathcal{D} := 1 - \frac{\beta^2\Gamma_N}{2\alpha} \quad \text{and} \quad \frac{\beta^2\Gamma_N}{2\alpha} < 1,$$

and $\lim_{g\to\infty} \|\widetilde{P}_g\|$ is bounded by

$$\lim_{g\to\infty} \|P_g\| \leq \frac{2\alpha}{\beta^2\Gamma_H} \left(\mathcal{D} - \sqrt{\mathcal{D}^2 - 4\frac{\beta^2\Gamma_H}{2\alpha}\frac{\beta^2\Gamma_B}{2\alpha}} \right) =: \mathcal{P}_\infty.$$

Moreover, since $\|\widetilde{P}_g\|$ is a non-decreasing function, the fixed-point iteration (33) also converges [42, Chap. 9].

Next, we consider the equation corresponding to $\widetilde{Q}_1$:

$$A^T \widetilde{Q}_1 + \widetilde{Q}_1 A + C^T C = 0,$$

which again can be rewritten as:

$$\widetilde{Q}_1 = \int_0^\infty e^{A^T t} C^T C e^{At} \mathrm{d}t$$

if A is stable. This implies

$$\|\widetilde{Q}_1\| \leq \beta^2 \Gamma_C \int_0^\infty e^{-2\alpha t} \mathrm{d}t = \beta^2 \frac{\Gamma_C}{2\alpha},$$

where $\Gamma_c := \|C^T C\|$. Next, we look at the equation corresponding to Q_g, that is,

$$A^T \widetilde{Q}_g + Q_g \widetilde{A} + \mathcal{H}^{(2)}(\widetilde{P}_\infty \otimes \widetilde{Q}_{g-1}) \left(\mathcal{H}^{(2)}\right)^T + \sum_{k=1}^{m} N_k^T Q_{g-1} N_k + C^T C = 0.$$

A similar analysis for $\|Q_g\|$ yields

$$\|Q_g\| \leq \frac{\beta^2}{2\alpha} \left(\left(\Gamma_N + \widetilde{\Gamma}_H \|\widetilde{P}_\infty\|\right) \widetilde{Q}_{g-1} + \Gamma_C \right) = \frac{\beta^2}{2\alpha} \left(\left(\Gamma_N + \widetilde{\Gamma}_H \mathcal{P}_\infty\right) \widetilde{Q}_{g-1} + \Gamma_C \right),$$

where $\widetilde{\Gamma}_H := \|\mathcal{H}^{(2)}\|$. An additional sufficient condition under which the above recurrence formula in $\|Q_g\|$ converges is as follows:

$$\frac{\beta^2}{2\alpha} \left(\Gamma_N + \widetilde{\Gamma}_H \mathcal{P}_\infty\right) < 1,$$

and $\lim_{g\to\infty} \|\widetilde{Q}_g\|$ is then bounded by

$$\lim_{g\to\infty} \|\widetilde{Q}_g\| \leq \frac{\beta^2}{2\alpha} \Gamma_C \left(1 - \frac{\beta^2}{2\alpha} \left(\Gamma_N + \widetilde{\Gamma}_H \mathcal{P}_\infty\right)\right)^{-1}.$$

Additionally, $\|\widetilde{Q}_g\|$ is a non-decreasing function; hence, the iterations also converge. This concludes the proof. □

4.2 MOR using truncated Gramians

In Section 2, we have proposed the Gramians and have shown that the reachable Gramian P satisfies a quadratic-type Lyapunov matrix equations. There have been several developments for solving linear matrix equations [43, 44]; however, quadratic-type matrix equations appear for the first time in this work, and new algorithm developments are required to solve such matrix equations efficiently. Furthermore, we

Algorithm 1 Balanced truncation for QB systems (truncated version).

Input: System matrices A, H, N_k, B and C, and the order of the reduced system $\widehat{n}$.
Output: The reduced system's matrices $\widehat{A}$, $\widehat{H}$, $\widehat{N}_k$, $\widehat{B}$, $\widehat{C}$.
1: Determine low-rank approximations of the truncated Gramians $P_{\mathcal{T}} \approx RR^T$ and $Q_{\mathcal{T}} \approx SS^T$.
2: Compute SVD of $S^T R$:
$S^T R = U \Sigma V = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \operatorname{diag}(\Sigma_1, \Sigma_2) \begin{bmatrix} V_1 & V_2 \end{bmatrix}^T$,
where Σ_1 contains the $\widehat{n}$ largest singular values of $S^T R$.
3: Construct the projection matrices $\mathcal{V}$ and $\mathcal{W}$:
$\mathcal{V} = SU_1 \Sigma_1^{-\frac{1}{2}}$ and $\mathcal{W} = RV_1 \Sigma_1^{-\frac{1}{2}}$.
4: Determine the reduced-order system's realization:
$\widehat{A} = \mathcal{W}^T A \mathcal{V}$, $\widehat{H} = \mathcal{W}^T H(\mathcal{V} \otimes \mathcal{V})$, $\widehat{N}_k = \mathcal{W}^T N_k \mathcal{V}$, $\widehat{B} = \mathcal{W}^T B$, $\widehat{C} = C\mathcal{V}$.

have discussed sufficient conditions in which these Gramians can be obtained using fixed-point iterations. However, these conditions are very conservative for large-scale models, particularly those coming from PDEs, and this is also what we observe in our numerical examples. On the other hand, in order to compute the truncated Gramians, there is no such convergence issue, and they capture information about the leading three integral-kernels, see Section 2. For weakly nonlinear quadratic systems, one can expect the a rapid convergence of the Volterra series, given in (5). In this case, the leading integral-kernels would contain most information about the system dynamics, and hence, a balancing scheme based on these truncated Gramians can already provide information about information subspaces. Therefore, in this work, we utilize the truncated Gramians to determine the reduced-order models, and we present the square-root balanced truncation for QB systems based on these truncated Gramians in Algorithm 1. Furthermore, we will see in Section 5 that these truncated Gramians also yield very good qualitative reduced-order systems for QB systems.

4.3 Stability preservation

We now discuss the stability of the reduced-order systems obtained by using Algorithm 1. For this, we consider only the autonomous part of the QB system as follows:

$$\dot{x}(t) = Ax(t) + H\ (x(t) \otimes x(t)), \tag{42}$$

where $x_{eq} = 0$ is a stable equilibrium.

In the following, we discuss Lyapunov stability of x_{eq}. For this, we first note the definition of local stability.

Definition 1 Consider a QB system with $u \equiv 0$ (42). If there exists a Lyapunov function $\mathcal{F} : \mathbb{R}^n \to \mathbb{R}$ such that

$$\mathcal{F}(x(t)) > 0 \quad \text{and} \quad \frac{d}{dt}\mathcal{F}(x(t)) > 0 \quad \forall\, x(t) \in \mathcal{B}_{0,r} \backslash \{0\}, \quad t \geq 0$$

along the trajectory of $\mathbf{x}(t)$, and $\mathcal{B}_{0,r}$ is a ball of radius r centered around 0, then $x_{eq} = 0$ is locally asymptotically stable.

However, many other notions of the stability of nonlinear systems are available in the literature, for instance, based on a certain dissipation inequality [45], which might be difficult to apply in a large-scale setting. In this paper, we stick to the notion of Lyapunov-based stability for reduced-order systems.

Theorem 5 *Consider the QB system (3) with a stable matrix A. Let $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ be its truncated reachability and observability Gramians, defined in Corollary 2.1, respectively. If the reduced-order system is determined as shown in Algorithm 1, then for the Lyapunov function $\mathcal{F}(\widehat{x}) = \widehat{x}^T \Sigma_1 \widehat{x}$, we have*

$$\mathcal{F}(\widehat{x}) > 0, \qquad \tfrac{d}{dt}(\mathcal{F}(\widehat{x})) < 0 \qquad \forall\, \widehat{x} \in \mathcal{B}_{0,r} \backslash \{0\},$$

along the trajectory of $x(t)$, and $r = \dfrac{\sigma_{\min}(\mathcal{V}^T \mathcal{G} \mathcal{V})}{2\|\Sigma_1\| \|\widehat{H}\|}$ and

$$\mathcal{G} = \mathcal{H}^{(2)}(P_1 \otimes Q_1)\left(\mathcal{H}^{(2)}\right)^T + \sum_{k=1}^{m} N_k^T Q_1 N_k + C^T C$$

with P_1 and Q_1 being the solutions of (23) and (24), respectively.

Proof First, we establish the relation between $\mathcal{V}$, $\mathcal{W}$, $Q_{\mathcal{T}}$, and Σ_1. For this, we consider

$$\begin{aligned} \mathcal{W}\Sigma_1 &= R V_1 \Sigma_1^{\frac{1}{2}} = R V_1 \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix}^T U^T U_1 \Sigma_1^{-\frac{1}{2}} = R V \Sigma U^T U_1 \Sigma_1^{-\frac{1}{2}} \\ &= R R^T S^T U_1 \Sigma_1^{-\frac{1}{2}} = Q_{\mathcal{T}} \mathcal{V}. \end{aligned}$$

Keeping in mind the above relation, we get

$$\begin{aligned} \widehat{A}^T \Sigma_1 + \Sigma_1 \widehat{A} + \mathcal{V}^T \mathcal{G} \mathcal{V} &= \mathcal{V}^T A^T \mathcal{W} \Sigma_1 + \Sigma_1 \mathcal{W}^T A \mathcal{V} + \mathcal{V}^T \mathcal{G} \mathcal{V} \\ &= \mathcal{V}^T A^T Q_{\mathcal{T}} \mathcal{V} + \mathcal{V}^T Q_{\mathcal{T}} A \mathcal{V} + \mathcal{V}^T \mathcal{G} \mathcal{V} = \mathcal{V}^T (A^T Q_{\mathcal{T}} + Q_{\mathcal{T}} A + \mathcal{G}) \mathcal{V} = 0. \end{aligned} \tag{43}$$

Since $\mathcal{G}$ is a positive semidefinite matrix and $\mathcal{V}$ has full column rank, $\mathcal{V}^T \mathcal{G} \mathcal{V}$ is also a positive semidefinite. This implies that $\eta(\widehat{A}) \leq 0$, where $\eta(\cdot)$ denotes the spectral abscissa of a matrix. Coming back to the Lyapunov function $\mathcal{F}(\widehat{x}) = \widehat{x}^T \Sigma_1 \widehat{x}$, which is always greater than 0 for all $\widehat{x} \neq 0$ due to Σ_1 being a positive definite matrix, we compute the derivative of the Lyapunov function as

$$\begin{aligned} \frac{d}{\mathrm{d}t}\mathcal{F}(\widehat{x}) &= \dot{\widehat{x}}^T \Sigma_1 \widehat{x} + \widehat{x}^T \Sigma_1 \dot{\widehat{x}} \\ &= \widehat{x}^T \widehat{A}^T \Sigma_1 \widehat{x} + (\widehat{x}^T \otimes \widehat{x}^T) \widehat{H}^T \Sigma_1 \widehat{x} + \widehat{x}^T \Sigma_1 \widehat{A} \widehat{x} + \widehat{x}^T \Sigma_1 \widehat{H} (\widehat{x} \otimes \widehat{x}) \\ &= \widehat{x}^T (\widehat{A}^T \Sigma_1 + \Sigma_1 \widehat{A}) \widehat{x} + (\widehat{x}^T \otimes \widehat{x}^T) \widehat{H}^T \Sigma_1 \widehat{x} + \widehat{x}^T \Sigma_1 \widehat{H} (\widehat{x} \otimes \widehat{x}). \end{aligned}$$

Substituting $\widehat{A}^T \Sigma_1 + \Sigma_1 \widehat{A} = -\mathcal{V}^T \mathcal{G} \mathcal{V}$ from (43) in the above equation yields

$$\tfrac{d}{\mathrm{d}t}\mathcal{F}(\widehat{x}) = -\widehat{x}^T \mathcal{V}^T \mathcal{G} \mathcal{V} \widehat{x} + 2\widehat{x}^T \Sigma_1 \widehat{H} (\widehat{x} \otimes \widehat{x}). \tag{44}$$

As

$$\widehat{x}^T \mathcal{V}^T \mathcal{G} \mathcal{V} \widehat{x} \geq \sigma_{\min}(\mathcal{V}^T \mathcal{G} \mathcal{V}) \|\widehat{x}\|^2,$$

implying

$$-\widehat{x}^T \mathcal{V}^T \mathcal{G} \mathcal{V} x \leq -\sigma_{\min}(\mathcal{V}^T \mathcal{G} \mathcal{V}) \|\widehat{x}\|^2,$$

inserting the above inequality in (44) leads to

$$\tfrac{d}{dt} \mathcal{F}(\widehat{x}) \leq -\sigma_{\min}(\mathcal{V}^T \mathcal{G} \mathcal{V}) \|\widehat{x}\|^2 + 2\|\widehat{x}\|^3 \|\Sigma_1\| \|\widehat{H}\|.$$

For local asymptotic stability of the reduced-order system, we require

$$\tfrac{d}{dt} \mathcal{F}(\widehat{x}) \leq -\sigma_{\min}(\mathcal{V}^T \mathcal{G} \mathcal{V}) \|\widehat{x}\|^2 + 2\|\widehat{x}\|^3 \|\Sigma_1\| \|\widehat{H}\| < 0,$$

which gives rise to the following bound on $\|\widehat{x}\|$:

$$\|\widehat{x}\| < \tfrac{\sigma_{\min}(\mathcal{V}^T \mathcal{G} \mathcal{V})}{2\|\Sigma_1\| \|\widehat{H}\|}.$$

This concludes the proof. □

5 Numerical experiments

In this section, we consider the MOR of several QB control systems and evaluate the efficiency of the proposed balanced truncation technique (Algorithm 1). For this, we need to solve a number of conventional Lyapunov equations. In our numerical experiments, we determine the low-rank factors of these Lyapunov equations by using the ADI method as proposed in [46]. We compare the proposed methodology with the existing moment-matching techniques for QB systems, namely one-sided moment-matching [12] and its extension to two-sided moment-matching [20]. These moment-matching methods aim at approximating the underlying generalized transfer functions of the system. Moreover, we need interpolation points in order to apply the moment-matching methods; thus, we choose l linear $\mathcal{H}_2$-optimal interpolation points, determined by using *IRKA* [47] on the corresponding linear part. This leads to a reduced QB system of order $\widehat{n} = 2l$.

5.1 Nonlinear RC ladder

As a first example, we discuss a nonlinear RC ladder. It is a well-known example and is used as one of the benchmark problems in the community of nonlinear model reduction; see, e.g., [12, 13, 15, 16, 48]. A detailed description of the dynamics can be found in the mentioned references; therefore, we omit it for the brevity of the paper. However, we like to comment on the nonlinearity present in the RC ladder. The nonlinearity arises from the presence of the diode I-V characteristic $i_D := e^{40v_D} - v_D - 1$,

where v_D denotes the voltage across the diode. As shown in [12], introducing some appropriate new variables allows us to write the system dynamics in the QB form of dimension $n = 2k$, where k is the number of capacitors in the ladder.

We consider 500 capacitors in the ladder, leading to a QB system of order $n = 1000$. For this particular example, the matrix A is a semi-stable matrix, i.e., $0 \in \sigma(A)$. As a result, the truncated Gramians of the system might not exist; therefore, we replace the matrix A by $A_s := A - 0.05I$, where I is the identity matrix, to determine these Gramians. However, note that we project the original system with the matrix A to compute a reduced-order system, but the projection matrices are computed using the Gramians obtained via the shifted matrix A_s. In Fig. 1, we show the decay of the singular values determined by the truncated Gramians (with the shifted A). We then compute the reduced system of order $\widehat{n} = 10$ by using balanced truncation. Also, we determine 5 $\mathcal{H}_2$-optimal linear interpolation points and compute reduced-order systems of order $\widehat{n} = 10$ via one-sided and two-sided projection methods.

To compare the quality of these approximations, we simulate these systems for the input signals $u_1(t) = 5\,(\sin(2\pi/10) + 1)$ and $u_2(t) = 10\left(t^2 \exp(-t/5)\right)$. Figure 2 presents the transient responses and relative errors of the output for these input signals, which shows that balanced truncation outperforms the one-sided interpolatory method; on the other hand, we see that balanced truncation is competitive to the two-sided interpolatory projection for this example.

Lastly, we present CPU time to simulate full-order and reduced-order models obtained using BT and time to compute reduced-order models using BT. Note that FOM simulations are done using the original unlifted systems, and ROMs are quadratic models of the lifted quadratic systems. The result is reported in Table 1. We note that the CPU time to compute reduced-order models using BT is much lower than even a single FOM for a given control input, and the obtained reduced-order models are significantly faster. For control design applications, where we are required to simulate FOM for many control scenarios, the designing cycle can be sped up significantly.

5.2 One-dimensional Chafee-Infante equation

As a second example, we consider the one-dimensional Chafee-Infante (Allen-Cahn) equation. This nonlinear system has been widely studied in the literature; see, e.g., [49, 50], and its model reduction related problem was recently considered in [20]. The

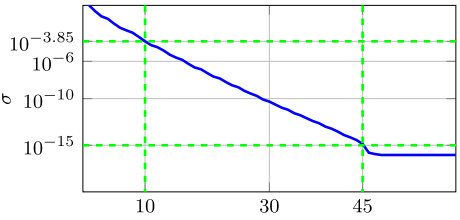
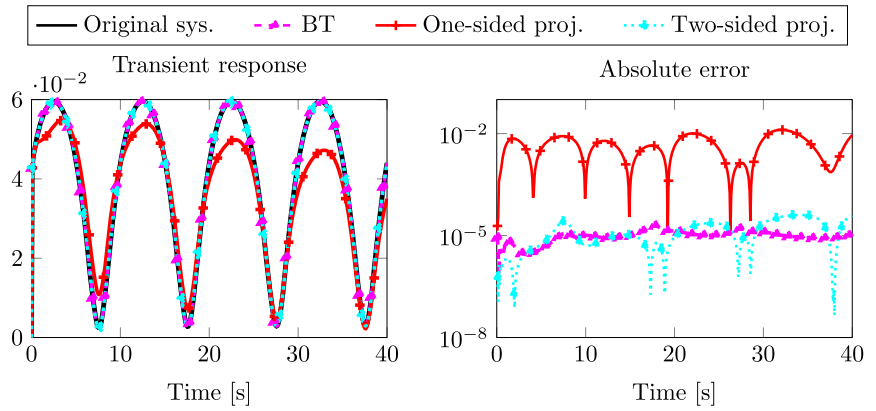
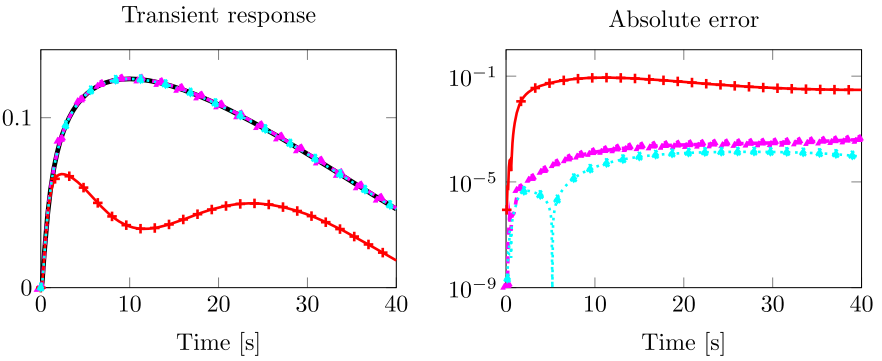


Fig. 1 A RC ladder: the decay of the normalized singular values based the truncated Gramians, and the dotted lines show the normalized singular value for $\widehat{n} = 10$ and the order of the reduced system corresponding to the normalized singular value $1e-15$



(a) Comparison of the original and the reduced-order systems for $u_1(t) = 5\,(\sin(2\pi/10) + 1)$.



(b) Comparison of the original and the reduced-order systems for $u_2(t) = 10\left(t^2 \exp(-t/5)\right)$.

Fig. 2 A RC ladder: comparison of reduced-order systems obtained by balanced truncation (BT) and moment-matching methods for two arbitrary control inputs

governing equation, subject to initial conditions and boundary control, has a cubic nonlinearity:

$$\begin{aligned} \dot{v} + v^3 &= v_{xx} + v, & (0, L) \times (0, T), \quad & v(0, \cdot) = u(t), & (0, T), \\ v_x(L, \cdot) &= 0, & (0, T), \quad & v(x, 0) = 0, & (0, L). \end{aligned} \tag{45}$$

Here, we make use of a finite difference scheme and consider k grid points in the spatial domain, leading to a semi-discretized nonlinear ODE. As shown in [20],

Table 1 A RC ladder: A comparison of CPU time to simulate full-order and reduced-order models

	FOM	ROM	Speed up (factor)	Time to compute ROM
input $u_1(t)$	7.74s	0.23s	∼33x	2.02s
input $u_2(t)$	6.34s	0.11s	∼57x	–

the smooth nonlinear system can be transformed into a QB system by introducing appropriate new state variables. Therefore, the system (45) with the cubic nonlinearity can be rewritten in the QB form by defining new variables $w_i = v_i^2$ with derivate $\dot{w}_i = 2v_i\dot{v}_i$. We observe the response at the right boundary $x = L$. We use $k = 500$ grid points, which results in a QB system of dimension $n = 2 \cdot 500 = 1000$ and set the length $L = 1$. In Fig. 3, we show the decay of the normalized singular values based on the truncated Gramians of the system.

We determine reduced-order systems of order $\widehat{n} = 20$ by using balanced truncation, and one-sided and two-sided interpolatory projection methods. To compare the quality of these reduced-order systems, we observe the outputs of the original and reduced-order systems for two arbitrary control inputs $u(t) = 5t\exp(-t)$ and $u(t) = 30(\sin(\pi t) + 1)$ in Fig. 4.

Figure 4a shows that the reduced-order systems obtained via balanced truncation and one-sided and two-sided interpolatory projection methods are almost of the same quality for input u_1. But for the input u_2, the reduced-order system obtained via the one-sided interpolatory projection method fails to capture the dynamics of the system. In contrast, balanced truncation and two-sided interpolatory projection can reproduce the system dynamics with a slight advantage of two-sided projection regarding accuracy.

Next, like the previous example, we present CPU time to simulate full-order and reduced-order models obtained using BT, and time to compute reduced-order models using BT. Here again, FOM simulations are done using the original unlifted cubic systems, which are efficient and fast, and ROMs are quadratic models of the lifted quadratic systems. The result is reported in Table 2. We note that the CPU time to compute reduced-order models using BT is similar to the FOM simulation for u_2, and the obtained reduced-order models are significantly faster. The designing cycle can be sped up significantly for control design applications, where we are required to simulate FOM for many control scenarios.

Lastly, for this example, we study the effect of reduced-order on the quality of reduced-order models, and the results are shown in Fig. 5. We notice that as we increase the order of the reduced-order system, the two-sided interpolatory projection method tends to produce unstable reduced-order systems. This is reflected in Fig. 5 as the missing values for two-sided interpolatory method indicate the order for which we obtain unstable reduced-order models. On the other hand, the accuracy of the reduced-order systems obtained by balanced truncation and one-sided moment-matching increases with the order of the reduced-order systems.

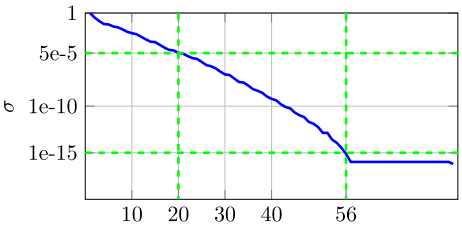
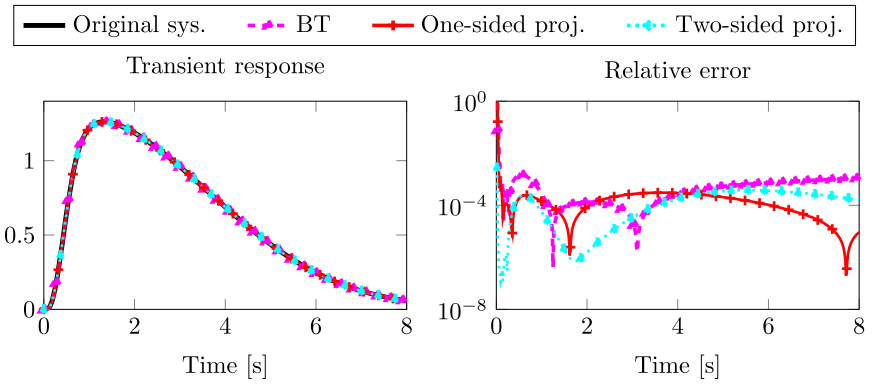
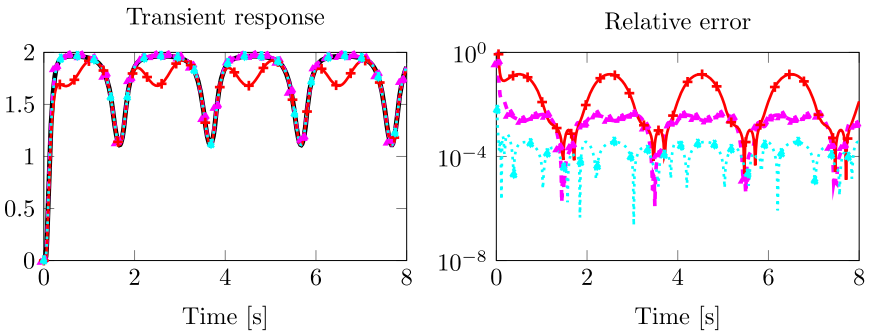


Fig. 3 Chafee-Infante equation: the decay of the normalized singular values based the truncated Gramians, and the dotted line shows the normalized singular value for $\widehat{n} = 20$ and the order of the reduced-order system corresponding to the normalized singular value $1e{-}15$



(a) Comparison of the original and the reduced-order systems for $u_1(t) = 5\,t\exp(-t)$.



(b) Comparison of the original and the reduced-order systems for $u_2(t) = 30\,(\sin(\pi t) + 1)$.

Fig. 4 Chafee-Infante equation: comparison of the reduced-order systems obtained via balanced truncation and moment-matching methods for the inputs $u_1(t) = 5\,(t\exp(-t))$ and $u_2(t) = 30\,(\sin(\pi t) + 1)$

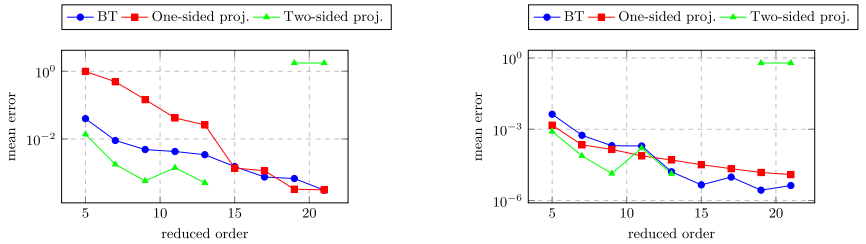
5.3 The FitzHugh-Nagumo (F-N) system

Lastly, we consider the F-N system, a simplified neuron model of the Hodgkin-Huxley model, which describes the activation and deactivation dynamics of a spiking neuron. This model has been considered in the framework of POD-based [7] and moment-matching model reduction techniques [51]. The dynamics of the system is governed by the following nonlinear coupled differential equations:

$$\begin{aligned} \epsilon v_t(x,t) &= \epsilon^2 v_{xx}(x,t) + f(v(x,t)) - w(x,t) + q, \\ w_t(x,t) &= hv(x,t) - \gamma w(x,t) + q \end{aligned} \tag{46}$$

Table 2 A RC ladder: A comparison of CPU time to simulate full-order and reduced-order models

	FOM	ROM	Speed up (factor)	Time to compute ROM
input $u_1(t)$	4.52s	0.23s	∼20x	1.93s
input $u_2(t)$	1.73s	0.09s	∼18x	–



(a) Mean errors for different order of reduced models and for $u_1(t) = 5\,t\exp(-t)$.

(b) Mean errors for different order of reduced models and for $u_2(t) = 30\,(\sin(\pi t) + 1)$.

Fig. 5 Chafee-Infante equation: Mean error with respect to different reduced models for two control inputs

with a nonlinear function $f(v(x,t)) = v(v-0.1)(1-v)$ and the initial and boundary conditions:

$$\begin{aligned} v(x,0) &= 0, & w(x,0) &= 0, & x &\in [0,L], \\ v_x(0,t) &= i_0(t), & v_x(1,t) &= 0, & t &\geq 0, \end{aligned} \tag{47}$$

where $\epsilon = 0.015$, $h = 0.5$, $\gamma = 2$, $q = 0.05$. We set the length $L = 0.2$. The stimulus i_0 acts as an actuator, taking the values $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$, and the variables v and w denote the voltage and recovery voltage, respectively. We also assume the same outputs of interest as considered in [51], which are $v(0,t)$ and $w(0,t)$. These outputs describe nothing but the limit cyclic at the left boundary. Using a finite difference discretization scheme, one can obtain a system with two inputs and two outputs of dimension $2k$ with cubic nonlinearities, where k is the number of degrees of freedom. Similar to the previous example, the F-H system can also be transformed into a QB system of dimension $n = 3k$ by introducing a new state variable $z_i = v_i^2$. We set $k = 500$, resulting in a QB system of order $n = 1500$. Figure 6 shows the decay of the singular values based on the truncated Gramians for the QB system.

Furthermore, we determine reduced-order systems of order $\widehat{n} = 20$ by using balanced truncation and moment-matching methods. We observe that the reduced-order systems, obtained via the moment-matching methods with linear $\mathcal{H}_2$-optimal interpolations, both one-sided and two-sided, fail to capture the dynamics and limit cycles. We made several attempts to adjust the order of the reduced-order systems, but, we were unable to determine a stable reduced-order system via these methods with linear $\mathcal{H}_2$-optimal points which could replicate the dynamics. Contrary to these methods,

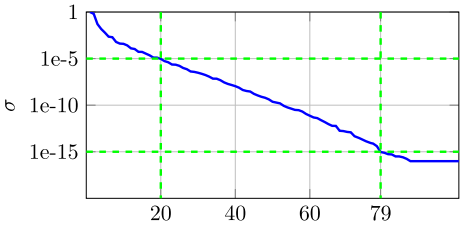
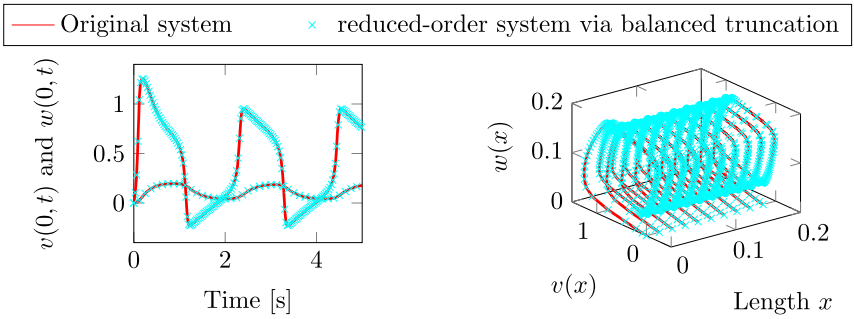


Fig. 6 xDecay of the normalized singular values based on the truncated Gramians of the system for the F-N example, and the dotted lines show the normalized singular value for $\widehat{n} = 20$ and the order of the reduced-order system corresponding to the normalized singular value $1e{-}15$



(a) The response $v(t)$ and $w(t)$ at the left boundary.

(b) Limit-cycles.

Fig. 7 FitzHugh-Nagumo system: comparison of the response at the left boundary and the limit cycle behavior of the original system and the reduced-order (balanced truncation) system. The reduced-order systems determined by moment-matching methods could not produce these limit cycles

balanced truncation replicates the dynamics of the system faithfully, as can be seen in Fig. 7a. Note that the reduced-order model reported in [51] was obtained using higher-order moments in a trial-and-error fashion but cannot be reproduced by an automated algorithm. As the dynamics of the system produce limit cycles for each spatial variable x, we, therefore, plot the solutions v and w over the spatial domain x, which is also captured by the reduced-order system very well.

Remark 4 Note that the proposed balanced truncation and POD were already compared to a $\mathcal{H}_2$-quasi-optimal scheme for QB systems in [21]. We do not repeat these experiments here but note that balanced truncation and $\mathcal{H}_2$-quasi-optimal scheme perform very similarly and outperform POD for inputs different from the training inputs used to produce the snapshots for POD.

6 Conclusions

In this paper, we have investigated balanced truncation for quadratic-bilinear (QB) control systems. We have proposed reachability and observability Gramians for QB systems based on the integral-kernels of their underlying Volterra series. Additionally, we have also introduced a truncated version of the Gramians. Furthermore, we investigated the connection between the Gramians and the reachability/observability of QB systems. We also discussed sufficient conditions for the existence of the Gramians. Also, we have discussed the advantages of the truncated version of the Gramians in the model reduction framework and studied the local Lyapunov stability of the reduced-order systems obtained via the square-root variant of balanced truncation. By means of various semi-discretized nonlinear PDEs, we have demonstrated the efficiency of the proposed balanced truncation method for QB systems and compared it with the exist-

ing moment-matching techniques. We have observed that balanced truncation yields more stable reduced-order models than the two-sided interpolation method. Another important advantage of balanced truncation is that we do not need to choose the order of the reduced model a priori. Additionally, for BT, a suitable order can be found by analyzing the decay of singular values. On the other hand, one-sided and two-sided interpolation methods require the number of interpolation points as an input, which fixes the order of reduced models, which is hard to know a priori.

Appendix A A convergence result

Lemma 1 *Consider a recurrence formula as follows:*

$$x_{k+1} = F(x_k), \quad \forall \quad k \geq 1, \tag{A1}$$

where $F(x) = ax^2 + bx + c$ *and* a, b, c *are real positive scaler numbers. Moreover, assume that* $x_1 = c$. *Then* $\lim_{k\to\infty} x_k =: x^*$ *is finite if*

$$b < 1, \quad \textit{and} \tag{A2a}$$

$$1 > (b-1)^2 - 4ac > 0. \tag{A2b}$$

Furthermore, x^* *is given by the smaller root of the the following quadratic equation:*

$$ax^2 + (b-1)x + c = 0, \quad \textit{i.e.,}$$

$$x^* = \frac{-(b-1) - \sqrt{(b-1)^2 - 4ac}}{2a}. \tag{A3}$$

Proof First, note that the sequence (A1) contains only real positive numbers. Thus, the equilibrium point must also be a real positive number. Furthermore, the equilibrium points solve the quadratic equation $F(x) - x = 0$, and we denote these equilibrium points by $x^{(1)}$ and $x^{(2)}$ with $x^{(1)} \leq x^{(2)}$. Since a, b and c all are positive, both equilibrium points either can be positive or negative depending on the value of b. To ensure the equilibrium points being positive, the minima of $F(x) - x$ must lie in the right half plane; thus, $b - 1 < 0$, leading to the condition (A2a).

Furthermore, we consider the derivative of $F(x)$, that is, $F'(x) := 2ax + b$. Since $F'(x)$ is an increasing function and $F'(x) \geq 0\, \forall x \in [c, x^{(1)}]$, we have for $y \in [c, x^{(1)}]$:

$$\begin{aligned} F'(y) &\leq F'(x^{(1)}) \\ &\leq 2ax^{(1)} + b = 2a\left(\frac{-(b-1) - \sqrt{(b-1)^2 - 4ac}}{2a}\right) + b \leq 1 - \sqrt{(b-1)^2 - 4ac}. \end{aligned}$$

Assuming $1 > (b-1)^2 - 4ac > 0$, we have $F'(y) < 1, \forall y \in [c, x^{(1)}]$. Thus, by the Banach fix-point theorem, $F(x)$ is a contraction on $[c, x^{(1)}]$, and the fixed point is given by $x^{(1)}$. □

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of Interest The authors declared that they have no conflict of interest.

References

1. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. SIAM Publications, Philadelphia, PA (2005)
2. Benner, P., Mehrmann, V., Sorensen, D.C.: Dimension Reduction of Large-Scale Systems. Lect. Notes Comput. Sci. Eng., vol. 45. Springer, Berlin/Heidelberg, Germany (2005)
3. Schilders, W.H.A., van der Vorst, H.A., Rommes, J.: Model Order Reduction: Theory. Research Aspects and Applications. Springer, Berlin, Heidelberg (2008)
4. Benner, P., Schilders, W., Grivet-Talocia, S., Quarteroni, A., Rozza, G., Miguel Silveira, L.: Model Order Reduction: Volume 3 Applications. De Gruyter, Berlin, Boston (2021)
5. Antoulas, A.C., Beattie, C.A., Gugercin, S.: Interpolatory Methods for Model Reduction. Computational Science & Engineering. Society for Industrial and Applied Mathematics, Philadelphia, PA (2020). https://doi.org/10.1137/1.9781611976083
6. Astrid, P., Weiland, S., Willcox, K., Backx, T.: Missing point estimation in models described by proper orthogonal decomposition. IEEE Trans. Autom. Control **53**(10), 2237–2251 (2008)
7. Chaturantabut, S., Sorensen, D.C.: Nonlinear model reduction via discrete empirical interpolation. SIAM J. Sci. Comput. **32**(5), 2737–2764 (2010)
8. Hinze, M., Volkwein, S.: Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: Error estimates and suboptimal control. In: Benner, P., Mehrmann, V., Sorensen, D.C. (eds.) Dimension Reduction of Large-Scale Systems. Lect. Notes Comput. Sci. Eng., vol. 45, pp. 261–306. Springer, Berlin/Heidelberg, Germany (2005)
9. Farhat, C., Grimberg, S., Manzoni, A., Quarteroni, A.: Computational bottlenecks for proms: precomputation and hyperreduction. In: Benner, P., Grivet-Talocia, S., Quarteroni, A., Rozza, G., Schilders, W., Silveira, L.M. (eds.) Model Order Reduction: Volume 2: Snapshot-Based Methods and Algorithms, pp. 181–244. De Gruyter, Berlin, Boston (2020)
10. Barrault, M., Maday, Y., Nguyen, N.C., Patera, A.T.: An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. C. R. Math. Acad. Sci. Paris **339**(9), 667–672 (2004)
11. Grepl, M.A., Maday, Y., Nguyen, N.C., Patera, A.T.: Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. ESAIM: Math. Model. Numer. Anal. **41**(3), 575–605 (2007)
12. Gu, C.: QLMOR: A projection-based nonlinear model order reduction approach using quadratic-linear representation of nonlinear systems. IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. **30**(9), 1307–1320 (2011)
13. Bai, Z.: Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems. Appl. Numer. Math. **43**(1), 9–44 (2002)
14. Feng, L., Zeng, X., Chiang, C., Zhou, D., Fang, Q.: Direct nonlinear order reduction with variational analysis. In: Proceedings of the Design, Automation and Test in Europe Conference and Exhibition, vol. 2, pp. 1316–1321 (2004)

15. Li, P., Pileggi, L.T.: Compact reduced-order modeling of weakly nonlinear analog and RF circuits. IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. **24**(2), 184–203 (2005)
16. Phillips, J.R.: Projection-based approaches for model reduction of weakly nonlinear, time-varying systems. IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. **22**(2), 171–187 (2003)
17. Ahmad, M.I., Benner, P., Feng, L.: Interpolatory model reduction for quadratic-bilinear systems using error estimators. Eng. Comput. **36**(1), 25–44 (2019). https://doi.org/10.1108/EC-04-2018-0162
18. Ahmad, M.I., Benner, P., Feng, L.: A new two-sided projection technique for model reduction of quadratic-bilinear descriptor systems. Int. J. Comput. Math. **96**(10), 1899–1909 (2019). https://doi.org/10.1080/00207160.2018.1542134
19. Ahmad, M.I., Benner, P., Jaimoukha, I.: Krylov subspace projection methods for model reduction of quadratic-bilinear systems. IET Control Theory Appl. **10**(16), 2010–2018 (2016). https://doi.org/10.1049/iet-cta.2016.0415
20. Benner, P., Breiten, T.: Two-sided projection methods for nonlinear model reduction. SIAM J. Sci. Comput. **37**(2), 239–260 (2015). https://doi.org/10.1137/14097255X
21. Benner, P., Goyal, P., Gugercin, S.: $\mathcal{H}_2$-quasi-optimal model order reduction for quadratic-bilinear control systems. SIAM J. Matrix Anal. Appl. **39**(2), 983–1032 (2018). https://doi.org/10.1137/16M1098280
22. Moore, B.C.: Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans. Autom. Control AC- **26**(1), 17–32 (1981)
23. Fujimoto, K., Scherpen, J.M.A.: Balanced realization and model order reduction for nonlinear systems based on singular value analysis. SIAM J. Cont. Optim. **48**(7), 4591–4623 (2010)
24. Gray, W.S., Scherpen, J.M.A.: On the nonuniqueness of singular value functions and balanced nonlinear realizations. Systems Control Lett. **44**(3), 219–232 (2001)
25. Scherpen, J.M.A.: Balancing for nonlinear systems. Systems Control Lett. **21**, 143–153 (1993)
26. Scherpen, J.M.A.: $\mathcal{H}_\infty$ balancing for nonlinear systems. Internat. J. Robust Nonlinear Control **6**(7), 645–668 (1996)
27. Scherpen, J.M.A., Schaft, A.J.: Normalized coprime factorizations and balancing for unstable nonlinear systems. Internat. J. Control **60**(6), 1193–1222 (1994)
28. Al-Baiyat, S.A., Bettayeb, M., Al-Saggaf, U.M.: New model reduction scheme for bilinear systems. Int. J. Syst. Sci. **25**(10), 1631–1642 (1994)
29. Benner, P., Damm, T.: Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems. SIAM J. Cont. Optim. **49**(2), 686–711 (2011). https://doi.org/10.1137/09075041X
30. Benner, P., Goyal, P., Redmann, M.: Truncated Gramians for bilinear systems and their advantages in model order reduction. In: Benner, P., Ohlberger, M., Patera, T., Rozza, G., Urban, K. (eds.) Model Reduction of Parametrized Systems. MS&A - Modeling, Simulation and Applications, vol. 17, pp. 285–300. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-58786-8_18
31. Condon, M., Ivanov, R.: Nonlinear systems-algebraic gramians and model reduction. COMPEL **24**(1), 202–219 (2005)
32. Gray, W.S., Mesko, J.: Energy functions and algebraic Gramians for bilinear systems. In: Preprints of the 4th IFAC Nonlinear Control Systems Design Symposium, Enschede, The Netherlands, pp. 103–108 (1998)
33. Benner, P., Goyal, P.: Balanced truncation model order reduction for quadratic-bilinear control systems. arXiv:1705.00160 (2017)
34. Kramer, B., Willcox, K.: Balanced truncation model reduction for lifted nonlinear systems. In: Realization and Model Reduction of Dynamical Systems - A Festschrift in Honor of the 70th Birthday of Thanos Antoulas, pp. 157–174. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-95157-3_9
35. Kramer, B.: Stability domains for quadratic-bilinear reduced-order models. SIAM J. Appl. Dyn. Syst. **20**(2), 981–996 (2021)
36. Zhang, G., Jiang, Y.-L., Xu, K.: Balanced truncation reduced models of quadratic-bilinear systems in time interval. Asian J. Control **25**(6), 4658–4666 (2023)
37. Bruni, C., Di Pillo, G., Koch, G., et al.: On the mathematical models of bilinear systems. Ric. Autom. **2**, 11–26 (1971)
38. Sastry, S.S.: Nonlinear Systems: Analysis, Stability, and Control. Springer, New York (1999)
39. Fujimoto, K., Scherpen, J.M.A., Gray, W.S.: Hamiltonian realizations of nonlinear adjoint operators. Automatica **38**(10), 1769–1775 (2002)

40. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009)
41. Goyal, P.K.: Interpolatory methods for model reduction of large-scale dynamical systems. Phd thesis, Otto-von-Guericke-Universität, Magdeburg, Germany (2018)
42. Rudin, W.: Principles of Mathematical Analysis, vol. 3. McGraw-hill, New York (1976)
43. Simoncini, V.: Computational methods for linear matrix equations. SIAM Rev. **58**(3), 377–441 (2016)
44. Benner, P., Saak, J.: Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey. GAMM-Mitteilungen **36**(1), 32–52 (2013)
45. Bond, B.N., Mahmood, Z., Li, Y., Sredojevic, R., Megretski, A., Stojanovi, V., Avniel, Y., Daniel, L.: Compact modeling of nonlinear analog circuits using system identification via semidefinite programming and incremental stability certification. IEEE Trans. Comput.-Aided Design Integr. Circuits Syst. **29**(8), 1149–1162 (2010)
46. Benner, P., Kürschner, P., Saak, J.: Self-generating and efficient shift parameters in ADI methods for large Lyapunov and Sylvester equations. Electron. Trans. Numer. Anal. **43**, 142–162 (2014)
47. Gugercin, S., Antoulas, A.C., Beattie, C.A.: $\mathcal{H}_2$ model reduction for large-scale dynamical systems. SIAM J. Matrix Anal. Appl. **30**(2), 609–638 (2008)
48. Breiten, T., Damm, T.: Krylov subspace methods for model order reduction of bilinear control systems. Systems Control Lett. **59**(10), 443–450 (2010)
49. Chafee, N., Infante, E.F.: A bifurcation problem for a nonlinear partial differential equation of parabolic type†. Appl. Anal. **4**(1), 17–37 (1974)
50. Hansen, E., Kramer, F., Ostermann, A.: A second-order positivity preserving scheme for semilinear parabolic problems. Appl. Numer. Math. **62**(10), 1428–1435 (2012)
51. Benner, P., Breiten, T.: Two-sided moment matching methods for nonlinear model reduction. Preprint MPIMD/12-12, MPI Magdeburg (2012). Available from http://www.mpi-magdeburg.mpg.de/preprints/