

BALANCED TRUNCATION MODEL ORDER REDUCTION FOR QUADRATIC-BILINEAR CONTROL SYSTEMS

PETER BENNER* AND PAWAN GOYAL†

Abstract. We discuss balanced truncation model order reduction for large-scale quadratic-bilinear (QB) systems. Balanced truncation 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 (1993), where it is shown that Gramians for nonlinear systems are the solutions of state-dependent nonlinear Hamilton-Jacobi equations. Therefore, they are not only difficult to compute for large-scale systems but also hard to utilize in the model reduction framework. In this paper, we propose algebraic Gramians for QB systems based on the underlying Volterra series representation of QB systems and their Hilbert adjoint systems. We then show their relations with a certain type of generalized quadratic Lyapunov equation. Furthermore, we present how these algebraic Gramians and energy functionals relate to each other. Moreover, we characterize the reachability and observability of QB systems based on the proposed algebraic Gramians. This allows us to find those states that are hard to control and hard to observe via an appropriate transformation based on the Gramians. Truncating such states yields reduced-order systems. Additionally, we present a truncated version of the Gramians for QB systems and discuss their advantages in the model reduction framework. We also investigate the Lyapunov stability of the reduced-order systems. We finally illustrate the efficiency of the proposed balancing-type model reduction for QB systems by means of various semi-discretized nonlinear partial differential equations and show its competitiveness with the existing moment-matching methods for QB systems.

Key words. Model order reduction, balanced truncation, Hilbert adjoint operator, tensor calculus, Lyapunov stability, energy functionals.

AMS subject classifications. 15A69, 34C20, 41A05, 49M05, 93A15, 93C10, 93C15.

1. Introduction. Numerical simulations are considered to be a primary tool in studying dynamical systems, e.g., in prediction and control 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 simulations of such high-fidelity systems, whose number of state variables can easily be of order $\mathcal{O}(10^5)$ – $\mathcal{O}(10^6)$, are 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 these large-scale dynamical systems by 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 instance, stability and passivity of the original system.

MOR techniques and strategies for linear systems are well-established and are widely applied in various application areas, see, e.g., [2, 12, 42]. In many applications, where the dynamics are governed by nonlinear PDEs, such as Navier-Stokes equations, Burgers' equations, a linear system can also be obtained via linearization of the system around a suitable expansion point, e.g., the steady-state solution. Notwithstanding 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 a general need to take nonlinear terms

*Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, 39106 Magdeburg, Germany (benner@mpi-magdeburg.mpg.de).

†Corresponding author. Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, 39106 Magdeburg, Germany (goyalp@mpi-magdeburg.mpg.de).

into consideration, thus resulting in a more accurate system. Consider a nonlinear system of the form

$$(1.1) \quad \begin{aligned} \dot{x}(t) &= f(x(t)) + g(x(t), u(t)), \\ y(t) &= h(x(t), u(t)), \quad x(0) = x_0, \end{aligned}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $g : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$ and $h : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^p$ are nonlinear state-input evolution 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. Also, we consider a fixed initial condition x_0 of the system. However, without loss of generality, we assume the zero initial condition, i.e., $x(0) = 0$. In case $x(0) \neq 0$, one can transform the system by defining new appropriate state variables as $\tilde{x}(t) = x(t) - x_0$ to ensure a zero initial condition of the transformed system, e.g., see [6]. The main goal of MOR is to construct a low-dimensional system, having a similar form as the system (1.1)

$$(1.2) \quad \begin{aligned} \hat{\dot{x}}(t) &= \hat{f}(\hat{x}(t)) + \hat{g}(\hat{x}(t), u(t)), \\ \hat{y}(t) &= \hat{h}(\hat{x}(t), u(t)), \quad \hat{x}(0) = 0, \end{aligned}$$

in which $\hat{f} : \mathbb{R}^{\hat{n}} \rightarrow \mathbb{R}^{\hat{n}}$, $\hat{g} : \mathbb{R}^{\hat{n}} \times \mathbb{R}^m \rightarrow \mathbb{R}^{\hat{n}}$ and $\hat{h} : \mathbb{R}^{\hat{n}} \times \mathbb{R}^m \rightarrow \mathbb{R}^p$ with $\hat{n} \ll n$ that fulfills our desired requirements.

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., [3, 17, 32]. 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 $x(t) \approx V\hat{x}(t)$. The nonlinear functions $\hat{f}(\hat{x})$ can be given as $\hat{f}(\hat{x}(t)) = V^T f(V\hat{x}(t))$, and similar expressions can also be derived for $\hat{g}(\hat{x}(t), u(t))$ and $\hat{h}(\hat{x}(t), u(t))$. This method preserves the structure of the original system in the reduced-order system, but still, the reduced-order system requires the computation of the nonlinear functions on the full grid. This may obstruct the success of MOR; however, there are many new advanced methodologies such as the empirical interpolation method (EIM), the discrete empirical interpolation method (DEIM), the best point interpolation method (BPIM), to perform the computation of the nonlinear functions cheaply and quite accurately. For details, we refer to [5, 17, 21, 28].

Another popular trajectory-based MOR technique is based on trajectory piecewise linearization (TPWL) [37], where nonlinear functions are replaced by a weighted combination of linear systems. These linear systems can then be reduced by applying well-established MOR techniques for linear systems such as balanced truncation or the interpolation-based iterative method (IRKA); see, e.g., [2, 30]. In recent years, reduced basis methods have been successfully applied to nonlinear systems to obtain reduced-order systems [5, 28]. In spite of all these, the trajectory-based MOR techniques have the drawback of being input dependent. This makes the obtained reduced-order systems inadequate to control applications, where the input function may vary significantly from any used training input.

In this article, we consider a certain class of nonlinear control systems, namely quadratic-bilinear (QB) control systems. The advantage of considering this special class of nonlinear systems is that systems, containing smooth mono-variate nonlinearities such as exponentials, polynomials, trigonometric functions, can also be rewritten in the QB form by introducing some new variables in the state vector [29]. Note that this transformation is exact, i.e., it requires no approximation and does not introduce any 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., [4, 22, 29, 34, 36], where a reduced system is determined by capturing the input-output behavior of the original system, given by generalized transfer functions. More recently, this has been extended to two-sided moment-matching in [8], ensuring more moments to be matched, for a given order of the reduced system. Despite these methods have evolved as an effective MOR technique for nonlinear systems in recent times, shortcomings of these methods are: how to choose an appropriate order of the reduced system and how to select good interpolation points. Moreover, the applicability of the two-sided moment-matching method [8] is limited to single-input single-output QB systems, and also the stability of the obtained reduced-order system is a major issue in this method.

In this article, our focus rather lies on balancing-type MOR techniques for QB systems. This technique mainly depends on controllability and observability energy functionals, or in other words, Gramians of the system. This method is presented for linear systems, e.g., in [2, 35], and later on, a theory of balancing for general nonlinear systems is developed in a sequence of papers [23, 27, 39, 40, 41]. 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 nonlinearities in the system. However, for some weakly nonlinear systems, the so-called bilinear systems, reachability and observability Gramians have been studied in [1, 9, 10, 18, 25], 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 [9, 10]

Moving in the direction of balancing-type MOR for QB systems, our first goal is to come up with reachability and observability Gramians for these systems, which are state-independent matrices and suitable for the MOR purpose. In addition to this, we need to show how the Gramians relate to the energy functionals of the QB systems and provide interpretations of reachability and observability of the system with respect to these Gramians. To this end, in the subsequent section, we review background material associated with energy functionals and a duality of the nonlinear systems, which serves as the basis for the rest of the paper. In Section 3, we propose the reachability Gramian and its truncated version for QB systems based on the underlying Volterra series of the system. Additionally, we determine the observability Gramian and its truncated version based on the dual system associate to the QB system. Furthermore, we establish relations between the solutions of a certain type of quadratic Lyapunov equations and these Gramians. In Section 4, we develop the connection between the proposed Gramians and the energy functionals of the QB systems and reveal their relations to reachability and observability of the system. Consequently, we utilize these Gramians for balancing of QB systems, allowing us to determine those states that are hard to control as well as hard to observe. Truncation of such states leads to reduced systems. In Section 5, we discuss the related computational issues and advantages of the truncated version of Gramians in the MOR framework. We further discuss the stability of these reduced systems. In Section 6, we test the efficiency of the proposed balanced truncation MOR technique for various semi-discretized nonlinear PDEs and compare it with the existing moment-matching techniques for the QB systems.

2. Preliminaries. We begin with recapitulation of energy functionals for nonlinear systems.

2.1. Energy functionals for nonlinear systems. In this subsection, we give a brief overview of energy functionals, namely controllability and observability energy functionals for nonlinear systems. For this, let us consider the following smooth, for example, C^∞ , nonlinear asymptotically stable input-affine nonlinear system of the form

$$(2.1) \quad \begin{aligned} \dot{x}(t) &= f(x) + g(x)u(t), \\ y(t) &= h(x), \quad x(0) = 0, \end{aligned}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ are the state, input and output vectors of the system, respectively, and also $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $g(x) : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times m}$ and $h(x) : \mathbb{R}^n \rightarrow \mathbb{R}^p$ are smooth nonlinear functions. Without loss of generality, we assume that 0 is an equilibrium of the system (2.1). The controllability and observability energy functionals for the general nonlinear systems first have been studied in the literature in [39]. In the following, we state the definitions of controllability and observability energy functionals for the system (2.1).

DEFINITION 2.1. [39] *The controllability energy functional is defined as the minimum amount of energy required to steer the system from $x(-\infty) = 0$ to $x(0) = x_0$:*

$$L_c(x_0) = \min_{\substack{u \in L_2^m(-\infty, 0] \\ x(-\infty)=0, x(0)=x_0}} \frac{1}{2} \int_{-\infty}^0 \|u(t)\|^2 dt.$$

DEFINITION 2.2. [39] *The observability energy functional can be defined as the energy generated by the nonzero initial condition $x(0) = x_0$ with zero control input:*

$$L_o(x_0) = \frac{1}{2} \int_0^\infty \|y(t)\|^2 dt.$$

We assume that the system (2.1) is controllable and observable. This implies that the system (2.1) can always be steered from $x(0) = 0$ to x_0 by using appropriate inputs. To define the observability energy functional (Definition 2.2), it is assumed that the nonlinear system (2.1) is a zero-state observable. It means that if $u(t) = 0$ and $y(t) = 0$ for $t \geq 0$, then $x(t) = 0 \forall t \geq 0$. However, as discussed in [26], for a nonlinear system such a condition can be very strong. As a result, therein, it is shown how this condition can be relaxed in the context of general input balancing, and a new definition for the observability functionals was proposed as follows:

DEFINITION 2.3. [26] *The observability energy functional can be defined as the energy generated by the nonzero initial condition $x(0) = x_0$ and by applying an L_2^m -bounded input:*

$$L_o(x_0) = \max_{\substack{u \in L_2^m[0, \infty), \|u\|_{L_2} \leq \alpha \\ x(0)=x_0, x(\infty)=0}} \frac{1}{2} \int_0^\infty \|y(t)\|^2 dt.$$

In an abstract way, the main idea of introducing Definition 2.3 to find the state component that contributes less from a state-to-output point of view for all possible L_2 -bounded inputs. The connections between these energy functionals and the solutions of the partial differential equations are established in [26, 39], which are outlined in the following theorem.

THEOREM 2.4. [26, 39] *Consider the nonlinear system (2.1), having $x = 0$ as an asymptotically stable equilibrium of the system in a neighborhood W_o of 0. Then, for all $x \in W_o$, the observability energy functional $L_o(x)$ can be determined by the following partial differential equation:*

$$(2.2) \quad \frac{\partial L_o}{\partial x} f(x) + \frac{1}{2} h^T(x) h(x) - \frac{1}{2} \mu^{-1} \frac{\partial L_o}{\partial x} g(x) g(x)^T \frac{\partial^T L_o}{\partial x} = 0, \quad L_o(0) = -\frac{1}{2} \mu,$$

assuming that there exists a smooth solution \bar{L}_o on W , and 0 is an asymptotically stable equilibrium of $\bar{f} := (f - \mu^{-1} g g^T \frac{\partial^T \bar{L}_o}{\partial x})$ on W with a negative real number $\mu := -\|g^T(\phi) \frac{\partial^T \bar{L}_o}{\partial x}(\phi)\|_{L_2}$, and $\dot{\phi} = \bar{f}(\phi)$ with $\phi(0) = x$. Moreover, for all $x \in W_c$, the controllability energy functional $L_c(x)$ is a unique smooth solution of the following Hamilton-Jacobi equation:

$$(2.3) \quad \frac{\partial L_c}{\partial x} f(x) + f(x) \frac{\partial L_c}{\partial x} + \frac{\partial L_c}{\partial x} g(x) g(x)^T \frac{\partial^T L_c}{\partial x} = 0, \quad L_c(0) = 0$$

under the assumption that (2.3) has a smooth solution \bar{L}_c on W_c , and 0 is an asymptotically stable equilibrium of $-\left(f(x) + g(x) g(x)^T \frac{\partial^T \bar{L}_c(x)}{\partial x}\right)$ on W_c .

Note that in Definition 2.3, the zero-state observable condition is relaxed by considering L_2 -bounded inputs. However, an alternative way to relax the zero-state observable condition by considering not only L_2 -bounded inputs but also L_∞ bounded inputs. We thus propose a new definition of the observability energy functional as follows:

DEFINITION 2.5. *The observability energy functional can be defined as the energy generated by the nonzero initial condition $x(0) = x_0$ and by applying an L_2 -bounded and L_∞ -bounded input:*

$$L_o(x_0) = \max_{\substack{u \in \mathcal{B}_{(\alpha, \beta)} \\ x(0)=x_0, x(\infty)=0}} \frac{1}{2} \int_0^\infty \|y(t)\|^2 dt,$$

where $\mathcal{B}_{(\alpha, \beta)} \stackrel{\text{def}}{=} \{u \in L_2^m[0, \infty), \|u\|_{L_2} \leq \alpha, \|u\|_{L_\infty} \leq \beta\}$. In this paper, we use the above definition to characterize the observability energy functional for QB systems.

2.2. Hilbert adjoint operator for nonlinear systems. The importance of the adjoint operator (dual system) can be seen, particularly, in the computation of the observability energy functional or Gramian. For general nonlinear systems, a duality between controllability and observability energy functionals is shown in [24] with the help of state-space realizations for nonlinear adjoint operators. In what follows, we briefly outline the state-space realizations for nonlinear adjoint operators of nonlinear systems. For this, we consider a nonlinear system of the form

$$(2.4) \quad \Sigma := \begin{cases} \dot{x}(t) = \mathcal{A}(x, u, t)x(t) + \mathcal{B}(x, u, t)u(t), \\ y(t) = \mathcal{C}(x, u, t)x(t) + \mathcal{D}(x, u, t)u(t), \end{cases} \quad x(0) = 0$$

in which $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ and $y(t) \in \mathbb{R}^p$ are the state, input and output vectors of the system, respectively, and $\mathcal{A}(x, u, t)$, $\mathcal{B}(x, u, t)$, $\mathcal{C}(x, u, t)$ and $\mathcal{D}(x, u, t)$ are appropriate size matrices. Also, we assume that the origin is a stable equilibrium of the system. The Hilbert adjoint operators for the general nonlinear systems have

been investigated in [24]. Therein, a connection between the state-space realization of the adjoint operators and port-control Hamiltonian systems is also discussed, leading to the state-space characterization of the nonlinear Hilbert adjoint operators of $\Sigma : L_2^m(\Omega) \rightarrow L_2^p(\Omega)$. In the following lemma, we summarize the state-space realization of the Hilbert adjoint operator of the nonlinear system.

LEMMA 2.6. [24] *Consider the system (2.4) with the initial condition $x(0) = 0$, and assume that the input-output mapping $u \rightarrow y$ is denoted by the operator $\Sigma : L_2^m(\Omega) \rightarrow L_2^p(\Omega)$. Then, the state-space realization of the nonlinear Hilbert adjoint operator $\Sigma^* : L_2^{m+p}(\Omega) \rightarrow L_2^m(\Omega)$ is given by*

$$(2.5) \quad \Sigma^*(u_d, u) := \begin{cases} \dot{x}(t) = \mathcal{A}(x, u, t)x(t) + \mathcal{B}(x, u, t)u(t), & x(0) = 0, \\ \dot{x}_d(t) = -\mathcal{A}(x, u, t)x_d(t) - \mathcal{C}^T(x, u, t)u_d(t), & x_d(\infty) = 0, \\ y_d(t) = \mathcal{B}^T(x, u, t)x_d(t) + \mathcal{D}^T(x, u, t)u_d(t), \end{cases}$$

where $x_d \in \mathbb{R}^n$, $u_d \in \mathbb{R}^p$ and $y_d \in \mathbb{R}^m$ can be interpreted as the dual state, dual input and dual output vectors of the system, respectively.

We will see in the subsequent section the importance of the dual system in determining the observability energy functional or observability Gramian for a QB system because a duality of the energy functionality holds.

3. Gramians for QB Systems. This section is devoted to determine algebraic Gramians for QB systems, which are also related to the energy functionals of the quadratic-bilinear systems as well. Let us consider QB systems of the form

$$(3.1a) \quad \dot{x}(t) = Ax(t) + H x(t) \otimes x(t) + \sum_{k=1}^m N_k x(t) u_k(t) + Bu(t),$$

$$(3.1b) \quad y(t) = Cx(t), \quad x(0) = 0,$$

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. Since the system (3.1) has a quadratic nonlinearity in the state vector $x(t)$ and also includes bilinear terms $N_k x(t) u_k(t)$, which are products of the state vector and inputs, the system is called a quadratic-bilinear (QB) system. We begin by deriving the reachability Gramian of the QB system and its connection with a certain type of quadratic Lyapunov equation.

3.1. Reachability Gramian for QB systems. In order to derive the reachability Gramian, we first formulate the Volterra series for the QB system (3.1). Before we proceed further, for ease we define the following short-hand notation:

$$u_{\sigma_1, \dots, \sigma_l}^{(k)}(t) := u_k(t - \sigma_1 \cdots - \sigma_l) \quad \text{and} \quad x_{\sigma_1, \dots, \sigma_l}(t) := x(t - \sigma_1 \cdots - \sigma_l).$$

We integrate both sides of the differential equation (3.1a) in the state variables with respect to time to obtain

$$(3.2) \quad x(t) = \int_0^t e^{A\sigma_1} B u_{\sigma_1}(t) d\sigma_1 + \sum_{k=1}^m \int_0^t e^{A\sigma_1} N_k x_{\sigma_1}(t) u_{\sigma_1}^{(k)}(t) d\sigma_1 \\ + \int_0^t e^{A\sigma_1} H (x_{\sigma_1}(t) \otimes x_{\sigma_1}(t)) d\sigma_1.$$

Based on the above equation, we obtain an expression for $x_{\sigma_1}(t)$ as follows:

$$\begin{aligned} x_{\sigma_1}(t) = & \int_0^{t-\sigma_1} e^{A\sigma_2} B u_{\sigma_1, \sigma_2}(t) d\sigma_2 + \sum_{k=1}^m \int_0^{t-\sigma_1} e^{A\sigma_2} N_k x_{\sigma_1, \sigma_2}(t) u_{\sigma_1, \sigma_2}^{(k)}(t) d\sigma_2 \\ & + \int_0^{t-\sigma_1} e^{A\sigma_2} H(x_{\sigma_1, \sigma_2}(t) \otimes x_{\sigma_1, \sigma_2}(t)) d\sigma_2 \end{aligned}$$

and substitute it in (3.2) to have

$$\begin{aligned} x(t) = & \int_0^t e^{A\sigma_1} B u_{\sigma_1}(t) d\sigma_1 + \sum_{k=1}^m \int_0^t \int_0^{t-\sigma_1} e^{A\sigma_1} N_k e^{A\sigma_2} B u_{\sigma_1}^{(k)}(t) u_{\sigma_1, \sigma_2}(t) d\sigma_1 d\sigma_2 \\ & + \int_0^t \int_0^{t-\sigma_1} \int_0^{t-\sigma_1} e^{A\sigma_1} H(e^{A\sigma_2} B \otimes e^{A\sigma_3} B) (u_{\sigma_1, \sigma_2}(t) \otimes u_{\sigma_1, \sigma_3}(t)) d\sigma_1 d\sigma_2 d\sigma_3 + \dots \end{aligned}$$

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

$$(3.3) \quad \bar{P} = [\bar{P}_1, \bar{P}_2, \bar{P}_3, \dots],$$

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

$$(3.4) \quad \begin{aligned} \bar{P}_1(t_1) &= e^{At_1} B, \\ \bar{P}_2(t_1, t_2) &= e^{At_2} [N_1, \dots, N_m] (I_m \otimes \bar{P}_1(t_1)), \\ &\vdots \\ \bar{P}_i(t_1, \dots, t_i) &= e^{At_i} \left[H[\bar{P}_1(t_1) \otimes \bar{P}_{i-2}(t_2, \dots, t_{i-1}), \bar{P}_2(t_1, t_2) \otimes \bar{P}_{i-3}(t_3, \dots, t_{i-1}), \dots, \bar{P}_{i-2}(t_1, \dots, t_{i-2}) \otimes \bar{P}_1(t_{i-1})], \right. \\ &\quad \left. [N_1, \dots, N_m] (I_m \otimes \bar{P}_{i-1}(t_1, \dots, t_{i-1})) \right], \forall i \geq 3. \end{aligned}$$

Using the mapping \bar{P} (3.3), we define the reachability Gramian P as

$$(3.5) \quad P = \sum_{i=1}^{\infty} P_i \quad \text{with} \quad P_i = \int_0^{\infty} \dots \int_0^{\infty} \bar{P}_i(t_1, \dots, t_i) \bar{P}_i^T(t_1, \dots, t_i) dt_1 \dots dt_i.$$

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

THEOREM 3.1. *Consider the QB system (3.1) with a stable matrix A . If the reachability Gramian P of the system defined as in (3.5) exists, then the Gramian P satisfies the generalized quadratic Lyapunov equation, given by*

$$(3.6) \quad AP + PA^T + H(P \otimes P)H^T + \sum_{k=1}^m N_k P N_k^T + BB^T = 0.$$

Proof. We begin by considering the first term in the summation (3.5). This is,

$$P_1 = \int_0^\infty \bar{P}_1 \bar{P}_1^T dt_1 = \int_0^\infty e^{At_1} BB^T e^{A^T t_1} dt_1.$$

As shown, e.g., in [2], P_1 satisfies the following Lyapunov equation, provided A is stable:

$$(3.7) \quad AP_1 + P_1 A^T + BB^T = 0.$$

Next, we consider the second term in the summation (3.5):

$$\begin{aligned} P_2 &= \int_0^\infty \int_0^\infty \bar{P}_2 \bar{P}_2^T dt_1 dt_2 \\ &= \int_0^\infty \int_0^\infty e^{At_2} [N_1, \dots, N_m] \left(I_m \otimes \left(e^{At_1} BB^T e^{A^T t_1} \right) \right) [N_1, \dots, N_m]^T e^{A^T t_2} dt_1 dt_2 \\ &= \sum_{k=1}^m \int_0^\infty e^{At_2} N_k \left(\int_0^\infty e^{At_1} BB^T e^{A^T t_1} dt_1 \right) N_k^T e^{A^T t_2} dt_1 dt_2 \\ &= \sum_{k=1}^m \int_0^\infty e^{At_2} N_k P_1 N_k^T e^{A^T t_2} dt_2. \end{aligned}$$

Again using the integral representation of the solution to Lyapunov equations [2], we see that P_2 is the solution of the following Lyapunov equation:

$$(3.8) \quad AP_2 + P_2 A^T + \sum_{k=1}^m N_k P_1 N_k^T = 0.$$

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

$$\begin{aligned} P_i &= \int_0^\infty \cdots \int_0^\infty \bar{P}_i \bar{P}_i^T dt_1 \cdots dt_i \\ &= \int_0^\infty e^{At_i} \left[H \left[\int_0^\infty \mathcal{F}(\bar{P}_1(t_1)) dt_1 \otimes \int_0^\infty \cdots \int_0^\infty \mathcal{F}(\bar{P}_{i-2}(t_2, \dots, t_{i-1})) dt_2 \cdots dt_{i-1} \right. \right. \\ &\quad \left. \left. + \cdots + \int_0^\infty \cdots \int_0^\infty \mathcal{F}(\bar{P}_{i-2}(t_1, \dots, t_{i-2})) dt_1 \cdots dt_{i-2} \otimes \int_0^\infty \mathcal{F}(\bar{P}_1(t_{i-1})) dt_{i-1} \right] H^T \right. \\ &\quad \left. + \sum_{k=1}^m N_k \left(\int_0^\infty \cdots \int_0^\infty \mathcal{F}(\bar{P}_{i-1}(t_1, \dots, t_{i-1})) \right) N_k^T \right] e^{A^T t_i} dt_i, \end{aligned}$$

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

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

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 $k = 1, \dots, i-1$:

$$(3.9) \quad AP_i + P_i A^T + H(P_1 \otimes P_{i-2} + \cdots + P_{i-2} \otimes P_1) H^T + \sum_{k=1}^m N_k P_{i-1} N_k^T = 0.$$

To the end, adding (3.7), (3.8) and (3.9) yields

$$A \sum_{i=1}^{\infty} P_i + \sum_{i=1}^{\infty} P_i A^T + H \left(\sum_{i=1}^{\infty} P_i \otimes \sum_{i=1}^{\infty} P_i \right) H^T + \sum_{k=1}^m N_k \left(\sum_{i=1}^{\infty} P_i \right) N_k^T + BB^T = 0.$$

This implies that $P = \sum_{i=1}^{\infty} P_i$ solves the generalized quadratic Lyapunov equation given by (3.6). \square

3.2. Dual system and observability Gramian for QB system. We first derive the dual system for the QB system; the dual system plays an important role in determining the observability Gramian for the QB system (3.1), and we aim at determining 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 of the dual system is the same as the reachability Gramian; here, we also consider the same analogy. If we compare the system (3.1) with the general nonlinear system as shown in (2.4), it turns out that for the system (3.1)

$$\mathcal{A}(x, u, t) = A + H(x \otimes I) + \sum_{k=1}^m N_k u_k, \quad \mathcal{B}(x, u, t) = B \quad \text{and} \quad \mathcal{C}(x, u, t) = C.$$

Using Lemma 2.6, we can write down the state-space realization of the adjoint operator of the QB system as follows:

$$(3.10a) \quad \dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)) + \sum_{k=1}^m N_k x(t) u_k(t) + Bu(t), \quad x(0) = 0,$$

$$(3.10b) \quad \dot{z}(t) = -A^T z(t) - (x(t)^T \otimes I) H^T z(t) - \sum_{k=1}^m N_k^T z(t) u_k(t) - C^T u_d(t),$$

$$z(\infty) = 0,$$

$$(3.10c) \quad y_d(t) = B^T z(t),$$

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. Next, we attempt to utilize the existing knowledge for the tensor multiplications and matricization to simplify the term $(x(t)^T \otimes I) H^T z(t)$ in the system (3.10) and to write it in the form of $x(t) \otimes z(t)$.

For this, we review some of the basic properties of tensor theory. Following [33], the *fiber* of a 3-dimensional tensor \mathcal{H} can be defined by fixing each index except one, e.g., $\mathcal{H}(:, j, k)$, $\mathcal{H}(j, :, k)$ and $\mathcal{H}(j, k, :)$. From the computational point of view, it is advantageous to consider the matrices associated with the tensor, which can be obtained via unfolding a tensor into a matrix. The process of unfolding a tensor into a matrix is called *matricization*, and the mode- μ matricization of the tensor \mathcal{H} is denoted by $\mathcal{H}^{(\mu)}$. For an l -dimensional tensor, there are l different possible ways to unfold the tensor into a matrix. We refer to [8, 33] for more detailed insights into matricization. Similar to matrix multiplications, one can carry out tensor multiplication using matricization of the tensor [33]. For instance, the mode- μ product of \mathcal{H} and a matrix $X \in \mathbb{R}^{n \times s}$ gives a tensor $\mathcal{F} \in \mathbb{R}^{s \times n \times n}$, satisfying

$$\mathcal{F} = \mathcal{H} \times_{\mu} X \quad \Leftrightarrow \quad \mathcal{F}^{(\mu)} = X \mathcal{H}^{(\mu)}.$$

Analogously, if we define a tensor-matrices product as:

$$\mathcal{F} = \mathcal{H} \times_1 X \times_2 Y \times_3 Z,$$

where $\mathcal{F} \in \mathbb{R}^{q_1 \times q_2 \times q_3}$, $X \in \mathbb{R}^{n \times q_1}$ and $Y \in \mathbb{R}^{n \times q_2}$ and $Z \in \mathbb{R}^{n \times q_3}$, then the following relations are fulfilled:

$$(3.11a) \quad \mathcal{F}^{(1)} = X^T \mathcal{H}^{(1)}(Y \otimes Z),$$

$$(3.11b) \quad \mathcal{F}^{(2)} = Z^T \mathcal{H}^{(2)}(Y \otimes X),$$

$$(3.11c) \quad \mathcal{F}^{(3)} = Y^T \mathcal{H}^{(3)}(Z \otimes X).$$

Coming back to the QB system, the matrix $H \in \mathbb{R}^{n \times n^2}$ in the system denotes a Hessian, which can be seen as an unfolding of a 3-dimensional tensor $\mathcal{H} \in \mathbb{R}^{n \times n \times n}$. Here, we choose the tensor $\mathcal{H} \in \mathbb{R}^{n \times n \times n}$ such that its mode-1 matricization is the same as the Hessian H , i.e., $H = \mathcal{H}^{(1)}$. Next, let us consider a tensor $\mathcal{T} \in \mathbb{R}^{1 \times n \times 1}$, whose mode-1 matricization $\mathcal{T}^{(1)}$ is given by

$$\mathcal{T}^{(1)} = z(t)^T H(x(t) \otimes I) = z(t)^T \mathcal{H}^{(1)}(x(t) \otimes I).$$

We then observe that the mode-1 matricization of the tensor \mathcal{T} is a transpose of the mode-2 matricization, i.e., $\mathcal{T}^{(1)} = (\mathcal{T}^{(2)})^T$, leading to

$$\mathcal{T}^{(1)} = \left(\mathcal{T}^{(2)} \right)^T = (x(t) \otimes z(t))^T (\mathcal{H}^{(2)})^T.$$

Therefore, we can rewrite the system (3.10) as:

$$(3.12a) \quad \dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)) + \sum_{k=1}^m N_k x(t) u_k(t) + Bu(t), \quad x(0) = 0,$$

$$(3.12b) \quad \dot{z}(t) = -A^T z(t) - \mathcal{H}^{(2)} x(t) \otimes z(t) - \sum_{k=1}^m N_k^T u_k(t) z(t) - C^T u_d(t), \quad z(\infty) = 0,$$

$$(3.12c) \quad y_d(t) = B^T z(t).$$

In the meantime, we like to point out that there are two possibilities to define $\mathcal{A}(x, u, t)$ in the case of a QB system. One is $\mathcal{A}(x, u, t) = A + H(x \otimes I) + \sum_{k=1}^m N_k u_k$, which we have used in the above discussion; however, there is another possibility to define $\mathcal{A}(x, u, t)$ as $\tilde{\mathcal{A}}(x, u, t) = A + H(I \otimes x) + \sum_{k=1}^m N_k u_k$, leading to the nonlinear Hilbert adjoint operator whose state-space realization is given as:

$$(3.13a) \quad \dot{x}(t) = Ax(t) + H(x(t) \otimes x(t)) + \sum_{k=1}^m N_k x(t) u_k(t) + Bu(t), \quad x(0) = 0,$$

$$(3.13b) \quad \dot{z}(t) = -A^T z(t) - \mathcal{H}^{(3)} x(t) \otimes z(t) - \sum_{k=1}^m N_k^T u_k(t) z(t) - C^T u_d(t), \quad z(\infty) = 0,$$

$$(3.13c) \quad y_d(t) = B^T z(t).$$

It can be noticed that the realizations (3.12) and (3.13) are the same, except the appearance of $\mathcal{H}^{(2)}$ in (3.12) instead of $\mathcal{H}^{(3)}$ in (3.13). Nonetheless, if one assumes that the Hessian H is symmetric, i.e., $H(u \otimes v) = H(v \otimes u)$ for $u, v \in \mathbb{R}^n$, then the mode-2 and mode-3 matricizations coincide, i.e., $\mathcal{H}^{(2)} = \mathcal{H}^{(3)}$. However, the Hessian H , obtained after discretization of the governing equations, may not be symmetric;

but as shown in [8] the Hessian can be modified in such a way that it becomes symmetric without any change in the system dynamics. Therefore, in the rest of the paper, without loss of generality, we assume that the Hessian H is symmetric.

Now, we turn our attention towards 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 as used for determining the reachability Gramian. Using the dual system (3.12), one can write the dual state $z(t)$ of the dual system at time t as follows:

$$\begin{aligned} z(t) &= \int_{\infty}^t e^{-A^T(t-\sigma_1)} C^T u_d(\sigma_1) d\sigma_1 + \sum_{k=1}^m \int_{\infty}^t e^{-A^T(t-\sigma_1)} N_k^T z(\sigma_1) u_k(\sigma_1) d\sigma_1, \\ &+ \int_{\infty}^t e^{-A^T(t-\sigma_1)} \mathcal{H}^{(2)}(x(\sigma_1) \otimes z(\sigma_1)) d\sigma_1, \end{aligned}$$

which after an appropriate change of variable leads to

$$\begin{aligned} (3.14) \quad z(t) &= \int_{\infty}^0 e^{A^T \sigma_1} C^T u^{(d)}(t + \sigma_1) d\sigma_1 + \sum_{k=1}^m \int_{\infty}^0 e^{A^T \sigma_1} N_k^T z(t + \sigma_1) u_k(t + \sigma_1) d\sigma_1 \\ &+ \int_{\infty}^0 e^{A^T \sigma_1} \mathcal{H}^{(2)}(x(t + \sigma_1) \otimes z(t + \sigma_1)) d\sigma_1. \end{aligned}$$

Equation (3.13a) gives the expression for $x(t + \sigma_1)$. This is

$$\begin{aligned} x(t + \sigma_1) &= \int_0^{t+\sigma_1} e^{A\sigma_2} B u(t + \sigma_1 - \sigma_2) d\sigma_2 + \sum_{k=1}^m \int_0^{t+\sigma_1} \left(e^{A\sigma_2} N_k x(t + \sigma_1 - \sigma_2) \right. \\ &\quad \left. \times u_k(t + \sigma_1 - \sigma_2) \right) d\sigma_2 + \int_0^{t+\sigma_1} e^{A\sigma_2} H(x(t + \sigma_1 - \sigma_2) \otimes x(t + \sigma_1 - \sigma_2)) d\sigma_2. \end{aligned}$$

We substitute for $x(t + \sigma_1)$ using the above equation, and $z(t + \sigma_1)$ using (3.14), which gives rise to the following expression:

$$\begin{aligned} (3.15) \quad z(t) &= \int_{\infty}^0 e^{A^T \sigma_1} C^T u_d(t + \sigma_1) d\sigma_1 + \sum_{k=1}^m \int_{\infty}^0 \int_{\infty}^0 e^{A^T \sigma_1} N_k^T \\ &\quad \times e^{A^T \sigma_2} C^T u_d(t + \sigma_1 + \sigma_2) u_k(t + \sigma_1) d\sigma_1 d\sigma_2 + \int_{\infty}^0 \int_0^{t+\sigma_1} \int_{\infty}^0 e^{A^T \sigma_1} \\ &\quad \times \mathcal{H}^{(2)}(e^{A\sigma_2} B \otimes e^{A^T \sigma_3} C^T) u(t + \sigma_1 - \sigma_2) u_d(t + \sigma_1 + \sigma_3) d\sigma_1 d\sigma_2 d\sigma_3 + \dots \end{aligned}$$

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. Carefully inspecting the kernels of the Volterra series of the dual system, we define the observability mapping \bar{Q} , similar to the reachability mapping, as follows:

$$(3.16) \quad \bar{Q} = [\bar{Q}_1, \bar{Q}_2, \bar{Q}_3, \dots],$$

in which

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

where $\bar{P}_i(t_1, \dots, t_i)$ are defined in (3.4). Based on the above observability mapping, we define the observability Gramian Q of the QB system as

$$(3.17) \quad Q = \sum_{i=1}^{\infty} Q_i \quad \text{with} \quad Q_i = \int_0^{\infty} \cdots \int_0^{\infty} \bar{Q}_i \bar{Q}_i^T dt_1 \cdots dt_i.$$

Analogous to the reachability Gramian, we next show a relation between the observability Gramian and the solution of a generalized Lyapunov equation.

THEOREM 3.2. *Consider the QB system (3.1) with a stable matrix A , and let Q , defined in (3.17), be the observability Gramian of the system and assume it exists. Then, the Gramian Q satisfies the following Lyapunov equation:*

$$(3.18) \quad A^T Q + Q A + \mathcal{H}^{(2)}(P \otimes Q)(\mathcal{H}^{(2)})^T + \sum_{k=1}^m N_k^T Q N_k + C^T C = 0,$$

where P is the reachability Gramian of the system, i.e., the solution of the generalized quadratic Lyapunov equation (3.5).

Proof. The proof of the above theorem is analogous to the proof of [Theorem 3.1](#); therefore, we skip it for the brevity of the paper. \square

REMARK 3.3. *As one would expect, the Gramians for QB systems reduce to the Gramians for bilinear systems [9] if the quadratic term is zero, i.e., $H = 0$.*

Furthermore, it will also be interesting to look at a truncated version of the Gramians of the QB system based on the leading kernels of the Volterra series. We call a truncated version of the Gramians *truncated* Gramians of QB systems. For this, let us consider approximate reachability and observability mappings as follows:

$$\tilde{P}_{\mathcal{T}} = [\tilde{P}_1, \tilde{P}_2, \tilde{P}_3], \quad \tilde{Q}_{\mathcal{T}} = [\tilde{Q}_1, \tilde{Q}_2, \tilde{Q}_3],$$

where

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

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

$$(3.19a) \quad P_{\mathcal{T}} = \sum_{i=1}^3 \widehat{P}_i, \quad \text{where} \quad \widehat{P}_i = \int_0^{\infty} \widetilde{P}_i(t_1, \dots, t_i) \widetilde{P}_i^T(t_1, \dots, t_i) dt_1 \cdots dt_i,$$

$$(3.19b) \quad Q_{\mathcal{T}} = \sum_{i=1}^3 \widehat{Q}_i, \quad \text{where} \quad \widehat{Q}_i = \int_0^{\infty} \widetilde{Q}_i(t_1, \dots, t_i) \widetilde{Q}_i^T(t_1, \dots, t_i) dt_1 \cdots dt_i,$$

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 3.4. *Let $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ be the truncated Gramians of the QB system as defined in (3.19). Then, $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ satisfy the following Lyapunov equations:*

$$(3.20a) \quad AP_{\mathcal{T}} + P_{\mathcal{T}}A^T + H(\widehat{P}_1 \otimes \widehat{P}_1)H^T + \sum_{k=1}^m N_k \widehat{P}_1 N_k^T + BB^T = 0, \quad \text{and}$$

$$(3.20b) \quad A^T Q_{\mathcal{T}} + Q_{\mathcal{T}}A + \mathcal{H}^{(2)}(\widehat{P}_1 \otimes \widehat{Q}_1)(\mathcal{H}^{(2)})^T + \sum_{k=1}^m N_k^T \widehat{Q}_1 N_k + C^T C = 0,$$

respectively, where P_1 and Q_1 are solutions to the following Lyapunov equations:

$$(3.21) \quad A\widehat{P}_1 + \widehat{P}_1A^T + BB^T = 0, \quad \text{and}$$

$$(3.22) \quad A^T \widehat{Q}_1 + \widehat{Q}_1A + C^T C = 0, \quad \text{respectively.}$$

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 (3.19a) and the truncated reachability Gramian $P_{\mathcal{T}}$ (3.5) 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 :

$$(3.23) \quad A\widehat{P}_1 + \widehat{P}_1A^T + BB^T = 0, \quad \text{and}$$

$$(3.24) \quad A\widehat{P}_2 + \widehat{P}_2A^T + \sum_{k=1}^m N_k \widehat{P}_1 N_k^T = 0.$$

Now, we consider the third term in the summation (3.19a). This is

$$\begin{aligned} P_3 &= \int_0^{\infty} \int_0^{\infty} \int_0^{\infty} \widetilde{P}_3(t_1, t_2, t_3) \widetilde{P}_3^T(t_1, t_2, t_3) dt_1 dt_2 dt_3 \\ &= \int_0^{\infty} \int_0^{\infty} \int_0^{\infty} e^{At_3} H(\widetilde{P}_1(t_1) \widetilde{P}_1^T(t_1) \otimes \widetilde{P}_1(t_2) \widetilde{P}_1^T(t_2)) H^T e^{A^T t_3} dt_1 dt_2 dt_3 \\ &= \int_0^{\infty} e^{At_3} H \left(\left(\int_0^{\infty} \widetilde{P}_1(t_1) \widetilde{P}_1^T(t_1) dt_1 \right) \otimes \left(\int_0^{\infty} \widetilde{P}_1(t_2) \widetilde{P}_1^T(t_2) dt_2 \right) \right) H^T e^{A^T t_3} dt_3 \\ &= \int_0^{\infty} e^{At_3} H \left(\widehat{P}_1 \otimes \widehat{P}_1 \right) H^T e^{A^T t_3} dt_3. \end{aligned}$$

Furthermore, we use the relation between the above integral representation and the solution of Lyapunov equation to show that \widehat{P}_3 solves:

$$(3.25) \quad A\widehat{P}_3 + \widehat{P}_3A^T + H(\widehat{P}_1 \otimes \widehat{P}_1)H^T = 0.$$

Summing (3.23), (3.24) and (3.25) yields

$$(3.26) \quad AP_{\mathcal{T}} + \mathcal{P}_{\mathcal{T}}A^T + H(\widehat{P}_1 \otimes \widehat{P}_1) + \sum_{k=1}^m N_k \widehat{P}_1 N_k + BB^T = 0.$$

Analogously, we can show that $Q_{\mathcal{T}}$ solves (3.20b), thus concluding the proof. \square

We will investigate the advantages of these truncated Gramians in the model reduction framework in the later part of the paper.

Next, we study the connection between the proposed Gramians for the QB system and energy functionals. Also, we show how the definiteness of the Gramians is related to reachability and observability of the QB systems. These all suggest us how to determine the state components that are hard to control as well as hard to observe.

4. Energy Functionals and MOR for QB systems. We start by establishing the conditions under which the Gramians approximate the energy functionals of the QB system, in the quadratic forms.

4.1. Comparison of energy functionals with Gramians. By using [Theorem 2.4](#), we obtain the following nonlinear partial differential equation, whose solution gives the controllability energy functional for the QB system:

$$(4.1) \quad \begin{aligned} & \frac{\partial L_c}{\partial x} (Ax + Hx \otimes x) + (Ax + Hx \otimes x)^T \frac{\partial L_c}{\partial x}^T \\ & + \frac{\partial L_c}{\partial x} ([N_1, \dots, N_m] (I_m \otimes x) + B) ([N_1, \dots, N_m] (I_m \otimes x) + B)^T \frac{\partial L_c}{\partial x}^T = 0. \end{aligned}$$

Unlike in the case of linear systems, the controllability energy functional $L_c(x)$ for nonlinear systems cannot be expressed as a simple quadratic form, i.e., $L_c(x) = x^T \widetilde{P}^{-1}x$, where \widetilde{P} is a constant matrix.

For nonlinear systems, the energy functionals are rather complicated nonlinear functions, depending on the state vector. Thus, we aim at providing some bounds between the quadratic form of the proposed Gramians for QB systems and energy functionals. For the controllability energy functional, we extend the reasoning given in [9, 10] for bilinear systems.

THEOREM 4.1. *Consider a controllable QB system (3.1) with a stable matrix A . Let $P > 0$ be its reachability Gramian which is the unique definite solution of the quadratic Lyapunov equation (3.6), and $L_c(x)$ denote the controllability energy functional of the QB system, solving (4.1). Then, there exists a neighborhood W of 0 such that*

$$L_c(x) \geq \frac{1}{2}x^T P^{-1}x, \text{ where } x \in W(0).$$

Proof. Consider a state x_0 and let a control input $u = u_0 : (-\infty, 0] \rightarrow \mathbb{R}^m$, which minimizes the input energy in the definition of $L_o(x_0)$ and steers the system from 0 to x_0 . Now, we consider the time-varying homogeneous nonlinear differential equation

$$(4.2) \quad \dot{\phi} = \left(A + H(\phi \otimes I) + \sum_{k=1}^m N_k u_k(t) \right) \phi =: A_u \phi(t),$$

and its fundamental solution $\Phi_u(t, \tau)$. The system (4.2) can thus be interpreted as

a time-varying system. The reachability Gramian of the time-varying control system [45, 47] $\dot{x} = A_u x(t) + Bu(t)$ is given by

$$P_u = \int_{-\infty}^0 \Phi(0, \tau) B B^T \Phi(0, \tau)^T d\tau.$$

The input u also steers the time-varying system from 0 to x_0 ; therefore, we have

$$\|u\|_{L_2}^2 \geq \frac{1}{2} x^T P_u^{-1} x.$$

An alternative way to determine P_u can be given by

$$P_u = \int_0^\infty \tilde{\Phi}(t, 0)^T B B^T \tilde{\Phi}(t, 0) dt,$$

where $\tilde{\Phi}$ is the fundamental solution of the following differential equation

$$(4.3) \quad \dot{\tilde{\Phi}} = \left(A^T + \mathcal{H}^{(2)}(x(-t) \otimes I) + \sum_{k=1}^m N_k^T u_k(-t) \right) \tilde{\Phi} \quad \text{with} \quad \tilde{\Phi}(t, t) = I,$$

and $x(t)$ is the solution of

$$\dot{x}(t) = Ax(t) + H(x \otimes x) + \sum_{k=1}^m N_k x(t) u_k(t) + Bu(t).$$

Then, we define $\eta(t)$, satisfying $\eta(t) = \tilde{\Phi}(t, 0)x_0$. Since it is assumed that the QB system is controllable, the state x_0 can be reached by using a finite input energy, i.e., $\|u\|_{L_2} < \infty$. Hence, the input $u(t)$ is a square-integrable function over $t \in (-\infty, 0]$ and so is $x(t)$. This implies that $\lim_{t \rightarrow \infty} \eta(t) \rightarrow 0$, provided A is stable. Thus, we have

$$\begin{aligned} x_0^T P x_0 &= - \int_0^\infty \frac{d}{dt} (\eta(t)^T P \eta(t)) dt \\ &= - \int_0^\infty \eta(t)^T \left(\left(A + H(x(-t) \otimes I) + \sum_{k=1}^m N_k u_k(-t) \right) P \right. \\ &\quad \left. + P \left(A^T + \mathcal{H}^{(2)}(x(-t) \otimes I) + \sum_{k=1}^m N_k^T u_k(-t) \right) \right) \eta(t) dt \\ &= - \int_0^\infty \eta(t)^T \left(AP + PA^T + H(P \otimes P)H^T + \sum_{k=1}^m N_k P N_k^T \right) \eta(t) \\ &\quad + \left(H(P \otimes P)H^T - H(x(-t) \otimes I)P - P\mathcal{H}^{(2)}(x(-t) \otimes I) \right. \\ &\quad \left. + \sum_{k=1}^m (N_k P N_k - P N_k^T u_k(-t) - N_k^T P u_k(-t)) \right) \eta(t) dt. \end{aligned}$$

Now, we have

$$\begin{aligned} - \int_0^\infty \eta(t)^T \left(AP + PA^T + H(P \otimes P)H^T + \sum_{k=1}^m N_k P N_k^T \right) \eta(t) \\ = \int_0^\infty \eta(t)^T B B^T \eta(t) = x_0^T P_u x_0. \end{aligned}$$

Hence, if

$$(4.4) \quad \int_0^\infty \eta(t)^T \left(H(P \otimes P)H^T - H(x(-t) \otimes I)P - P\mathcal{H}^{(2)}(x(-t) \otimes I) \right. \\ \left. + \sum_{k=1}^m (N_k P N_k - P N_k^T u_k(-t) - N_k^T P u_k(-t)) \right) \eta(t) dt \geq 0,$$

then $x_0^T P x_0 \geq x_0^T P_u x_0$. Further, if x_0 lies in a small ball W in the neighborhood of the origin, i.e., $x_0 \in W(0)$, then a small input u is sufficient to steer the system from 0 to x_0 and $x(t) \in W(0)$ for $t \in (-\infty, 0]$ which ensures that the relation (4.4) holds for all $x_0 \in W(0)$. Therefore, we have $x_0^T P^{-1} x_0 \leq x_0^T P_u^{-1} x_0$ if $x_0 \in W(0)$. \square

Similarly, we next show an upper bound for the observability energy functional for the QB system in terms of the observability Gramian (in the quadratic form).

THEOREM 4.2. *Consider the QB system (3.1) with $B \equiv 0$ and an initial condition x_0 , and let L_o be the observability energy functional. Let $P > 0$ and $Q \geq 0$ be solutions to the generalized Lyapunov equations (3.6) and (3.18), respectively. Then, there exists a neighborhood \widetilde{W} of the origin such that*

$$L_o(x_0) \leq \frac{1}{2} x^T Q x, \quad \text{where } x \in \widetilde{W}(0).$$

Proof. Using the definition of the observability energy functional, see [Definition 2.5](#), we have

$$(4.5) \quad L_o(x_0) = \max_{\substack{u \in \mathcal{B}_{(\alpha, \beta)} \\ x(0)=x_0, x(\infty)=0}} \frac{1}{2} \int_0^\infty \widetilde{L}_o(x_0, u) dt,$$

where $\mathcal{B}_{(\alpha, \beta)} \stackrel{\text{def}}{=} \{u \in L_2^m[0, \infty), \|u\|_{L_2} \leq \alpha, \|u\|_{L_\infty} \leq \beta\}$ and $\widetilde{L}_o(x_0, u) := \|y(t)\|_2$. Thus, we have

$$\widetilde{L}_o(x_0, u) = \|y(t)\|_2 = \|Cx(t)\|_2 = x(t)^T C^T C x(t).$$

Substituting for $C^T C$ from (3.18), we obtain

$$\widetilde{L}_o(x_0, u) = -2x(t)^T Q A x(t) - x(t)^T \mathcal{H}^{(2)} P \otimes Q (\mathcal{H}^{(2)})^T x(t) - \sum_{k=1}^m x(t)^T N_k^T Q N_k x(t).$$

Next, we substitute for Ax from (3.1) (with $B = 0$) to have

$$\begin{aligned} \widetilde{L}_o(x_0, u) = -2x(t)^T Q \dot{x}(t) + 2x(t)^T Q H x(t) \otimes x(t) + 2 \sum_{k=1}^m x(t)^T Q N_k x(t) u_k(t) \\ - x(t)^T \mathcal{H}^{(2)} (P \otimes Q) (\mathcal{H}^{(2)})^T x(t) - \sum_{k=1}^m x(t)^T N_k^T Q N_k x(t) \end{aligned}$$

$$\begin{aligned}
&= -\frac{d}{dt}(x(t)^T Q x(t)) + x(t)^T \left(QH(I \otimes x(t)) + QH(x(t) \otimes I) \right. \\
&\quad \left. + \sum_{k=1}^m (QN_k + N_k^T Q)u_k(t) - \mathcal{H}^{(2)}(P \otimes Q)(\mathcal{H}^{(2)})^T - \sum_{k=1}^m N_k^T Q N_k \right) x(t).
\end{aligned}$$

This gives

$$\begin{aligned}
L_o(x_0) &= \max_{u \in \mathcal{B}_{(\alpha, \beta)}} \frac{1}{2} \int_0^\infty \tilde{L}_o(x_0, u) dt, \\
&\quad x(0)=x_0, x(\infty)=0 \\
&= \frac{1}{2} x_0^T Q x_0 + \max_{u \in \mathcal{B}_{(\alpha, \beta)}} \frac{1}{2} \int_0^\infty x(t)^T \left(R_H(x, u) + \sum_{k=1}^m R_{N_k}(x, u) \right) x(t) dt, \\
&\quad x(0)=x_0, x(\infty)=0
\end{aligned}$$

where

$$\begin{aligned}
R_H(x, u) &:= QH(I \otimes x) + QH(x \otimes I) - \mathcal{H}^{(2)}(P \otimes Q)(\mathcal{H}^{(2)})^T, \\
R_{N_k}(x, u) &:= (QN_k u_k + N_k^T Q u_k - N_k^T Q N_k).
\end{aligned}$$

First, note that if for a vector v , $v^T N_k^T Q N_k v = 0$, then $Q N_k v = 0$. Therefore, there exist inputs u for which $\|u\|_{L_\infty}$ is small, ensuring $R_{N_k}(x, u)$ is a negative semidefinite. Similarly, if for a vector w , $w^T \mathcal{H}^{(2)}(P \otimes Q)(\mathcal{H}^{(2)})^T w = 0$ and $P > 0$, then $(I \otimes Q)(\mathcal{H}^{(2)})^T w = 0$. Using (3.11), it can be shown that $QH(w \otimes I) = QH(I \otimes w) = 0$. Now, we consider an initial condition x_0 lies in the small neighborhood of the origin and $u \in \mathcal{B}_{(\alpha, \beta)}$ ensuring that the resulting trajectory $x(t)$ for all time t is such that $R_H(x, u)$ is a negative semi-definite. Finally, we get

$$L_o(x_0) - \frac{1}{2} x_0^T Q x_0 \leq 0,$$

for x_0 lies in the neighborhood of the origin and for the inputs u , having small L_2 and L_∞ norms and $x_0 \in \tilde{W}(0)$. This concludes the proof. \square

Until this point, we have proven that in the neighborhood of the origin, the energy functionals of the QB system can be approximated by the Gramians in the quadratic form. However, one can also prove similar bounds for the energy functionals using the truncated Gramians for QB systems (defined in Corollary 3.4). We summarize this in the following corollary.

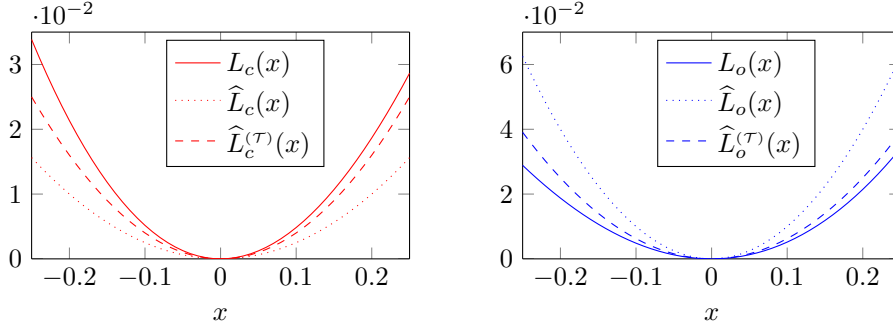
COROLLARY 4.3. *Consider the system (3.1), having a stable matrix A , to be locally reachable and observable. Let $L_c(x)$ and $L_o(x)$ be controllability and observability energy functionals of the system, respectively, and the truncated Gramians $P_{\mathcal{T}} > 0$ and $Q_{\mathcal{T}} > 0$ be solutions to the Lyapunov equations as shown in Corollary 3.4. Then,*

(i) *there exists a neighborhood $W_{\mathcal{T}}$ of the origin such that*

$$L_c(x) \geq \frac{1}{2} x^T P_{\mathcal{T}}^{-1} x, \text{ where } x \in W_{\mathcal{T}}(0).$$

(ii) *Moreover, there also exists a neighborhood $\tilde{W}_{\mathcal{T}}$ of the origin, where*

$$L_o(x) \leq \frac{1}{2} x^T Q_{\mathcal{T}} x, \text{ where } x \in \tilde{W}_{\mathcal{T}}(0).$$



(a) Comparison of the controllability energy functional and its approximations. (b) Comparison of the observability energy functional and its approximations.

Figure 4.1: Comparison of exact energy functionals with approximated energy functionals via the Gramians and truncated Gramians.

In what follows, we illustrate the above bounds using Gramians and truncated Gramians by considering a scalar dynamical system, where A, H, N, B, C are scalars, and are denoted by a, h, n, b, c , respectively.

EXAMPLE 4.4. Consider a scalar system (a, h, n, b, c) , where $a < 0$ (stability) and nonzero h, b, c . For simplicity, we take $n = 0$ so that we can easily obtain analytic expressions for the controllability and observability energy functionals, denoted by $L_c(x)$ and $L_o(x)$, respectively. Assume that the system is reachable on \mathbb{R} . Then, $L_c(x)$ and $L_o(x)$ can be determined via solving partial differential equations (2.2) and (2.3) (with $g(x) = 0$), respectively. These are:

$$L_c(x) = -\left(ax^2 + \frac{2}{3}hx^3\right) \frac{1}{b^2}, \quad L_o(x) = -\frac{c^2}{2h} \left(x - \frac{a}{h} \log\left(\frac{a+hx}{a}\right)\right),$$

respectively. The quadratic approximations of these energy functionals by using the Gramians, are:

$$\begin{aligned} \widehat{L}_c(x) &= \frac{x^2}{2P} & \text{with } P &= -\frac{-a - \sqrt{a^2 - h^2b^2}}{h^2}, \\ \widehat{L}_o(x) &= \frac{Qx^2}{2} & \text{with } Q &= -\frac{c^2}{2a + h^2P}, \end{aligned}$$

and the approximations in terms of the truncated Gramians are:

$$\begin{aligned} \widehat{L}_c^{(\tau)}(x) &= \frac{x^2}{2P_\tau} & \text{with } P_\tau &= -\frac{h^2b^4 + 4a^2b^2}{8a^3}, \\ \widehat{L}_o^{(\tau)}(x) &= \frac{Q_\tau x^2}{2} & \text{with } Q_\tau &= -\frac{h^2b^2c^2 + 4a^2c^2}{8a^3}. \end{aligned}$$

In order to compare these functionals, we set $a = -2, b = c = 2$ and $h = 1$ and plot the resulting energy functionals in Figure 4.1.

Clearly, Figure 4.1 illustrates the lower and upper bounds for the controllability and observability energy functionals, respectively at least locally. Moreover, we observe

that the bounds for the energy functionals, given in terms of truncated Gramians are closer to the actual energy functionals of the system in the small neighborhood of the origin.

So far, we have shown the bounds for the energy functionals in terms of the Gramians of the QB system. In order to prove those bounds, it is assumed that P is a positive definite. However, this assumption might not be fulfilled for many QB systems, especially arising from semi-discretization of nonlinear PDEs. Therefore, our next objective is to provide another interpretation of the proposed Gramians and truncated Gramians, that is, the connection of Gramians and truncated Gramians with reachability and observability of the system. For the observability energy functional, we consider the output y of the following *homogeneous* QB system:

$$(4.6) \quad \begin{aligned} \dot{x}(t) &= Ax + Hx(t) \otimes x(t) + \sum_{k=1}^m N_k x(t) u_k(t), \\ y(t) &= Cx(t), \quad x(0) = x_0, \end{aligned}$$

as considered for bilinear systems in [9, 25]. However, it might also be possible to consider an *inhomogeneous* system by setting the control input u completely zero, as shown in [39]. We first investigate how the proposed Gramians are related to reachability and observability of the QB systems, analogues to derivation for bilinear systems in [9].

THEOREM 4.5.

- (a) Consider the QB system (3.1), and assume the reachability Gramian P to be the solution of (3.6). If the system is steered from 0 to x_0 , where $x_0 \notin \text{Im}P$, then $L_c(x_0) = \infty$ for all input functions u .
- (b) Furthermore, consider the homogeneous QB system (4.6) and assume $P > 0$ and Q to be the reachability and observability Gramians of the QB system which are solutions of (3.6) and (3.18), respectively. If the initial state satisfies $x_0 \in \text{Ker}Q$, then $L_o(x_0) = 0$.

Proof.

- (a) By assumption, P satisfies

$$(4.7) \quad AP + PA^T + H(P \otimes P)H^T + \sum_{k=1}^m N_k P N_k^T + BB^T = 0.$$

Next, we consider a vector $v \in \text{Ker}P$ 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 $B^T v = 0$, $P N_k^T v = 0$ and $(P \otimes P)H^T v = 0$. From (4.7), we thus obtain $PA^T v = 0$. Now we consider an arbitrary state vector $x(t)$, which is the solution of (3.1) at time t for any given input function u . If $x(t) \in \text{Im}P$

for some t , then we have

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

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

- (b) Following the above discussion, we can show that $(I \otimes Q) (\mathcal{H}^{(2)})^T \text{Ker}Q = 0$, $Q N_k \text{Ker}Q = 0$, $Q A \text{Ker}Q = 0$, and $C \text{Ker}Q = 0$. Let $x(t)$ denote the solution of the homogeneous system at time t . If $x(t) \in \text{Ker}Q$ and a vector $\tilde{v} \in \text{Im}Q$, then we have

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

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

The above theorem suggests that the state components, belonging to $\text{Ker}P$ or $\text{Ker}Q$, do not play a major role as far as the system dynamics are concerned. This shows that the states which belong to $\text{Ker}P$, are uncontrollable, and similarly, the states, lying in $\text{Ker}Q$ are unobservable once the uncontrollable states are removed. Furthermore, we have shown in [Theorems 4.1](#) and [4.2](#) the lower and upper bounds for the controllability and observability energy functions in the quadratic form of the Gramians P and Q of QB systems (at least in the neighborhood of the origin). This coincides with the concept of balanced truncation model reduction which aims at eliminating weakly controllable and weakly observable state components. Such states are corresponding to zero or small singular values of P and Q . In order to find these states simultaneously, we utilize the balancing tools similar to the linear case; see, e.g., [\[1, 2\]](#). 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

$$(4.8) \quad \begin{aligned} \dot{\tilde{x}}(t) &= \tilde{A} \tilde{x}(t) + \tilde{H} \tilde{x}(t) \otimes \tilde{x}(t) + \sum_{k=1}^m \tilde{N}_k \tilde{x}(t) u_k(t) + \tilde{B} u(t), \\ y(t) &= \tilde{C} \tilde{x}(t), \quad \tilde{x}(0) = 0 \end{aligned}$$

with

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

Then, the above transformed system [\(4.8\)](#) is a balanced system, as the Gramians \tilde{P} and \tilde{Q} of the system [\(4.8\)](#) are equal and diagonal, i.e., $\tilde{P} = \tilde{Q} = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_n)$.

The attractiveness of the balanced system is that it allows us to find state components corresponding to small singular values of both \tilde{P} and \tilde{Q} . If $\sigma_{\hat{n}} > \sigma_{\hat{n}+1}$, for some $\hat{n} \in \mathbb{N}$, then it is easy to see that states related to $\{\sigma_{\hat{n}+1}, \dots, \sigma_n\}$ are not only hard to control but also hard to observe; hence, they can be eliminated. In order to determine a reduced system of order \hat{n} , we partition $T = [T_1 \ T_2]$ and $T^{-1} = [S_1^T \ S_2^T]^T$, where $T_1, S_1^T \in \mathbb{R}^{n \times \hat{n}}$, and define the reduced-order system's realization as follows:

$$(4.9) \quad \hat{A} = S_1 A T_1, \quad \hat{H} = S_1 H (T_1 \otimes T_1), \quad \hat{N}_k = S_1 N_k T_1, \quad \hat{B} = S_1 B, \quad \hat{C} = C T_1,$$

which is generally a locally good approximate of the original system; though 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.

Based on the above discussions, we propose the following corollary, showing how the truncated Gramians of a QB system relate to reachability and observability of the system.

COROLLARY 4.6.

- (a) Consider the QB system (3.1), and let $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ be the truncated Gramians of the system, which are solutions of the Lyapunov equations as in (3.20). If the system is steered from 0 to x_0 where, $x_0 \notin \text{Im} P_{\mathcal{T}}$, then $L_c(x_0) = \infty$ for all input functions u .
- (b) Assume the QB system (3.1) is locally controllable around the origin, i.e., (A, B) is controllable. Then, for the homogeneous QB system (4.6), if the initial state $x_0 \in \text{Ker} Q_{\mathcal{T}}$, then $L_o(x_0) = 0$.

The above corollary can be proven, along of the lines of the proof for Theorem 4.5, keeping in mind that if $\gamma \in \text{Ker} P_{\mathcal{T}}$, then γ also belongs to $\text{Ker} P_1$, where P_1 is the solution to (3.21). Similarly, if $\xi \in \text{Ker} Q_{\mathcal{T}}$, then ξ also lies in $\text{Ker} Q_1$, where Q_1 is the solution to (3.22). This can easily be verified using simple linear algebra. Having noted this, Corollary 4.6 also suggests that $\text{Ker} P_{\mathcal{T}}$ is uncontrollable, and $\text{Ker} Q_{\mathcal{T}}$ is also unobservable if the system is locally controllable. Moreover, these truncated Gramians also bound the energy functions for QB systems in the quadratic form, see Corollary 4.3. Based on these, we conclude that the truncated Gramians are also a good candidate to use for balancing the system and to compute the reduced-order systems.

5. Computational Issues and Advantages of Truncated Gramians. Up to now, we have proposed the Gramians for the QB systems and showed their relations to energy functionals of the system which allows us to determine the reduced-order systems. Here, we discuss computational issues and the advantages of considering this truncated Gramians in the MOR framework. Towards this end, we address stability issues of the reduced-order systems, obtained by using the truncated Gramians.

5.1. Computational issues. One of the major concerns in applying balanced truncation MOR is that it requires the solutions of two Lyapunov equations (3.6) and (3.18). These equations are quadratic in nature, which are not trivial to solve, and they appear to be computationally expensive. So far, it is not clear how to solve these generalized quadratic Lyapunov equation efficiently; however, under some assumptions, a fix point iteration scheme can be employed, which is based on the theory of convergent splitting presented in [20, 43]. This has been studied for solving generalized Lyapunov equation for bilinear systems in [19], wherein the proposed

stationary method is as follows:

$$(5.1) \quad \mathcal{L}(X_i) = \mathcal{N}(X_{i-1}) - BB^T, \quad i = 1, 2, \dots,$$

with $\mathcal{L}(X) = AX + XA^T$ and $\mathcal{N}(X_i) = -\sum_{k=1}^m N_k X_i N_k^T$. To perform this stationary iteration, we require the solution of a conventional Lyapunov equation at each iteration. Assuming $\sigma(A) \subset \mathbb{C}^-$ and spectral radius of $\mathcal{L}^{-1}\mathcal{N} < 1$, the iteration (5.1) linearly converges to a positive semidefinite solution X of the generalized Lyapunov equation for bilinear systems, which is

$$AX + XA^T + \sum_{k=1}^m N_k X N_k^T + BB^T = 0.$$

Later on, the efficiency of this iterative method was improved in [44] by utilizing tools for inexact solution of $Ax = b$. The main idea was to determine a low-rank factor of $\mathcal{N}(X_{i-1}) - BB^T$ by truncating the columns, corresponding to small singular values and to increase the accuracy of the low-rank solution of the linear Lyapunov equation (5.1) with each iteration. In total, this results in an efficient method to determine a low-rank solution of the generalized Lyapunov equation for bilinear systems with the desired tolerance. For detailed insights, we refer to [44].

One can utilize the same tools to determine the solutions of generalized quadratic-type Lyapunov equations. We begin with the inexact form equation, which on convergence gives the reachability Gramian; this is,

$$(5.2) \quad \mathcal{L}(X_i) = \Pi(X_{i-1}) - BB^T, \quad i = 1, 2, \dots$$

where $\mathcal{L}(X) = AX + XA^T$ and $\Pi(X) = -H(X \otimes X)H^T - \sum_{k=1}^m N_k X N_k^T$. Similar to the bilinear case, if $\sigma(A) \subset \mathbb{C}^-$ and the spectral radius of $\mathcal{L}^{-1}\Pi < 1$, then the iteration (5.2) converges to a positive semidefinite solution of the generalized quadratic Lyapunov equation. Next, we determine a low-rank approximation of $\Pi(X) = -H(X \otimes X)H^T - \sum_{k=1}^m N_k X N_k^T$. For this, let us assume a low-rank product $X := FDF^T$, where $F \in \mathbb{R}^{n \times k}$ and a QR decomposition of $F := Q_F R_F$. We then perform an eigenvalue decomposition of the relatively small matrix $R_F D R_F^T := U \Sigma U^T$, where $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_k)$ with $\sigma_j \geq \sigma_{j+1}$. Assuming there exists a scalar β such that

$$\sqrt{\sigma_{\beta+1}^2 + \dots + \sigma_k^2} \leq \tau \sqrt{\sigma_1^2 + \dots + \sigma_k^2},$$

where τ is a given tolerance, this gives us a low-rank approximation of X as:

$$X \approx \tilde{F} \tilde{D} \tilde{F}^T,$$

where $\tilde{F} = Q_F \tilde{U}$ and $\tilde{D} = \text{diag}(\sigma_1, \dots, \sigma_\beta)$. Following the short-hand notation, we denote $\tilde{Z} = \mathcal{T}_\tau(Z)$ which gives the low-rank approximation of ZZ^T with the tolerance τ , i.e., $ZZ^T \approx \tilde{Z}\tilde{Z}^T$. Considering a low-rank factor of $X_{k-1} \approx Z_{k-1} Z_{k-1}^T$, the right side of (5.2)

$$\begin{aligned} \Pi(X_{k-1}) - BB^T &\approx -[H(Z_{k-1} \otimes Z_{k-1}), [N_1, \dots, N_m] Z_{k-1}, B] \\ &\quad \times [H(Z_{k-1} \otimes Z_{k-1}), [N_1, \dots, N_m] Z_{k-1}, B]^T \end{aligned}$$

can be replaced with its truncated version $\mathcal{T}_\tau(\Pi(X_{k-1}) - BB^T) =: -\mathbb{F}_k \mathbb{F}_k^T$ with the desired tolerance. This indicates that we now need to solve the following linear Lyapunov equation at each step:

$$(5.3) \quad AX_k + X_k A = -\mathbb{F}_k \mathbb{F}_k^T,$$

Algorithm 5.1 Iterative scheme to determine Gramians for QB systems.

Input: System matrices $A, H, N_1, \dots, N_m, B, C$ and tolerance τ .

Output: Low-rank factors of the Gramians: Z_k ($P \approx Z_k Z_k^T$) and X_k ($Q \approx X_k X_k^T$).

- 1: Solve approximately $AM + MA^T + BB^T = 0$ for $P_1 \approx Z_1 Z_1^T$.
 - 2: Solve approximately $A^T G + GA + C^T C = 0$ for $Q_1 \approx X_1 X_1^T$.
 - 3: **for** $k = 2, 3, \dots$ **do**
 - 4: Determine low-rank factors:

$$\mathbb{B}_k = \mathcal{T}_\tau([H(Z_{k-1} \otimes Z_{k-1}), N_1 Z_{k-1}, \dots, N_m Z_{k-1}, B]),$$

$$\mathbb{C}_k = \mathcal{T}_\tau([\mathcal{H}^{(2)}(Z_{k-1} \otimes X_{k-1}), N_1^T X_{k-1}, \dots, N_m^T X_{k-1}, C^T]).$$
 - 5: Solve approximately $AM + MA^T + \mathbb{B}_k \mathbb{B}_k^T = 0$ for $P_k \approx Z_k Z_k^T$.
 - 6: Solve approximately $A^T G + GA + \mathbb{C}_k \mathbb{C}_k^T = 0$ for $Q_k \approx X_k X_k^T$.
 - 7: **if** solutions are sufficiently accurate **then** stop.
 - 8: **end if**
 - 9: **end for**
-

which can be solved very efficiently by using any of the recently developed low-rank solvers for Lyapunov equations; see, e.g., [13, 46]. In the following, we outline all the necessary steps in Algorithm 5.1 to determine the Gramians by summarizing the all above discussed ingredients.

REMARK 5.1. *At step 7 of Algorithm 5.1, one can check the accuracy of solutions by measuring the relative changes in the solutions, i.e., $\frac{\|P_k - P_{k-1}\|}{\|P_k\|}$ and $\frac{\|Q_k - Q_{k-1}\|}{\|Q_k\|}$. When these relative changes are smaller than a tolerance level, e.g. the machine precision, then one can stop the iterations to have sufficiently accurate solutions of the quadratic Lyapunov equations.*

REMARK 5.2. *In order to employ Algorithm 5.1, the right side of the conventional Lyapunov equation (see step 4) requires the computation of $H(Z_i \otimes Z_i) =: \Gamma$ at each step, which is also computationally and memory-wise expensive. If $Z_i \in \mathbb{R}^{n \times n_z}$, then the direct multiplication of $Z_i \otimes Z_i$ would have complexity of $\mathcal{O}(n^2 \cdot n_z^2)$, leading to an unmanageable task for large-scale systems, even on modern computer architectures. However, it is shown in [8] that Γ can be determined efficiently by making use of the tensor multiplication properties, which are also reported in the previous section. In the following, we provide the procedure to compute Γ efficiently:*

- Determine $\mathcal{Y} \in \mathbb{R}^{n_z \times n \times n}$ such that $\mathcal{Y}^{(2)} = Z_i^T \mathcal{H}^{(2)}$.
- Determine $\mathcal{K} \in \mathbb{R}^{n \times n_z \times n_z}$ such that $\mathcal{K}^{(3)} = Z_i^T \mathcal{Y}^{(3)}$.
- Then, $\Gamma = \mathcal{K}^{(1)}$.

This way, we can avoid determining the full matrix $Z_i \otimes Z_i$. Analogously, we can also compute the term $\mathcal{H}^{(2)}(Z_i \otimes X_i)$.

Next, we discuss the existence of the solutions of quadratic type generalized Lyapunov equations. As noted Algorithm 5.1, one can determine the solution of these Lyapunov equations using fixed point iterations. In the following, we discuss sufficient conditions under which these iterations converge to finite solutions.

THEOREM 5.3. *Consider a QB system as defined in (3.1) and let P and Q be its reachability and observability Gramians, respectively. Assume that the Gramians P and Q are determined using fixed point iterations as shown in Algorithm 5.1. Then,*

the Gramian P converges to a positive semidefinite solution if

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

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

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

and $\|P\|$ is bounded by

$$(5.4) \quad \|P\| \leq \frac{2\alpha}{\beta^2 \Gamma_H} \left(\mathcal{D} - \sqrt{\mathcal{D}^2 - 4 \frac{\beta^2 \Gamma_H}{2\alpha} \frac{\beta^2 \Gamma_B}{2\alpha}} \right) =: \mathcal{P}_\infty.$$

Furthermore, the Gramian Q also converges to a positive semidefinite solution if in addition to the above conditions (i)–(iii), the following condition satisfies

$$(5.5) \quad \frac{\beta^2}{2\alpha} \left(\Gamma_N + \tilde{\Gamma}_H \mathcal{P}_\infty \right) < 1,$$

where $\tilde{\Gamma}_H := \|\mathcal{H}^{(2)}\|^2$. Moreover, $\|Q\|$ is bounded by

$$(5.6) \quad \|Q\| \leq \frac{\beta^2}{2\alpha} \Gamma_C \left(1 - \frac{\beta^2}{2\alpha} \left(\Gamma_N + \tilde{\Gamma}_H \mathcal{P}_\infty \right) \right)^{-1},$$

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

Proof. Let us first consider the equation corresponding to P_1 :

$$(5.7) \quad AP_1 + AP_1 + BB^T = 0.$$

Alternatively, if A is stable, we can write P_1 in the integral form as

$$(5.8) \quad P_1 = \int_0^\infty e^{At} BB^T e^{A^T t} dt,$$

implying

$$(5.9) \quad \|P_1\| \leq \beta^2 \|BB^T\| \int_0^\infty e^{-2\alpha t} dt = \frac{\beta^2 \Gamma_B}{2\alpha},$$

where $\Gamma_B := \|BB^T\|$. Next, we look at the equation corresponding to P_k , which is given in terms of P_{k-1} :

$$(5.10) \quad AP_k + P_k A^T + H(P_{k-1} \otimes P_{k-1})H^T + \sum_{k=1}^m N_k P_{k-1} N_k + BB^T = 0.$$

We can also write P_k in an integral form, provided A is stable:

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

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 P_{k-1} in order to provide an upper bound for P_k and apply [Lemma A.1](#), then we know that $\lim_{k \rightarrow \infty} \|P_k\|$ is bounded if

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

and $\lim_{k \rightarrow \infty} \|P_k\|$ is bounded by

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

Now, we consider the equation corresponding to Q_1 :

$$A^T Q_1 + A^T Q_1 + C^T C = 0,$$

which again can be rewritten as:

$$Q_1 = \int_0^\infty e^{A^T t} C^T C e^{A t} dt$$

if A is stable. This implies

$$\|Q_1\| \leq \beta^2 \Gamma_C \int_0^\infty e^{-2\alpha t} dt = \beta^2 \frac{\Gamma_C}{2\alpha},$$

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

$$A^T Q_k + Q_k A + \mathcal{H}^{(2)}(P_{k-1} \otimes Q_{k-1}) \left(\mathcal{H}^{(2)} \right)^T + \sum_{k=1}^m N_k^T Q_{k-1} N_k + C^T C = 0.$$

A similar analysis for Q_k yields

$$\|Q_k\| \leq \frac{\beta^2}{2\alpha} \left(\left(\Gamma_N + \tilde{\Gamma}_H \|P_{k-1}\| \right) \|Q_{k-1}\| + \Gamma_C \right),$$

where $\tilde{\Gamma}_H := \|\mathcal{H}^{(2)}\|$. Since $\|P_{k-1}\| \leq \mathcal{P}_\infty$ for all $k \geq 1$, we further have

$$\|Q_k\| \leq \frac{\beta^2}{2\alpha} \left(\left(\Gamma_N + \tilde{\Gamma}_H \mathcal{P}_\infty \right) \|Q_{k-1}\| + \Gamma_C \right).$$

An additional sufficient condition under which the above recurrence formula in $\|Q_k\|$ converges is as follows:

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

and $\lim_{k \rightarrow \infty} \|Q_k\|$ is then bounded by

$$\lim_{k \rightarrow \infty} \|Q_k\| \leq \frac{\beta^2}{2\alpha} \Gamma_C \left(1 - \frac{\beta^2}{2\alpha} \left(\Gamma_N + \tilde{\Gamma}_H \mathcal{P}_\infty \right) \right)^{-1}.$$

This concludes the proof. □

Algorithm 5.2 Balanced truncation for QB systems (truncated version).

Input: System matrices A, H, N_k, B and C , and the order of the reduced system \hat{n} .

Output: The reduced system's matrices $\hat{A}, \hat{H}, \hat{N}_k, \hat{B}, \hat{C}$.

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 $S^T R$:

$$S^T R = U \Sigma V = [U_1 \ U_2] \text{diag}(\Sigma_1, \Sigma_2) [V_1 \ V_2]^T,$$

where Σ_1 contains the \hat{n} largest singular values of $S^T R$.

3: Construct the projection matrices \mathcal{V} and \mathcal{W} :

$$\mathcal{V} = S U_1 \Sigma_1^{-\frac{1}{2}} \text{ and } \mathcal{W} = R V_1 \Sigma_1^{-\frac{1}{2}}.$$

4: Determine the reduced-order system's realization:

$$\hat{A} = \mathcal{W}^T A \mathcal{V}, \quad \hat{H} = \mathcal{W}^T H (\mathcal{V} \otimes \mathcal{V}), \quad \hat{N}_k = \mathcal{W}^T N_k \mathcal{V}, \quad \hat{B} = \mathcal{W}^T B, \quad \hat{C} = C \mathcal{V}.$$

REMARK 5.4. In [Algorithm 5.1](#), we propose to determine the low-rank solutions of the Lyapunov equation at each intermediate step with the same tolerance. However, one can also consider to increase the tolerance adaptively for computing the low-rank solution of the Lyapunov equation with each iteration which probably can speed up even more, see [\[44\]](#) for the generalized Lyapunov equations for bilinear systems.

5.2. MOR using truncated Gramians. As noted in [Section 4](#), the quadratic forms of both actual Gramians and its truncated versions (truncated Gramians) impose bounds for the energy functionals of QB systems, at least in the neighborhood of the origin, and we also provide the interpretation of reachability and observability of the system in terms of Gramians and truncated Gramians. We have seen in the previous subsection that determining Gramians P and Q is very challenging task for large-scale settings. Moreover, the convergence of [Algorithm 5.1](#) highly depends on the spectral radius condition $\mathcal{L}^{-1}\Pi$, which should be less than 1. This condition might not be satisfied for large H and N_k ; thus, [Algorithm 5.1](#) may not converge. On the other hand, in order to compute the truncated Gramians, there is no such convergence issue. Furthermore, it can also be observed that the bounds for energy functionals using the truncated Gramians can be much better (in the neighborhood of the origin), for example see [Figure 4.1](#).

Additionally, if we remove those states that are completely uncontrollable and completely unobservable, then the truncated Gramians may provide reduced systems which are of smaller orders as compared to using the Gramians of QB systems. This is due to the fact that $P \geq P_{\mathcal{T}}$ and $Q \geq Q_{\mathcal{T}}$. This motivates us to use 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 5.2](#). Furthermore, we will see in [Section 6](#) as well that these truncated Gramians also yield very good qualitative reduced-order systems for QB systems.

5.3. Stability Preservation. We now discuss the stability of the reduced-order systems, obtained by using [Algorithm 5.2](#). For this, we consider only the autonomous part of the QB system as follows:

$$(5.11) \quad \dot{x}(t) = Ax(t) + H x(t) \otimes x(t),$$

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 the latter stability.

DEFINITION 5.5. Consider a QB system with $u \equiv 0$ (5.11). If there exists a Lyapunov function $\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{R}$ such that

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

where $\mathcal{B}_{0,r}$ is a ball of radius r centered around 0, then $x_{eq} = 0$ is a 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 [14], which might be difficult to apply in the large-scale setting. In this paper, we stick to the notion of the Lyapunov-based stability for the reduced-order systems.

THEOREM 5.6. Consider the QB system (3.1) with a stable matrix A . Let $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ be its truncated reachability and observability Gramians, defined in Corollary 4.6, respectively. If the reduced-order system is determined as shown in Algorithm 5.2, then for a Lyapunov function $\mathcal{F}(\hat{x}) = \hat{x}^T \Sigma_1 \hat{x}$, we have

$$\mathcal{F}(\hat{x}) > 0, \quad \frac{d}{dt}(\mathcal{F}(\hat{x})) < 0 \quad \forall \hat{x} \in \mathcal{B}_{0,r} \setminus \{0\},$$

where $r = \frac{\sigma_{\min}(\mathcal{V}^T \mathcal{G} \mathcal{V})}{2\|\Sigma_1\| \|\hat{H}\|}$ and $\mathcal{G} = \mathcal{H}^{(2)}(P_1 \otimes Q_1) (\mathcal{H}^{(2)})^T + \sum_{k=1}^m N_k^T Q_1 N_k + C^T C$ with P_1 and Q_1 being the solutions of (3.21) and (3.22), respectively.

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

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

Keeping in mind the above relation, we get

$$\begin{aligned} (5.12) \quad \hat{A}^T \Sigma_1 + \Sigma_1 \hat{A} + \mathcal{V}^T \mathcal{G} \mathcal{V} &= \mathcal{V}^T A^T \mathcal{W} \Sigma_1 + \Sigma_1 \mathcal{W}^T A \mathcal{V} + \mathcal{V}^T \mathcal{G} \mathcal{V} \\ &= \mathcal{V}^T A^T Q_{\mathcal{T}} \mathcal{V} + \mathcal{V}^T Q_{\mathcal{T}} A \mathcal{V} + \mathcal{V}^T \mathcal{G} \mathcal{V} = \mathcal{V}^T (A^T Q_{\mathcal{T}} + Q_{\mathcal{T}} A + \mathcal{G}) \mathcal{V} = 0. \end{aligned}$$

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

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

Substituting $\hat{A}^T \Sigma_1 + \Sigma_1 \hat{A} = -\mathcal{V}^T \mathcal{G} \mathcal{V}$ from (5.12) in the above equation yields

$$(5.13) \quad \frac{d}{dt} \mathcal{F}(\hat{x}) = -\hat{x}^T \mathcal{V}^T \mathcal{G} \mathcal{V} \hat{x} + 2\hat{x}^T \Sigma_1 \hat{H}(\hat{x} \otimes \hat{x}).$$

As

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

implying

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

inserting the above inequality in (5.13) leads to

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

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

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

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

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

This concludes the proof. \square

6. Numerical Experiments. In this section, we consider MOR of several QB control systems and evaluate the efficiency of the proposed balanced truncation technique (Algorithm 5.2). For this, we need to solve a number of conventional Lyapunov equations. In our numerical experiments, we determine low-rank factors of these Lyapunov equations by using the ADI method as proposed in [11]. We compare the proposed methodology with the existing moment-matching techniques for QB systems, namely one-sided moment-matching [29] and its recent extension to two-sided moment-matching [8]. 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* [30] on the corresponding linear part. This leads to a reduced QB system of order $\hat{n} = 2l$. All the simulations are done on MATLAB[®] Version 8.0.0.783(R2012b)64-bit(glnxa64) on a board with 4 Intel[®] Xeon[®] E7-8837 CPUs with a 2.67-GHz clock speed, 8 Cores each and 1TB of total RAM, openSUSE Linux 12.04.

6.1. 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., [4, 15, 29, 34, 36]. 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 [29], 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

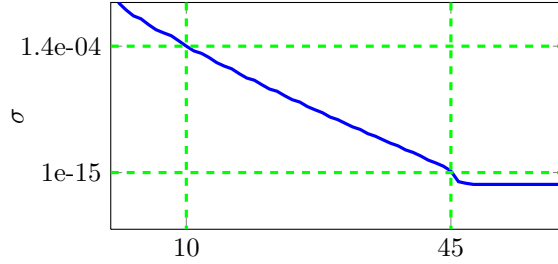


Figure 6.1: A RC ladder: decay of the normalized singular values based the truncated Gramians, and the dotted lines show the normalized singular value for $\hat{n} = 10$ and the order of the reduced system corresponding to the normalized singular value $1e-15$.

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 Figure 6.1, 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 $\hat{n} = 10$ by using balanced truncation. Also, we determine 5 \mathcal{H}_2 -optimal linear interpolation points and compute reduced-order systems of order $\hat{n} = 10$ 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 $u_2(t) = 10(t^2 \exp(-t/5))$. Figure 6.2 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.

6.2. 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., [16, 31], and its model reduction related problem was recently considered in [8]. The governing equation, subject to initial conditions and boundary control, have a cubic nonlinearity:

$$(6.1) \quad \begin{aligned} \dot{v} + v^3 &= v_{xx} + v, & (0, L) \times (0, T), & & v(0, \cdot) &= u(t), & (0, T), \\ v_x(L, \cdot) &= 0, & (0, T), & & v(x, 0) &= 0, & (0, L). \end{aligned}$$

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 [8], the smooth nonlinear system can be transformed into a QB system by introducing appropriate new state variables. Therefore, the system (6.1) 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 at $x = L$. We use the number of grid points $k = 500$, which results in a QB system of dimension $n = 2 \cdot 500 = 1000$ and set the length $L = 1$. In Figure 6.3, we show the decay of the normalized singular values based on the truncated Gramians of the system.

We determine reduced systems of order $\hat{n} = 20$ 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 Figure 6.4.

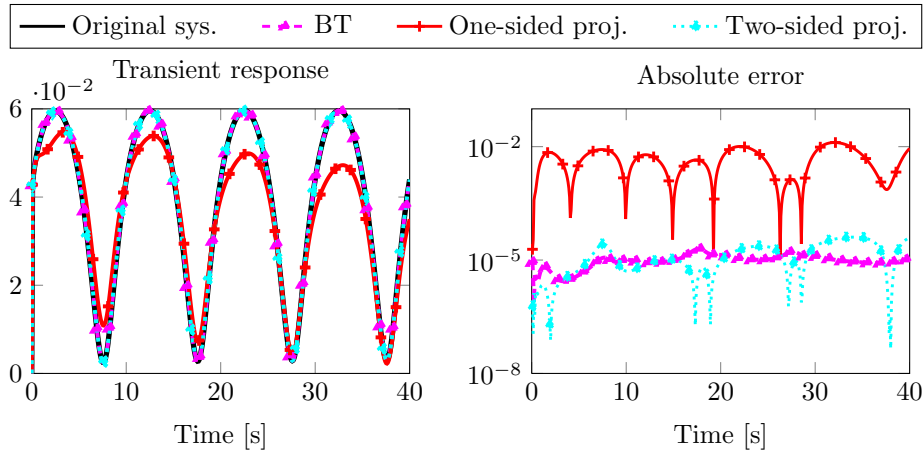
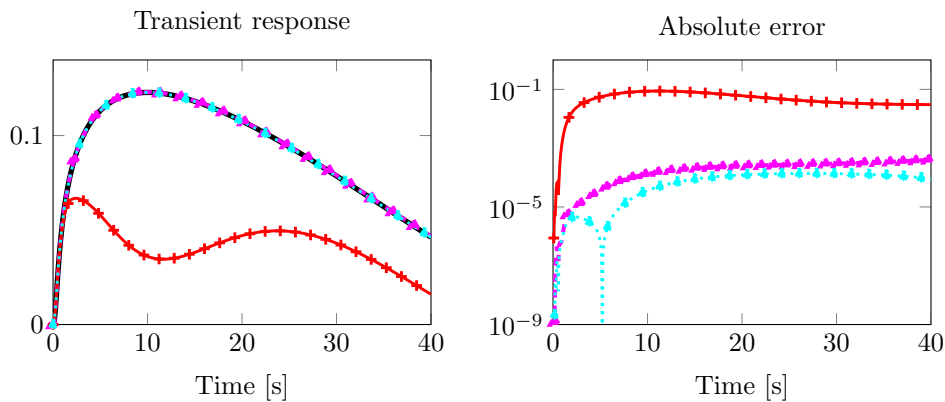
(a) Comparison of the original and the reduced-order systems for $u_1(t) = 5(\sin(2\pi/10) + 1)$.(b) Comparison of the original and the reduced-order systems for $u_2(t) = 10(t^2 \exp(-t/5))$.

Figure 6.2: A RC ladder: comparison of reduced-order systems obtained by balanced truncation (BT) and moment-matching methods for two arbitrary control inputs.

Figure 6.4a shows that the reduced 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 system obtained via the one-sided interpolatory projection method completely fails to capture the dynamics of the system, while balanced truncation and two-sided interpolatory projection can reproduce the system dynamics with a slight advantage of two-sided projection regarding accuracy.

However, it is worthwhile to mention that as we increase the order of the reduced system, the two-sided interpolatory projection method tends to produce unstable reduced-order systems. On the other hand, the accuracies of the reduced-order systems obtained by balanced truncation and one-sided moment-matching increase with the order of the reduced systems.

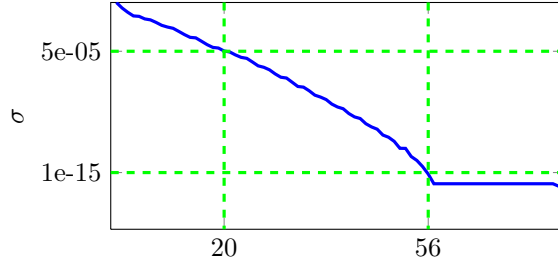


Figure 6.3: Chafee-Infante equation: decay of the normalized singular values based on the truncated Gramians, and dotted line shows the normalized singular value for $\hat{n} = 20$ and the order of the reduced-order system corresponding to the normalized singular value $1e-15$.

6.3. The FitzHugh-Nagumo (F-N) system. Lastly, we consider the F-N system, a simplified neuron model of the Hodgkin-Huxley model, which describes activation and deactivation dynamics of a spiking neuron. This model has been considered in the framework of POD-based [17] and moment-matching model reduction techniques [7]. The dynamics of the system is governed by the following nonlinear coupled differential equations:

$$(6.2) \quad \begin{aligned} \epsilon v_t(x, t) &= \epsilon^2 v_{xx}(x, t) + f(v(x, t)) - w(x, t) + q, \\ w_t(x, t) &= hv(x, t) - \gamma w(x, t) + q \end{aligned}$$

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

$$(6.3) \quad \begin{aligned} v(x, 0) &= 0, & w(x, 0) &= 0, & x &\in [0, L], \\ v_x(0, t) &= i_0(t), & v_x(1, t) &= 0, & t &\geq 0, \end{aligned}$$

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 [7], which are $v(0, t)$ and $w(0, t)$. These outputs describe nothing but the limit cycle 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 6.5 shows the decay of the singular values based on the truncated Gramians for the QB system.

Furthermore, we determine reduced-order systems of order $\hat{n} = 20$ 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 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, the balanced truncation replicates the dynamics of the system faithfully as can be seen

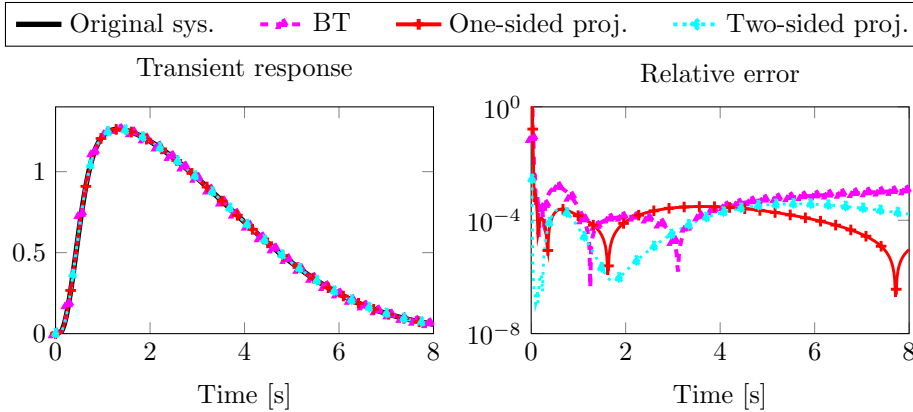
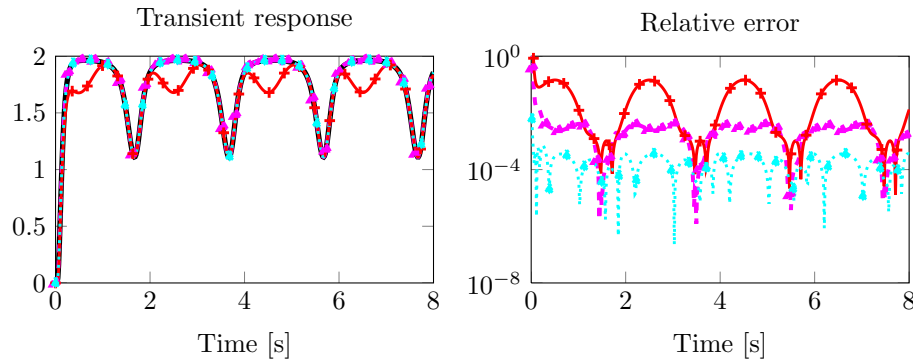
(a) Comparison of the original and the reduced-order systems for $u_1(t) = 5 t \exp(-t)$.(b) Comparison of the original and the reduced-order systems for $u_2(t) = 30 (\sin(\pi t) + 1)$.

Figure 6.4: 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 $u_2(t) = 30 (\sin(\pi t) + 1)$.

in [Figure 6.6a](#). Note that the reduced-order model reported in [7] 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 produces 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.

7. Conclusions. In this paper, we have investigated balanced truncation model reduction for QB control systems. We have proposed reachability and observability Gramians for QB systems based on the kernels of their underlying Volterra series. Additionally, we have also introduced a truncated version of the Gramians. We, furthermore, have compared the controllability and observability energy functionals of the QB system with the quadratic forms of the proposed Gramians for the system and also investigated the connection between the Gramians and reachability/observability of the QB system. Also, we have discussed the advantages of the truncated version of Gramians in the MOR framework and studied local Lyapunov stability of the

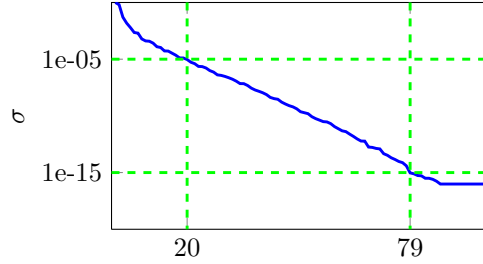
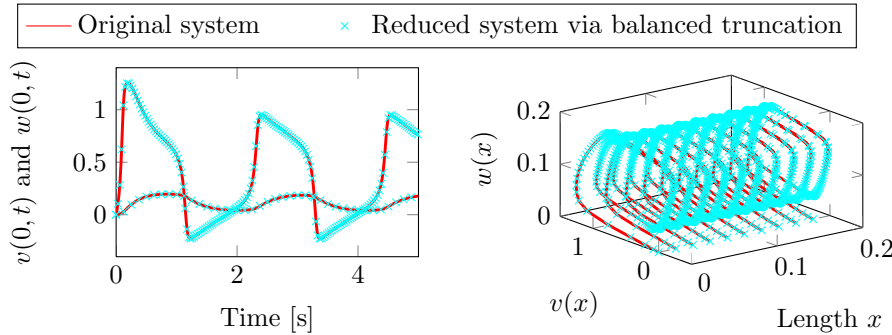


Figure 6.5: Decay of the normalized singular values based on the truncated Gramians of the system for the F-N example, and the dotted lines show the normalized singular value for $\hat{n} = 20$ and the order of the reduced system corresponding to the normalized singular value $1e-15$.



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

(b) Limit-cycles.

Figure 6.6: 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 were unable to produce these limit cycles.

reduced-order systems, obtained via the square-root balanced truncation. By means of various semi-discretized nonlinear PDEs, we have demonstrated the efficiency of the proposed balanced truncation methods for QB systems and compared it with the existing moment-matching techniques.

REFERENCES

- [1] S. A. AL-BAIYAT, M. BETTAYEB, AND U. M. AL-SAGGAF, *New model reduction scheme for bilinear systems*, Int. J. Syst. Sci., 25 (1994), pp. 1631–1642.
- [2] A. C. ANTOUNAS, *Approximation of Large-Scale Dynamical Systems*, SIAM Publications, Philadelphia, PA, 2005.
- [3] P. ASTRID, S. WEILAND, K. WILLCOX, AND T. BACKX, *Missing point estimation in models described by proper orthogonal decomposition*, IEEE Trans. Autom. Control, 53 (2008), pp. 2237–2251.
- [4] Z. BAI, *Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems*,

- Appl. Numer. Math., 43 (2002), pp. 9–44.
- [5] M. BARRAULT, Y. MADAY, N. C. NGUYEN, AND A. T. PATERA, *An ‘empirical interpolation’ method: application to efficient reduced-basis discretization of partial differential equations*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 667–672.
 - [6] U. BAUR, P. BENNER, AND L. FENG, *Model order reduction for linear and nonlinear systems: A system-theoretic perspective*, Arch. Comput. Methods Eng., 21 (2014), pp. 331–358.
 - [7] P. BENNER AND T. BREITEN, *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/>.
 - [8] ———, *Two-sided projection methods for nonlinear model reduction*, SIAM J. Sci. Comput., 37 (2015), pp. B239–B260.
 - [9] P. BENNER AND T. DAMM, *Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems*, SIAM J. Cont. Optim., 49 (2011), pp. 686–711.
 - [10] P. BENNER, P. GOYAL, AND M. REDMANN, *Truncated Gramians for bilinear systems and their advantages in model order reduction*, in P. Benner, M. Ohlberger, T. Patera, G. Rozza, K. Urban (Eds.), *Model Reduction of Parametrized Systems, MS&A - Modeling, Simulation and Applications*, Springer International Publishing, Cham., 2016. Accepted.
 - [11] P. BENNER, P. KÜRSCHNER, AND J. SAAK, *Self-generating and efficient shift parameters in ADI methods for large Lyapunov and Sylvester equations*, Electron. Trans. Numer. Anal., 43 (2014), pp. 142–162.
 - [12] P. BENNER, V. MEHRMANN, AND D. C. SORENSEN, *Dimension Reduction of Large-Scale Systems*, vol. 45 of Lect. Notes Comput. Sci. Eng., Springer-Verlag, Berlin/Heidelberg, Germany, 2005.
 - [13] P. BENNER AND J. SAAK, *Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey*, GAMM-Mitteilungen, 36 (2013), pp. 32–52.
 - [14] B. N. BOND, Z. MAHMOOD, Y. LI, R. SREDOJEVIC, A. MEGRETSKI, V. STOJANOVI, Y. AVNIEL, AND L. DANIEL, *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 (2010), pp. 1149–1162.
 - [15] T. BREITEN AND T. DAMM, *Krylov subspace methods for model order reduction of bilinear control systems*, Systems Control Lett., 59 (2010), pp. 443–450.
 - [16] N. CHAFEE AND E. F. INFANTE, *A bifurcation problem for a nonlinear partial differential equation of parabolic type*, Appl. Anal., 4 (1974), pp. 17–37.
 - [17] S. CHATURANTABUT AND D. C. SORENSEN, *Nonlinear model reduction via discrete empirical interpolation*, SIAM J. Sci. Comput., 32 (2010), pp. 2737–2764.
 - [18] M. CONDON AND R. IVANOV, *Nonlinear systems-algebraic gramians and model reduction*, COMPEL, 24 (2005), pp. 202–219.
 - [19] T. DAMM, *Direct methods and ADI-preconditioned Krylov subspace methods for generalized Lyapunov equations*, Numer. Lin. Alg. Appl., 15 (2008), pp. 853–871.
 - [20] T. DAMM AND D. HINRICHSEN, *Newton’s method for a rational matrix equation occurring in stochastic control*, Linear Algebra Appl., 332 (2001), pp. 81–109.
 - [21] Z. DRMAČ AND S. GUGERCIN, *A new selection operator for the discrete empirical interpolation method—improved a priori error bound and extensions*, SIAM J. Sci. Comput., 38 (2016), pp. A631–A648.
 - [22] L. FENG, X. ZENG, C. CHIANG, D. ZHOU, AND Q. FANG, *Direct nonlinear order reduction with variational analysis*, in Proc. of the Design, Automation and Test in Europe Conference and Exhibition, vol. 2, 2004, pp. 1316–1321.
 - [23] K. FUJIMOTO AND J. M. A. SCHERPEN, *Balanced realization and model order reduction for nonlinear systems based on singular value analysis*, SIAM J. Cont. Optim., 48 (2010), pp. 4591–4623.
 - [24] K. FUJIMOTO, J. M. A. SCHERPEN, AND W. S. GRAY, *Hamiltonian realizations of nonlinear adjoint operators*, Automatica, 38 (2002), pp. 1769–1775.
 - [25] W. S. GRAY AND J. MESKO, *Energy functions and algebraic Gramians for bilinear systems*, in Preprints of the 4th IFAC Nonlinear Control Systems Design Symposium, Enschede, The Netherlands, 1998, pp. 103–108.
 - [26] W. S. GRAY AND J. P. MESKO, *Controllability and observability functions for model reduction of nonlinear systems*, in Proc. of the Conf. on Information Sci. and Sys., Princeton, NJ, 1996, pp. 1244–1249.
 - [27] W. S. GRAY AND J. M. A. SCHERPEN, *On the nonuniqueness of singular value functions and balanced nonlinear realizations*, Systems Control Lett., 44 (2001), pp. 219–232.
 - [28] M. A. GREPL, Y. MADAY, N. C. NGUYEN, AND A. T. PATERA, *Efficient reduced-basis treatment*

of nonaffine and nonlinear partial differential equations, ESAIM: Math. Model. Numer. Anal., 41 (2007), pp. 575–605.

[29] C. GU, *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 (2011), pp. 1307–1320.

[30] S. GUGERCIN, A. C. ANTOULAS, AND C. A. BEATTIE, *\mathcal{H}_2 model reduction for large-scale dynamical systems*, SIAM J. Matrix Anal. Appl., 30 (2008), pp. 609–638.

[31] E. HANSEN, F. KRAMER, AND A. OSTERMANN, *A second-order positivity preserving scheme for semilinear parabolic problems*, Appl. Numer. Math., 62 (2012), pp. 1428–1435.

[32] M. HINZE AND S. VOLKWEIN, *Proper orthogonal decomposition surrogate models for nonlinear dynamical systems: Error estimates and suboptimal control*, in Dimension Reduction of Large-Scale Systems, P. Benner, V. Mehrmann, and D. Sorensen, eds., vol. 45 of Lect. Notes Comput. Sci. Eng., Springer-Verlag, Berlin/Heidelberg, Germany, 2005, pp. 261–306.

[33] T. G. KOLDA AND B. W. BADER, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

[34] P. LI AND L. T. PILEGGI, *Compact reduced-order modeling of weakly nonlinear analog and RF circuits*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 24 (2005), pp. 184–203.

[35] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability, and model reduction*, IEEE Trans. Autom. Control, AC-26 (1981), pp. 17–32.

[36] J. R. PHILLIPS, *Projection-based approaches for model reduction of weakly nonlinear, time-varying systems*, IEEE Trans. Comput.-Aided Design Integr. Circuits Syst., 22 (2003), pp. 171–187.

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

[38] S. S. SASTRY, *Nonlinear Systems: Analysis, Stability, and Control*, Springer, 1999.

[39] J. M. A. SCHERPEN, *Balancing for nonlinear systems*, Systems Control Lett., 21 (1993), pp. 143–153.

[40] J. M. A. SCHERPEN, *\mathcal{H}_∞ balancing for nonlinear systems*, Internat. J. Robust Nonlinear Control, 6 (1996), pp. 645–668.

[41] J. M. A. SCHERPEN AND A. J. VAN DER SCHAFT, *Normalized coprime factorizations and balancing for unstable nonlinear systems*, Internat. J. Control, 60 (1994), pp. 1193–1222.

[42] W. H. A. SCHILDERS, H. A. VAN DER VORST, AND J. ROMMES, *Model Order Reduction: Theory, Research Aspects and Applications*, Springer-Verlag, Berlin, Heidelberg, 2008.

[43] H. SCHNEIDER, *Positive operators and an inertia theorem*, Numerische Mathematik, 7 (1965), pp. 11–17.

[44] S. D. SHANK, V. SIMONCINI, AND D. B. SZYLD, *Efficient low-rank solution of generalized Lyapunov equations*, Numer. Math., 134 (2016), pp. 327–342.

[45] S. SHOKOHI, L. M. SILVERMAN, AND P. VAN DOOREN, *Linear time-variable systems: Balancing and model reduction*, IEEE Trans. Autom. Control, 28 (1983), pp. 810–822.

[46] V. SIMONCINI, *Computational methods for linear matrix equations*, SIAM Rev., 58 (2016), pp. 377–441.

[47] E. VERRIEST AND T. KAILATH, *On generalized balanced realizations*, IEEE Trans. Autom. Control, 28 (1983), pp. 833–844.

Appendix A. A Convergence Result.

LEMMA A.1. Consider a recurrence formula as follows:

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

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

$$(A.2a) \quad b < 1, \quad \text{and}$$

$$(A.2b) \quad 1 > (b - 1)^2 - 4ac > 0.$$

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

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

$$(A.3) \quad x^* = \frac{-(b - 1) - \sqrt{(b - 1)^2 - 4ac}}{2a}.$$

Proof. First, note that the sequence (A.1) 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 (A.2a).

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 Banach fix-point theorem, $F(x)$ is a contraction on $[c, x^{(1)}]$, and the fixed point is given by $x^{(1)}$. \square