SIAM J. MATRIX ANAL. APPL.
Vol. 39, No. 2, pp. 983–1032

$\mathcal{H}_2$-QUASI-OPTIMAL MODEL ORDER REDUCTION FOR QUADRATIC-BILINEAR CONTROL SYSTEMS*

PETER BENNER†, PAWAN GOYAL‡, AND SERKAN GUGERCIN§

Abstract. We investigate the optimal model reduction problem for large-scale quadratic-bilinear (QB) control systems. Our contributions are threefold. First, we discuss the variational analysis and the Volterra series formulation for QB systems. We then define the $\mathcal{H}_2$-norm for a QB system based on the *kernels* of the underlying Volterra series and propose a truncated $\mathcal{H}_2$-norm as well. Next, we derive first-order necessary conditions for an optimal approximation, where optimality is measured in terms of the truncated $\mathcal{H}_2$-norm of the error system. We then propose an iterative model reduction algorithm, which upon convergence yields a reduced-order system that *approximately* satisfies the newly derived optimality conditions. We also discuss an efficient computation of the reduced Hessian, using the special Kronecker structure of the Hessian of the system. We illustrate the efficiency of the proposed method by means of several numerical examples resulting from semidiscretized nonlinear partial differential equations and show its competitiveness with existing model reduction schemes such as moment-matching and balanced truncation for QB systems by comparing accuracy in the time-domain simulations and in the truncated $\mathcal{H}_2$-norm.

Key words. model order reduction, quadratic-bilinear control systems, tensor calculus, $\mathcal{H}_2$-norm, Sylvester equations, nonlinear partial differential equations

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

DOI. 10.1137/16M1098280

1. Introduction.

Numerical simulation is a fundamental tool in the analysis of dynamical systems and is required repeatedly, for example, in control, design, optimization, and uncertainty quantification. Dynamical systems are generally governed by partial differential equations (PDEs), ordinary differential equations (ODEs), or a combination of both. A high-fidelity approximation of the underlying physical phenomena requires a finely discretized mesh over the spatial domain of interest, leading to complex dynamical systems with a high-dimensional state space. The simulation of such large-scale systems, however, imposes a huge computational burden. This inspires *model order reduction* (MOR), which aims at constructing simple and reliable surrogate models such that their input-output behavior approximates that of the original large-scale system accurately. These surrogate models can then be used in engineering studies, which make numerical simulations faster and efficient.

In recent decades, numerous theoretical and computational aspects of MOR for linear systems have been investigated; see, e.g., [2, 6, 15, 43]. MOR methods have

*Received by the editors October 11, 2016; accepted for publication (in revised form) by D. Szyld February 6, 2018; published electronically June 5, 2018.
http://www.siam.org/journals/simax/39-2/M109828.html
Funding: This work was supported by a research grant of the "International Max Planck Research School (IMPRS) for Advanced Methods in Process and System Engineering (Magdeburg)." The first author's work was supported by the DFG Research Training Group RTG 2297/1 "Mathematical Complexity Reduction." The third author's work was supported in part by NSF through grant DMS-1522616 and by the Alexander von Humboldt Foundation.

†Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, 39106 Magdeburg, Germany, and Fakultät für Mathematik, Otto-von-Guericke-Universität 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).
§Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123 (gugercin@math.vt.edu).

been successfully applied in various fields, e.g., optimal control, PDE-constrained optimization, and uncertainty quantification; see, for example, [17, 29]. In recent years, however, MOR of nonlinear systems has gained significant interest with the goal of extending the input-independent, optimal MOR techniques from linear systems to nonlinear ones. For example, MOR techniques for linear systems such as balanced truncation (BT) [2, 36], or the iterative rational Krylov algorithm (IRKA) [24], have been extended to a special class of nonlinear systems, the so-called *bilinear systems*, in which nonlinear terms arise from the product of the state and input [8, 11, 22, 48]. In this article, we address another vital class of nonlinear systems, called *quadratic-bilinear* (QB) systems. These are of the form

$$\Sigma : \begin{cases} \dot{x}(t) = Ax(t) + H\left(x(t) \otimes x(t)\right) + \sum_{k=1}^{m} N_k x(t) u_k(t) + Bu(t), \\ y(t) = Cx(t), \quad x(0) = 0, \end{cases} \tag{1.1}$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and $y(t) \in \mathbb{R}^p$ are the states, inputs, and outputs of the system at time t, respectively; u_k is the kth component of u; and n is the state dimension. Furthermore, $A, N_k \in \mathbb{R}^{n\times n}$ for $k \in \{1, \ldots, m\}$, $H \in \mathbb{R}^{n\times n^2}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$.

There is a variety of applications where the system inherently contains a quadratic nonlinearity, which can be modeled in the QB form (1.1) e.g., spatial discretizations of the Burgers' equation, the Allen–Cahn or Chafee–Infante equation, and many other models from engineering and physics. Moreover, a large class of smooth nonlinear systems, involving combinations of elementary functions like exponential, trigonometric, polynomial functions, etc., can be equivalently rewritten as QB systems (1.1) as shown in [10, 23]. This is achieved by introducing some new appropriate state variables to simplify the nonlinearities present in the underlying control system and by deriving differential equations corresponding to the newly introduced variables, or by using appropriate algebraic constraints. When algebraic constraints are introduced in terms of the state and the newly defined variables, the system contains algebraic equations along with differential equations. Such systems are called *differential-algebraic equations* (DAEs) or *descriptor systems* [34]. MOR procedures for DAEs become inevitably more complicated, even in the linear and bilinear settings; e.g., see [12, 25]. In this article, we restrict ourselves to QB ODE systems and leave MOR for QB descriptor systems as a future research topic.

For a given QB system (1.1) Σ of order n, our aim is to construct a reduced-order system

$$\widehat{\Sigma} : \begin{cases} \dot{\widehat{x}}(t) = \widehat{A}\widehat{x}(t) + \widehat{H}\left(\widehat{x}(t) \otimes \widehat{x}(t)\right) + \sum_{k=1}^{m} \widehat{N}_k \widehat{x}(t) u_k(t) + \widehat{B}u(t), \\ \widehat{y}(t) = \widehat{C}\widehat{x}(t), \quad \widehat{x}(0) = 0, \end{cases} \tag{1.2}$$

where $\widehat{A}, \widehat{N}_k \in \mathbb{R}^{r\times r}$ for $k \in \{1, \ldots, m\}$, $\widehat{H} \in \mathbb{R}^{r\times r^2}$, $\widehat{B} \in \mathbb{R}^{r\times m}$, and $\widehat{C} \in \mathbb{R}^{p\times r}$ with $r \ll n$ such that the outputs of the system (1.1) and (1.2), y and $\widehat{y}$, are close in a proper norm for all admissible inputs $u_k \in L^2[0, \infty[$.

Similar to the linear and bilinear cases, we construct the reduced-order system (1.2) via projection. Toward this goal, we construct two model reduction basis matrices $V, W \in \mathbb{R}^{n\times r}$ such that $W^T V$ is invertible. Then, the reduced matrices in (1.2) are given by

$$\widehat{A} = (W^TV)^{-1}W^TAV, \qquad \widehat{N}_k = (W^TV)^{-1}W^TN_kV \quad \text{for} \quad k \in \{1, \dots, m\},$$
$$\widehat{H} = (W^TV)^{-1}W^TH(V \otimes V), \qquad \widehat{B} = (W^TV)^{-1}W^TB, \quad \text{and} \quad \widehat{C} = CV.$$

It can be easily seen that the quality of the reduced-order system depends on the choice of the reduction subspaces spanned by the columns of V and W, respectively. There exist various MOR approaches in the literature to determine these subspaces. One of the earlier and popular methods for nonlinear systems is proper orthogonal decomposition (POD); see, e.g., [3, 19, 28, 33]. POD relies on the Galerkin projection $\mathcal{P} = \mathcal{V}\mathcal{V}^T$, where $\mathcal{V}$ is determined based on extracting the dominant modes of the system dynamics from a selection of snapshots of the solution trajectories computed using some training input. A Petrov–Galerkin-type projection can be obtained using the dual/adjoining system in either time or frequency domain [40, 46]. Another widely used method for nonlinear systems is the trajectory piecewise linear method; e.g., see [39]. For this method, the nonlinear system is replaced by a weighted sum of linear systems; these linear systems can then be reduced by using well-known methods for linear systems such as BT, or interpolation methods; e.g., see [2]. However, the abovementioned methods require some snapshots or solution trajectories of the original systems for particular inputs. This indicates that the resulting reduced-order system depends on the choice of inputs, which may make the reduced-order system inadequate in many applications such as control and optimization, where the variation of the input is inherent to the problem.

MOR methods based on interpolation or moment-matching have been extended from linear systems to QB systems, with the aim of capturing the input-output behavior of the underlying system independent of a training input. One-sided interpolatory projection for QB systems is studied in [4, 23, 37, 38], and has been recently extended to a two-sided interpolatory projection in [9, 10]. These methods result in reduced-order systems that do not rely on the training data for a control input; see also the survey [6] for some related approaches. Thus, the resulting reduced-order systems can be used in input-varying applications. In the aforementioned references, the authors have shown how to construct an interpolating reduced-order system for a given set of interpolation points. But it is still an open problem how to choose these interpolation points optimally with respect to an appropriate norm. Furthermore, the two-sided interpolatory projection method [10] is only applicable to single-input single-output (SISO) systems, which is very restrictive, and additionally, the stability of the resulting reduced-order systems also remains another major issue. We note here that the method proposed in this paper does not resolve this issue. It remains an open problem, even in the case of linear systems, to give general conditions for stability preservation of two-sided projection methods.

Very recently, BT has been extended from linear/bilinear systems to QB systems [13]. This method first determines the states which are hard to control and observe, and constructs the reduced model by truncating those states. Importantly, balance truncation yields locally Lyapunov stable reduced-order systems, and an appropriate order of the reduced-order system can be determined based on the singular values of the Gramians of the system. But unlike in the linear case, the resulting reduced-order systems do not retain other desirable properties such as an a priori error bound. Moreover, in order to apply BT to QB systems, we require the solutions of four conventional Lyapunov equations, which could be computationally cumbersome in large-scale settings, although there have been many advancements in recent times related to computing the low-rank solutions of Lyapunov equations [16, 44].

Another popular input-independent MOR approach for linear and bilinear systems is based on computing models that satisfy optimality conditions for the best approximation in the $\mathcal{H}_2$ system norm. Therefore, in this paper, we study the $\mathcal{H}_2$-optimal approximation problem for QB systems. Precisely, we show how to choose the model reduction bases in a two-sided projection framework for QB systems so that the reduced-order system approximately minimizes the cost encoding the approximation error in the $\mathcal{H}_2$-norm. Our main contributions are threefold. In section 2, we derive various expressions and formulas related to Kronecker products, which are later heavily utilized in deriving the optimality conditions. In section 3, we first define the $\mathcal{H}_2$-norm of the QB system (1.1) based on the *kernels* of its Volterra series (input/output mapping), and also derive an expression for a truncated $\mathcal{H}_2$-norm for QB systems. Subsequently, based on the truncated $\mathcal{H}_2$-norm of the error system, we derive first-order necessary conditions for optimal model reduction of QB systems. We then propose an iterative algorithm to construct reduced-order systems that *approximately* satisfy the newly derived optimality conditions. Furthermore, we discuss an efficient alternative way to compute reduced Hessians as compared with the one proposed in [10]. In section 4, we illustrate the efficiency of the proposed method for various semidiscretized nonlinear PDEs and compare it with existing methods such as BT [13] as well as the one-sided and two-sided interpolatory projection methods for QB systems [10, 23]. We conclude the paper with a short summary and potential future directions in section 5.

Notation: Throughout the paper, we make use of the following notation:

- I_q denotes the identity matrix of size $q \times q$, and its pth column is denoted by e_p^q.
- $\operatorname{vec}(\cdot)$ denotes vectorization of a matrix, and $\mathcal{I}_m$ denotes $\operatorname{vec}(I_m)$.
- $\otimes$ denotes the Kronecker product of two matrices (including vectors as special cases).
- $\operatorname{tr}(\cdot)$ refers to the trace of a matrix.
- Using MATLAB® notation, we denote the jth column of the matrix A by $A(:, j)$.
- $\mathbf{0}$ is a zero matrix of appropriate size.
- We denote the full-order system (1.1) and reduced-order system (1.2) by Σ and $\widehat{\Sigma}$, respectively.
- $\operatorname{orth}(A)$ returns an orthonormal basis for the range of the matrix A.

2. Tensor matricizations and their properties. We first review some basic concepts from tensor algebra. First, we note the following important properties of the $\operatorname{vec}(\cdot)$ operator:

$$\operatorname{tr}\left(X^T Y\right) = \operatorname{vec}(X)^T \operatorname{vec}(Y) = \operatorname{vec}(Y)^T \operatorname{vec}(X) \text{ and} \tag{2.1a}$$

$$\operatorname{vec}(XYZ) = (Z^T \otimes X) \operatorname{vec}(Y). \tag{2.1b}$$

Next, we review the concepts of matricization of a tensor. Since the Hessian H of the QB system in (1.1) is a third-order tensor, we focus on three-way tensors $\mathcal{X}^{n\times n\times n}$. However, most of the concepts can be extended to general kth-order tensors. Similar to how rows and columns are defined for a matrix, one can define a fiber of $\mathcal{X}$ by fixing all indices but one, e.g., $\mathcal{X}(:, i, j), \mathcal{X}(i, :, j)$, and $\mathcal{X}(i, j, :)$. Mathematical operations involving tensors are easier to perform using their corresponding matrix representations. For this purpose, there exists a very well-known process of unfolding a tensor into a matrix, called *matricization* of a tensor. For a third-order tensor, there are three different ways to unfold it, depending on the mode-μ fibers that are used for

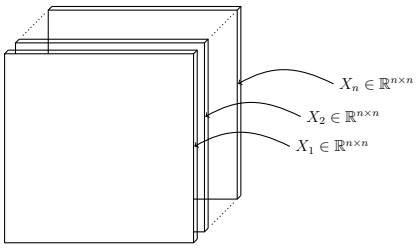

FIG. 2.1. *Representation of a tensor using frontal slices* [32].

the unfolding. If the tensor is unfolded using its mode-μ fibers, it is called the mode-μ matricization of $\mathcal{X}$. We refer to [30, 32] for more details on these basic concepts of tensor theory.

In the following example, we illustrate how a third-order tensor $\mathcal{X} \in \mathbb{R}^{n\times n\times n}$ can be unfolded into different matrices.

Example 2.1. Consider a third-order tensor $\mathcal{X}^{n\times n\times n}$ whose frontal slices are given by matrices $X_i \in \mathbb{R}^{n\times n}$, as shown in Figure 2.1. Then, its mode-μ matricizations, $\mu \in \{1, 2, 3\}$, are given by

$$\mathcal{X}^{(1)} = [X_1, X_2, \ldots, X_n], \quad \mathcal{X}^{(2)} = [X_1^T, X_2^T, \ldots, X_n^T], \text{ and}$$
$$\mathcal{X}^{(3)} = [\mathrm{vec}\,(X_1), \mathrm{vec}\,(X_2), \ldots, \mathrm{vec}\,(X_n)]^T.$$

Similar to the matrix-matrix product, one can also perform a tensor-matrix or tensor-tensor multiplication. Of particular interest in this paper are tensor-matrix multiplications, which can be performed by means of matricizations; see, e.g., [32]. For a given tensor $\mathcal{X} \in \mathbb{R}^{n\times n\times n}$ and a matrix $\mathcal{A} \in \mathbb{R}^{n_1\times n}$, the μ-mode matrix product is denoted by $\mathcal{X} \times_\mu \mathcal{A} =: \mathcal{Y}$, i.e., $\mathcal{Y} \in \mathbb{R}^{n_1\times n\times n}$ for $\mu = 1$. In the case of the μ-mode matrix multiplication, the mode-μ fiber is multiplied with the matrix $\mathcal{A}$, which can be written as

$$\mathcal{Y} = \mathcal{X} \times_\mu \mathcal{A} \Leftrightarrow \mathcal{Y}^{(\mu)} = \mathcal{A}\mathcal{X}^{(\mu)}.$$

Furthermore, if a tensor is given as

$$\mathcal{Z} = \mathcal{X} \times_1 \mathcal{A} \times_2 \mathcal{B} \times_3 \mathcal{C}, \tag{2.2}$$

where $\mathcal{A} \in \mathbb{R}^{n_1\times n}$, $\mathcal{B} \in \mathbb{R}^{n_2\times n}$, and $\mathcal{C} \in \mathbb{R}^{n_3\times n}$, then the mode-$\mu$ matriciziations of $\mathcal{Z}$ satisfy

$$\mathcal{Z}^{(1)} = \mathcal{A}\mathcal{X}^{(1)}(\mathcal{C}\otimes\mathcal{B})^T,\ \mathcal{Z}^{(2)} = \mathcal{B}\mathcal{X}^{(2)}(\mathcal{C}\otimes\mathcal{A})^T,\ \mathcal{Z}^{(3)} = \mathcal{C}\mathcal{X}^{(3)}(\mathcal{B}\otimes\mathcal{A})^T. \tag{2.3}$$

Using these properties of the tensor products, we now introduce our first result on tensor matricizations.

LEMMA 2.2. *Consider tensors $\mathcal{X}, \mathcal{Y} \in \mathbb{R}^{n\times n\times n}$ and let $\mathcal{X}^{(i)}$ and $\mathcal{Y}^{(i)}$ denote, respectively, their mode-i matricizations. Then,*

$$\mathrm{tr}\left(\mathcal{X}^{(1)}(\mathcal{Y}^{(1)})^T\right) = \mathrm{tr}\left(\mathcal{X}^{(2)}(\mathcal{Y}^{(2)})^T\right) = \mathrm{tr}\left(\mathcal{X}^{(3)}(\mathcal{Y}^{(3)})^T\right).$$

Proof. We begin by denoting the ith frontal slice of $\mathcal{X}$ and $\mathcal{Y}$ by X_i and Y_i, respectively; see Figure 2.1. Thus,

$$\begin{aligned}\operatorname{tr}\left(\mathcal{X}^{(1)}(\mathcal{Y}^{(1)})^T\right) &= \operatorname{tr}\left(\left[X_1, X_2, \ldots, X_n\right]\left[Y_1, Y_2, \ldots, Y_n\right]^T\right)\\ &= \sum_{i=1}^{n} \operatorname{tr}\left(X_i Y_i^T\right) = \sum_{i=1}^{n} \operatorname{tr}\left(X_i^T Y_i\right)\\ &= \operatorname{tr}\left(\left[X_1^T, X_2^T, \ldots, X_n^T\right]\left[Y_1^T, Y_2^T, \ldots, Y_n^T\right]^T\right) = \operatorname{tr}\left(\mathcal{X}^{(2)}(\mathcal{Y}^{(2)})^T\right).\end{aligned}$$

Furthermore, since $\operatorname{tr}\left(X^T Y\right) = \operatorname{vec}(X)^T \operatorname{vec}(Y)$, this allows us to write

$$\operatorname{tr}\left(\mathcal{X}^{(1)}(\mathcal{Y}^{(1)})^T\right) = \sum_{i=1}^{n} \operatorname{tr}\left(X_i^T Y_i\right) = \sum_{i=1}^{n} \operatorname{vec}(X_i)^T \operatorname{vec}(Y_i).$$

Since the ith rows of $\mathcal{X}^{(3)}$ and $\mathcal{Y}^{(3)}$ are given by $\operatorname{vec}(X_i)^T$ and $\operatorname{vec}(Y_i)^T$, respectively, it holds that $\sum_{i=1}^{n} \operatorname{vec}(X_i)^T \operatorname{vec}(Y_i) = \operatorname{tr}\left(\mathcal{X}^{(3)}(\mathcal{Y}^{(3)})^T\right)$. This concludes the proof. ∎

Recall that the Hessian H in the QB system (1.1) is of size $n \times n^2$; thus, it can be interpreted as an unfolding of a tensor $\mathcal{H}^{n\times n\times n}$. Without loss of generality, we assume the Hessian H to be the mode-1 matricization of $\mathcal{H}$, i.e., $H = \mathcal{H}^{(1)}$. Also, we assume $\mathcal{H}$ to be symmetric. This means that for given vectors u and v,

$$H(u \otimes v) = \mathcal{H}^{(1)}(u \otimes v) = \mathcal{H}^{(1)}(v \otimes u) = H(v \otimes u). \tag{2.4}$$

This provides the additional information that the other two matricization modes of $\mathcal{H}$ are the same, i.e.,

$$\mathcal{H}^{(2)} = \mathcal{H}^{(3)}. \tag{2.5}$$

In general, it is not necessary that the Hessian H (mode-1 matrization of the tensor $\mathcal{H}$) obtained from the discretization of the governing PDEs satisfies (2.4). However, as shown in [10], the Hessian H can be modified in such a way that the modified Hessian $\widetilde{H}$ satisfies (2.4) without any change in the dynamics of the system; thus, for the rest of the paper, without loss of generality, we assume that the tensor $\mathcal{H}$ is symmetric.

The additional property that the Hessian is symmetric will allow us to derive some new relationships between matricizations and matrices that will prove to be crucial ingredients in simplifying the expressions arising in the derivation of optimality conditions in section 3.

LEMMA 2.3. *Let $\mathcal{H} \in \mathbb{R}^{n\times n\times n}$ be a third-order tensor, satisfying* (2.4) *and* (2.5)*, and consider matrices $\mathcal{A}, \mathcal{B}, \mathcal{C} \in \mathbb{R}^{n\times n}$. Then,*

$$\mathcal{H}^{(1)}(\mathcal{B} \otimes \mathcal{C})\left(\mathcal{H}^{(1)}\right)^T = \mathcal{H}^{(1)}(\mathcal{C} \otimes \mathcal{B})\left(\mathcal{H}^{(1)}\right)^T \tag{2.6}$$

and

$$\begin{aligned}(\operatorname{vec}(\mathcal{B}))^T \operatorname{vec}\left(\mathcal{H}^{(2)}(\mathcal{C} \otimes \mathcal{A})(\mathcal{H}^{(2)})^T\right) &= (\operatorname{vec}(\mathcal{C}))^T \operatorname{vec}\left(\mathcal{H}^{(2)}(\mathcal{B} \otimes \mathcal{A})(\mathcal{H}^{(2)})^T\right)\\ &= (\operatorname{vec}(\mathcal{A}))^T \operatorname{vec}\left(\mathcal{H}^{(1)}(\mathcal{C} \otimes \mathcal{B})(\mathcal{H}^{(1)})^T\right).\end{aligned}$$

Proof. We begin by proving the relation in (2.6). The order in the Kronecker product can be changed via pre- and post-multiplication of appropriate permutation matrices; see [27, sect. 3]. Thus,

$$\mathcal{B} \otimes \mathcal{C} = S(\mathcal{C} \otimes \mathcal{B})S^T,$$

where S is the permutation matrix $S = \sum_{i=1}^{n}((e_i^n)^T \otimes I_n \otimes e_i^n)$. We can then write

$$\mathcal{H}^{(1)}(\mathcal{B} \otimes \mathcal{C}) \left(\mathcal{H}^{(1)}\right)^T = \mathcal{H}^{(1)} S(\mathcal{C} \otimes \mathcal{B}) \left(\mathcal{H}^{(1)} S\right)^T. \tag{2.7}$$

We now manipulate the term $\mathcal{H}^{(1)}S$:

$$\mathcal{H}^{(1)} S = \sum_{i=1}^{n} \mathcal{H}^{(1)}((e_i^n)^T \otimes I_n \otimes e_i^n). \tag{2.8}$$

Furthermore, we can write I_n as the Kronecker product

$$I_n = \sum_{j=1}^{n} (e_j^n)^T \otimes e_j^n, \tag{2.9}$$

and since for a vector $f \in \mathbb{R}^q$, $f^T \otimes f = ff^T$, we can write (2.9) in another form as

$$I_n = \sum_{j=1}^{n} e_j^n (e_j^n)^T. \tag{2.10}$$

Substituting these relations in (2.8) leads to

$$\begin{aligned}
\mathcal{H}^{(1)} S &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{H}^{(1)}((e_i^n)^T \otimes (e_j^n)^T \otimes e_j^n \otimes e_i^n) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{H}^{(1)} \left(e_j^n \otimes e_i^n\right) \left((e_i^n)^T \otimes (e_j^n)^T\right) \qquad \left(\because \text{for } f \in \mathbb{R}^q, f^T \otimes f = ff^T\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathcal{H}^{(1)}(e_i^n \otimes e_j^n)((e_i^n)^T \otimes (e_j^n)^T). \qquad (\because \text{the relation (2.4)}) \qquad (2.11)
\end{aligned}$$

Next, we use a tensor-multiplication property in the above equation, namely,

$$(\mathcal{P} \otimes \mathcal{Q})(\mathcal{R} \otimes \mathcal{S}) = (\mathcal{P}\mathcal{R} \otimes \mathcal{Q}\mathcal{S}), \tag{2.12}$$

where $\mathcal{P}, \mathcal{Q}, \mathcal{R}$, and $\mathcal{S}$ are of compatible dimensions. Using the Kronecker product property (2.12) in (2.11), we obtain

$$\begin{aligned}
\mathcal{H}^{(1)} S &= \mathcal{H}^{(1)} \left(\sum_{i=1}^{n} e_i^n (e_i^n)^T \otimes \sum_{j=1}^{n} e_j^n (e_j^n)^T\right) \\
&= \mathcal{H}^{(1)}(I_n \otimes I_n) = \mathcal{H}^{(1)}. \qquad \text{(from (2.10))}
\end{aligned}$$

Substituting the above relation in (2.7) proves (2.6). For the second part, we utilize the trace property (2.1a) to obtain

$$(\operatorname{vec}(\mathcal{B}))^T \operatorname{vec}\left(\mathcal{H}^{(2)}(\mathcal{C} \otimes \mathcal{A})(\mathcal{H}^{(2)})^T\right) = \operatorname{tr}\left(\underbrace{\mathcal{B}^T \mathcal{H}^{(2)}(\mathcal{C} \otimes \mathcal{A})}_{=\mathcal{L}^{(2)}}(\mathcal{H}^{(2)})^T\right),$$

where $\mathcal{L}^{(2)} \in \mathbb{R}^{n\times n^2}$ can be considered as a mode-2 matricization of a tensor $\mathcal{L}^{n\times n\times n}$. Using Lemma 2.2 and the relations (2.3), we obtain

$$\begin{aligned}\operatorname{tr}\left(\mathcal{L}^{(2)}\left(\mathcal{H}^{(2)}\right)^T\right) &= \operatorname{tr}\left(\mathcal{L}^{(3)}\left(\mathcal{H}^{(3)}\right)^T\right) = \operatorname{tr}\left(\mathcal{C}^T\mathcal{H}^{(3)}(\mathcal{B}\otimes\mathcal{A})\left(\mathcal{H}^{(3)}\right)^T\right)\\ &= \operatorname{tr}\left(\mathcal{C}^T\mathcal{H}^{(2)}(\mathcal{B}\otimes\mathcal{A})\left(\mathcal{H}^{(2)}\right)^T\right) \qquad \text{(using (2.5))}\\ &= (\operatorname{vec}(\mathcal{C}))^T \operatorname{vec}\left(\mathcal{H}^{(2)}(\mathcal{B}\otimes\mathcal{A})\left(\mathcal{H}^{(2)}\right)^T\right).\end{aligned}$$

Furthermore, we also have

$$\begin{aligned}\operatorname{tr}\left(\mathcal{L}^{(2)}\left(\mathcal{H}^{(2)}\right)^T\right) &= \operatorname{tr}\left(\mathcal{L}^{(1)}\left(\mathcal{H}^{(1)}\right)^T\right) = \operatorname{tr}\left(\mathcal{A}^T\mathcal{H}^{(1)}(\mathcal{C}\otimes\mathcal{B})\left(\mathcal{H}^{(1)}\right)^T\right)\\ &= (\operatorname{vec}(\mathcal{A}))^T \operatorname{vec}\left(\mathcal{H}^{(1)}(\mathcal{C}\otimes\mathcal{B})\left(\mathcal{H}^{(1)}\right)^T\right),\end{aligned}$$

which completes the proof. ∎

Next, we prove the connection of a certain permutation matrix to the Kronecker product.

LEMMA 2.4. *Consider two matrices $X, Y \in \mathbb{R}^{n\times m}$. Define the permutation matrix $T_{(n,m)} \in \{0,1\}^{n^2m^2\times n^2m^2}$ as*

$$T_{(n,m)} = I_m \otimes \left[I_m \otimes e_1^n, \ldots, I_m \otimes e_n^n\right] \otimes I_n. \tag{2.13}$$

Then,

$$\operatorname{vec}(X\otimes Y) = T_{(n,m)}\left(\operatorname{vec}(X)\otimes\operatorname{vec}(Y)\right).$$

Proof. Let us denote the ith columns of X and Y by x_i and y_i, respectively. We can then write

$$\operatorname{vec}(X\otimes Y) = \begin{bmatrix}\operatorname{vec}(x_1\otimes Y)\\ \vdots \\ \operatorname{vec}(x_m\otimes Y)\end{bmatrix}. \tag{2.14}$$

Now we concentrate on the ith block row of $\operatorname{vec}(X\otimes Y)$, which, using (2.1b) and (2.12), can be written as

$$\begin{aligned}\operatorname{vec}(x_i\otimes Y) &= \operatorname{vec}((x_i\otimes I_n)Y) = (I_m\otimes x_i\otimes I_n)\operatorname{vec}(Y)\\ &= \left(I_m\otimes\left[x_i^{(1)}e_1^n + \cdots + x_i^{(n)}e_n^n\right]\otimes I_n\right)\operatorname{vec}(Y),\end{aligned} \tag{2.15}$$

where $x_i^{(j)}$ is the (j,i)th entry of the matrix X. An alternative way to write (2.15) is

$$\begin{aligned}\operatorname{vec}(x_i\otimes Y) &= \left[I_m\otimes e_1^n\otimes I_n, \ldots, I_m\otimes e_n^n\otimes I_n\right](x_i\otimes I_{nm})\operatorname{vec}(Y)\\ &= \left(\left[I_m\otimes e_1^n, \ldots, I_m\otimes e_n^n\right]\otimes I_n\right)(x_i\otimes\operatorname{vec}(Y)).\end{aligned}$$

This yields

$$\begin{aligned}\operatorname{vec}(X \otimes Y) &= \begin{bmatrix} ([I_m \otimes e_1^n, \ldots, I_m \otimes e_n^n] \otimes I_n)\,(x_1 \otimes \operatorname{vec}(Y)) \\ \vdots \\ ([I_m \otimes e_1^n, \ldots, I_m \otimes e_n^n] \otimes I_n)\,(x_m \otimes \operatorname{vec}(Y)) \end{bmatrix} \\ &= (I_m \otimes [I_m \otimes e_1^n, \ldots, I_m \otimes e_n^n] \otimes I_n) \begin{bmatrix} x_1 \otimes \operatorname{vec}(Y) \\ \vdots \\ x_m \otimes \operatorname{vec}(Y) \end{bmatrix} \\ &= (I_m \otimes [I_m \otimes e_1^n, \ldots, I_m \otimes e_n^n] \otimes I_n) \left(\begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \otimes \operatorname{vec}(Y) \right) \\ &= (I_m \otimes [I_m \otimes e_1^n, \ldots, I_m \otimes e_n^n] \otimes I_n)\,(\operatorname{vec}(X) \otimes \operatorname{vec}(Y)),\end{aligned}$$

which proves the assertion. ∎

Lemma 2.4 will be utilized in simplifying the error expressions in the next section.

3. $\mathcal{H}_2$-Norm for QB systems and optimality conditions. In this section, we first define the $\mathcal{H}_2$-norm for the QB systems (1.1) and its truncated version. Then, based on the truncated $\mathcal{H}_2$ measure, we derive first-order necessary conditions for optimal model reduction. These optimality conditions will naturally lead to a numerical algorithm to construct quasi-optimal reduced models for QB systems that are independent of training data. The proposed model reduction framework extends the optimal $\mathcal{H}_2$ methodology from linear [24] and bilinear systems [8, 22] to QB nonlinear systems.

3.1. $\mathcal{H}_2$-norm of QB systems. In order to define the $\mathcal{H}_2$-norm for QB systems and its truncated version, we first require the input/output representation for QB systems. In other words, we aim at obtaining the solution of QB systems with the help of *Volterra* series as has been done for bilinear systems, e.g., as in [41, sect. 3.1]. For this, one can utilize *variational analysis* [41, sect. 3.4]. Since the QB system falls under the class of linear-analytic systems, for a scalar α, we can write the solution $x(t)$ of (1.1) for an input $\alpha u(t)$ as

$$x(t) = \sum_{s=1}^{\infty} \alpha^s x_s(t),$$

where $x_s(t) \in \mathbb{R}^n$. Thus, we obtain

$$\begin{aligned}\sum_{s=1}^{\infty} \alpha^s \dot{x}_s(t) &= A\left(\sum_{s=1}^{\infty} \alpha^s x_s(t)\right) + H\left(\left(\sum_{s=1}^{\infty} \alpha^s x_s(t)\right) \otimes \left(\sum_{s=1}^{\infty} \alpha^s x_s(t)\right)\right) \\ &\quad + \sum_{k=1}^{m} \alpha N_k \sum_{s=1}^{\infty} (\alpha^s x_s(t))\, u_k(t) + \alpha B u(t). \qquad (3.1)\end{aligned}$$

Since the expression (3.1) holds for arbitrary α, the coefficients of α^i, $i \in \{1, 2, \ldots\}$, can be equated on both sides of (3.1), leading to

$$\begin{aligned}
\dot{x}_1(t) &= Ax_1(t) + Bu(t),\\
\dot{x}_2(t) &= Ax_2(t) + H\left(x_1(t) \otimes x_1(t)\right) + \sum_{k=1}^{m} N_k x_1 u_k(t), \text{ and}\\
\dot{x}_s(t) &= Ax_s(t) + \sum_{\substack{i,j\geq 1\\ i+j=s}} H\left(x_i(t) \otimes x_j(t)\right) + \sum_{k=1}^{m} N_k x_{s-1}(t) u_k(t), \quad s \geq 3.
\end{aligned} \tag{3.2}$$

Then, let $\alpha = 1$ so that $x(t) = \sum_{s=1}^{\infty} x_s(t)$, where $x_s(t)$ solves the coupled linear differential equation (3.2). The equation for $x_1(t)$ corresponds to a linear system, thus allowing us to write the expression for $x_1(t)$ as a convolution:

$$x_1(t) = \int_0^t e^{At_1} B u(t - t_1) dt_1. \tag{3.3}$$

Using the expression for $x_1(t)$, we can obtain an explicit expression for $x_2(t)$:

$$\begin{aligned}
x_2(t) = &\sum_{k=1}^{m} \int_0^t \int_0^{t-t_2} e^{At_2} N_k e^{At_1} B u(t - t_1 - t_2) u_k(t - t_2) dt_1 dt_2\\
&+ \int_0^t \int_0^{t-t_3} \int_0^{t-t_3} e^{At_3} H\left(e^{At_2} B \otimes e^{At_1} B\right) u(t - t_2 - t_3) \otimes u(t - t_1 - t_3) dt_1 dt_2 dt_3.
\end{aligned}$$

Similarly, one can write down explicit expressions for $x_s(t), s \geq 3$, as well, but the notation and expression become tedious, and we skip them for brevity. Then, we can write the output $y(t)$ of the QB system as $y(t) = \sum_{s=1}^{\infty} C x_s(t)$, leading to the input/output relation of the QB system (1.1)

$$\begin{aligned}
y(t) = &\int_0^t C e^{At_1} B u(t - t_1) dt_1\\
&+ \int_0^t \int_0^{t-t_2} C e^{At_2} \left[N_1, \ldots, N_m\right] \left(I_m \otimes e^{At_1} B\right) \left(u(t - t_2) \otimes u(t - t_1 - t_2)\right) dt_1 dt_2\\
&+ \int_0^t \int_0^{t-t_3} \int_0^{t-t_3} C e^{At_3} H\left(e^{At_2} B \otimes e^{At_1} B\right)\\
&\times u(t - t_2 - t_3) \otimes u(t - t_1 - t_3) dt_1 dt_2 dt_3 + \cdots.
\end{aligned} \tag{3.4}$$

Examining the structure of (3.4) reveals that the *kernels* $f_i(t_1, \ldots, t_i)$ of (3.4) are given by the recurrence formula

$$f_i(t_1, \ldots, t_i) = C g_i(t_1, \ldots, t_i), \tag{3.5}$$

where

$$\begin{aligned}
g_1(t_1) &= e^{At_1} B,\\
g_2(t_1, t_2) &= e^{At_2} \left[N_1, \ldots, N_m\right] \left(I_m \otimes e^{At_1} B\right),
\end{aligned}$$

$$g_i(t_1,\dots,t_i) = e^{At_i}\Big[H\left[g_1(t_1)\otimes g_{i-2}(t_2,\dots,t_{i-1}),\dots,g_{i-2}(t_1,\dots,t_{i-2})\otimes g_1(t_{i-1})\right],$$
$$\left[N_1,\dots,N_m\right](I_m\otimes g_{i-1})\Big], \quad i\geq 3. \tag{3.6}$$

As shown in [48], the $\mathcal{H}_2$-norm of a bilinear system can be defined in terms of an infinite series of kernels, corresponding to its input/output mapping. Inspired by this definition, next we introduce the $\mathcal{H}_2$-norm of a QB system based on these kernels.

DEFINITION 3.1. *Consider the QB system* (1.1) *with its Volterra kernels, defined in* (3.5)*. Then, we define the $\mathcal{H}_2$-norm of the QB system by*

$$\|\Sigma\|_{\mathcal{H}_2} := \sqrt{\operatorname{tr}\left(\sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} f_i(t_1,\dots,t_i)f_i^T(t_1,\dots,t_i)dt_1\dots dt_i\right)}. \tag{3.7}$$

Even though this definition naturally extends the $\mathcal{H}_2$-norm to QB system, it is not suitable for computation. Fortunately, we can find an alternative way to compute the norm in a numerically efficient way using matrix equations. We know from the cases of linear and bilinear systems that the $\mathcal{H}_2$-norms of these systems can be computed in terms of certain system Gramians. We next show that this is also the case for QB systems. The algebraic Gramians for QB systems have recently been studied in [13]. So, in the following, we extend such relations between the $\mathcal{H}_2$-norm (see Definition 3.1) and the systems Gramians to QB systems.

LEMMA 3.2. *Consider a QB system with a stable matrix A, and let P and Q, respectively, be the controllability and observability Gramians of the system, which are the unique positive semidefinite solutions of the following quadratic-type Lyapunov equations:*

$$AP + PA^T + H(P\otimes P)H^T + \sum_{k=1}^{m} N_kPN_k^T + BB^T = 0 \quad \textit{and} \tag{3.8}$$

$$A^TQ + QA + \mathcal{H}^{(2)}(P\otimes Q)\left(\mathcal{H}^{(2)}\right)^T + \sum_{k=1}^{m} N_k^TQN_k + C^TC = 0. \tag{3.9}$$

Assuming the $\mathcal{H}_2$-norm of the QB system exists, i.e., the infinite series in (3.7) *converges, then the $\mathcal{H}_2$-norm of the QB system can be computed as*

$$\|\Sigma\|_{\mathcal{H}_2} := \sqrt{\operatorname{tr}\left(CPC^T\right)} = \sqrt{\operatorname{tr}\left(B^TQB\right)}. \tag{3.10}$$

Proof. We begin with the definition of the $\mathcal{H}_2$-norm of a QB system, that is,

$$\|\Sigma\|_{\mathcal{H}_2} = \sqrt{\operatorname{tr}\left(\sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} f_i(t_1,\dots,t_i)f_i^T(t_1,\dots,t_i)dt_1\dots dt_i\right)}$$
$$= \sqrt{\operatorname{tr}\left(C\left(\sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} g_i(t_1,\dots,t_i)g_i^T(t_1,\dots,t_i)dt_1\dots dt_i\right)C^T\right)},$$

where $f_i(t_1,\dots,t_i)$ and $g_i(t_1,\dots,t_i)$ are defined in (3.5) and(3.6), respectively. It is shown in [13] that

$$\sum_{i=1}^{\infty}\int_0^{\infty}\cdots\int_0^{\infty} g_i(t_1,\dots,t_i)g_i^T(t_1,\dots,t_i)dt_1\dots dt_i = P, \tag{3.11}$$

where P solves (3.8) if the series in (3.11) converges. Thus,

$$\|\Sigma\|_{\mathcal{H}_2} = \sqrt{\operatorname{tr}(CPC^T)}.$$

Next, we prove that $\operatorname{tr}\left(CPC^T\right) = \operatorname{tr}\left(B^TQB\right)$, where Q solves (3.9). Making use of the Kronecker product properties (2.1), we can write $\operatorname{tr}\left(CPC^T\right)$ as

$$\operatorname{tr}\left(CPC^T\right) = \mathcal{I}_p^T(C \otimes C)\operatorname{vec}(P).$$

Vectorizing both sides of (3.8) yields

(3.12)
$$\left(A \otimes I_n + I_n \otimes A + \sum_{k=1}^{m} N_k \otimes N_k\right)\operatorname{vec}(P) + (H \otimes H)\operatorname{vec}(P \otimes P) + (B \otimes B)\mathcal{I}_m = 0.$$

Using Lemma 2.4 in the above equation and performing some simple manipulations yield an expression for $\operatorname{vec}(P)$ as

$$\operatorname{vec}(P) = \mathcal{G}^{-1}(B \otimes B)\mathcal{I}_m =: P_v,$$

where

$$\mathcal{G} = -\left(A \otimes I_n + I_n \otimes A + \sum_{k=1}^{m} N_k \otimes N_k + (H \otimes H)T_{(n,n)}(I_{n^2} \otimes \operatorname{vec}(P))\right).$$

Thus,

$$\operatorname{tr}\left(CPC^T\right) = \mathcal{I}_p^T(C \otimes C)\mathcal{G}^{-1}(B \otimes B)\mathcal{I}_m = \mathcal{I}_m^T(B^T \otimes B^T)\mathcal{G}^{-T}(C^T \otimes C^T)\mathcal{I}_p. \tag{3.13}$$

Now, let $Q_v = \mathcal{G}^{-T}(C^T \otimes C^T)\mathcal{I}_p^T$. As a result, we obtain

$$\begin{aligned}(C^T \otimes C^T)\mathcal{I}_p^T &= \operatorname{vec}\left(C^TC\right) = \mathcal{G}^TQ_v \\ &= -\left(A^T \otimes I_n + I_n \otimes A^T + \sum_{k=1}^{m} N_k^T \otimes N_k^T\right)Q_v \\ &\quad + \left((H \otimes H)\,T_{(n,n)}\,(I_{n^2} \otimes P_v)\right)^T Q_v.\end{aligned}$$

Next, we consider a matrix $\widetilde{Q}$ such that $(\widetilde{Q}) = Q_v$, which further simplifies the above equation as

(3.14)
$$\operatorname{vec}\left(C^TC\right) = -\operatorname{vec}\left(A^T\widetilde{Q} + \widetilde{Q}A + \sum_{k=1}^{m} N_k^T\widetilde{Q}N_k\right) - \left((H \otimes H)T_{(n,n)}(I_{n^2} \otimes P_v)\right)^T Q_v.$$

Now, we focus on the transpose of the second part of (3.14), that is,

$$\begin{aligned}&Q_v^T(H \otimes H)T_{(n,n)}\,(I_{n^2} \otimes \operatorname{vec}(P)) \\ &\quad = Q_v^T(H \otimes H)T_{(n,n)}\left[e_1^{n^2} \otimes \operatorname{vec}(P), \ldots, e_{n^2}^{n^2} \otimes \operatorname{vec}(P)\right] \\ &\quad = Q_v^T(H \otimes H)\left[\operatorname{vec}(\Psi_1 \otimes P), \ldots, \operatorname{vec}(\Psi_{n^2} \otimes P)\right] =: \Xi, \qquad \text{(using Lemma 2.4)}\end{aligned}$$

where $\Psi_i \in \mathbb{R}^{n \times n}$ is such that $e_i^{n^2} = \operatorname{vec}(\Psi_i)$. Using (2.1) and Lemma 2.3, we further analyze the above equation:

$$
\begin{aligned}
\Xi &= \operatorname{vec}\left(\widetilde{Q}\right)^T \left[\operatorname{vec}\left(H\left(\Psi_1 \otimes P\right)H^T\right), \ldots, \operatorname{vec}\left(H\left(\Psi_{n^2} \otimes P\right)H^T\right)\right] \\
&= \operatorname{vec}\left(\widetilde{Q}\right)^T \left[\operatorname{vec}\left(H\left(P \otimes \Psi_1\right)H^T\right), \ldots, \operatorname{vec}\left(H\left(P \otimes \Psi_{n^2}\right)H^T\right)\right] \\
&= \Big[\operatorname{vec}\left(\Psi_1\right)^T \operatorname{vec}\left(\mathcal{H}^{(2)}\left(P \otimes \widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right), \ldots, \\
&\qquad\qquad \operatorname{vec}\left(\Psi_{n^2}\right)^T \operatorname{vec}\left(\mathcal{H}^{(2)}\left(P \otimes \widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right)\Big] \\
&= \left[\left(e_1^{n^2}\right)^T \operatorname{vec}\left(\mathcal{H}^{(2)}\left(P \otimes \widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right), \ldots, \left(e_{n^2}^{n^2}\right)^T \operatorname{vec}\left(\mathcal{H}^{(2)}\left(P \otimes \widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right)\right] \\
&= \left(\operatorname{vec}\left(\mathcal{H}^{(2)}\left(P \otimes \widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right)\right)^T.
\end{aligned}
$$

Substituting this relation into (3.14) yields

$$
\operatorname{vec}\left(C^TC\right) = -\operatorname{vec}\left(A^T\widetilde{Q} + \widetilde{Q}A + \sum_{k=1}^{m} N_k^T\widetilde{Q}N_k + \mathcal{H}^{(2)}\left(P \otimes \widetilde{Q}\right)(\mathcal{H}^{(2)})^T\right),
$$

which shows that $\widetilde{Q}$ solves (3.9) as well. Since it is assumed that Eq. (3.9) has a unique solution, we get $\widetilde{Q} = Q$. Replacing $\mathcal{G}^{-T}(C^T \otimes C^T)\mathcal{I}_p^T$ by $\operatorname{vec}(Q)$ in (3.13) and using (2.1) results in

$$
\operatorname{tr}\left(CPC^T\right) = \mathcal{I}_m^T(B^T \otimes B^T)\operatorname{vec}(Q) = \operatorname{tr}\left(B^TQB\right).
$$

This concludes the proof. ∎

It can be seen that if H is zero, the expression (3.10) boils down to the $\mathcal{H}_2$-norm of bilinear systems, and if all N_k are also set to zero then it provides us the $\mathcal{H}_2$-norm of stable linear systems as one would expect.

Remark 3.3. In Lemma 3.2, we have assumed that the solutions of (3.8) and (3.9) exist and that they are unique and positive semidefinite. Equivalently, the series appearing in the definition of the $\mathcal{H}_2$-norm is finite (see Definition 3.1); hence, the $\mathcal{H}_2$-norm exists. Naturally, the stability of the matrix A is necessary for the existence of Gramians, and a detailed study of the solutions of (3.8) and (3.9) has been carried out in [13]. However, as for bilinear systems, these Gramians may not have the desired properties such as uniqueness and positive semidefiniteness when $\|N_k\|$ and $\|H\|$ are large.

Nonetheless, from a MOR viewpoint, a solution of these problems can be obtained via rescaling of the system as has been done in the bilinear case [20]. For this, we need to rescale the input variable $u(t)$ as well as the state vector $x(t)$. More precisely, in (1.1), we can replace $x(t)$ and $u(t)$ with $x(t) =: \gamma\widetilde{x}(t)$ and $u(t) =: \gamma\widetilde{u}(t)$, respectively. This leads to

$$
\begin{aligned}
\gamma\dot{\widetilde{x}}(t) &= \gamma A\widetilde{x}(t) + \gamma^2 H\left(\widetilde{x}(t) \otimes \widetilde{x}(t)\right) + \gamma^2 \sum_{k=1}^{m} N_k\widetilde{x}(t)\widetilde{u}_k(t) + \gamma B\widetilde{u}(t), \\
y(t) &= \gamma C\widetilde{x}(t), \quad \widetilde{x}(0) = 0.
\end{aligned} \tag{3.15}
$$

For $\gamma \neq 0$, we get a scaled system as follows:

$$
\begin{aligned}
\dot{\widetilde{x}}(t) &= A\widetilde{x}(t) + (\gamma H)\left(\widetilde{x}(t) \otimes \widetilde{x}(t)\right) + \sum_{k=1}^{m} (\gamma N_k)\widetilde{x}(t)\widetilde{u}_k(t) + B\widetilde{u}(t), \\
\widetilde{y}(t) &= C\widetilde{x}(t), \quad \widetilde{x}(0) = 0,
\end{aligned} \tag{3.16}
$$

where $\widetilde{y}(t) = y(t)/\gamma$. Comparing the systems (1.1) and (3.16) shows that the input/output mappings differ by the scaling factor γ. Hence, we can use the system (3.16) as an auxiliary system during the MOR process; more precisely, to compute the model reduction bases. However, note that the reduced-order system is constructed by applying Petrov–Galerkin projection applied to the original, unscaled matrices in (1.1).

Our primary aim is to determine a reduced-order system that minimizes the $\mathcal{H}_2$-norm of the error system. From the derived $\mathcal{H}_2$-norm expression for the QB system, it is clear that the true $\mathcal{H}_2$-norm has a complicated structure as defined in (3.7) and does not lend itself well to deriving necessary conditions for optimality. Therefore, to simplify the problem, we focus only on the first three leading terms of the series (3.4). The main reason for considering the three terms is that it is the minimum number of terms containing contributions from all the system matrices (A, H, N_k, B, C); in other words, linear, bilinear, and quadratic terms are already contained in these first three terms. Our approach is also inspired by [22], where a truncated $\mathcal{H}_2$-norm is defined for bilinear systems and used to construct high-fidelity reduced-order models minimizing corresponding error measures. Therefore, based on these three leading terms, we define a truncated $\mathcal{H}_2$-norm for QB systems, denoted by $\|\Sigma\|_{\mathcal{H}_2^{(\mathcal{T})}}$. Precisely, the truncated norm can be defined as follows:

$$\|\Sigma\|_{\mathcal{H}_2^{(\mathcal{T})}} := \sqrt{\operatorname{tr}\left(\sum_{i=1}^{3}\int_0^\infty \cdots \int_0^\infty \widetilde{f}_i(t_1,\ldots,t_i)\left(\widetilde{f}_i(t_1,\ldots,t_i)\right)^T dt_1 \cdots dt_i\right)}, \tag{3.17}$$

where

$$\widetilde{f}_i(t_1,\ldots,t_i) = C\widetilde{g}_i(t_1,\ldots,t_i), \qquad i \in \{1,2,3\}, \tag{3.18}$$

and

$$\widetilde{g}_1(t_1) = e^{At_1}B, \qquad \widetilde{g}_2(t_1,t_2) = e^{At_2}\left[N_1,\ldots,N_m\right]\left(I_m \otimes e^{At_1}B\right),$$
$$\widetilde{g}_3(t_1,t_2,t_3) = e^{At_3}H(e^{At_2}B \otimes e^{At_1}B).$$

Analogous to the $\mathcal{H}_2$-norm of the QB system, a truncated $\mathcal{H}_2$-norm of QB systems can be determined by truncated controllability and observability Gramians associated with the QB system, denoted by $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$, respectively [13]. If the matrix A is stable, these truncated Gramians exist (see [13] for the integral form) and are the unique and positive semidefinite solutions of the following Lyapunov equations:

$$AP_{\mathcal{T}} + P_{\mathcal{T}}A^T + \sum_{k=1}^{m} N_k P_l N_k^T + H(P_l \otimes P_l)H^T + BB^T = 0, \tag{3.19a}$$

$$A^TQ_{\mathcal{T}} + Q_{\mathcal{T}}A + \sum_{k=1}^{m} N_k^T Q_l N_k + \mathcal{H}^{(2)}(P_l \otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T + C^TC = 0, \tag{3.19b}$$

where $\mathcal{H}^{(2)}$ is the mode-2 matricization of the QB Hessian and P_l and Q_l are the unique solutions of the following Lyapunov equations:

$$AP_l + P_lA^T + BB^T = 0, \tag{3.20a}$$

$$A^TQ_l + Q_lA + C^TC = 0. \tag{3.20b}$$

In what follows, we show the connection between the truncated $\mathcal{H}_2$-norm and the defined truncated Gramians for QB systems.

LEMMA 3.4. *Let Σ be the QB system (1.1) with a stable A matrix. Then the truncated $\mathcal{H}_2$-norm based on the first three terms of the Volterra series is given by*

$$\|\Sigma\|_{\mathcal{H}_2^{(\mathcal{T})}} = \sqrt{\operatorname{tr}\left(CP_{\mathcal{T}}C^T\right)} = \sqrt{\operatorname{tr}\left(B^TQ_{\mathcal{T}}B\right)},$$

where $P_{\mathcal{T}}$ and $Q_{\mathcal{T}}$ are truncated controllability and observability Gramians of the system, satisfying (3.19).

Proof. First, we note that (3.19) and (3.20) are standard Lyapunov equations. As A is assumed to be stable, these equations have unique solutions [5]. Next, let $\mathcal{R}_i$ be

$$\mathcal{R}_i = \int_0^\infty \cdots \int_0^\infty \widetilde{f}_i(t_1,\ldots,t_i)\left(\widetilde{f}_i(t_1,\ldots,t_i)\right)^T dt_1\cdots dt_i,$$

where $\widetilde{f}_i(t_1,\ldots,t_i)$ are as defined in (3.18). Thus, $\|\Sigma\|^2_{\mathcal{H}_2^{(\mathcal{T})}} = \operatorname{tr}\left(C(\sum_{i=1}^3 \mathcal{R}_i)C^T\right)$. It is shown in [13] that $\sum_{i=1}^3 \mathcal{R}_i = P_{\mathcal{T}}$ solves the Lyapunov equation (3.19a). Hence,

$$\|\Sigma\|^2_{\mathcal{H}_2^{(\mathcal{T})}} = \operatorname{tr}\left(CP_{\mathcal{T}}C^T\right).$$

Next, we show that $\operatorname{tr}\left(CP_{\mathcal{T}}C^T\right) = \operatorname{tr}\left(B^TQ_{\mathcal{T}}B\right)$. For this, we use the trace property (2.1b) to obtain

$$\operatorname{tr}\left(CP_{\mathcal{T}}C^T\right) = \left(\mathcal{I}_p\right)^T(C\otimes C)\operatorname{vec}\left(P_{\mathcal{T}}\right) \text{ and } \operatorname{tr}\left(B^TQ_{\mathcal{T}}B\right) = \left(\operatorname{vec}\left(Q_{\mathcal{T}}\right)\right)^T(B\otimes B)\mathcal{I}_m.$$

Applying $\operatorname{vec}(\cdot)$ to both sides of (3.19) results in

$$\operatorname{vec}\left(P_{\mathcal{T}}\right) = \mathcal{L}^{-1}\left((B\otimes B)\mathcal{I}_m + \sum_{k=1}^m (N_k\otimes N_k)\mathcal{L}^{-1}(B\otimes B)\mathcal{I}_m + \operatorname{vec}\left(H\left(P_l\otimes P_l\right)H^T\right)\right) \text{ and}$$

$$\operatorname{vec}\left(Q_{\mathcal{T}}\right) = \mathcal{L}^{-T}\left((C\otimes C)^T\mathcal{I}_p + \sum_{k=1}^m (N_k\otimes N_k)^T\mathcal{L}^{-T}(C\otimes C)^T\mathcal{I}_p + \operatorname{vec}\left(\mathcal{H}^{(2)}(P_l\otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T\right)\right),$$

where $\mathcal{L} = -(A\otimes I_n + I_n\otimes A)$ and P_l and Q_l solve (3.20). Thus,

$$\operatorname{tr}\left(B^TQ_{\mathcal{T}}B\right) = \left(\left(\mathcal{I}_p\right)^T(C\otimes C) + \left(\mathcal{I}_p\right)^T(C\otimes C)\mathcal{L}^{-1}\sum_{k=1}^m (N_k\otimes N_k) + \left(\operatorname{vec}\left(\mathcal{H}^{(2)}(P_l\otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T\right)\right)^T\right)\mathcal{L}^{-1}(B\otimes B)\mathcal{I}_m. \tag{3.21}$$

Since P_l and Q_l are the unique solutions of (3.20a) and (3.20b), this gives $\operatorname{vec}\left(P_l\right) = \mathcal{L}^{-1}(B\otimes B)\mathcal{I}_m$ and $\operatorname{vec}\left(Q_l\right) = \mathcal{L}^{-T}(C\otimes C)^T\mathcal{I}_p$. This implies that

$$\begin{aligned}\left(\operatorname{vec}\left(\mathcal{H}^{(2)}(P_l\otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T\right)\right)^T &\operatorname{vec}(P_l)\\ &= \operatorname{vec}(P_l)^T \operatorname{vec}\left(\mathcal{H}^{(2)}(P_l\otimes Q_l)\left(\mathcal{H}^{(2)}\right)^T\right)\\ &= \operatorname{vec}(Q_l)^T \operatorname{vec}\left(H(P_l\otimes P_l)H^T\right) \qquad \text{(using Lemma 2.2)}\\ &= (\mathcal{I}_p)^T (C\otimes C)\mathcal{L}^{-1}\operatorname{vec}\left(H(P_l\otimes P_l)H^T\right).\end{aligned}$$

Substituting the above relation in (3.21) yields

$$\begin{aligned}\operatorname{tr}\left(B^TQ_{\mathcal{T}}B\right) &= (\mathcal{I}_p)^T(C\otimes C)\mathcal{L}^{-1}\Big((B\otimes B)\mathcal{I}_m+\sum_{k=1}^m (N_k\otimes N_k)\mathcal{L}^{-1}(B\otimes B)\mathcal{I}_m\\ &\qquad +\operatorname{vec}\left(H(P_l\otimes P_l)H^T\right)\Big)\\ &= (\mathcal{I}_p)^T(C\otimes C)\operatorname{vec}(P_{\mathcal{T}}) = \operatorname{tr}\left(CP_{\mathcal{T}}C^T\right).\end{aligned}$$

This concludes the proof. ∎

Remark 3.5. One can consider the first $\mathcal{M}$ terms of the corresponding Volterra series and, based on these $\mathcal{M}$ kernels, another truncated $\mathcal{H}_2$-norm can be defined. However, this significantly increases the complexity of the problem. In this paper, we stick to the truncated $\mathcal{H}_2$-norm for the QB system that depends on the first three terms of the input/output mapping. We intend to construct reduced-order systems (1.2) such that this truncated $\mathcal{H}_2$-norm of the error system is minimized. Another motivation for the derived truncated $\mathcal{H}_2$-norm for QB systems is that for bilinear systems, the authors in [22] showed that the $\mathcal{H}_2$-optimal model reduction based on a truncated $\mathcal{H}_2$-norm (with only two terms of the Volterra series of a bilinear system) also mimics the accuracy of the true $\mathcal{H}_2$-optimal approximation very closely.

3.2. Optimality conditions based on the truncated $\mathcal{H}_2$-norm. We now derive necessary conditions for optimal model reduction based on the truncated $\mathcal{H}_2$-norm of the error system. First, we define the QB error system. For the full QB model Σ in (1.1) and the reduced QB model $\widehat{\Sigma}$ in (1.2), we can write the error system as

(3.22)
$$\begin{aligned}\begin{bmatrix}\dot{x}(t)\\ \dot{\widehat{x}}(t)\end{bmatrix} &= \underbrace{\begin{bmatrix}A & \mathbf{0}\\ \mathbf{0} & \widehat{A}\end{bmatrix}}_{A^e}\underbrace{\begin{bmatrix}x(t)\\ \widehat{x}(t)\end{bmatrix}}_{x^e(t)} + \begin{bmatrix}H\left(x(t)\otimes x(t)\right)\\ \widehat{H}\left(\widehat{x}(t)\otimes\widehat{x}(t)\right)\end{bmatrix} + \sum_{k=1}^m \underbrace{\begin{bmatrix}N_k & \mathbf{0}\\ \mathbf{0} & \widehat{N}_k\end{bmatrix}}_{N_k^e}\begin{bmatrix}x(t)\\ \widehat{x}(t)\end{bmatrix}u_k(t) + \underbrace{\begin{bmatrix}B\\ \widehat{B}\end{bmatrix}}_{B^e}u(t),\\ y^e(t) &= y(t)-\widehat{y}(t) = \underbrace{\begin{bmatrix}C & -\widehat{C}\end{bmatrix}}_{C^e}\begin{bmatrix}x^T(t) & \widehat{x}^T(t)\end{bmatrix}^T, \quad x^e(0)=0.\end{aligned}$$

It can be seen that the error system (3.22) is not in the conventional QB form due to the absence of the quadratic term $x^e(t)\otimes x^e(t)$. However, we can rewrite the system (3.22) into a regular QB form by using an appropriate Hessian of the error system (3.22) as follows:

(3.23)
$$\Sigma^e := \begin{cases}\dot{x}^e(t) = A^ex^e(t) + H^e\left(x^e(t)\otimes x^e(t)\right) + \sum_{k=1}^m N_k^e x^e(t)u_k(t) + B^eu(t),\\ y^e(t) = C^ex^e(t), \quad x^e(0)=0,\end{cases}$$

where $H^e = \left[\begin{smallmatrix} H\mathcal{F} \\ \widehat{H}\widehat{\mathcal{F}} \end{smallmatrix}\right]$ with $\mathcal{F} = \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix} \otimes \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix}$ and $\widehat{\mathcal{F}} = \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix}$. Next, we consider the truncated $\mathcal{H}_2$-norm, as defined in Lemma 3.4, for the error system (3.23). For the existence of this norm for the system (3.23), it is necessary to assume that the matrix A^e is stable, i.e., the matrices A and $\widehat{A}$ are stable. Further, we assume that the matrix $\widehat{A}$ is diagonalizable. Then, by performing basic algebraic manipulations and making use of Lemma 2.4, we obtain the expression for the error functional $\mathcal{E}$ based on the truncated $\mathcal{H}_2$-norm of the error system (3.23) as shown next.

COROLLARY 3.6. *Let Σ be the original system, having a stable matrix A, and let $\widehat{\Sigma}$ be the reduced-order system, having a stable and diagonalizable matrix $\widehat{A}$. Then*

$$\begin{aligned} \mathcal{E}^2 := \|\Sigma^e\|^2_{\mathcal{H}_2^{(\mathcal{T})}} = (\mathcal{I}_p)^T (C^e \otimes C^e)(-A^e \otimes I_{n+r} - I_{n+r} \otimes A^e)^{-1} \Big((B^e \otimes B^e)\mathcal{I}_m \\ + \sum_{k=1}^{m} (N_k^e \otimes N_k^e)\,\mathrm{vec}\,(P_l^e) + \mathrm{vec}\left(H^e (P_l^e \otimes P_l^e)\,(H^e)^T \right) \Big), \end{aligned} \tag{3.24}$$

where P_l^e solves

$$A^e P_l^e + P_l^e (A^e)^T + B^e (B^e)^T = 0.$$

Furthermore, let $\widehat{A} = \widehat{R}\widehat{\Lambda}\widehat{R}^{-1}$ be the spectral decomposition of $\widehat{A}$, and define $\widetilde{B} = \widehat{R}^{-1}\widehat{B}$, $\widetilde{C} = \widehat{C}\widehat{R}$, $\widetilde{N}_k = \widehat{R}^{-1}\widehat{N}_k\widehat{R}$, and $\widetilde{H} = \widehat{R}^{-1}\widehat{H}(\widehat{R} \otimes \widehat{R})$. Then, the error can be rewritten as

$$\begin{aligned} \mathcal{E}^2 = (\mathcal{I}_p)^T \left(\widetilde{C}^e \otimes \widetilde{C}^e \right) \left(-\widetilde{A}^e \otimes I_{n+r} - I_{n+r} \otimes \widetilde{A}^e \right)^{-1} \left(\left(\widetilde{B}^e \otimes \widetilde{B}^e \right) \mathcal{I}_m \right. \\ \left. + \sum_{k=1}^{m} \left(\widetilde{N}_k^e \otimes \widetilde{N}_k^e \right) \mathcal{P}_l + \left(\widetilde{H}^e \otimes \widetilde{H}^e \right) T_{(n+r,n+r)} (\mathcal{P}_l \otimes \mathcal{P}_l) \right), \end{aligned} \tag{3.25}$$

where

$$\widetilde{A}^e = \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & \Lambda \end{bmatrix}, \widetilde{N}_k^e = \begin{bmatrix} N_k & \mathbf{0} \\ \mathbf{0} & \widetilde{N}_k \end{bmatrix}, \widetilde{H}^e = \begin{bmatrix} H\mathcal{F} \\ \widetilde{H}\widehat{\mathcal{F}} \end{bmatrix}, \widetilde{B}^e = \begin{bmatrix} B \\ \widetilde{B} \end{bmatrix}, \widetilde{C}^e = \begin{bmatrix} C^T \\ -\widetilde{C}^T \end{bmatrix}^T,$$

$$\mathcal{P}_l = \begin{bmatrix} \mathcal{P}_l^{(1)} \\ \mathcal{P}_l^{(2)} \end{bmatrix} = \begin{bmatrix} \left(-A \otimes I_{n+r} - I_n \otimes \widetilde{A}^e \right)^{-1} \left(B \otimes \widetilde{B}^e \right) \mathcal{I}_m \\ \left(-\Lambda \otimes I_{n+r} - I_r \otimes \widetilde{A}^e \right)^{-1} \left(\widetilde{B} \otimes \widetilde{B}^e \right) \mathcal{I}_m \end{bmatrix}, \text{ and} \tag{3.26}$$

$$T_{(n+r,n+r)} = I_{n+r} \otimes \left[I_{n+r} \otimes e_1^{n+r}, \ldots, I_{n+r} \otimes e_{n+r}^{n+r} \right] \otimes I_{n+r}.$$

The above spectral decomposition for $\widehat{A}$ is computationally useful in simplifying the expressions, as we will see later. It reduces the number of optimization variables by $r(r-1)$ since Λ becomes a diagonal matrix without changing the value of the cost function (this is a state-space transformation of the reduced model, which does not change the input-output mapping). Even though it limits the reduced-order systems to those only having diagonalizable $\widehat{A}$, as observed in the linear [24] and bilinear cases [8, 22], it is extremely rare in practice that the optimal $\mathcal{H}_2$ models will have a nondiagonalizable $\widehat{A}$; therefore, this diagonalizability assumption does not incur any restriction from a practical perspective.

Our aim is to choose the optimization variables Λ, $\widetilde{B}$, $\widetilde{C}$, $\widetilde{N}_k$, and $\widetilde{H}$ such that the $\|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2^{(\mathcal{T})}}$, i.e., equivalently the error expression (3.25), is minimized. Before

we proceed further, we introduce a particular permutation matrix

$$M_{pqr} = \begin{bmatrix} I_p \otimes \begin{bmatrix} I_q \\ \mathbf{0} \end{bmatrix} & I_p \otimes \begin{bmatrix} \mathbf{0} \\ I_r \end{bmatrix} \end{bmatrix}, \tag{3.27}$$

which will prove helpful in simplifying the expressions related to the Kronecker product of block matrices. For example, consider matrices $\mathcal{A} \in \mathbb{R}^{p\times p}$, $\mathcal{B} \in \mathbb{R}^{q\times q}$, and $\mathcal{C} \in \mathbb{R}^{r\times r}$. Then, the following relation holds:

$$M_{pqr}^T \left(\mathcal{A} \otimes \begin{bmatrix} \mathcal{B} & \mathbf{0} \\ \mathbf{0} & \mathcal{C} \end{bmatrix} \right) M_{pqr} = \begin{bmatrix} \mathcal{A} \otimes \mathcal{B} & \mathbf{0} \\ \mathbf{0} & \mathcal{A} \otimes \mathcal{C} \end{bmatrix}.$$

Similar block structures can be found in the error expression $\mathcal{E}$ in Corollary 3.6, which can be simplified analogously. Moreover, due to the presence of many Kronecker products, it will be convenient to derive necessary conditions for optimality in the Kronecker product formulation itself. Furthermore, these conditions can be easily translated into a theoretically equivalent framework of Sylvester equations, which are more concise, are more easily interpretable, and, more importantly, automatically lead to an effective numerical algorithm for model reduction. To this end, let $V_i \in \mathbb{R}^{n\times r}$ and $W_i \in \mathbb{R}^{n\times r}$, $i \in \{1, 2\}$ be the solutions of the following standard Sylvester equations:

$$V_1(-\Lambda) - AV_1 = B\widetilde{B}^T, \tag{3.28a}$$

$$W_1(-\Lambda) - A^T W_1 = C^T \widetilde{C}, \tag{3.28b}$$

$$V_2(-\Lambda) - AV_2 = \sum_{k=1}^{m} N_k V_1 \widetilde{N}_k^T + H(V_1 \otimes V_1)\widetilde{H}^T, \quad \text{and} \tag{3.28c}$$

$$W_2(-\Lambda) - A^T W_2 = \sum_{k=1}^{m} N_k^T W_1 \widetilde{N}_k + 2 \cdot \mathcal{H}^{(2)}(V_1 \otimes W_1)(\widetilde{\mathcal{H}}^{(2)})^T, \tag{3.28d}$$

where $\Lambda, \widetilde{N}_k, \widetilde{B}$, and $\widetilde{C}$ are as defined in Corollary 3.6. Furthermore, we define trial and test basis matrices $V \in \mathbb{R}^{n\times r}$ and $W \in \mathbb{R}^{n\times r}$ as

$$V = V_1 + V_2 \quad \text{and} \quad W = W_1 + W_2. \tag{3.29}$$

We also define $\widehat{V} \in \mathbb{R}^{r\times r}$ and $\widehat{W} \in \mathbb{R}^{r\times r}$ (which will appear in the optimality conditions as we see later) as follows:

$$\widehat{V} = \widehat{V}_1 + \widehat{V}_2 \quad \text{and} \quad \widehat{W} = \widehat{W}_1 + \widehat{W}_2, \tag{3.30}$$

where $\widehat{V}_i \in \mathbb{R}^{r\times r}$, $\widehat{W}_i \in \mathbb{R}^{r\times r}$, $i \in \{1, 2\}$, are the solutions of the set of equations in (3.28) but with the original system's state-space matrices being replaced with the reduced-order system ones; for example, A with $\widehat{A}$ and B with $\widehat{B}$, etc. Next, we present first-order necessary conditions for optimality, which aim at minimizing the error expression (3.25). The following theorem extends the truncated $\mathcal{H}_2$ optimal conditions from the bilinear case to the much more general QB nonlinearities.

THEOREM 3.7. *Let Σ and $\widehat{\Sigma}$ be the original and reduced-order systems as defined in* (1.1) *and* (1.2)*, respectively. Let $\widehat{\Lambda} = \widehat{R}^{-1}\widehat{A}\widehat{R}$ be the spectral decomposition of $\widehat{A}$, and define $\widetilde{H} = \widehat{R}^{-1}\widehat{H}(\widehat{R} \otimes \widehat{R})$, $\widetilde{N}_k = \widehat{R}^{-1}\widehat{N}_k\widehat{R}$, $\widetilde{C} = \widehat{C}\widehat{R}$, $\widetilde{B} = \widehat{R}^{-1}\widehat{B}$. If $\widehat{\Sigma}$ is a reduced-order system that minimizes the truncated $\mathcal{H}_2$-norm of the error system* (3.23) *subject to $\widehat{A}$ being diagonalizable, then $\widehat{\Sigma}$ satisfies the following conditions:*

(3.31a)
$$\operatorname{tr}\left(CVe_i^r\left(e_j^p\right)^T\right) = \operatorname{tr}\left(\widehat{C}\widehat{V}e_i^r\left(e_j^p\right)^T\right), \qquad i \in \{1,\dots,r\},\quad j \in \{1,\dots,p\},$$

(3.31b)
$$\operatorname{tr}\left(B^TWe_i^r\left(e_j^m\right)^T\right) = \operatorname{tr}\left(\widehat{B}^T\widehat{W}e_i^r\left(e_j^m\right)^T\right), \qquad i \in \{1,\dots,r\},\quad j \in \{1,\dots,m\},$$

(3.31c)
$$(W_1(:,i))^TN_kV_1(:,j) = (\widehat{W}_1(:,i))^T\widehat{N}_k\widehat{V}_1(:,j), \qquad i,j \in \{1,\dots,r\},\quad k \in \{1,\dots,m\},$$

(3.31d)
$$(W_1(:,i))^TH(V_1(:,j)\otimes V_1(:,l)) = (\widehat{W}_1(:,i))^T\widehat{H}(\widehat{V}_1(:,j)\otimes \widehat{V}_1(:,l)), \qquad i,j,l \in \{1,\dots,r\},$$

(3.31e)
$$(W_1(:,i))^TV(:,i) + (W_2(:,i))^TV_1(:,i) = (\widehat{W}_1(:,i))^T\widehat{V}(:,i) + \left(\widehat{W}_2(:,i)\right)^T\widehat{V}_1(:,i), \qquad i \in \{1,\dots,r\}.$$

Proof. The proof is given in Appendix B. ∎

3.3. Truncated QB iterative rational Krylov algorithm. The remaining challenge is now to develop a numerically efficient model reduction algorithm to construct a reduced QB system satisfying the first-order optimality conditions in Theorem 3.7. However, as in the linear [24] and bilinear [8, 22] cases, since the optimality conditions involve the matrices $V, W, \widehat{V}, \widehat{W}$, which depend on the reduced-order system matrices we are trying to construct, it is not a straightforward task to determine a reduced-order system directly that satisfies all the necessary conditions for optimality, i.e., (3.31a)–(3.31e). We propose Algorithm 3.1, which upon convergence leads to reduced-order systems that *approximately* satisfy the first-order necessary conditions for optimality given in Theorem 3.7. Throughout the paper, we denote the algorithm by truncated QB-IRKA or TQB-IRKA.

Remark 3.8. Ideally, *upon convergence* implies that the reduced-order quantities $\widehat{A}$, $\widehat{H}$, $\widehat{N}_k$, $\widehat{B}$, $\widehat{C}$ in Algorithm 3.1 stagnate. In a numerical implementation, one can check the stagnation based on the change of eigenvalues of the reduced matrix $\widehat{A}$ and terminate the algorithm once the relative change in the eigenvalues of $\widehat{A}$ is of the order of the machine precision. However, in all of our numerical experiments, we run TQB-IRKA until the relative change in the eigenvalues of $\widehat{A}$ is less than 10^{-5}. We observe that the quality of reduced-order systems does not change significantly thereafter, as in the cases of IRKA, B-IRKA, and TB-IRKA.

Our next goal is to show how the reduced-order system resulting from TQB-IRKA upon convergence relates to the first-order optimality conditions (3.31). As a first step, we provide explicit expressions showing how far away the resulting reduced-order system is from satisfying the optimality conditions. Later, based on these expressions, we discuss how far the reduced-order systems, obtained from TQB-IRKA for weakly nonlinear QB systems, satisfy the optimality condition with small perturbations. We also illustrate using our numerical examples in section 4 that, in practice, the reduced-order system seemingly often satisfies optimality conditions quite accurately.

THEOREM 3.9. *Let Σ be a QB system* (1.1) *and let $\widehat{\Sigma}$ be the reduced-order QB system* (1.2)*, computed by TQB-IRKA upon convergence. Let V_i, W_i, for $i \in \{1,2\}$, be the matrices that solve* (3.28)*, and let V and W be the matrices defining the projection*

Algorithm 3.1. TQB-IRKA for QB systems.

Input: The system matrices: $A, H, N_1, \ldots, N_m, B, C$.
Output: The reduced matrices: $\widehat{A}, \widehat{H}, \widehat{N}_1, \ldots, \widehat{N}_m, \widehat{B}, \widehat{C}$.

1: Symmetrize the Hessian H and determine its mode-2 matricization $\mathcal{H}^{(2)}$.
2: Make an initial guess for the reduced matrices $\widehat{A}, \widehat{H}, \widehat{N}_1, \ldots, \widehat{N}_m, \widehat{B}, \widehat{C}$ with $\widehat{A}$ being diagonalizable.
3: **while** not converged **do**
4: Perform the spectral decomposition of $\widehat{A}$ and define:
$$\widehat{\Lambda} = \widehat{R}^{-1}\widehat{A}\widehat{R}, \widetilde{N}_k = \widehat{R}^{-1}\widehat{N}_k\widehat{R}, \widetilde{H} = \widehat{R}^{-1}\widehat{H}\left(\widehat{R}\otimes\widehat{R}\right), \widetilde{B} = \widehat{R}^{-1}\widehat{B}, \widetilde{C} = \widehat{C}\widehat{R}.$$
5: Compute mode-2 matricization $\widetilde{\mathcal{H}}^{(2)}$.
6: Solve for V_1 and V_2:
$$-V_1\Lambda - AV_1 = B\widetilde{B}^T,$$
$$-V_2\Lambda - AV_2 = H(V_1\otimes V_1)\widetilde{H}^T + \sum_{k=1}^{m} N_k V_1 \widetilde{N}_k^T.$$
7: Solve for W_1 and W_2:
$$-W_1\Lambda - A^TW_1 = C^T\widetilde{C},$$
$$-W_2\Lambda - A^TW_2 = 2\cdot\mathcal{H}^{(2)}(V_1\otimes W_1)(\widetilde{\mathcal{H}}^{(2)})^T + \sum_{k=1}^{m} N_k^T W_1 \widetilde{N}_k.$$
8: Compute V and W:
$$V := V_1 + V_2, \qquad W := W_1 + W_2.$$
9: $V = \operatorname{orth}(V)$, $W = \operatorname{orth}(W)$.
10: Determine the reduced matrices:
$$\widehat{A} = (W^TV)^{-1}W^TAV, \qquad \widehat{H} = (W^TV)^{-1}W^TH(V\otimes V),$$
$$\widehat{N}_k = (W^TV)^{-1}W^TN_kV, \qquad \widehat{B} = (W^TV)^{-1}W^TB, \qquad \widehat{C} = CV.$$
11: **end while**

used for MOR, as defined in (3.29). *Similarly, let* $\widehat{V}_i, \widehat{W}_i$, *for* $i \in \{1,2\}$, *be the matrices that solve* (3.28), *where the original system's state-space matrices are being replaced with their reduced-order counterparts. Moreover, let* $\widehat{V}$ *and* $\widehat{W}$ *be the matrices defined in* (3.30). *Assume that* $\sigma(\widehat{A}) \cap \sigma(-\Pi A) = \emptyset$ *and* $\sigma(\widehat{A}) \cap \sigma(-\Pi^T A^T) = \emptyset$, *where* $\Pi = V(W^TV)^{-1}W^T$ *and* $\sigma(\cdot)$ *denotes the eigenvalue spectrum of a matrix. Furthermore, assume* $\Pi_v = V_1(W^TV_1)^{-1}W^T$ *and* $\Pi_w = W_1(V^TW_1)^{-1}V^T$ *exist. Then, the reduced-order system* $\widehat{\Sigma}$ *satisfies the following relations:*

(3.32a)
$$\operatorname{tr}\left(CVe_i^r\left(e_j^p\right)^T\right) = \operatorname{tr}\left(\widehat{C}\widehat{V}e_i^r\left(e_j^p\right)^T\right) + \epsilon_C^{(i,j)}, \qquad i\in\{1,\ldots,r\},\quad j\in\{1,\ldots,p\},$$

(3.32b)
$$\operatorname{tr}\left(B^TWe_i^r\left(e_j^m\right)^T\right) = \operatorname{tr}\left(\widehat{B}^T\widehat{W}e_i^r\left(e_j^m\right)^T\right) + \epsilon_B^{(i,j)}, \qquad i\in\{1,\ldots,r\},\quad j\in\{1,\ldots,m\},$$

(3.32c)
$$(W_1(:,i))^TN_kV_1(:,j) = (\widehat{W}_1(:,i))^T\widehat{N}_k\widehat{V}_1(:,j) + \epsilon_N^{(i,j,k)}, \qquad i,j\in\{1,\ldots,r\},\quad k\in\{1,\ldots,m\},$$

(3.32d)
$$(W_1(:,i))^TH(V_1(:,j)\otimes V_1(:,l)) = (\widehat{W}_1(:,i))^T\widehat{H}(\widehat{V}_1(:,j)\otimes\widehat{V}_1(:,l)) + \epsilon_H^{(i,j,l)}, \qquad i,j,l\in\{1,\ldots,r\},$$

(3.32e)

$$(W_1(:,i))^T V(:,i) + (W_2(:,i))^T V_1(:,i) = (\widehat{W}_1(:,i))^T \widehat{V}(:,i) + \left(\widehat{W}_2(:,i)\right)^T \widehat{V}_1(:,i) + \epsilon_\lambda^{(i)}, \quad i \in \{1,\dots,r\},$$

where

$$\begin{aligned}
\epsilon_C^{(i,j)} &= -\operatorname{tr}\left(CV\Gamma_v e_i^r \left(e_j^p\right)^T\right),\\
\epsilon_B^{(i,j)} &= -\operatorname{tr}\left(B^T W (W^T V)^{-T} \Gamma_w e_i^r \left(e_j^m\right)^T\right),\\
\epsilon_N^{(i,j,k)} &= (\epsilon_w(:,i))^T N_k (V_1(:,j) - \epsilon_v(:,j)) + (W_1(:,i))^T N_k(\epsilon_v(:,j)),\\
\epsilon_H^{(i,j,l)} &= (W_1(:,i) - \epsilon_w(:,i))^T H(\epsilon_v(:,j) \otimes (V_1(:,l) - \epsilon_v(:,l)) + V_1(:,j) \otimes \epsilon_v(:,l))\\
&\quad + (\epsilon_w(:,i))^T H((V(:,j) - \epsilon_v(:,j)) \otimes (V_1(:,l) - \epsilon_v(:,l))), \text{ and}\\
\epsilon_\lambda^{(i)} &= -\left(\widehat{W}(:,i)\right)^T \Gamma_v(:,i) - (\Gamma_w(:\,.i))^T \left(\widehat{V}(:,i) - \Gamma_v(:,i)\right)\\
&\quad - (W_2(:,i))^T V_2(:,i) + (\widehat{W}_2(:,i))^T \widehat{V}_2(:,i),
\end{aligned}$$

in which ϵ_v, ϵ_w, Γ_v, *and* Γ_w, *respectively, solve*

$$\epsilon_v \Lambda + \Pi A \epsilon_w = (\Pi - \Pi_v)(AV_1 + B\widetilde{B}^T), \tag{3.33a}$$

$$\epsilon_w \Lambda + (A\Pi)^T \epsilon_w = (\Pi^T - \Pi_w)(A^T W_1 + C^T \widetilde{C}), \tag{3.33b}$$

$$\Gamma_v \Lambda + \widehat{A}\Gamma_v = -(W^T V)^{-1} W^T \Bigg(\sum_{k=1}^m N_k \epsilon_v \widetilde{N}_k^T + H(\epsilon_v \otimes (V_1 + \epsilon_v) + V_1 \otimes \epsilon_v)\widetilde{H}^T \Bigg), \tag{3.33c}$$

$$\Gamma_w \Lambda + \widehat{A}^T \Gamma_w = V^T \Bigg(\sum_{k=1}^m N_k^T \epsilon_w \widetilde{N}_k + \mathcal{H}^{(2)}(\epsilon_v \otimes (W_1 + \epsilon_w), + V_1 \otimes \epsilon_w) \left(\mathcal{H}^{(2)}\right)^T \Bigg). \tag{3.33d}$$

Proof. The proof is given in Appendix C. ∎

Remark 3.10. In Theorem 3.9, we have presented measures, e.g., the distance between $\operatorname{tr}\left(CVe_i^r(e_j^p)^T\right)$ and $\operatorname{tr}\left(\widehat{C}\widehat{V}e_i^r(e_j^p)^T\right)$, denoted by $\epsilon_C^{(i,j)}$, with which the reduced-order system via TQB-IRKA satisfies the optimality conditions (3.31). But Theorem 3.9 in general does not provide a guarantee for the smallness of these distances. However, we provide intuition for the weakly nonlinear QB systems, i.e., QB systems for which $\|H\|$ and $\|N_k\|$ are small with respect to $\|B\|$ and $\|C\|$. Recall that V_1 and V_2 solve the Sylvester equations (3.28a) and (3.28c), respectively, and the right-hand side for V_2 is quadratic in H and N_k. Therefore, for a weakly nonlinear QB system, $\|V_2\|$ will be relatively small compared with $\|V_1\|$. Hence, V is expected to be close to V_1. Thus, one could anticipate that the projectors $\Pi = V(W^TV)W^T$ and $\Pi_v = V_1(W^TV_1)W^T$ will be close to each other. As a result, the right-hand side of the Sylvester equation (3.33a) will be small, and hence so is ϵ_v. In a similar way, one can argue that ϵ_w in (3.33b) will be small. Therefore, it can be shown that in the case of weakly nonlinear QB systems (1.1), all ϵ's in (3.32) such as $\epsilon_C^{(i,j)}$ should be small.

Indeed, the situation in practice proves much better. We observe in our numerical results (see section 4) that even for strongly nonlinear QB systems, i.e., $\|H\|$ and $\|N_k\|$ are comparable or even much larger than $\|B\|$ and $\|C\|$, Algorithm 3.1 still yields reduced-order systems which satisfy the optimality conditions (3.31) almost exactly with negligible perturbations.

Remark 3.11. Algorithm 3.1 can be seen as an extension of the *truncated* B-IRKA with truncation index 2 [22, Algo. 2] from bilinear systems to QB systems. In [22], the truncation index $\mathcal{N}$, which denotes the number of terms in the underlying Volterra series for bilinear systems, is free, and as $\mathcal{N} \to \infty$, all the perturbations go to zero. However, it is shown in [22] that in most cases a small $\mathcal{N}$, for example 2 or 3, is enough to satisfy all optimality conditions closely. In our case, a similar convergence will occur if we let the number of terms in the underlying Volterra series of the QB system grow; however, this is not numerically feasible since the subsystems in the QB case become rather complicated after the first three terms. Indeed, because of this, [23], [10], [1] have considered the interpolation of multivariate transfer functions corresponding to only the first two subsystems. Moreover, even in the case of balanced truncation for QB systems [13], it is shown by means of numerical examples that the truncated Gramians for QB systems based on the first three terms of the underlying Volterra series produce quantitatively accurate reduced-order systems. Our numerical examples show that this is the case here as well.

Remark 3.12. So far, in all of our discussions we have assumed that the reduced matrix $\widehat{A}$ is diagonalizable. This is a reasonable assumption since nondiagonalizable matrices lie in a set of Lebesgue measure zero. The probability of entering this set by any numerical algorithm including TQB-IRKA is zero with respect to the Lebesgue measure. Thus, TQB-IRKA can be considered safe in this regard.

Furthermore, throughout the analysis, it has been assumed that the reduced matrix $\widehat{A}$ is Hurwitz. However, in case $\widehat{A}$ is not Hurwitz, then the truncated $\mathcal{H}_2$-norm of the error system will be unbounded; thus the reduced-order systems indeed cannot be (locally) optimal. In general, the stability of $\widehat{A}$ obtained from iterative schemes to compute $\mathcal{H}_2$-suboptimal approximations is still under investigation, even for linear systems. Guaranteeing asymptotic stability for a quadratic-nonlinear system is also an open question except for the special and simple case of $A = A^*$ and A is negative definite; in this case, a Galerkin projection preserves stability. However, a simple fix to this problem is to reflect the unstable eigenvalues of $\widehat{A}$ in every step back to the left-half plane. Also, see [31] for a more involved approach to stabilize a reduced-order system.

Theorem 3.9 assumes that TQB-IRKA has converged. As stated in Remark 3.11, TQB-IRKA extends IRKA, B-IRKA, and TB-IRKA to the kind of QB systems we consider. Even for the linear case, i.e., for IRKA, convergence cannot be theoretically guaranteed despite overwhelming numerical evidence that IRKA (and (T)B-IRKA), in most cases, converge rapidly to a local minimum. Convergence of IRKA can be guaranteed theoretically only for the symmetric case [21]. Moreover, in [7] and [21], variants of IRKA with guaranteed global convergence have been introduced; however, due to the success of regular IRKA and its simple implementation, these modifications have not been as widely used. Therefore, guaranteed theoretical convergence in this iterative setting is an open issue even for the linear and bilinear cases, and naturally for the QB case as well.

Remark 3.13. As mentioned above, so far the analysis is based on the assumption that the reduced matrix $\widehat{A}$ is diagonalizable. For a reduced matrix $\widehat{A}$ with Jordan

blocks, one would need to extend the derivation of the Sylvester-equation–based $\mathcal{H}_2$ optimality conditions in [47], where Wilson [47] differentiates the $\mathcal{H}_2$ error with respect to the reduced matrix $\widehat{A}$ as opposed to individual eigenvalues $\{\lambda_i\}$ as we do here. An interpolation interpretation of the Jordan blocks in the linear case has also been established; see [45]. However, since the Jordan blocks in the optimal reduced models so far have never been observed in practice, extensions of the $\mathcal{H}_2$ theory to the bilinear case have focused on the diagonalizabilty assumption; thus, we keep the same assumption here. However, based on how the Sylvester-equation based conditions for the linear case appear, for QB systems with nondiagonalizable $\widehat{A}$, one can reasonably expect an algorithm similar to Algorithm 3.1, where the steps 6 and 7 are replaced by solving consecutively for V_1, V_2, W_1, W_2 in the following Sylvester equations:

$$\begin{aligned}
-V_1\widehat{A} - AV_1 &= B\widehat{B}^T,\\
-V_2\widehat{A} - AV_2 &= H(V_1 \otimes V_1)\widehat{H}^T + \textstyle\sum_{k=1}^{m} N_k V_1 \widehat{N}_k^T,\\
-W_1\widehat{A}^T - A^T W_1 &= C^T\widehat{C},\\
-W_2\widehat{A}^T - A^T W_2 &= 2\cdot\mathcal{H}^{(2)}(V_1 \otimes W_1)(\widehat{H}^{(2)})^T + \textstyle\sum_{k=1}^{m} N_k^T W_1 \widehat{N}_k.
\end{aligned}$$

Note that with this formulation, $\widehat{A}$ enters into the algorithm directly without diagonalization. However, due to the reasons listed before, we leave this theoretical development for future work.

Remark 3.14. Thus far, we have used $E = I$ in front of $\dot{x}(t)$ in the QB system (1.1); however, in the case of $E \neq I$, but nonetheless being nonsingular, we can still employ Algorithm 3.1. One obvious way is to invert E, but this is inadmissible in the large-scale setting. Moreover, the resulting matrices may be dense, making the algorithm computationally expensive. Nevertheless, Algorithm 3.1 can be employed without inverting E. For this, we need to modify steps 6 and 7 in Algorithm 3.1 as follows:

$$\begin{aligned}
-EV_1\Lambda - AV_1 &= B\widetilde{B}^T,\\
-EV_2\Lambda - AV_2 &= H(V_1 \otimes V_1)\widetilde{H}^T + \textstyle\sum_{k=1}^{m} N_k V_1 \widetilde{N}_k^T,\\
-E^T W_1\Lambda - A^T W_1 &= C^T\widetilde{C},\\
-E^T W_2\Lambda - A^T W_2 &= 2\cdot\mathcal{H}^{(2)}(V_1 \otimes W_1)(\widetilde{\mathcal{H}}^{(2)})^T + \textstyle\sum_{k=1}^{m} N_k^T W_1 \widetilde{N}_k,
\end{aligned}$$

and replace $(W^TV)^{-1}$ with $(W^TEV)^{-1}$, assuming W^TEV is invertible while determining the reduced-order system matrices in step 9 of Algorithm 3.1. Then, the modified iterative algorithm with the matrix E also provides a reduced-order system, *approximately* satisfying optimality conditions subject to the structure of a reduced-order system as in (1.2) and the matrix $\widehat{A}$ to be diagonalizable. We skip the rigorous proof for the $E \neq I$ case, but it can be proven along the lines of $E = I$. Indeed, under the assumption of W^TEV being invertible, one does not even need to invert W^TEV by letting the reduced QB system have a reduced E term as W^TEV, and the spectral decomposition of $\widehat{A}$ is replaced by a generalized eigenvalue decomposition of $\widehat{E}$ and $\widehat{A}$. But to keep the notation of Algorithm 3.1 simple, we omit these details.

As noted, we assume that the reduced-order system has the structure as in (1.2). However, one can consider a general matrix $\widehat{E}$ including being singular in front of $\dot{\widehat{x}}(t)$ in the reduced-order system (1.2) and derive the Wilson-type conditions by taking

Algorithm 3.2. Computation of the Hessian of the reduced QB system [10].

1: Determine $\mathcal{Y} \in \mathbb{R}^{r\times n\times n}$, such that $\mathcal{Y}^{(1)} = W^T H$.
2: Determine $\mathcal{Z} \in \mathbb{R}^{r\times r\times n}$, such that $\mathcal{Z}^{(2)} = V^T \mathcal{Y}^{(2)}$.
3: Determine $\mathcal{X} \in \mathbb{R}^{r\times r\times r}$, such that $\mathcal{X}^{(3)} = V^T \mathcal{Z}^{(3)}$.
4: Then, the reduced Hessian is $\widehat{H} = \mathcal{X}^{(1)}$.

derivatives of the truncated $\mathcal{H}_2$-norm with respect to the reduced matrices such as $\widehat{E}$, $\widehat{A}$, etc. This is worth investigating problem for future work.

3.4. Computational issues. The main bottleneck in applying TQB-IRKA is the computation of the reduced matrices, especially the computational cost related to $\widehat{H} := W^T H(V \otimes V)$ that needs to be evaluated at each iteration. Regarding this, there is an efficient method, proposed in [10], utilizing the properties of tensor matricizations, which we summarize in Algorithm 3.2.

Algorithm 3.2 avoids the highly undesirable explicit formulation of $V \otimes V$ for large-scale systems to compute the reduced Hessian, and the algorithm does not rely on any particular structure of the Hessian. However, a QB system resulting from semi-discretization of PDEs usually leads to a Hessian which has a particular structure related to that particular PDE and the choice of the discretization method.

Therefore, we propose another efficient way to compute $\widehat{H}$ that utilizes a particular sparsity structure of the Hessian, arising from the governing PDEs or ODEs. Generally, the term $H(x \otimes x)$ in the QB system (1.1) can be written as

$$H(x \otimes x) = \sum_{j=1}^{p} (\mathcal{A}^{(j)} x) \circ (\mathcal{B}^{(j)} x),$$

where $\circ$ denotes the Hadamard product; $\mathcal{A}^{(j)}$ and $\mathcal{B}^{(j)}$ are sparse matrices, depending on the nonlinear operators in the underlying PDE and the discretization scheme; and p is generally a very small integer; for instance, it is equal to 1 in case of Burgers' equations. Furthermore, using the ith rows of $\mathcal{A}^{(j)}$ and $\mathcal{B}^{(j)}$, we can construct the ith row of the Hessian:

$$H(i,:) = \sum\nolimits_{j=1}^{p} \mathcal{A}^{(j)}(i,:) \otimes \mathcal{B}^{(j)}(i,:),$$

where $H(i,:)$, $\mathcal{A}^{(j)}(i,:)$, and $\mathcal{B}^{(j)}(i,:)$ represent the ith rows of the matrices H, $\mathcal{A}^{(j)}$, and $\mathcal{B}^{(j)}$, respectively. This clearly shows that there is a particular Kronecker structure of the Hessian H, which can be used in order to determine $\widehat{H}$. Using the Chafee–Infante equation as an example, we illustrate how the structure of the Hessian (Kronecker product structure) can be exploited to determine $\widehat{H}$ efficiently.

Example 3.15. Here, we consider the Chafee–Infante equation, which is discretized over the spatial domain via a finite difference scheme. The MOR problem for this example is considered in subsection 4.1, where one can also find the governing equations and boundary conditions. For this particular example, the Hessian (after having rewritten the system into the QB form) is given by

$$\begin{aligned} H(i,:) &= -\frac{1}{2} e_i^n \otimes e_{k+i}^n - \frac{1}{2} e_{k+i}^n \otimes e_i^n, \quad i \in \{1,\dots,k\}, \\ H(i,:) &= -2(e_i^n \otimes e_i^n) + e_{i-k}^n \otimes \begin{bmatrix} X(i-k,:) & \mathbf{0} \end{bmatrix} + \begin{bmatrix} X(i-k,:) & \mathbf{0} \end{bmatrix} \otimes e_{i-k}^n, \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i \in \{k+1,\dots,n\}, \end{aligned}$$

Algorithm 3.3. Computation of the reduced Hessian for Chafee–Infante example.

Input: $V, W \in \mathbb{R}^{2k\times r}, X \in \mathbb{R}^{k\times k}$(as defined in (3.34))
Compute $V_x := XV(1{:}k,:)$, where $V(1{:}k,:)$ denotes the first k row vectors of V.
for $i = 1 : k$ **do**

$$H_v(i,:) = -\frac{1}{2}V(i,:)\otimes V(k+i,:) - \frac{1}{2}V(i,:)\otimes V(k+i,:),$$
$$H_v(k+i,:) = -2\left(V(i,:)\otimes V(i,:)\right) + V(i,:)\otimes V_x(i,:) + V_x(i,:)\otimes V(i,:),$$

where $H_v(q,:)$ is the qth row vector of H_v and the same holds for other matrices.
end for
Then, the reduced Hessian is $\widehat{H} = W^T H_v$.

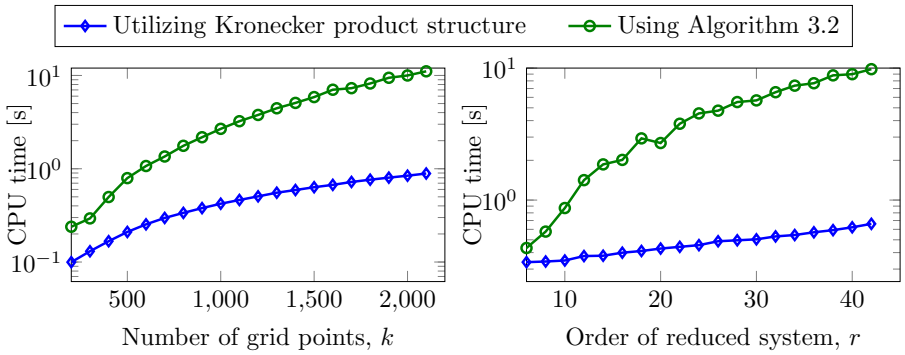

FIG. 3.1. *The left figure shows the computational time for* $\widehat{H} := W^T H(V\otimes V)$ *by varying the number of grid points in the spatial domain by fixing the order of the reduced-order system to* $r = 20$. *In the right figure, we show the computational time for different orders of the reduced-order system using a fixed number of grid points,* $k = 1000$.

where k is the number of grid points, $n = 2k$, and $H(i,:)$ is the ith row vector of the matrix H. $X(i,:)$ also denotes the ith row vector of the matrix $X \in \mathbb{R}^{k\times k}$

$$X = \begin{bmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix}. \tag{3.34}$$

The Kronecker product representation of each row of the matrix H allows us to compute the rows of $H_v := H(V\otimes V)$ by selecting only the required rows of V. This way, we can determine H_v efficiently in large-scale, sparse settings, and then multiply with W^T to obtain the desired reduced Hessian. We describe this procedure in Algorithm 3.3 that shows how one can determine the reduced Hessian for the Chafee–Infante example.

In order to show the effectiveness of the proposed methodology that uses the special Kronecker product structure of the Hessian H, we compute $\widehat{H} = W^T H(V\otimes V)$ for different orders of original and reduced-order systems and show the required CPU-time to compute it in Figure 3.1. The simulations were performed on a board with 4 Intel® Xeon® E7-8837 CPUs with a 2.67-GHz clock speed using MATLAB 8.0.0.783 (R2012b).

Figure 3.1 illustrates that the computational cost for constructing the reduced Hessian by using the proposed method, which exploits the Kronecker product struc-

ture of the Hessian H, grows much slower than the cost in Algorithm 3.2. Therefore, we conclude here that it is worth exploiting the Kronecker product structure of the Hessian of the system for an efficient computation of $\widehat{H}$ in large-scale settings.

4. Numerical results. In this section, we illustrate the behavior of the proposed model reduction method TQB-IRKA for QB systems, using several semidiscretized nonlinear PDEs and compare its performance with existing MOR techniques, such as one-sided and two-sided subsystem-based interpolatory projection methods [10, 23, 38], BT for QB systems [13], and POD [28, 33], in terms of the accuracy of the time-domain performance and the truncated $\mathcal{H}_2$-norm. We iterate Algorithm 3.1 until the relative change in the eigenvalues of $\widehat{A}$ becomes smaller than a given tolerance, which we set to 10^{-5}. Moreover, we determine the interpolation points for the one-sided and two-sided interpolatory projection methods applying IRKA [24] to the corresponding linear part, which appear to be a good set of interpolation points as shown in [10]. All the simulations were done on a board with 4 Intel® Xeon® E7-8837 CPUs with a 2.67-GHz clock speed using MATLAB 8.0.0.783 (R2012b). Some more details related to the numerical examples are as follows.

1. For all time-domain simulations, the original and reduced-order systems are integrated by the routine `ode15s` in MATLAB with a relative error tolerance of 10^{-8} and an absolute error tolerance of 10^{-10}.
2. We measure the output at 500 equidistant points within the time interval $[0, T]$, where T is defined in each numerical example.
3. In order to employ BT, we need to solve four standard Lyapunov equations. For this, we use `mess_lyap.m` from M.E.S.S.-1.0.1 [42] which is based on one of the latest ADI methods proposed in [14].
4. We initialize TQB-IRKA (Algorithm 3.1) by choosing an arbitrary reduced system using the `rand` command in MATLAB, while ensuring $\widehat{A}$ is Hurwitz and diagonalizable.
5. Since POD can be applied to a general nonlinear system, we apply POD to the original nonlinear system, without transforming it into a QB system as we observe that this way, POD yields better reduced systems.
6. One of the aims of the numerical examples is to determine the residuals in Theorem 3.9. For this, we first define $\Phi_C^e \in \mathbb{R}^{r\times p}$, $\Phi_B^e \in \mathbb{R}^{r\times m}$, $\Phi_N^e \in \mathbb{R}^{r\times r\times m}$, $\Phi_H^e \in \mathbb{R}^{r\times r\times r}$, and $\Phi_\Lambda^e \in \mathbb{R}^r$ such that $\epsilon_C^{(i,j)}$ is the (i,j)th entry of Φ_C^e, $\epsilon_B^{(i,j)}$ is the (i,j)th entry of Φ_B^e, $\epsilon_N^{(i,j,k)}$ is the (i,j,k)th entry of Φ_N^e, $\epsilon_H^{(i,j,k)}$ is the (i,j,k)th entry of Φ_H^e, and $\epsilon_\Lambda^{(i)}$is ith entry of Φ_Λ^e.
   Furthermore, we define Φ_C, Φ_B, Φ_N, Φ_H, and Φ_Λ to be the terms on the left-hand side of (3.32a)–(3.32e) in Theorem 3.9, e.g., the (i,j)th entry of Φ_C is tr $(CVe_i^r(e_j^p)^T)$. As a result, we define relative perturbation measures as follows:
   (4.1)
$$\mathcal{E}_C = \frac{\|\Phi_C^e\|_2}{\|\Phi_C\|_2},\ \mathcal{E}_B = \frac{\|\Phi_B^e\|_2}{\|\Phi_B\|_2},\ \mathcal{E}_N = \frac{\|\Phi_N^{e(1)}\|_2}{\|\Phi_N^{(1)}\|_2},\ \mathcal{E}_H = \frac{\|\Phi_H^{e(1)}\|_2}{\|\Phi_H^{(1)}\|_2},\ \mathcal{E}_\Lambda = \frac{\|\Phi_\Lambda^e\|_2}{\|\Phi_\Lambda\|_2},$$
   where $\Phi_{\{N,H\}}^{(1)}$ and $\Phi_{\{N,H\}}^{e(1)}$ are mode-1 matricizations of the tensors $\Phi_{\{N,H\}}$ and $\Phi_{\{N,H\}}^e$, respectively.
7. We also address a numerical issue which one might face while employing Algorithm 3.1. In step 8 of Algorithm 3.1, we need to take the sum of the two matrices V_1 and V_2. If $\|H\|$ and $\|N_k\|$ are too large, then the norm of

V_2 can be much larger than that of V_1. Thus, a direct sum might reduce the effect of V_1. As a remedy we propose using a scaling factor γ for H and N_k, resulting in matrices V_1 and V_2 such that $\frac{\|V_2\|}{\|V_1\|} \in \mathcal{O}\left(10^0 - 10^2\right)$. We have already noted in Remark 3.3 that this scaling just scales the input-output mapping. Once again we emphasize that we just compute the model reduction bases V and W using the scaled system, but we project the original, unscaled system to construct the reduced-order system.

4.1. One–dimensional Chafee–Infante equation. Here, we consider the one-dimensional Chafee–Infante (Allen–Cahn) equation whose governing equation, initial condition, and boundary controls are given by

$$\begin{aligned} \dot{v} + v^3 &= v_{xx} + v, & (0,L)\times(0,T), \qquad & v(0,\cdot) = u(t), & (0,T),\\ v_x(L,\cdot) &= 0, & (0,T), \qquad & v(x,0) = 0, & (0,L). \end{aligned} \tag{4.2}$$

MOR for this system has been considered in various articles; see, e.g., [10, 13]. The governing equation (4.2) contains a cubic nonlinearity, which can then be rewritten into QB form as shown in [10]. For more details on the system, we refer to [18, 26]. Next, we utilize a finite difference scheme by using k equidistant points over the length, resulting in a semidiscretized QB system of order $2k$. The output of our interest is the response at the right boundary, i.e., $v(L,t)$, and we set the number of grid points to $k = 500$, leading to an order $n = 1000$ QB system.

We construct reduced-order systems of order $r = 10$ using TQB-IRKA, BT, one-sided and two-sided interpolatory projection methods, and POD. Having initialized TQB-IRKA randomly, it takes nine iterations to converge, and for this example we choose the scaling factor $\gamma = 10^{-3}$. We compute the reduced Hessian as shown in Algorithm 3.3. For the POD-based approximation, we collect 500 snapshots of the true solution for the training input $u^{(1)}(t) = (1 + \sin(\pi t))\exp(-t/5)$ and compute the projection by taking the 10 dominant basis vectors.

In order to compare the quality of these reduced-order systems with respect to the original system, we first simulate them using the same training input used to construct the POD basis, i.e., $u^{(1)}(t) = (1 + \sin(\pi t))\exp(-t/5)$. We plot the transient responses and relative output errors for this input in Figure 4.1. As expected, since we are comparing the reduced models for the same forcing term used to train POD, Figure 4.1 shows that the POD approximation outperforms the other methods for the input $u^{(1)}$. However, the interpolatory methods also provide adequate reduced-order systems for $u^{(1)}$ even though the reduction is performed without any knowledge of $u^{(1)}(t)$.

To test the robustness of the reduced systems, we compare the time-domain simulations of the reduced systems with the original one in Figure 4.2 for a slightly different input, namely $u^{(2)}(t) = 25\,(1 + \sin(\pi t))$. First, observe that the POD approximation fails to reproduce the system's dynamics for the input $u^{(2)}(t)$ accurately as POD is input-dependent. Moreover, the one-sided interpolatory projection method also performs worse for the input $u^{(2)}(t)$. On the other hand, TQB-IRKA, BT, and the two-sided interpolatory projection method all yield very accurate reduced-order systems of comparable qualities; TQB-IRKA produces marginally better reduced systems. Once again it is important to emphasize that neither $u^{(1)}(t)$ nor $u^{(2)}(t)$ has entered the model reduction procedure in TQB-IRKA. To give a quantitative comparison of the reduced systems for both inputs, $u^{(1)}(t)$ and $u^{(2)}$, we report the mean relative errors in Table 4.1 as well, which also provides us a similar information.

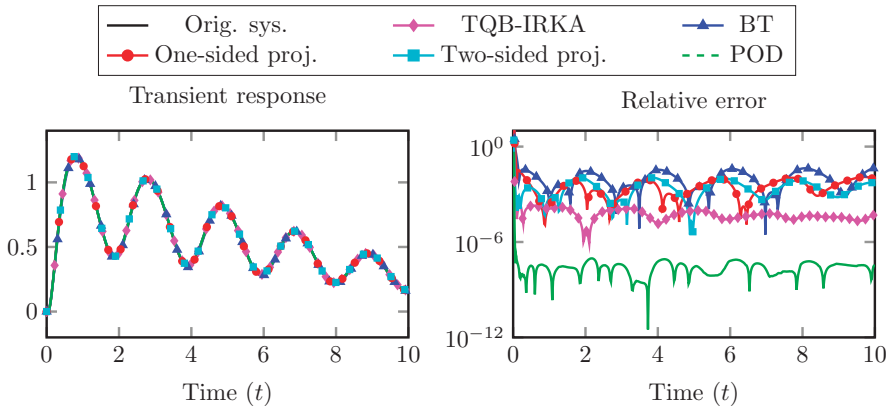

FIG. 4.1. *Chafee–Infante: comparison of responses for the boundary control input* $u^{(1)}(t) = (1 + \sin(\pi t)) \exp(-t/5)$.

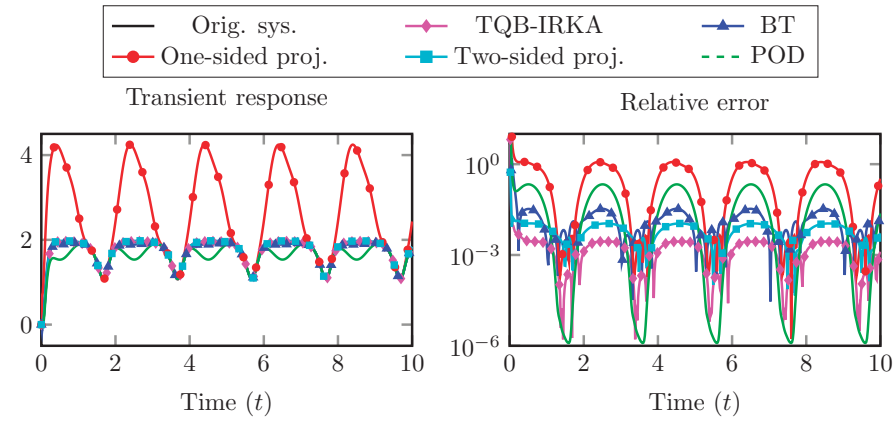

FIG. 4.2. *Chafee–Infante: comparison of responses for the boundary control input* $u^{(2)}(t) = 25(1 + \sin(\pi t))$.

TABLE 4.1
Chafee–Infante: The mean relative errors of the output.

Input	TQB-IRKA	BT	One-sided	Two-sided	POD
$u^{(1)}(t)$	$6.54 \cdot 10^{-5}$	$1.40 \cdot 10^{-2}$	$4.30 \cdot 10^{-3}$	$3.51 \cdot 10^{-3}$	$2.87 \cdot 10^{-8}$
$u^{(2)}(t)$	$1.63 \cdot 10^{-3}$	$1.43 \cdot 10^{-2}$	$4.59 \cdot 10^{-1}$	$6.65 \cdot 10^{-3}$	$6.70 \cdot 10^{-2}$

Furthermore, we study the impact of the scaling factor γ, as discussed in Remark 3.3, on the performance reduced-order systems obtained via TQB-IRKA. For the same inputs $u^{(i)}$, $i \in \{1, 2\}$, we plot the relative errors in the time-domain responses for different values of the scaling factor in Figure 4.3. For this example, we observe that for $\gamma = 10^{-3}$, TQB-IRKA produces a slightly better reduced-order system in terms of the accuracy of the time-domain simulations than for all other tested values of γ; however, all scaling factors $\gamma \in \{10^0, 10^{-1}, \ldots, 10^{-4}\}$ produce comparable reduced-order systems. For very small values of γ such as $\gamma = \{10^{-5}, 10^{-6}\}$, TQB-IRKA yields very poor reduced-order systems. This is expected since by choosing

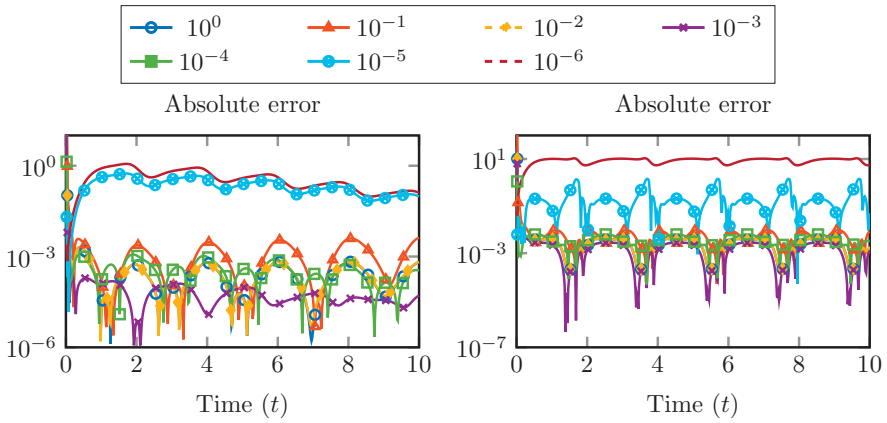

FIG. 4.3. *Chafee–Infante: absolute error between the original and reduced-order systems ($r = 10$) obtained by using TQB-IRKA for different scaling factors γ for inputs $u^{(1)}$ and $u^{(2)}$.*

TABLE 4.2
Chafee–Infante: Perturbations to the optimality conditions.

Method	$\mathcal{E}_C$	$\mathcal{E}_B$	$\mathcal{E}_N$	$\mathcal{E}_H$	$\mathcal{E}_\lambda$
TQB-IRKA	$2.64 \cdot 10^{-8}$	$4.75 \cdot 10^{-12}$	$1.24 \cdot 10^{-17}$	$2.40 \cdot 10^{-12}$	$7.62 \cdot 10^{-12}$
BT	$6.60 \cdot 10^{-4}$	$6.23 \cdot 10^{-5}$	$8.62 \cdot 10^{-16}$	$1.24 \cdot 10^{-11}$	$7.62 \cdot 10^{-4}$
One–Sided	$2.43 \cdot 10^{1}$	$6.90 \cdot 10^{-3}$	$1.09 \cdot 10^{-10}$	$1.02 \cdot 10^{-5}$	$8.40 \cdot 10^{-3}$
Two–Sided	$1.99 \cdot 10^{-5}$	$2.38 \cdot 10^{-7}$	$1.45 \cdot 10^{-16}$	$1.25 \cdot 10^{-11}$	$1.71 \cdot 10^{-4}$

a very small scaling factor, the effect of the quadratic and bilinear terms is reduced significantly and the model reduction basis matrices almost correspond to the linear term only; hence, poor reduced-order systems result. We have observed that if a scaling factor is chosen such that $\frac{\|V_2\|}{\|V_1\|} \approx \mathcal{O}\left(10^0\text{–}10^2\right)$, then TQB-IRKA not only provides a better reduced-order system but also converges faster, although we do not have a theoretical justification for this observation yet. Therefore, as future work, it would be interesting to investigate the influence of the scaling factor γ on the quality of the obtained reduced-order systems also from a theoretical point of view.

In Theorem 3.9, we presented the quantities, denoted by ϵ_C, ϵ_B, ϵ_λ, ϵ_N, and ϵ_H, which measure how far a reduced-order system is from satisfying the optimality conditions (3.31). We list these quantities for the reduced-order systems obtained via TQB-IRKA, BT, and the one-/two-sided interpolatory projection methods. These quantities are computed as in (4.1), and are listed in Table 4.2, showing that the reduced-order system obtained by TQB-IRKA satisfies the optimality conditions best among all the considered methods.

In Remark 3.10, we argued that for a weakly nonlinear QB system, we expect these quantities to be small. However, even for this example with strong nonlinearity, i.e., $\|H\|$ and $\|N_k\|$ are not small at all, the reduced-order system computed by TQB-IRKA satisfies the optimality conditions (3.31) very accurately. This result also strongly supports the discussion of Remark 3.11 that a small truncation index is expected to be enough in many cases.

Furthermore, since TQB-IRKA approximately minimizes the truncated $\mathcal{H}_2$-norm of the error system, i.e., $\|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2^{(\mathcal{T})}}$, we also compare the truncated $\mathcal{H}_2$-norm of

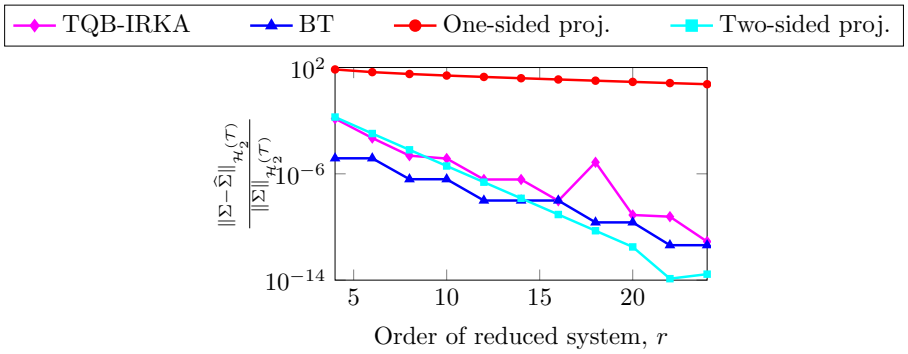

FIG. 4.4. *Chafee–Infante: comparison of the truncated $\mathcal{H}_2$-norm of the error system, having obtained reduced systems of different orders via different methods.*

the error system in Figure 4.4, where the reduced model $\widehat{\Sigma}$ is constructed by various methods of different orders. As mentioned before, the reduced-order systems obtained via POD preserve the structure of the original nonlinearities; therefore, the truncated $\mathcal{H}_2$-norm definition, given in Lemma 3.4, does not apply.

Figure 4.4 indicates that the reduced-order systems obtained via one-sided interpolatory projection perform worst in the truncated $\mathcal{H}_2$-norm measure. Moreover, while BT performs better than TQB-IRKA and the two-sided interpolatory method for small reduced orders with respect to the truncated $\mathcal{H}_2$-norm, for higher reduced orders, the two-sided interpolatory method yields the best reduced systems. However, it is important to emphasize that unlike in the case of linear dynamical systems, the $\mathcal{H}_2$-norm and the L^∞-norm of the output for nonlinear systems, including QB systems, are not as strongly connected as in the linear case. This can be seen in Figure 4.4; for reduced order $r = 10$, even though BT yields the smallest truncated $\mathcal{H}_2$ error, in the time-domain simulations for inputs $u^{(1)}$ and $u^{(2)}$, it is not the best in terms of the L^∞-norm of the output. Furthermore, it can also be noted that for $r = 10$, TQB-IRKA yields a reduced-order system, which satisfies the optimality conditions most accurately (see Table 4.2), but in the truncated $\mathcal{H}_2$-norm, it does not perform the best, as illustrated in Figure 4.4.

Nevertheless, we believe that the truncated $\mathcal{H}_2$-norm of the error system is a robust indicator for the quality of the reduced system, because this norm is defined by the *kernels*, which define the mapping from the input to the output. Thus, if the kernels are ensured to be close enough, then one can expect an accurate approximation of the output.

4.2. Nonlinear RC ladder. We consider a nonlinear RC ladder, which consists of capacitors and nonlinear I–V diodes. The characteristics of the I–V diodes are governed by exponential nonlinearities, which can also be rewritten in the QB form. For a detailed description of the dynamics of this electronic circuit, we refer to [4, 23, 35, 37, 38]. We set the number of capacitors in the ladder to $k = 500$, resulting in a QB system of order $n = 1000$. Note that the matrix A of the resulting QB system has eigenvalues at zero; therefore, the truncated $\mathcal{H}_2$-norm may not exist. Moreover, BT also cannot be employed as we need to solve Lyapunov equations that require a stable A matrix. Thus, we shift the matrix A to $A_s := A - 0.01I_n$ to determine the projection matrices for TQB-IRKA and BT, but we project the original system matrices.

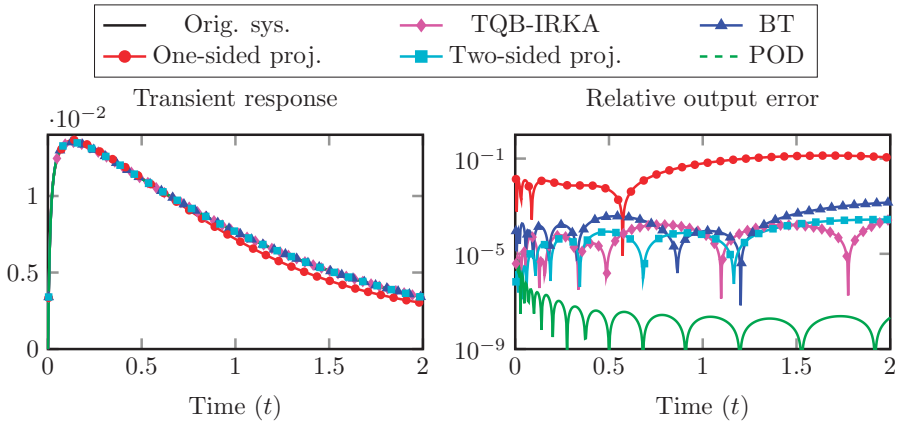

FIG. 4.5. *RC circuit: comparison of responses for the input* $u^{(1)}(t) = e^{-t}$.

TABLE 4.3
RC circuit: The mean absolute errors of the output.

Input	TQB-IRKA	BT	One-sided	Two-sided	POD
$u^{(1)}(t)$	$8.82 \cdot 10^{-5}$	$3.67 \cdot 10^{-4}$	$6.50 \cdot 10^{-2}$	$1.01 \cdot 10^{-4}$	$7.24 \cdot 10^{-8}$
$u^{(2)}(t)$	$1.12 \cdot 10^{-3}$	$2.15 \cdot 10^{-3}$	$2.32 \cdot 10^{-1}$	$7.80 \cdot 10^{-4}$	$7.8 \cdot 10^{-3}$

We construct reduced-order systems of order $r = 10$ using all five different methods. In this example as well, we initialize TQB-IRKA randomly and it converges after 27 iterations. We choose the scaling factor $\gamma = 0.01$. For this example, we determine the reduced Hessian by exploiting the particular structure of the Hessian. In order to compute a reduced-order system via POD, we first obtain 500 snapshots of the true solution for the training input $u^{(1)}(t) = e^{-t}$ and then use the 10 dominant modes to determine the projection.

We first compare the accuracy of these reduced systems for the same training input $u^{(1)}(t) = e^{-t}$, which is also used to compute the POD basis. Figure 4.5 shows the transient responses and relative errors of the output for the input $u^{(1)}$. As one would expect, POD outperforms all other methods since the control input $u^{(1)}$ is the same as the training input for POD. Nonetheless, TQB-IRKA, BT, and two-sided interpolatory projection also yield very good reduced-order systems, considering they are obtained without any prior knowledge of the input.

We also test the reduced-order systems for an input different from the training input, precisely, $u^{(2)}(t) = 2.5\,(\sin(\pi t/5) + 1)$. Figure 4.6 shows the transient responses and relative errors of the output for the input $u^{(2)}$. We observe that POD does perform almost as well as the other methods, such as TQB-IRKA, BT, and two-sided interpolatory projection methods even for this input, and the one-sided interpolatory projection method completely fails to capture the system dynamics for the input $u^{(2)}$. This can also be observed from Table 4.3, where the mean relative errors of the outputs are reported.

Further, we compute the quantities as defined in (4.1) using the reduced system of order $r = 10$ obtained from the various investigated methods, and list them in Table 4.4. As in the previous example, the reduced-order system obtained using TQB-IRKA satisfies the optimality conditions (3.31) most accurately among all the considered methods.

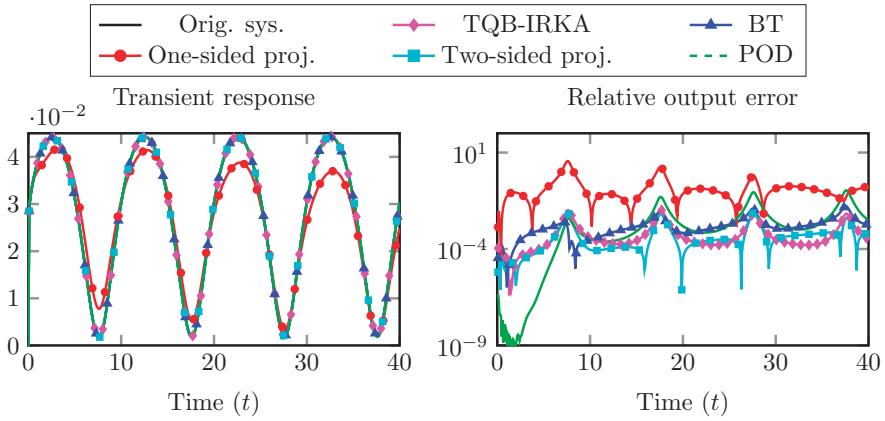

FIG. 4.6. *RC circuit: comparison of responses for the input* $u^{(2)}(t) = 2.5\,(\sin(\pi t/5) + 1)$.

TABLE 4.4
RC circuit: Perturbations to the optimality conditions.

Method	$\mathcal{E}_C$	$\mathcal{E}_B$	$\mathcal{E}_N$	$\mathcal{E}_H$	$\mathcal{E}\lambda$
TQB-IRKA	$2.50 \cdot 10^{-8}$	$3.88 \cdot 10^{-6}$	$3.92 \cdot 10^{-7}$	$3.37 \cdot 10^{-8}$	$3.91 \cdot 10^{-8}$
BT	$5.32 \cdot 10^{-6}$	$2.95 \cdot 10^{-5}$	$6.39 \cdot 10^{-6}$	$3.23 \cdot 10^{-6}$	$1.36 \cdot 10^{-5}$
One-sided	$1.00 \cdot 10^{-2}$	$4.40 \cdot 10^{-2}$	$2.98 \cdot 10^{-2}$	$1.43 \cdot 10^{-2}$	$4.26 \cdot 10^{-2}$
Two-sided	$6.29 \cdot 10^{-4}$	$2.30 \cdot 10^{-3}$	$7.33 \cdot 10^{-4}$	$5.00 \cdot 10^{-4}$	$3.57 \cdot 10^{-4}$

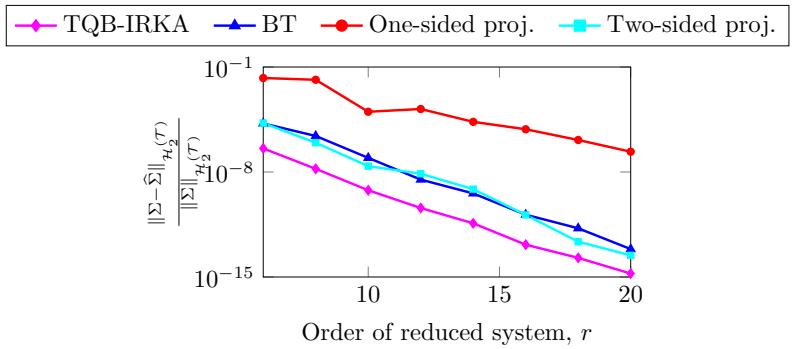

FIG. 4.7. *RC circuit: comparison of the truncated* $\mathcal{H}_2$*-norm of the error system obtained via different methods of various orders.*

Next, we also compare the truncated $\mathcal{H}_2$-norm of the error system, i.e., $\|\Sigma - \widehat{\Sigma}\|_{\mathcal{H}_2^{(\mathcal{T})}}$, in Figure 4.7, where the reduced models are constructed by various methods of different orders. The figure shows that TQB-IRKA yields the best reduced-order systems with respect to the truncated $\mathcal{H}_2$-norm among the investigated methods as well.

Note that we apply POD to the original system with exponential nonlinearities; therefore, we cannot compute the truncated $\mathcal{H}_2$-norm defined in Lemma 3.4 for the POD approximation. Hence, POD is omitted in Figure 4.7.

4.3. The FitzHugh–Nagumo (F–N) system. This example considers the F-N system, describing activation and deactivation dynamics of spiking neurons. This model is a simplification of the Hodgkin–Huxley neuron model. The dynamics of the system are governed by the following nonlinear coupled PDEs:

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

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

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

where $\epsilon = 0.015$, $h = 0.5$, $\gamma = 2$, $q = 0.05$, $L = 0.3$, and $i_0(t)$ is an actuator, acting as a control input. The voltage and recovery voltage are denoted by v and w, respectively. MOR for this model has been considered in [9, 13, 19]. Furthermore, we also consider the same output as considered in [9, 13], which is the limit-cycle at the left boundary, i.e., $x = 0$. The system can be considered as having two inputs, namely q and $i_0(t)$; it has also two outputs, which are $v(0,t)$ and $w(0,t)$. This means that the system is a multi-input multi-output system (MIMO) as opposed to the two previous examples. We discretize the governing equations using a finite difference scheme. This leads to an ODE system, having cubic nonlinearity, which can then be transformed into the QB form. We consider $k = 300$ grid points, resulting in a QB system of order $3k = 900$.

We next determine reduced systems of order $r = 35$ using TQB-IRKA, BT, and POD. We choose the scaling factor $\gamma = 1$ in TQB-IRKA, and it requires 26 iterations to converge. For this example, we also utilize the Kronecker product structure of the Hessian to perform an efficient computation of the reduced Hessian. In order to apply POD, we first collect 500 snapshots of the original system for the time interval $(0, 10]$ using $i_0(t) = 50(\sin(2\pi t) - 1)$ and then determine the projection based on the 35 dominant modes. The one-sided and two-sided subsystem-based interpolatory projection methods have major disadvantages in the MIMO QB case. The one-sided interpolatory projection approach of [23] can be applied to MIMO QB systems; however, the dimension of the subspace V, and thus the dimension of the reduced model, increases quadratically due to the $V \otimes V$ term. As we mentioned in section 1, two-sided interpolatory projection is only applicable to SISO QB systems. When the number of inputs and outputs are the same, which is the case in this example, one can still employ [10, Alg. 1] to construct a reduced system. This is exactly what we did here. However, it is important to note that even though the method can be applied numerically, it no longer ensures the theoretical subsystem interpolation property. Despite these drawbacks, for completeness of the comparison, we still construct reduced models using both one-sided and two-sided subsystem-based interpolatory projections.

Since the F-N system has two inputs and two outputs, each interpolation point yields 6 columns of the projection matrices V and W. Thus, to apply the two-sided projection, we use 6 linear $\mathcal{H}_2$-optimal points and determine the reduced system of order 35 by taking the 35 dominant vectors. We do the same for the one-sided interpolatory projection method to compute the reduced-order system.

Next, we compare the quality of the reduced-order systems and plot the transient responses and the absolute errors of the outputs in Figure 4.8 for the training input $i_0(t) = 50(\sin(2\pi t) - 1)$.

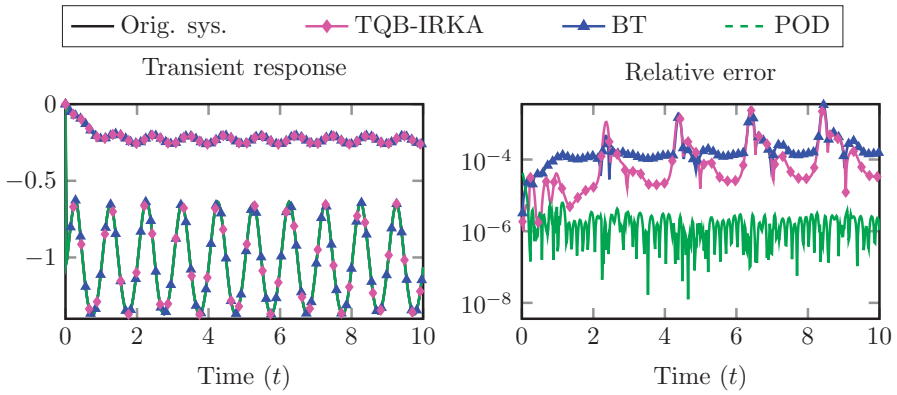

FIG. 4.8. *The FitzHugh–Nagumo system: comparison of the limit-cycle at the left boundary,* $x = 0$ *for* $i_0(t) = 50(\sin(2\pi t) - 1)$.

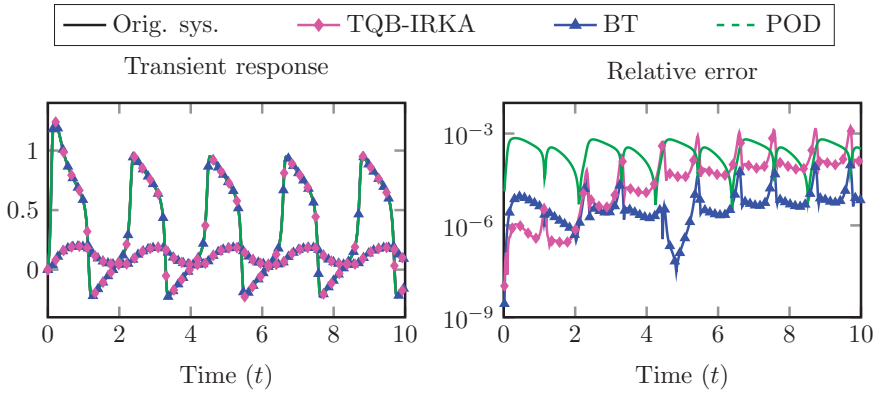

FIG. 4.9. *The FitzHugh–Nagumo system: comparison of the limit-cycle at the left boundary,* $x = 0$ *for* $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$.

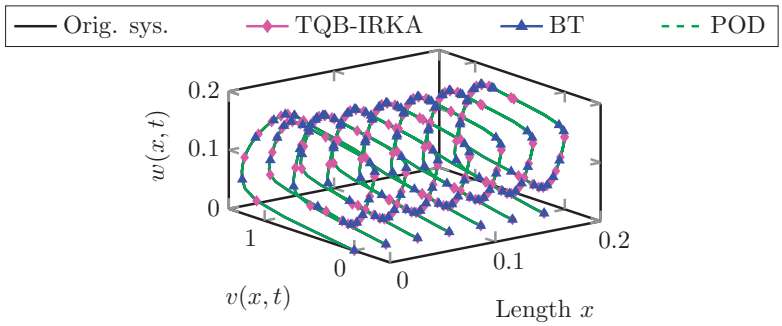

FIG. 4.10. *The FitzHugh–Nagumo system: limit-cycle behavior of the original and reduced-order systems in the spatial domain.*

As anticipated, POD provides a very good reduced-order system since the POD basis is constructed by using the same trajectory. Note that despite not reporting CPU times for the offline phases in this paper, due to the very different levels of the

TABLE 4.5
The FitzHugh–Nagumo system: Perturbations to the optimality conditions.

Method	$\mathcal{E}_C$	$\mathcal{E}_B$	$\mathcal{E}_N$	$\mathcal{E}_H$	$\mathcal{E}_\lambda$
TQB-IRKA	$8.76 \cdot 10^{-8}$	$7.35 \cdot 10^{-9}$	$1.78 \cdot 10^{-11}$	$4.27 \cdot 10^{-9}$	$9.14 \cdot 10^{-10}$
BT	$1.10 \cdot 10^{-6}$	$1.73 \cdot 10^{-8}$	$3.36 \cdot 10^{-9}$	$3.56 \cdot 10^{-5}$	$5.49 \cdot 10^{-9}$
One-sided	$4.10 \cdot 10^{-3}$	$7.40 \cdot 10^{-3}$	$9.26 \cdot 10^{-17}$	$5.10 \cdot 10^{-3}$	$6.36 \cdot 10^{-4}$
Two-sided	$1.62 \cdot 10^{3}$	$4.10 \cdot 10^{-3}$	$1.65 \cdot 10^{-10}$	$1.94 \cdot 10^{9}$	$2.57 \cdot 10^{0}$

implementations used for the various methods, we would like to mention that in this example the construction of the POD basis with the fairly sophisticated MATLAB integrator `ode15s` takes roughly 1.5 more CPU time than constructing the TQB-IRKA reduced-order model with our vanilla implementation.

Comparing TQB-IRKA and BT, TQB-IRKA gives a marginally better reduced-order system as compared with BT for $i_0(t) = 50(\sin(2\pi t) - 1)$, but still both are very competitive. In contrast, the one-sided and two-sided interpolatory projection methods produce unstable reduced-order systems and are therefore omitted from the figures.

To test the robustness of the obtained reduced-order systems, we choose a different control input, $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$, and compare the transient responses in Figure 4.9. In this figure, we observe that BT performs the best among all methods for $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$, and POD and TQB-IRKA produce reduced-order systems of almost the same quality. One-sided and two-sided projections result in unstable reduced-order systems for $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$ as well. Furthermore, we also show the limit-cycles on the full space obtained from the original and reduced-order systems in Figure 4.10 for $i_0(t) = 5 \cdot 10^4 t^3 \exp(-15t)$, and observe that the reduced-order systems obtained using POD, TQB-IRKA, and BT enable us to reproduce the limit-cycles, which is a typical neuronal dynamics, as shown in Figures 4.8 and 4.10.

As shown in [9], for particular interpolation points and higher-order moments, it might be possible to construct reduced-order systems via one-sided and two-sided interpolatory projection methods, which can reconstruct the limit-cycles. But as discussed in [9], stability of the reduced-order systems is highly sensitive to these specific choices, and even slight modifications may lead to unstable systems. For the $\mathcal{H}_2$ linear optimal interpolation points selection we made here, the one-sided and two-sided approaches were not able to reproduce the limit-cycles, thus motivating the usage of TQB-IRKA and BT once again, especially for the MIMO case.

Moreover, we report how far the reduced systems of order $r = 35$ due to TQB-IRKA, BT, one-sided and two-sided projection methods are from satisfying the optimality conditions (3.31). For this, we compute the perturbations (4.1) and list them in Table 4.5. Following the trend in the first two examples, the reduced-order system obtained by using TQB-IRKA satisfies the optimality conditions most accurately. Moreover, it can also be noticed that two-sided interpolatory projection method satisfies the optimality conditions very poorly, indicating a poor reduced-order model.

Lastly, we measure the truncated $\mathcal{H}_2$-norm of the error systems, using the reduced-order systems obtained via different methods of various orders. We plot the relative truncated $\mathcal{H}_2$-norm of the error systems in Figure 4.11. We observe that TQB-IRKA produces better reduced-order systems with respect to the truncated $\mathcal{H}_2$-norm as compared with BT and one-sided interpolatory projection as well. Furthermore, since we require stability of the matrix $\widehat{A}$ in the reduced QB system (1.2) to be able to compute the truncated $\mathcal{H}_2$-norm of the error systems, we could not achieve this in the case of

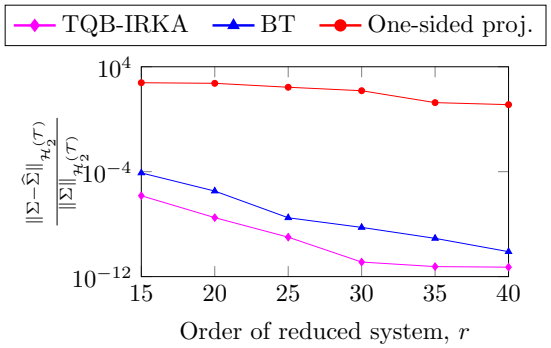

FIG. 4.11. *The FitzHugh–Nagumo system: comparison of the truncated $\mathcal{H}_2$-norm of the error system, having obtained reduced systems of different orders using various methods.*

two-sided interpolatory projection. For POD, we preserve the cubic nonlinearity in the reduced-order system; hence, the truncated $\mathcal{H}_2$-norm definition in Lemma 3.4 does not apply. Thus, we cannot compute the truncated $\mathcal{H}_2$-norm of the error system in the cases of the two-sided interpolatory projection and POD; thereby these methods are not included in Figure 4.11.

5. Conclusions. In this paper, we have investigated the optimal model reduction problem for QB control systems. We have first defined the $\mathcal{H}_2$-norm for QB systems based on the kernels of the underlying Volterra series and introduced a truncated $\mathcal{H}_2$-norm. We have then derived the first-order necessary conditions to be satisfied by a minimizer of the newly defined truncated $\mathcal{H}_2$-norm of the error system. These optimality conditions lead to the proposed model reduction algorithm (TQB-IRKA), which iteratively constructs reduced-order models that *approximately* satisfy the optimality conditions. We have also discussed the efficient computation of the reduced Hessian, utilizing the Kronecker structure of the Hessian of the QB system. Via several numerical examples, we have shown that TQB-IRKA outperforms the one-sided interpolation method, performs better than the two-sided projection in the majority of the cases, and is comparable with BT. Furthermore, unlike POD, since TQB-IRKA only depends on the state space quantities and not a specific choice of input, it outperforms POD for input functions that were not in the training set. Even for inputs which are used to train POD, TQB-IRKA still yields satisfactory performance, but is not better than POD as expected. Especially for MIMO QB systems, TQB-IRKA and BT are the preferred methods of choice to construct reduced systems since the current framework of two-sided subspace interpolatory projection method is only applicable to SISO systems and the extension of the one-sided interpolatory projection method to MIMO QB systems yields reduced models whose dimension increases quadratically with the number of inputs. Moreover, our numerical experiments reveal that in terms of stability, the reduced systems via TQB-IRKA and BT are more robust as compared with the one-sided and two-sided interpolatory projection methods although we do not have any theoretical justification of this observation yet. Additionally, even though a stable random initialization of TQB-IRKA has performed well in all of our numerical examples, a more educated but cheaper initial guess, for example via the two-sided interpolatory method [10], can further improve the convergence of TQB-IRKA and the quality of the obtained reduced-order systems.

Even though we have investigated the efficient computation of the reduced Hessian by utilizing the Kronecker product structure of the Hessian of the QB system, further

research in this direction using even more sophisticated tools from tensor theory would prove significant in accelerating the iteration steps in TQB-IRKA. Furthermore, it is worthwhile to further investigate the convergence of $\mathcal{H}_2$ iterative schemes such as TQB-IRKA, and the asymptotic stability of the reduced systems upon convergence.

Appendix A. Important relations of the Kronecker products. In this section, we provide some relations between Kronecker products, which will simplify the optimality conditions in Appendix B.

LEMMA A.1 (see [8, Lemma A.1]). *Consider $f(x) \in \mathbb{R}^{s\times n}$, $A(y) \in \mathbb{R}^{n\times n}$, $G \in \mathbb{R}^{n\times q}$ with $x, y \in \mathbb{R}$, and let $\mathcal{L}(y)$ be defined as*

$$\mathcal{L}(y) = -A(y) \otimes I_n - I_n \otimes A(y).$$

If the functions f and A are differentiable with respect to x and y, respectively, then

$$\frac{\partial}{\partial x}\left[(\mathcal{I}_s)^T (f(x) \otimes f(x))\, \mathcal{L}^{-1}(y)(G \otimes G)\mathcal{I}_q\right] = 2(\mathcal{I}_s)^T \left(\left(\frac{\partial}{\partial x} f(x)\right) \otimes f(x)\right) \mathcal{L}^{-1}(y)(G \otimes G)\mathcal{I}_q.$$

Moreover, let $X, Y \in \mathbb{R}^{n\times n}$ be symmetric matrices. Then,

$$\frac{\partial}{\partial y}\left[\operatorname{vec}(X)^T \mathcal{L}^{-1}(y) \operatorname{vec}(Y)\right] = 2 \cdot \operatorname{vec}(X)^T \mathcal{L}^{-1}(y) \left(\frac{\partial}{\partial y} A(y) \otimes I_n\right) \mathcal{L}^{-1}(y) \operatorname{vec}(Y).$$

LEMMA A.2. *Let $\mathcal{F}$ and $\widehat{\mathcal{F}}$ be defined as follows:*

$$\mathcal{F} = \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix} \otimes \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix} \quad \text{and} \quad \widehat{\mathcal{F}} = \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix},$$

and consider a permutation matrix

$$M = \begin{bmatrix} M_{nnr} & \mathbf{0} \\ \mathbf{0} & M_{rnr} \end{bmatrix}, \tag{A.1}$$

where M_{pqr} is defined in (3.27). *Moreover, let the two column vectors x and y be partitioned as*

$$x = \begin{bmatrix} x_1^T & x_2^T & x_3^T & x_4^T \end{bmatrix}^T \quad \text{and} \quad y = \begin{bmatrix} y_1^T & y_2^T & y_3^T & y_4^T \end{bmatrix}^T,$$

where $x_1, y_1 \in \mathbb{R}^{n^2}$, $x_{\{2,3\}}, y_{\{2,3\}} \in \mathbb{R}^{nr}$, and $x_4, y_4 \in \mathbb{R}^{r^2}$. Then, the following relations hold:

$$(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)} (M \otimes M)(x \otimes y) = T_{(n,r)}(x_3 \otimes y_3) \quad \text{and} \tag{A.2}$$

$$(\widehat{\mathcal{F}} \otimes \widehat{\mathcal{F}}) T_{(n+r,n+r)} (M \otimes M)(x \otimes y) = T_{(r,r)}(x_4 \otimes y_4), \tag{A.3}$$

where $T_{(n,m)}$ is also a permutation matrix given by

$$T_{(n,m)} = I_m \otimes \left[I_m \otimes e_1^n, \ldots, I_m \otimes e_n^n\right] \otimes I_n.$$

Proof. Let us begin by considering the following equation:

$$(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)} = \left[\mathbf{0} \quad I_r \otimes \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \mathcal{F}\right] (I_{n+r} \otimes \mathcal{G}),$$

where $\mathcal{G} = \begin{bmatrix} I_{n+r} \otimes e_1^{n+r}, \ldots, I_{n+r} \otimes e_{n+r}^{n+r} \end{bmatrix} \otimes I_{n+r}$. Next, we split I_{n+r} as $I_{n+r} = \left[\begin{smallmatrix} I_n & \mathbf{0} \\ \mathbf{0} & I_r \end{smallmatrix}\right]$, leading to

$$\begin{aligned}(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)} &= \begin{bmatrix} \mathbf{0} & I_r \otimes \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \mathcal{F} \end{bmatrix} \begin{bmatrix} I_n \otimes \mathcal{G} & \mathbf{0} \\ \mathbf{0} & I_r \otimes \mathcal{G} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \left(I_r \otimes \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \mathcal{F}\right) (I_r \otimes \mathcal{G}) \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & I_r \otimes \Big(\left(\begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \mathcal{F} \right) \mathcal{G} \Big) \end{bmatrix}. \end{aligned} \tag{A.4}$$

Now, we investigate the following equation (a component of the previous equation):

$$\left(\begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \mathcal{F} \right) \mathcal{G}_i =: \mathcal{L}_i,$$

where $\mathcal{G}_i$ is the ith block column of the matrix $\mathcal{G}$ given by $\mathcal{G}_i = I_{n+r} \otimes e_i^{n+r} \otimes I_{n+r}$. This yields

$$\begin{aligned}\mathcal{L}_i &= \left(\begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \mathcal{F} \right) \left(I_{n+r} \otimes e_i^{n+r} \otimes I_{n+r} \right) \\ &= \left(\begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} I_{n+r} \right) \otimes \left(\mathcal{F}(e_i^{n+r} \otimes I_{n+r}) \right) = \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \left(\mathcal{F}(e_i^{n+r} \otimes I_{n+r}) \right). \end{aligned}$$

Assuming that $1 \le i \le n$, we can write $\mathcal{L}_i$ as

$$\begin{aligned}\mathcal{L}_i &= \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \left(\mathcal{F} \left(\begin{bmatrix} e_i^n \\ \mathbf{0} \end{bmatrix} \otimes I_{n+r} \right) \right) = \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \left(\begin{bmatrix} I_n \otimes \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix} & \mathbf{0} \end{bmatrix} \begin{bmatrix} e_i^n \otimes I_{n+r} \\ \mathbf{0} \end{bmatrix} \right) \\ &= \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \begin{bmatrix} e_i^n \otimes \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \mathbf{0} & I_r \otimes \left(e_i^n \otimes \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix} \right) \end{bmatrix}. \end{aligned}$$

Subsequently, we assume $n + r \ge i > n$, which leads to

$$\begin{aligned}\mathcal{L}_i &= \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \left(\mathcal{F} \left(\begin{bmatrix} \mathbf{0} \\ e_{i-n}^r \end{bmatrix} \otimes I_{n+r} \right) \right) \\ &= \begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \left(\begin{bmatrix} I_n \otimes \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ e_{i-n}^r \otimes I_{n+r} \end{bmatrix} \right) = \mathbf{0}. \end{aligned}$$

Thus,

$$\left(\begin{bmatrix} \mathbf{0} & I_r \end{bmatrix} \otimes \mathcal{F} \right) \mathcal{G} = \begin{bmatrix} \mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_n, \mathbf{0} \end{bmatrix} =: \mathcal{L}.$$

Inserting the above expression in (A.4) yields

$$(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)} = \begin{bmatrix} \mathbf{0} & I_r \otimes \mathcal{L} \end{bmatrix}.$$

Now, we are ready to investigate the following term:

$$\begin{aligned}(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)} (M \otimes M) &= \begin{bmatrix} \mathbf{0} & I_r \otimes \mathcal{L} \end{bmatrix} \begin{bmatrix} M_{nnr} \otimes M & \mathbf{0} \\ \mathbf{0} & M_{rnr} \otimes M \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & I_r \otimes \mathcal{L} \end{bmatrix} \begin{bmatrix} M_{nnr} \otimes M & \mathbf{0} \\ \mathbf{0} & M_{rnr} \otimes M \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & (I_r \otimes \mathcal{L}) (M_{rnr} \otimes M) \end{bmatrix}. \end{aligned}$$

Further, we consider the second block column of the above relation and substitute for M_{nnr} and M_{rnr} using (3.27) to get

$$\begin{aligned}(I_r \otimes \mathcal{L}) (M_{rnr} \otimes M) &= (I_r \otimes \mathcal{L}) \begin{bmatrix} I_r \otimes \begin{bmatrix} I_n \\ \mathbf{0} \end{bmatrix} \otimes M & I_r \otimes \begin{bmatrix} \mathbf{0} \\ I_r \end{bmatrix} \otimes M \end{bmatrix} \\ &= \begin{bmatrix} (I_r \otimes \mathcal{L}) \left(I_r \otimes \begin{bmatrix} I_n \\ \mathbf{0} \end{bmatrix} \otimes M \right) & (I_r \otimes \mathcal{L}) \left(I_r \otimes \begin{bmatrix} \mathbf{0} \\ I_r \end{bmatrix} \otimes M \right) \end{bmatrix}. \end{aligned} \tag{A.5}$$

Our following task is to examine each block column of (A.5). We begin with the first block; this is

$$(I_r \otimes \mathcal{L})\left(I_r \otimes \begin{bmatrix} I_n \\ \mathbf{0} \end{bmatrix} \otimes M\right) = I_r \otimes \left(\mathcal{L}\left(\begin{bmatrix} I_n \\ \mathbf{0} \end{bmatrix} \otimes M\right)\right) = I_r \otimes \left(\mathcal{L}\begin{bmatrix} I_n \otimes M \\ \mathbf{0} \end{bmatrix}\right)$$
$$= I_r \otimes \left[\mathcal{L}_1 M, \ldots, \mathcal{L}_n M\right].$$

We next aim to simplify the term $\mathcal{L}_i M$, which appears in the previous equation:

$$\begin{aligned} \mathcal{L}_i M &= \begin{bmatrix} \mathbf{0} & I_r \otimes \left[e_j^n \otimes \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix}\right] \end{bmatrix} \begin{bmatrix} M_{nnr} & \mathbf{0} \\ \mathbf{0} & M_{rnr} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \left(I_r \otimes \left[e_j^n \otimes \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix}\right]\right) M_{rnr} \end{bmatrix} \\ &= \left[\mathbf{0} \quad \left(I_r \otimes e_j^n \otimes \begin{bmatrix} I_n & \mathbf{0} \end{bmatrix}\right) \left[I_r \otimes \begin{bmatrix} I_n \\ \mathbf{0} \end{bmatrix} \quad I_r \otimes \begin{bmatrix} \mathbf{0} \\ I_r \end{bmatrix}\right]\right] \\ &= \begin{bmatrix} \mathbf{0} & \left(I_r \otimes e_j^n \otimes I_n\right) & \mathbf{0} \end{bmatrix} := \mathcal{X}_i. \end{aligned} \tag{A.6}$$

The second block column of (A.5) can be studied in a similar fashion, and it can be shown that

$$(I_r \otimes \mathcal{L})\left(I_r \otimes \begin{bmatrix} \mathbf{0} \\ I_r \end{bmatrix} \otimes M\right) = \mathbf{0}.$$

Summing up all these expressions, we obtain

$$(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)}(M \otimes M) = \begin{bmatrix} \mathbf{0} & \left(I_r \otimes \left[\mathcal{X}_1, \ldots, \mathcal{X}_n\right]\right) & \mathbf{0} \end{bmatrix},$$

where $\mathcal{X}_i$ is defined in (A.6). This gives

$$\begin{aligned} (\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)}(M \otimes M)(x \otimes y) &= \begin{bmatrix} \mathbf{0} & I_r \otimes \left[\mathcal{X}_1, \ldots, \mathcal{X}_n\right] & \mathbf{0} \end{bmatrix} (x \otimes y) \\ &= \left(I_r \otimes \left[\mathcal{X}_1, \ldots, \mathcal{X}_n\right]\right)(x_3 \otimes y). \end{aligned} \tag{A.7}$$

Next, we define another permutation

$$\mathcal{Q} = \left[\underbrace{I_r \otimes I_n \otimes \begin{bmatrix} I_{n^2} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}}_{\mathcal{Q}_1} \quad \underbrace{I_r \otimes I_n \otimes \begin{bmatrix} \mathbf{0} \\ I_{nr} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}}_{\mathcal{Q}_2} \quad \underbrace{I_r \otimes I_n \otimes \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ I_{nr} \\ 0 \end{bmatrix}}_{\mathcal{Q}_3} \quad \underbrace{I_r \otimes I_n \otimes \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ I_{r^2} \end{bmatrix}}_{\mathcal{Q}_4}\right],$$

which allows us to write

$$(x_3 \otimes y) = \mathcal{Q} \begin{bmatrix} x_3 \otimes y_1 \\ x_3 \otimes y_2 \\ x_3 \otimes y_3 \\ x_3 \otimes y_4 \end{bmatrix}.$$

Substituting this into (A.7) results in

$$\begin{aligned} &(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)}(M \otimes M)(x \otimes y) \\ &\qquad = \left(I_r \otimes \left[\mathcal{X}_1, \ldots, \mathcal{X}_n\right]\right) \begin{bmatrix} \mathcal{Q}_1 & \mathcal{Q}_2 & \mathcal{Q}_3 & \mathcal{Q}_4 \end{bmatrix} \begin{bmatrix} x_3 \otimes y_1 \\ x_3 \otimes y_2 \\ x_3 \otimes y_3 \\ x_3 \otimes y_4 \end{bmatrix}. \end{aligned}$$

Now, it can be easily verified that $\left(I_r \otimes \left[\mathcal{X}_1, \ldots, \mathcal{X}_n\right]\right)\left[\mathcal{Q}_1 \quad \mathcal{Q}_2 \quad \mathcal{Q}_4\right] = 0$. Thus, we obtain

$$\begin{aligned}(\widehat{\mathcal{F}} \otimes \mathcal{F}) T_{(n+r,n+r)}(M \otimes M)(x \otimes y) &= \left(I_r \otimes \left[\mathcal{X}_1, \ldots, \mathcal{X}_n\right]\right) \mathcal{Q}_3 (x_3 \otimes y_3) \\ &= \left(I_r \otimes \left[\mathcal{X}_1, \ldots, \mathcal{X}_n\right]\right) \left(I_r \otimes I_n \otimes \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ I_{nr} \\ \mathbf{0} \end{bmatrix} \right) (x_3 \otimes y_3) \\ &= \left(I_r \otimes \left[I_r \otimes e_1^n \otimes I_n, \ldots, I_r \otimes e_1^n \otimes I_n\right]\right) (x_3 \otimes y_3) = T_{(n,r)}(x_3 \otimes y_3).\end{aligned}$$

One can prove the relation (A.2) in a similar manner. However, for brevity, we omit it. This concludes the proof. ∎

We will find similar expressions as (A.2) and (A.3) in Appendix B, where we then make use of Lemma A.2 to simplify them.

Appendix B. Proof of Theorem 3.7.

Optimality conditions with respect to $\widetilde{C}$. We start with deriving the optimality conditions by taking the derivative of the error functional $\mathcal{E}$ (3.25) with respect to $\widetilde{C}$. By using Lemma A.1, we obtain

$$\begin{aligned}\frac{\partial \mathcal{E}^2}{\partial \widetilde{C}_{ij}} &= 2(\mathcal{I}_p)^T \left(\left[\mathbf{0} \ -e_i^p (e_j^r)^T\right] \otimes \widetilde{C}^e \right) \left(-\widetilde{A}^e \otimes I_{n+r} - I_{n+r} \otimes \widetilde{A}^e \right)^{-1} \\ &\quad \left(\left(\widetilde{B}^e \otimes \widetilde{B}^e \right) \mathcal{I}_m + \sum_{k=1}^{m} \left(\widetilde{N}_k^e \otimes \widetilde{N}_k^e \right) \mathcal{P}_l + \left(\widetilde{H}^e \otimes \widetilde{H}^e \right) T_{(n+r,n+r)} (\mathcal{P}_l \otimes \mathcal{P}_l) \right),\end{aligned}$$

where $\mathcal{P}_l$ is defined in (3.26). Simplifying this expression, we get

$$\begin{aligned}\frac{\partial \mathcal{E}^2}{\partial \widetilde{C}_{ij}} &= 2(\mathcal{I}_p)^T \left(-e_i^p (e_j^r)^T \otimes \widetilde{C}^e \right) \left(-\Lambda \otimes I_{n+r} - I_r \otimes \widetilde{A}^e \right)^{-1} \left(\left(\widetilde{B} \otimes \widetilde{B}^e \right) \mathcal{I}_m \right. \\ &\quad \left. + \sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e \right) \mathcal{P}_l^{(2)} + \left(\widetilde{H} \widehat{\mathcal{F}} \otimes \widetilde{H}^e \right) T_{(n+r,n+r)} (\mathcal{P}_l \otimes \mathcal{P}_l) \right), \\ &= 2(\mathcal{I}_p)^T \left(-e_i^p (e_j^r)^T \otimes \widetilde{C}^e \right) \left(M_{rnr} \left(-\mathcal{J}_\Lambda - \mathcal{J}_A \right) M_{rnr}^T \right)^{-1} \left(\left(\widetilde{B} \otimes \widetilde{B}^e \right) \mathcal{I}_m \right. \\ &\quad \left. + \sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e \right) \mathcal{P}_l^{(2)} + \left(\widetilde{H} \widehat{\mathcal{F}} \otimes \widetilde{H}^e \right) T_{(n+r,n+r)} (\mathcal{P}_l \otimes \mathcal{P}_l) \right), \qquad \text{(B.1)}\end{aligned}$$

where

$$\mathcal{J}_\Lambda = \begin{bmatrix} \Lambda \otimes I_n & \mathbf{0} \\ \mathbf{0} & \Lambda \otimes I_r \end{bmatrix}, \qquad \mathcal{J}_A = \begin{bmatrix} I_r \otimes A & \mathbf{0} \\ \mathbf{0} & I_r \otimes \Lambda \end{bmatrix}, \quad \text{and}$$

$\mathcal{P}_l^{(2)}$ is the lower block row of $\mathcal{P}_l$ as shown in (3.26). Furthermore, since M_{rnr} is a permutation matrix, this implies $M_{rnr} M_{rnr}^T = I$. Using this relation in (B.1), we obtain

$$
\begin{aligned}
\frac{\partial \mathcal{E}^2}{\partial \widetilde{C}_{ij}} &= 2(\mathcal{I}_p)^T \left(-e_i^p (e_j^r)^T \otimes \begin{bmatrix} C & -\widetilde{C} \end{bmatrix}\right) M_{rnr} \left(-\mathcal{J}_\Lambda - \mathcal{J}_A\right)^{-1} \left(M_{rnr}^T \left(\widetilde{B} \otimes \widetilde{B}^e\right) \mathcal{I}_m \right. \\
&\quad \left. + M_{rnr}^T \sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e\right) \mathcal{P}_1^{(2)} + M_{rnr}^T \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}^e\right) T_{(n+r,n+r)} (\mathcal{P}_l \otimes \mathcal{P}_l)\right) \\
&= 2(\mathcal{I}_p)^T \left(\begin{bmatrix} -e_i^p (e_j^r)^T \otimes C & e_i e_j^T \otimes \widetilde{C} \end{bmatrix}\right) \left(-\mathcal{J}_\Lambda - \mathcal{J}_A\right)^{-1} \left(\begin{bmatrix} \widetilde{B} \otimes B \\ \widetilde{B} \otimes \widetilde{B} \end{bmatrix} \mathcal{I}_m \right. \\
&\quad + \sum_{k=1}^{m} \begin{bmatrix} \widetilde{N}_k \otimes N_k & \mathbf{0} \\ \mathbf{0} & \widetilde{N}_k \otimes \widetilde{N}_k \end{bmatrix} M_{rnr}^T \mathcal{P}_1^{(2)} \\
&\quad \left. + \begin{bmatrix} \left(\widetilde{H}\widehat{\mathcal{F}} \otimes H\mathcal{F}\right) T_{(n+r,n+r)} (M \otimes M)(M^T \otimes M^T)(\mathcal{P}_l \otimes \mathcal{P}_l) \\ \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}\widehat{\mathcal{F}}\right) T_{(n+r,n+r)} (M \otimes M)(M^T \otimes M^T)(\mathcal{P}_l \otimes \mathcal{P}_l) \end{bmatrix} \right), \qquad \text{(B.2)}
\end{aligned}
$$

where M is the permutation matrix defined in (A.1). The multiplication of M^T and $\mathcal{P}_l$ yields

$$
M^T \mathcal{P}_l = \begin{bmatrix} M_{nnr} \mathcal{P}_l^{(1)} \\ M_{rnr} \mathcal{P}_1^{(2)} \end{bmatrix} = \begin{bmatrix} p_1^T & p_2^T & p_3^T & p_4^T \end{bmatrix}^T =: \widetilde{\mathcal{P}}_l,
$$

where

(B.3)

$$
\begin{aligned}
p_1 &= \left(-A \otimes I_n - I_n \otimes A\right)^{-1} (B \otimes B) \mathcal{I}_m, \quad p_2 = \left(-A \otimes I_r - I_n \otimes \Lambda\right)^{-1} \left(B \otimes \widetilde{B}\right) \mathcal{I}_m, \\
p_3 &= \left(-\Lambda \otimes I_n - I_r \otimes A\right)^{-1} \left(\widetilde{B} \otimes B\right) \mathcal{I}_m, \quad p_4 = \left(-\Lambda \otimes I_r - I_r \otimes \Lambda\right)^{-1} \left(\widetilde{B} \otimes \widetilde{B}\right) \mathcal{I}_m.
\end{aligned}
$$

Moreover, note that $p_3 = \operatorname{vec}(V_1)$, where V_1 solves (3.28a). Applying the result of Lemma A.2 in (B.2) yields

$$
\begin{aligned}
\frac{\partial \mathcal{E}^2}{\partial \widetilde{C}_{ij}} &= 2(\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes C\right) \left(-\Lambda \otimes I_n - I_r \otimes A\right)^{-1} \left((\widetilde{B} \otimes B) \mathcal{I}_m + \sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) p_3 \right. \\
&\quad \left. + (\widetilde{H} \otimes H) T_{(n,r)} (p_3 \otimes p_3)\right) - 2(\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes \widetilde{C}\right) \left(-\Lambda \otimes I_n - I_r \otimes \widetilde{A}\right)^{-1} \\
&\quad \times \left((\widetilde{B} \otimes \widetilde{B}) \mathcal{I}_m + \sum_{k=1}^{m} (\widetilde{N}_k \otimes \widetilde{N}_k) p_4 + (\widetilde{H} \otimes \widetilde{H}) T_{(r,r)} (p_4 \otimes p_4)\right) \\
&= 2(\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes C\right) \left(-\Lambda \otimes I_n - I_r \otimes A\right)^{-1} \left((\widetilde{B} \otimes B) \mathcal{I}_m + \sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) p_3 \right. \\
&\quad \left. + (\widetilde{H} \otimes H) T_{(n,r)} (p_3 \otimes p_3)\right) - 2(\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes \widehat{C}\right) \left(-\Lambda \otimes I_n - I_r \otimes \widehat{A}\right)^{-1} \\
&\quad \times \left((\widetilde{B} \otimes \widehat{B}) \mathcal{I}_m + \sum_{k=1}^{m} (\widetilde{N}_k \otimes \widehat{N}_k) \widehat{p}_4 + (\widetilde{H} \otimes \widehat{H}) T_{(r,r)} (\widehat{p}_4 \otimes \widehat{p}_4)\right), \qquad \text{(B.4)}
\end{aligned}
$$

where $\widehat{p}_4 = (-\Lambda \otimes I_r - I_r \otimes \widehat{A})^{-1}(\widetilde{B} \otimes \widehat{B})\mathcal{I}_m = \operatorname{vec}(\widehat{V}_1)$, where $\widehat{V}_1$ is as defined in (3.30). Setting (B.4) equal to zero results in a necessary condition with respect to $\widetilde{C}$ as follows:

$$\begin{aligned}
&(\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes C\right)(-\Lambda \otimes I_n - I_r \otimes A)^{-1}\Big((\widetilde{B} \otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3 \\
&\qquad + (\widetilde{H} \otimes H)T_{(n,r)}(p_3 \otimes p_3)\Big) \\
&= (\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes \widehat{C}\right)\left(-\Lambda \otimes I_n - I_r \otimes \widehat{A}\right)^{-1} \\
&\qquad \times \Big((\widetilde{B} \otimes \widehat{B})\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes \widehat{N}_k)\widehat{p}_4 + (\widetilde{H} \otimes \widehat{H})T_{(r,r)}(\widehat{p}_4 \otimes \widehat{p}_4)\Big).
\end{aligned} \tag{B.5}$$

Now, we first manipulate the left-hand side of (B.5). Using Lemma 2.4 and (2.1), we get

$$\begin{aligned}
&(\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes C\right)(-\Lambda \otimes I_n - I_r \otimes A)^{-1}\left((\widetilde{B} \otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3\right. \\
&\qquad \left. + (\widetilde{H} \otimes H)T_{(n,r)}(p_3 \otimes p_3)\right) \\
&= (\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes C\right)(-\Lambda \otimes I_n - I_r \otimes A)^{-1}\left(\operatorname{vec}\left(B\widetilde{B}^T\right) + \sum_{k=1}^{m}\operatorname{vec}\left(N_k V_1 \widetilde{N}_k^T\right)\right. \\
&\qquad \left. + (\widetilde{H} \otimes H)\operatorname{vec}(V_1 \otimes V_1)\right) \\
&= (\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes C\right)(-\Lambda \otimes I_n - I_r \otimes A)^{-1}\left(\operatorname{vec}\left(B\widetilde{B}^T + \sum_{k=1}^{m} N_k V_1 \widetilde{N}_k^T\right)\right. \\
&\qquad \left. + \operatorname{vec}\left(H(V_1 \otimes V_1)\widetilde{H}^T\right)\right) \\
&= (\mathcal{I}_p)^T \left(e_i^p (e_j^r)^T \otimes C\right)(\operatorname{vec}(V_1) + \operatorname{vec}(V_2)) = \operatorname{tr}\left(C(V_1 + V_2)e_j^r (e_i^p)^T\right) \\
&= \operatorname{tr}\left(CV e_j^r (e_i^p)^T\right),
\end{aligned}$$

where V_2 solves (3.28c) and $V = V_1 + V_2$. Using the similar steps, we can show that the right-hand side of (B.5) is equal to $\operatorname{tr}(\widehat{C}\widehat{V}e_j^r(e_i^p)^T)$, where $\widehat{V}$ is defined in (3.30). Therefore, (B.5) is the same as (3.31a).

Necessary conditions with respect to Λ. By utilizing Lemma A.1, we aim at deriving the necessary condition with respect to the ith diagonal entry of Λ. We differentiate $\mathcal{E}$ with respect to λ_i to obtain

$$\begin{aligned}
\frac{\partial \mathcal{E}^2}{\partial \lambda_i} &= 2(\mathcal{I}_p)^T \left(\widetilde{C}^e \otimes \widetilde{C}^e\right)\mathcal{L}_e^{-1}\mathbb{E}\mathcal{L}_e^{-1}\Big(\left(\widetilde{B}^e \otimes \widetilde{B}^e\right)\mathcal{I}_m + \sum_{k=1}^{m}\left(\widetilde{N}_k^e \otimes \widetilde{N}_k^e\right)\mathcal{P}_l \\
&\quad + \left(\widetilde{H}^e \otimes \widetilde{H}^e\right)T_{(n+r,n+r)}(\mathcal{P}_l \otimes \mathcal{P}_l)\Big) + (\mathcal{I}_p)^T \left(\widetilde{C}^e \otimes \widetilde{C}^e\right)\mathcal{L}_e^{-1} \\
&\quad \times \left(2\sum_{k=1}^{m}\left(\widetilde{N}_k^e \otimes \widetilde{N}_k^e\right)\mathcal{L}_e^{-1}\mathbb{E}\mathcal{P}_l + 4\left(\widetilde{H}^e \otimes \widetilde{H}^e\right)T_{(n+r,n+r)}\left((\mathcal{L}_e^{-1}\mathbb{E}\mathcal{P}_l) \otimes \mathcal{P}_l\right)\right),
\end{aligned}$$

where

$$\mathcal{L}_e = -\left(\widetilde{A}^e \otimes I_{n+r} + I_{n+r} \otimes \widetilde{A}^e\right) \quad \text{and} \quad \mathbb{E} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & e_i^r (e_i^r)^T \end{bmatrix} \otimes I_{n+r}.$$

Performing some algebraic calculations gives rise to the following expression:

$$\begin{aligned}\frac{\partial \mathcal{E}^2}{\partial \lambda_i} &= 2(\mathcal{I}_p)^T \left(-\widetilde{C} \otimes \widetilde{C}^e\right) \mathcal{Z}_e^{-1} \Xi_{n+r} \mathcal{Z}_e^{-1} \Big(\left(\widetilde{B} \otimes \widetilde{B}^e\right) \mathcal{I}_m + \sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e\right) \mathcal{P}_1^{(2)} \\ &\quad + \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}^e\right) T_{(n+r,n+r)} (\mathcal{P}_l \otimes \mathcal{P}_l) \Big) + 2(\mathcal{I}_p)^T \left(-\widetilde{C} \otimes \widetilde{C}^e\right) \mathcal{Z}_e^{-1} \\ &\quad \times \left(\sum_{k=1}^{m} \left(\widetilde{N}_k \otimes \widetilde{N}_k^e\right) \mathcal{Z}_e^{-1} \Xi_{n+r} \mathcal{P}_1^{(2)} + 2 \left(\widetilde{H}\widehat{\mathcal{F}} \otimes \widetilde{H}^e\right) T_{(n+r,n+r)} (\mathcal{L}_e^{-1} \mathbb{E} \mathcal{P}_l \otimes \mathcal{P}_l) \right),\end{aligned}$$

where $\mathcal{Z}_e := -(\Lambda \otimes I_{n+r} + I_r \otimes A^e)$ and $\Xi_m := (e_i^r (e_i^r)^T \otimes I_m)$. Next, we utilize Lemma A.2 and use the permutation matrix M (as done while deriving the necessary conditions with respect to $\widetilde{C}$) to obtain

$$\begin{aligned}&\frac{\partial \mathcal{E}^2}{\partial \lambda_i} \\ &= 2(\mathcal{I}_p)^T \mathcal{S} \left((\widetilde{B} \otimes B) \mathcal{I}_m + \sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) p_3 + (\widetilde{H} \otimes H) T_{(n,r)} (p_3 \otimes p_3) \right) \\ &\quad - 2(\mathcal{I}_p)^T \widetilde{\mathcal{S}} \left((\widetilde{B} \otimes \widetilde{B}) \mathcal{I}_m + \sum_{k=1}^{m} (\widetilde{N}_k \otimes \widetilde{N}_k) p_4 + (\widetilde{H} \otimes \widetilde{H}) T_{(r,r)} (p_4 \otimes p_4) \right) \\ &\quad + 2(\mathcal{I}_p)^T \left(\widetilde{C} \otimes C\right) L^{-1} \left(\sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) L^{-1} \Xi_n p_3 + 2(\widetilde{H} \otimes H) T_{(n,r)} (L^{-1} \Xi_n p_3 \otimes p_3) \right) \\ &\quad - 2(\mathcal{I}_p)^T \left(\widetilde{C} \otimes \widetilde{C}\right) \widetilde{L}^{-1} \left(\sum_{k=1}^{m} (\widetilde{N}_k \otimes \widetilde{N}_k) \widetilde{L}^{-1} \Xi_r p_4 + 2(\widetilde{H} \otimes \widetilde{H}) T_{(r,r)} (\widetilde{L}^{-1} \Xi_r (p_4 \otimes p_4)) \right) \\ &= 2(\mathcal{I}_p)^T \mathcal{S} \left((\widetilde{B} \otimes B) \mathcal{I}_m + \sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) p_3 + (\widetilde{H} \otimes H) T_{(n,r)} (p_3 \otimes p_3) \right) \\ &\quad - 2(\mathcal{I}_p)^T \widehat{\mathcal{S}} \left((\widetilde{B} \otimes \widehat{B}) \mathcal{I}_m + \sum_{k=1}^{m} (\widetilde{N}_k \otimes \widehat{N}_k) \widehat{p}_4 + (\widetilde{H} \otimes \widehat{H}) T_{r,r} (\widehat{p}_4 \otimes \widehat{p}_4) \right) \\ &\quad + 2(\mathcal{I}_p)^T \left(\widetilde{C} \otimes C\right) L^{-1} \left(\sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k) L^{-1} \Xi_n p_3 + 2(\widetilde{H} \otimes H) T_{(n,r)} (L^{-1} \Xi_n (p_3 \otimes p_3)) \right) \\ &\quad - 2(\mathcal{I}_p)^T \left(\widetilde{C} \otimes \widehat{C}\right) \widehat{L}^{-1} \left(\sum_{k=1}^{m} (\widetilde{N}_k \otimes \widehat{N}_k) \widehat{L}^{-1} \Xi_r \widehat{p}_4 + 2(\widetilde{H} \otimes \widehat{H}) T_{(r,r)} (\widehat{L}^{-1} \Xi_r (\widehat{p}_4 \otimes \widehat{p}_4) \right),\end{aligned}$$

where p_3 and p_4 are the same as defined in (B.3), and

$$\begin{aligned}&\mathcal{S} := \left(\widetilde{C} \otimes C\right) L^{-1} (e_i^r (e_i^r)^T \otimes I_n) L^{-1}, \quad \widetilde{\mathcal{S}} := \left(\widetilde{C} \otimes \widetilde{C}\right) \widetilde{L}^{-1} (e_i^r (e_i^r)^T \otimes I_r) \widetilde{L}^{-1}, \\ &\widehat{\mathcal{S}} := \left(\widetilde{C} \otimes \widehat{C}\right) \widehat{L}^{-1} (e_i^r (e_i^r)^T \otimes I_r) \widehat{L}^{-1}, \quad L := -(\Lambda \otimes I_n + I_r \otimes A), \\ &\widetilde{L} := -(\Lambda \otimes I_r + I_r \otimes \Lambda), \qquad\qquad \widehat{L} := -\left(\Lambda \otimes I_r + I_r \otimes \widehat{A}\right).\end{aligned}$$

By using the properties derived in Lemma 2.2, we can simplify the above equation:

$$\begin{aligned}\frac{\partial \mathcal{E}^2}{\partial \lambda_i} &= 2(\mathcal{I}_p)^T \mathcal{S}\Big((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3 + (\widetilde{H}\otimes H)T_{(n,r)}(p_3\otimes p_3)\Big)\\ &\quad - 2(\mathcal{I}_p)^T \widehat{\mathcal{S}}\Big((\widetilde{B}\otimes \widehat{B})\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes \widehat{N}_k)p_4 + (\widetilde{H}\otimes \widehat{H})T_{(r,r)}(\widehat{p}_4\otimes \widehat{p}_4)\Big)\\ &\quad + 2(\mathcal{I}_m)^T\left(\widetilde{B}\otimes B\right)L^{-T}\Xi_n L^{-T}\left(\sum_{k=1}^{m}(\widetilde{N}_k\otimes N_k)^T q_3 + 2(\widetilde{\mathcal{H}}^{(2)}\otimes \mathcal{H}^{(2)})T_{(n,r)}(p_3\otimes q_3)\right)\\ &\quad - 2(\mathcal{I}_m)^T\left(\widetilde{B}\otimes \widehat{B}\right)\widehat{L}^{-T}\Xi_r \widehat{L}^{-T}\left(\sum_{k=1}^{m}(\widetilde{N}_k\otimes \widehat{N}_k)^T \widehat{q}_4 + 2(\widetilde{\mathcal{H}}^{(2)}\otimes \widehat{\mathcal{H}}^{(2)})T_{(r,r)}(\widehat{p}_4\otimes \widehat{q}_4)\right),\end{aligned}$$

where

$$q_3 = (-\Lambda\otimes I_n - I_r\otimes A)^{-T}\left(\widetilde{C}\otimes C\right)\mathcal{I}_p \text{ and } \widehat{q}_4 = (-\Lambda\otimes I_r - I_r\otimes A_r)^{-T}\left(\widetilde{C}\otimes \widehat{C}\right)\mathcal{I}_p.$$

Once again, we determine an interpolation-based necessary condition with respect to Λ_i by setting the last equation equal to zero:

(B.6)

$$\begin{aligned}&(\mathcal{I}_p)^T \mathcal{S}\Big((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3 + (\widetilde{H}\otimes H)T_{(n,r)}(p_3\otimes p_3)\Big)\\ &\quad + (\mathcal{I}_m)^T\left(\widetilde{B}\otimes B\right)L^{-T}\Xi_n L^{-T}\left(\sum_{k=1}^{m}(\widetilde{N}_k\otimes N_k)^T q_3 + 2(\widetilde{\mathcal{H}}^{(2)}\otimes \mathcal{H}^{(2)})T_{(n,r)}(p_3\otimes q_3)\right)\\ &= (\mathcal{I}_p)^T \widehat{\mathcal{S}}\left((\widetilde{B}\otimes \widehat{B})\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes \widehat{N}_k)p_4 + (\widetilde{H}\otimes \widehat{H})T_{(r,r)}(\widehat{p}_4\otimes \widehat{p}_4)\right)\\ &\quad + (\mathcal{I}_m)^T\left(\widetilde{B}\otimes \widehat{B}\right)\widehat{L}^{-T}\Xi_r \widehat{L}^{-T}\left(\sum_{k=1}^{m}(\widetilde{N}_k\otimes \widehat{N}_k)^T \widehat{q}_4 + 2(\widetilde{\mathcal{H}}^{(2)}\otimes \widehat{\mathcal{H}}^{(2)})T_{(r,r)}(\widehat{p}_4\otimes \widehat{q}_4)\right).\end{aligned}$$

Now, we first simplify the left-hand side of the above equation using Lemma 2.4 and (2.1). We first focus on the first part of the left-hand side of (B.6). This yields

$$\begin{aligned}&(\mathcal{I}_p)^T \mathcal{S}\Big((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3 + (\widetilde{H}\otimes H)T_{(n,r)}(p_3\otimes p_3)\Big)\\ &\quad = (\mathcal{I}_p)^T\left(\widetilde{C}\otimes C\right)L^{-1}(e_i^r(e_i^r)^T\otimes I_n)\\ &\qquad \times L^{-1}\Big((\widetilde{B}\otimes B)\mathcal{I}_m + \sum_{k=1}^{m}(\widetilde{N}_k \otimes N_k)p_3 + (\widetilde{H}\otimes H)T_{(n,r)}(p_3\otimes p_3)\Big)\\ &\quad = \underbrace{(\mathcal{I}_p)^T\left(\widetilde{C}\otimes C\right)L^{-1}}_{=(\mathrm{vec}(W_1))^T}(e_i^r(e_i^r)^T\otimes I_n)\,\mathrm{vec}\,(V) = \mathrm{tr}\left(Ve_i^r(e_i^r)^T W_1^T\right)\\ &\quad = (V_1(:,i))^T W(:,i) = (W_1(:,i))^T V(:,i),\end{aligned}$$

where W_1 solves (3.28b). Analogously, we can show that

(B.7)

$$\begin{aligned}&(\mathcal{I}_m)^T\left(\widetilde{B}\otimes B\right)L^{-T}\Xi_n L^{-T}\Big(\sum_{k=1}^{m}(\widetilde{N}_k\otimes N_k)^T q_3 + 2(\widetilde{\mathcal{H}}^{(2)}\otimes \mathcal{H}^{(2)})T_{(n,r)}(p_3\otimes q_3)\Big)\\ &\qquad = (W(:,i))^T V_1(:,i).\end{aligned}$$

Thus, the left-hand side of (B.6) is equal to $(W(:,i))^T V_1(:,i) + (W_1(:,i))^T V(:,i)$. Using similar steps, we can also show that the right-hand side of (B.6) is equal to $(\widehat{W}(:,i))^T \widehat{V}_1(:,i) + (\widehat{W}_1(:,i))^T \widehat{V}(:,i)$. Thus, we obtain the optimality conditions with respect to Λ given in (3.31e).

The necessary conditions with respect to $\widetilde{B}$, $\widetilde{N}$, and $\widetilde{H}$ can also be determined in a similar manner as for $\widetilde{C}$ and λ_i. For brevity of the paper, we skip detailed derivations; however, we state the final optimality conditions. A necessary condition for optimality with respect to the (i,j)th entry of $\widetilde{N}_k$ is

$$(\mathcal{I}_p)^T \left(\widetilde{C} \otimes C\right) L^{-1} \left((e_i^r (e_j^r)^T \otimes N_k) p_3\right) = (\mathcal{I}_p)^T \left(\widetilde{C} \otimes \widehat{C}\right) \widehat{L}^{-1} \left((e_i^r (e_j^r)^T \otimes \widehat{N}_k) \widehat{p}_4\right),$$

which then yields (3.31c) in the Sylvester equation form. A similar optimality condition with respect to the (i,j)th entry of $\widetilde{H}$ is given by

$$\begin{aligned}(\mathcal{I}_p)^T \left(\widetilde{C} \otimes C\right) L^{-1} &\left((e_i^r (e_j^{r^2})^T \otimes H) T_{(n,r)} (p_3 \otimes p_3)\right) \\ &= (\mathcal{I}_p)^T \left(\widetilde{C} \otimes \widehat{C}\right) \widehat{L}^{-1} \left((e_i^r (e_j^{r^2})^T \otimes \widehat{H}) T_{(r,r)} (\widehat{p}_4 \otimes \widehat{p}_4)\right),\end{aligned}$$

which can be equivalently described as (3.31d). Finally, the necessary condition appearing with respect to the (i,j)th entry of $\widetilde{B}$ is

$$\begin{aligned}(\mathcal{I}_m)^T \left(e_i^r (e_j^m)^T \otimes B\right) L^{-T} \Big(&(\widetilde{C} \otimes C) \mathcal{I}_p + \sum_{k=1}^{m} (\widetilde{N}_k \otimes N_k)^T q_3 \\ &+ 2(\widetilde{\mathcal{H}}^{(2)} \otimes \mathcal{H}^{(2)}) T_{(n,r)} (p_3 \otimes q_3)\Big), \\ = (\mathcal{I}_m)^T \left(e_i^r (e_j^m)^T \otimes \widehat{B}\right) \widehat{L}^{-T} \Big(&(\widetilde{C} \otimes \widehat{C}) \mathcal{I}_p + \sum_{k=1}^{m} (\widetilde{N}_k \otimes \widehat{N}_k)^T \widehat{q}_4 \\ &+ 2(\widetilde{\mathcal{H}}^{(2)} \otimes \widehat{\mathcal{H}}^{(2)}) T_{(r,r)} (\widehat{p}_4 \otimes \widehat{q}_4)\Big),\end{aligned}$$

which gives rise to (3.31b).

Appendix C. Proof of Theorem 3.9. We begin by establishing a relationship between $V_1 \in \mathbb{R}^{n\times r}, \widehat{V}_1 \in \mathbb{R}^{r\times r}$, and $V \in \mathbb{R}^{n\times r}$. For this, consider the Sylvester equation related to V_1

$$-V_1\Lambda - AV_1 = B\widetilde{B}^T, \tag{C.1}$$

and the oblique projector $\Pi_v := V_1(W^T V_1)^{-1} W^T$. Then, we apply the projector Π_v to the Sylvester equation (C.1) from the left to obtain

$$\begin{aligned}-V_1\Lambda - \Pi_v A V_1 &= \Pi_v B \widetilde{B}^T \text{ and} \\ -V_1\Lambda - \Pi A V_1 &= (\Pi_v - \Pi) A V_1 + \Pi_v B \widetilde{B}^T,\end{aligned} \tag{C.2}$$

where $\Pi := V(W^T V)^{-1} W^T$. Now, recall that $\widehat{V}_1$ satisfies the Sylvester equation

$$-\widehat{V}_1\Lambda - \widehat{A}\widehat{V}_1 = \widehat{B}\widetilde{B}^T.$$

We next multiply it by V from the left and substitute for $\widehat{A}$ and $\widehat{B}$ to obtain

$$-V\widehat{V}_1\Lambda - \Pi A V \widehat{V}_1 = \Pi B \widetilde{B}^T. \tag{C.3}$$

Subtracting (C.2) from (C.3) yields

$$(V_1 - V\widehat{V}_1)\Lambda + \Pi A(V_1 - V\widehat{V}_1) = (\Pi - \Pi_v)\left(AV_1 + B\widetilde{B}^T\right).$$

Since it is assumed that $\sigma(\widehat{A}) \cap \sigma(-\Pi A) = \emptyset$, this implies that $\Lambda \otimes I_n + I_r \otimes (\Pi A)$ is invertible. Therefore, we can write

$$V_1 = V\widehat{V}_1 + \epsilon_v, \tag{C.4}$$

where ϵ_v solves the Sylvester equation

$$\epsilon_v\Lambda + \Pi_v A\epsilon_v = (\Pi - \Pi_v)\left(AV_1 + B\widetilde{B}^T\right). \tag{C.5}$$

Similarly, one can show that

$$W_1 = W(W^TV)^{-T}\widehat{W}_1 + \epsilon_w, \tag{C.6}$$

where ϵ_w solves

$$\epsilon_w\Lambda + \Pi^T A^T\epsilon_w = (\Pi^T - \Pi_w)(A^TW_1 + C^T\widetilde{C}),$$

in which $\Pi_w := W_1(V^TW)V^T$. Using (C.4) and (C.6), we obtain

$$\begin{aligned}
\widehat{W}_1(:,i)^T\widehat{N}_k\widehat{V}_1(:,j) &= \widehat{W}_1(:,i)^T(W^TV)^{-1}W^TN_kV\widehat{V}_1(:,j)\\
&= \left(W_1(:,i) - \epsilon_w(:,i)\right)^T N_k\left(V_1(:,j) - \epsilon_v(:,j)\right)\\
&= W_1(:,i)^TN_kV_1(:,j) - \left(\epsilon_w(:,i)\right)^T N_k(V_1(:,j) - \epsilon_v(:,j))\\
&\quad - \left(W_1(:,i)\right)^T N_k(\epsilon(:,j)),
\end{aligned}$$

which is (3.32c) in Theorem 3.7. Similarly, one can prove (3.32d). To prove (3.32a), we consider the following Sylvester equation for V:

$$V(-\Lambda) - AV = B\widetilde{B}^T + \sum_{k=1}^{m} N_kV_1\widetilde{N}_k^T + H(V_1 \otimes V_1)\widetilde{H}^T. \tag{C.7}$$

Applying Π to both sides of the above Sylvester equation yields

$$V\left(I_r(-\Lambda) - \widehat{A}I_r\right) = V\left(\widehat{B}\widetilde{B}^T + \mathcal{Y}\right), \tag{C.8}$$

where $\mathcal{Y} = (W^TV)^{-1}W^T(\sum_{k=1}^m N_kV_1\widetilde{N}_k^T + H(V_1 \otimes V_1)\widetilde{H}^T)$. This implies that

$$I_r(-\Lambda) - \widehat{A}I_r = \widehat{B}\widetilde{B}^T + \mathcal{Y}. \tag{C.9}$$

Next, we consider the Sylvester equation for $\widehat{V}$,

$$\widehat{V}(-\Lambda) - \widehat{A}\widehat{V} = \widehat{B}\widetilde{B}^T + \sum_{k=1}^{m} \widehat{N}_k\widehat{V}_1\widetilde{N}_k^T + \widehat{H}(\widehat{V}_1 \otimes \widehat{V}_1)\widetilde{H}^T. \tag{C.10}$$

We then subtract (C.10) and (C.9) to obtain

$$\begin{aligned}
(I_r - \widehat{V})(-\Lambda) - \widehat{A}(I_r - \widehat{V}) = &\sum_{k=1}^{m}(W^TV)^{-1}W^TN_k\left(V_1 - V\widehat{V}_1\right)\widetilde{N}_k^T\\
&+ (W^TV)^{-1}W^TH\left(V_1 \otimes V_1 - (V\widehat{V}_1 \otimes V\widehat{V}_1)\right)\widetilde{H}^T.
\end{aligned}$$

Substituting $V\widehat{V}_1$ from (C.4) gives

$$(I_r - \widehat{V})(-\Lambda) - \widehat{A}(I_r - \widehat{V}) = \sum_{k=1}^{m} (W^T V)^{-1} W^T N_k \epsilon_v \widetilde{N}_k^T + (W^T V)^{-1} W^T H \left(\epsilon_v \otimes V_1 + V_1 \otimes \epsilon_v + \epsilon_v \otimes \epsilon_v\right) \widetilde{H}^T.$$

Since Λ contains the eigenvalues of $\widehat{A}$ and $\widehat{A}$ is stable, Λ and $-\widehat{A}$ cannot have any common eigenvalues. Hence, the matrix $\Lambda \otimes I_r + I_r \otimes \widehat{A}$ is invertible. Therefore the above Sylvester equations for $\Gamma := \widehat{V} - I_r$ exists and have a unique solution and can be written as

$$\Gamma_v \Lambda + \widehat{A}\Gamma_v = \sum_{k=1}^{m} (W^T V)^{-1} W^T N_k \epsilon_v \widetilde{N}_k^T + (W^T V)^{-1} W^T H \left(\epsilon_v \otimes V_1 + V_1 \otimes \epsilon_v + \epsilon_v \otimes \epsilon_v\right) \widetilde{H}^T.$$

To prove (3.32a), we observe that

$$\begin{aligned} \operatorname{tr}\left(\widehat{C}\widehat{V} e_i^r \left(e_j^p\right)^T\right) &= \operatorname{tr}\left(CV(I_r + \Gamma_v) e_i^r \left(e_j^p\right)^T\right) \\ &= \operatorname{tr}\left(CV e_i^r \left(e_j^p\right)^T\right) + \operatorname{tr}\left(CV\Gamma_v e_i^r \left(e_j^p\right)^T\right). \end{aligned}$$

Thus,

$$\operatorname{tr}\left(CV e_i^r \left(e_j^p\right)^T\right) = \operatorname{tr}\left(\widehat{C}\widehat{V} e_i^r \left(e_j^p\right)^T\right) + \epsilon_C^{(i,j)}.$$

Analogously, we can prove that there exists Γ_w such that $\widehat{W} = (W^T V)^T + \Gamma_w$ and that it satisfies

$$\Gamma_w \Lambda + \widehat{A}^T \Gamma_w = V^T \left(\sum_{k=1}^{m} N_k^T \epsilon_w \widetilde{N}_k + \mathcal{H}^{(2)}(\epsilon_v \otimes (W_1 + \epsilon_w) + V_1 \otimes \epsilon_w) \left(\mathcal{H}^{(2)}\right)^T\right).$$

To prove (3.32b), we observe that

$$\operatorname{tr}\left(\widehat{B}^T \widehat{W} e_i^r \left(e_j^m\right)^T\right) = \operatorname{tr}\left(B^T W (W^T V)^{-T} ((W^T V)^T + \Gamma_v) e_i^r \left(e_j^p\right)^T\right).$$

Thus,

$$\operatorname{tr}\left(\widehat{B}^T \widehat{W} e_i^r \left(e_j^m\right)^T\right) = \operatorname{tr}\left(B^T W^T + B^T W (W^T V)^{-T} \Gamma_w) e_i^r \left(e_j^p\right)^T\right).$$

Since we now know that $\widehat{V} = I_r + \Gamma_v$ and $\widehat{W} = (W^T V)^T + \Gamma_w$, we get

$$V\widehat{V} = V + V\Gamma_v \quad \text{and} \quad W(W^T V)^{-T}\widehat{W} = W + W(W^T V)^{-T}\Gamma_w. \tag{C.11}$$

We make use of (C.11) to prove (3.32e) in the following:

$$\begin{aligned} &(W_1(:,i))^T V(:,i) + (W_2(:,i))^T V_1(:,i) \\ &\quad = (W(:,i))^T V(:,i) - (W_2(:,i))^T V_2(:,i) \\ &\quad = \left(W(W^T V)^{-T}\left(\widehat{W}(:,i) - \Gamma_w(:\,.i)\right)\right)^T V \left(\widehat{V}(:,i) - \Gamma_v(:,i)\right) \\ &\qquad - (W_2(:,i))^T V_2(:,i) \\ &\quad = \left(\widehat{W}(:,i) - \Gamma_w(:\,.i)\right)^T \left(\widehat{V}(:,i) - \Gamma_v(:,i)\right) - (W_2(:,i))^T V_2(:,i) \end{aligned}$$

$$
\begin{aligned}
&= \left(\widehat{W}(:,i)\right)^T \widehat{V}(:,i) - \left(\widehat{W}(:,i)\right)^T \Gamma_v(:,i) - (\Gamma_w(:\,.i))^T \left(\widehat{V}(:,i) - \Gamma_v(:,i)\right) \\
&\quad - (W_2(:,i))^T V_2(:,i) \\
&= (\widehat{W}_1(:,i))^T \widehat{V}(:,i) + \left(\widehat{W}_2(:,i)\right)^T \widehat{V}_1(:,i) + \epsilon_\lambda^{(i)},
\end{aligned}
$$

where

$$
\begin{aligned}
\epsilon_\lambda^{(i)} = &- \left(\widehat{W}(:,i)\right)^T \Gamma_v(:,i) - (\Gamma_w(:\,.i))^T \left(\widehat{V}(:,i) - \Gamma_v(:,i)\right) \\
&- (W_2(:,i))^T V_2(:,i) + (\widehat{W}_2(:,i))^T \widehat{V}_2(:,i).
\end{aligned}
$$

This completes the proof.

Acknowledgments. The authors would like to thank Dr. Tobias Breiten for providing the numerical examples, and MATLAB implementations for one-sided and two-sided interpolatory projection methods. We would also like to thank the anonymous referees for their constructive comments.

REFERENCES

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

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

[3] P. Astrid, S. Weiland, K. Willcox, and T. Backx, *Missing point estimation in models described by proper orthogonal decomposition*, IEEE Trans. Automat. 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] R. H. Bartels and G. W. Stewart, *Solution of the matrix equation $AX+XB = C$*, Commun. ACM, 15 (1972), pp. 820–826.

[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] C. A. Beattie and S. Gugercin, *A trust region method for optimal H_2-model reduction*, in Proceedings of the Joint 48th IEEE Conference on Decision and Control, and 28th Chinese Control Conference, IEEE, 2009, pp. 5370–5375.

[8] P. Benner and T. Breiten, *Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems*, SIAM J. Matrix Anal. Appl., 33 (2012), pp. 859–885.

[9] 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/.

[10] P. Benner and T. Breiten, *Two-sided projection method for nonlinear model reduction*, SIAM J. Sci. Comput., 37 (2015), pp. B239–B260, https://doi.org/10.1137/14097255X.

[11] P. Benner and T. Damm, *Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems*, SIAM J. Control Optim., 49 (2011), pp. 686–711, https://doi.org/10.1137/09075041X.

[12] P. Benner and P. Goyal, *Multipoint interpolation of Volterra series and $\mathcal{H}_2$-model reduction for a family of bilinear descriptor systems*, Systems Control Lett., 97 (2016), pp. 1–11.

[13] P. Benner and P. Goyal, *Balanced Truncation Model Order Reduction for Quadratic-Bilinear Control Systems*, e-prints, arXiv:1705.00160, https://arxiv.org/abs/1705.00160 [math.OC], 2017.

[14] 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.

[15] P. Benner, V. Mehrmann, and D. C. Sorensen, *Dimension Reduction of Large-Scale Systems*, Lect. Notes Comput. Sci. Eng. 45, Springer-Verlag, Berlin/Heidelberg, Germany, 2005.

[16] 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-Mitt., 36 (2013), pp. 32–52.

[17] P. Benner, E. Sachs, and S. Volkwein, *Model order reduction for PDE constrained optimization*, in Trends in PDE Constrained Optimization, Internat. Ser. Numer. Math. 165, G. Leugering, P. Benner, S. Engell, A. Griewank, H. Harbrecht, M. Hinze, R. Rannacher, and S. Ulbrich, eds., Springer International Publishing, Birkhäuser, Cham, 2014, pp. 303–326, https://doi.org/10.1007/978-3-319-05083-6_19.

[18] 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.

[19] S. Chaturantabut and D. C. Sorensen, *Nonlinear model reduction via discrete empirical interpolation*, SIAM J. Sci. Comput., 32 (2010), pp. 2737–2764.

[20] M. Condon and R. Ivanov, *Nonlinear systems-algebraic gramians and model reduction*, COMPEL, 24 (2005), pp. 202–219.

[21] G. Flagg, C. Beattie, and S. Gugercin, *Convergence of the iterative rational krylov algorithm*, Systems Control Lett., 61 (2012), pp. 688–691.

[22] G. Flagg and S. Gugercin, *Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems*, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 549–579.

[23] 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.

[24] 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.

[25] S. Gugercin, T. Stykel, and S. Wyatt, *Model reduction of descriptor systems by interpolatory projection methods*, SIAM J. Sci. Comput., 35 (2013), pp. B1010–B1033, https://doi.org/10.1137/130906635.

[26] 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.

[27] H. V. Henderson and S. R. Searle, *The vec-permutation matrix, the vec operator and Kronecker products: A review*, Linear Multilinear Algebra, 9 (1981), pp. 271–288.

[28] 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. C. Sorensen, eds., Springer-Verlag, Berlin/Heidelberg, Germany, 2005, pp. 261–306.

[29] M. Hinze and S. Volkwein, *Error estimates for abstract linear-quadratic optimal control problems using proper orthogonal decomposition*, Comput. Optim. Appl., 39 (2008), pp. 319–345.

[30] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1990.

[31] M. Köhler, *On the closest stable descriptor system in the respective spaces RH_2 and RH_∞*, Linear Algebra Appl., 443 (2014), pp. 34–49.

[32] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Rev., 51 (2009), pp. 455–500.

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

[34] P. Kunkel and V. Mehrmann, *Differential-algebraic equations: Analysis and numerical solution*, European Mathematical Society, Zürich, Switzerland, 2006.

[35] 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.

[36] B. C. Moore, *Principal component analysis in linear systems: Controllability, observability, and model reduction*, IEEE Trans. Automat. Control, AC-26 (1981), pp. 17–32.

[37] J. R. Phillips, *Projection frameworks for model reduction of weakly nonlinear systems*, in Proceedings 37th Design Automation Conference, Los Angeles, CA, 2000, pp. 184–189.

[38] 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.

[39] M. J. Rewieński, *A Trajectory Piecewise-Linear Approach to Model Order Reduction of Nonlinear Dynamical Systems*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 2003.

[40] C. W. Rowley, *Model reduction for fluids, using balanced proper orthogonal decomposition*, Int. J. Bifurcation Chaos Appl. Sci. Eng., 15 (2005), pp. 997–1013.

[41] W. J. Rugh, *Nonlinear System Theory*, The Johns Hopkins University Press, Baltimore, MD, 1981.

[42] J. Saak, M. Köhler, and P. Benner, *M-M.E.S.S.-1.0.1–The Matrix Equations Sparse Solvers library*. doi:10.5281/zenodo.50575, 2016; see also: www.mpi-magdeburg.mpg.de/projects/mess.

[43] 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.

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

[45] P. van Dooren, K. A. Gallivan, and P.-A. Absil, *$\mathcal{H}_2$-optimal model reduction with higher-order poles*, SIAM J. Matrix Anal. Appl., 31 (2010), pp. 2738–2753.

[46] K. Willcox and J. Peraire, *Balanced model reduction via the proper orthogonal decomposition*, AIAA J., 40 (2002), pp. 2323–2330.

[47] D. Wilson, *Optimum solution of model-reduction problem*, Proc. IEE, 117 (1970), pp. 1161–1165.

[48] L. Zhang and J. Lam, *On H_2 model reduction of bilinear systems*, Automatica J. IFAC, 38 (2002), pp. 205–216.