$\mathcal{H}_2$ optimal model reduction of linear systems with multiple quadratic outputs

Sean Reiter[†] Igor Pontes Duff[⋆] Ion Victor Gosea[*] Serkan Gugercin[‡]

[†] *Department of Mathematics, Virginia Tech, Blacksburg, VA 24061, USA.*
Email: seanr7@vt.edu, ORCID: 0000-0002-7510-1530
[⋆] *Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany.*
Email: pontes@mpi-magdeburg.mpg.de, ORCID: 0000-0001-6433-6142
[*] *Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106 Magdeburg, Germany.*
Email: gosea@mpi-magdeburg.mpg.de, ORCID: 0000-0003-3580-4116
[‡] *Department of Mathematics and Division of Computational Modeling and Data Analytics, Academy of Data Science, Virginia Tech, Blacksburg, VA 24061, USA.*
Email: gugercin@vt.edu, ORCID: 0000-0003-4564-5999

Abstract: In this work, we consider the $\mathcal{H}_2$ optimal model reduction of dynamical systems that are linear in the state equation and up to quadratic nonlinearity in the output equation. As our primary theoretical contributions, we derive gradients of the squared $\mathcal{H}_2$ system error with respect to the reduced model quantities and, from the stationary points of these gradients, introduce Gramian-based first-order necessary conditions for the $\mathcal{H}_2$ optimal approximation of a linear quadratic output (LQO) system. The resulting $\mathcal{H}_2$ optimality framework neatly generalizes the analogous Gramian-based optimality framework for purely linear systems. Computationally, we show how to enforce the necessary optimality conditions using Petrov-Galerkin projection; the corresponding projection matrices are obtained from a pair of Sylvester equations. Based on this result, we propose an iteratively corrected algorithm for the $\mathcal{H}_2$ model reduction of LQO systems, which we refer to as LQO-TSIA (linear quadratic output two-sided iteration algorithm). Numerical examples are included to illustrate the effectiveness of the proposed computational method against other existing approaches.

Keywords: linear quadratic output systems, $\mathcal{H}_2$ optimal, model-order reduction, Gramians, two-sided iteration, optimality conditions, gradients

Mathematics subject classification: 15A24, 46N10, 49K15, 93A15 93C10, 93C80,

1 Introduction

The focus of this work is a class of weakly nonlinear dynamical systems; those that are linear time-invariant in the state equation with up to quadratic terms in the output equation. We refer to these systems as *linear quadratic output* (LQO) systems. Dynamical systems with quadratic output functions arise naturally whenever one's interests lie in observing quantities computed as the product of time or frequency domain components of the state [27]. Examples include applications where energy or power are considered; e.g., the internal energy functional of a system [18, 21], or the objective cost function in optimal quadratic control problems [13]. Other examples include observables that capture the variance or deviation of the state coordinates from a point of reference; for instance, the root mean squared displacement of several spatial coordinates about a point of excitation [4, 24, 27], or the variance of a random variable in stochastic modeling [22].

The accurate modeling of complex physical phenomena often requires dynamical systems having a very large state-space dimension, e.g., on the order of 10^6 or even greater.

In this large-scale setting, direct calculations involving the full-order model (FOM) are prohibitively costly, and system approximation becomes desirable. *Model order reduction* (MOR) is the procedure by which one approximates a large-scale dynamical system with a comparatively low-order surrogate model. The resultant *reduced-order model* (ROM) should capture the significant input-to-output response characteristics of the original system in an appropriate sense, and can thus be used as a computationally efficient and faithful proxy for the FOM in applications of, e.g., prediction and simulation, or control.

In the *purely* linear setting (that is, both the internal dynamics *and* output equation depend only linearly on the state coordinates) there are a variety of model reduction techniques at one's disposal; we refer the reader to [1,2,9,10] and to the collection of references therein for an overview of approximation methods for linear dynamical systems. In recent years, there has been an increased amount of study dedicated towards the model reduction of LQO systems; cf., [7, 13, 15, 16, 20–22, 24–27]. Some of these works tackle the LQO-MOR problem by rewriting the governing equations of a multi-input, single-output LQO system as a *wholly linear* multi-input, multi-output system that emulates the quadratic quantity of interest via multiple linear outputs [26, 27]. Any of the well-known model reduction techniques available for linear systems can then be applied to the intermediate model in order to determine suitable projection subspaces. However, a drawback of this approach is that it will produce an intermediate model with a very large number of outputs; this is undesirable in the design of reduced models since systems with large input/output dimensions tend to have slowly decaying Hankel singular values, and are thus more difficult to approximate [3].

Other works [7, 13, 15, 16, 20, 21, 25] devise approaches that leverage the quadratic output structure *directly* (that is, without any intermediate lifting or linearization) in order to determine suitable approximation subspaces for order reduction. Notably, the work [7] introduces a so-called *quadratic output system Gramian* based upon the nonlinear state-to-output kernels of an LQO system, and related $\mathcal{H}_2$ system norm. These authors propose a balanced truncation model reduction algorithm that balances the linear reachability and quadratic output observability Gramians; the approach guarantees asymptotic stability and provides an *a posteriori* $\mathcal{H}_2$ error bound. Extensions of this methodology to handle systems of differential-algebraic equations and frequency or time-limited model reduction were proposed in [20] and [25], respectively. The works [13,15,16,24] all consider model reduction based on the rational interpolation of transfer functions in the frequency domain.

The focus of this paper is the $\mathcal{H}_2$ *optimal model reduction problem* for LQO systems. Our interest in the $\mathcal{H}_2$ model reduction problem is motivated by the fact that the $\mathcal{H}_2$ system error provides an upper bound on the $\mathcal{L}_\infty$ output error [7, Theorem 3.4]. So, if the design objective in order reduction is ensuring that the 'worst case' error in the approximate output is uniformly small over all inputs, then one should aim to make the $\mathcal{H}_2$ system error as small as possible. Outside of [7], use of the $\mathcal{H}_2$ norm as a direct performance measure in LQO-MOR has not been considered in the current literature. For fully linear systems, the $\mathcal{H}_2$ model reduction problem is well studied; cf., [2, 17, 19, 28–30], and the references therein. $\mathcal{H}_2$ model reduction has also been considered for other classes of weakly nonlinear dynamical systems; see, e.g., [5, 6, 14, 32]. The work [15] proposes an algorithm heuristically motivated by *linear* $\mathcal{H}_2$ optimal model reduction, but without an explicit connection to the $\mathcal{H}_2$ optimal model reduction of LQO systems. Even in the linear setting, finding a global minimizer of the $\mathcal{H}_2$ error is a non-convex optimization problem. As a consequence, the most common practice in the $\mathcal{H}_2$ landscape typically relies on the identification of ROMs that satisfy *first-order necessary conditions* (FONCs) for *local* $\mathcal{H}_2$ optimality. The two most well-known optimality frameworks for $\mathcal{H}_2$ model reduction of fully linear systems are derived from the interpolation-based FONCs of Meier and Leunberger [17, 19], and the Gramian-based FONCs of Wilson [28, 29]. The major theoretical result of our work is the establishment of *Gramian-based* FONCs for the $\mathcal{H}_2$ optimal model reduction of LQO systems. Namely, our significant contributions are:

1. Theorem 3.1, which derives gradients of the squared $\mathcal{H}_2$ system error with respect to the system matrices of the LQO-ROM as parameters. The stationary points of these gradients directly yield Gramian-based FONCs for $\mathcal{H}_2$ optimality, which are presented in Theorem 3.2. These results generalize the analogous Gramian-based FONCs for linear $\mathcal{H}_2$ optimal model [28, 29] to the LQO setting.

2. Also in Theorem 3.2, we show that a $\mathcal{H}_2$ optimal LQO-ROM is necessarily defined by Petrov-Galerkin projection. The relevant projection matrices are obtainable as solutions to a pair of Sylvester equations.

3. Based on this theoretical optimality framework, we propose an iteratively corrected algorithm for $\mathcal{H}_2$ optimal LQO-MOR in Algorithm 1; we call this the linear quadratic output two-sided iteration algorithm (LQO-TSIA). The algorithm produces (upon convergence) LQO-ROMs that satisfy the Gramian-based FONCs from Theorem 3.2.

The particular organization of the manuscript is as follows: In Section 2, we review the necessary systems theory background of linear systems with quadratic output functions and introduce several different characterizations of the $\mathcal{H}_2$ system norm. These characterizations are central to the theoretical results developed in Section 3, where we present the main theoretical contributions of this work. These are the derivation of gradients of the squared $\mathcal{H}_2$ system error and resulting Gramian-based FONCs for $\mathcal{H}_2$ optimal LQO-MOR. Section 4 presents LQO-TSIA, an iterative computational algorithm for LQO-MOR based on the previously established $\mathcal{H}_2$ optimality framework. The algorithm requires the solutions to a pair of Sylvester equations. To validate the proposed approach, in Section 5 we test our method on an example where a quadratic output naturally occurs from the discretization of a quadratic cost function. Section 6 concludes the work and looks toward future research endeavors.

2 Background and preliminaries

2.1 Linear systems with quadratic outputs

Throughout this work, we consider multi-input, multi-output (MIMO) dynamical systems that are linear in the state equation with a quadratic output (QO) term. In its state-space formulation, such a system is described by the following equations

$$\mathcal{S}: \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), & \mathbf{x}(0) = \mathbf{0}, \\ \mathbf{y}(t) = \underbrace{\mathbf{C}\mathbf{x}(t)}_{:=\mathbf{y}_1(t)} + \underbrace{\begin{bmatrix} \mathbf{x}(t)^\mathsf{T}\mathbf{M}_1\mathbf{x}(t) \\ \vdots \\ \mathbf{x}(t)^\mathsf{T}\mathbf{M}_p\mathbf{x}(t) \end{bmatrix}}_{:=\mathbf{y}_2(t)}, \end{cases} \tag{1}$$

where $\mathbf{A}$, $\mathbf{M}_1, \ldots, \mathbf{M}_p \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, and $\mathbf{C} \in \mathbb{R}^{p\times n}$. In (1) $\mathbf{x}\colon [0,\infty) \to \mathbb{R}^n$ contains the state coordinates; $\mathbf{u}\colon [0,\infty) \to \mathbb{R}^m$ and $\mathbf{y}\colon [0,\infty) \to \mathbb{R}^p$ are the inputs and outputs of the system, respectively. We assume henceforth that $\mathcal{S}$ is *asymptotically stable*, i.e., the eigenvalues of $\mathbf{A}$ denoted $\lambda(\mathbf{A})$ lie in the open left half of the complex plane, and without loss of generality that $\mathbf{M}_k = \mathbf{M}_k^\mathsf{T}$ for all $k = 1, \ldots, p$. We refer to a system of the form (1) as a *linear quadratic output* (LQO) system; we represent such a system using the notation $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{M}_1, \ldots, \mathbf{M}_p)$. The quadratic output term $\mathbf{y}_2(t)$ in (1) can be re-written via a Kronecker product of the state

$$\mathbf{y}_2(t) = \mathbf{M}\left(\mathbf{x}(t) \otimes \mathbf{x}(t)\right), \quad \mathbf{M} = \begin{bmatrix} \operatorname{vec}(\mathbf{M}_1)^\mathsf{T} \\ \vdots \\ \operatorname{vec}(\mathbf{M}_p)^\mathsf{T} \end{bmatrix} \in \mathbb{R}^{p\times n^2}, \tag{2}$$

where $\operatorname{vec}\colon \mathbb{R}^{n\times n} \to \mathbb{R}^{n^2}$ denotes the *vectorization* operator. For compactness and ease of theoretical development, we switch freely between these representations as necessary. Depending on the application of interest, in (1) we allow for $\mathbf{C} = \mathbf{0}_{p\times n}$, the $p \times n$ zero matrix, in which case the output equation is *purely* quadratic. In this instance, we denote the LQO system in (1) by $\mathcal{S}_2$ since only $\mathbf{y}_2$ appears in the output $\mathbf{y}$. On the other hand, if $\mathbf{M}_k = \mathbf{0}_{n\times n}$ for all $k = 1, \ldots, p$, the system (1) reduces to a standard linear time-invariant (LTI) system. We denote the system (1) by $\mathcal{S}_1$ in this instance.

Remark 2.1. *We illustrate why we may assume without loss of generality that $\mathbf{M}_k$ is symmetric for each $k = 1, \ldots, p$. For an arbitrary (not necessarily symmetric) matrix $\mathbf{M}_k \in \mathbb{R}^{n\times n}$, the associated quadratic output in (1) can always be expressed as*

$$\mathbf{x}(t)^\mathsf{T}\mathbf{M}_k\mathbf{x}(t) = \mathbf{x}(t)^\mathsf{T}\left(\frac{1}{2}\left(\mathbf{M}_k + \mathbf{M}_k^\mathsf{T}\right)\right)\mathbf{x}(t).$$

In other words, $\mathbf{M}_k$ can always be replaced by its symmetric part $\frac{1}{2}\left(\mathbf{M}_k + \mathbf{M}_k^\mathsf{T}\right)$.

In practical applications, the state dimension n of the system in (1) may be large enough so that any repeated computation involving the full-order model is prohibitively expensive. Model order reduction (MOR) seeks to replace the original large-scale model with a lower dimensional reduced-order model (ROM) that can be used as a cheap-to-evaluate surrogate in numerical simulation, devising control laws, solving optimization tasks, etc. In this work, we consider the construction of reduced models that retain the LQO structure, i.e.,

$$\mathcal{S}_r: \begin{cases} \dot{\mathbf{x}}_r(t) = \mathbf{A}_r\mathbf{x}_r(t) + \mathbf{B}_r\mathbf{u}(t), \\ \mathbf{y}_r(t) = \mathbf{C}_r\mathbf{x}_r(t) + \begin{bmatrix} \mathbf{x}(t)^\mathsf{T}\mathbf{M}_{1,r}\mathbf{x}(t) \\ \vdots \\ \mathbf{x}(t)^\mathsf{T}\mathbf{M}_{p,r}\mathbf{x}(t) \end{bmatrix}, \end{cases} \tag{3}$$

where the matrices $\mathbf{A}_r, \mathbf{M}_{1,r}, \ldots, \mathbf{M}_{p,r} \in \mathbb{R}^{r\times r}$, $\mathbf{B}_r \in \mathbb{R}^{r\times m}$, and $\mathbf{C}_r \in \mathbb{R}^{p\times r}$ are of a significantly reduced dimension $1 \le r \ll n$; $\mathbf{x}_r\colon [0,\infty) \to \mathbb{R}^r$ is the reduced state vector,

and $\mathbf{y}_r\colon [0,\infty) \to \mathbb{R}^p$ is the approximate output. We would like the surrogate model (3) to be a good approximation in the sense that it accurately recreates the input-to-output response of the original full-order system in (1). In other words, the reduced output $\mathbf{y}_r$ should be a faithful replication of $\mathbf{y}$ in the sense that $\|\mathbf{y}-\mathbf{y}_r\|$ is small in an appropriate norm $\|\cdot\|$ over a range of admissible inputs $\mathbf{u}$.

The general framework we consider is that of model reduction using *projection*. Consider a linear quadratic output system as in (1). In projection-based model reduction, the problem resolves to choosing left and right reduction bases $\mathbf{W}_r \in \mathbb{R}^{n\times r}$ and $\mathbf{V}_r \in \mathbb{R}^{n\times r}$ so that: $\mathbf{x} \approx \mathbf{V}_r\mathbf{x}_r$, the matrix $\mathbf{W}_r^{\mathsf{T}}\mathbf{V}_r$ is nonsingular, and the Petrov-Galerkin condition

$$\mathbf{W}_r{}^{\mathsf{T}}\left(\mathbf{V}_r\dot{\mathbf{x}}_r(t) - \mathbf{A}\mathbf{V}_r\mathbf{x}_r(t) - \mathbf{B}\mathbf{u}(t)\right) = 0,$$

is satisfied. The resulting order-r LQO reduced model of the form (3) is determined by the reduced order matrices

$$\begin{aligned} \mathbf{A}_r &= \left(\mathbf{W}_r^{\mathsf{T}}\mathbf{V}_r\right)^{-1}\mathbf{W}_r^{\mathsf{T}}\mathbf{A}\mathbf{V}_r &&\in \mathbb{R}^{r\times r},\\ \mathbf{B}_r &= \left(\mathbf{W}_r^{\mathsf{T}}\mathbf{V}_r\right)^{-1}\mathbf{W}_r^{\mathsf{T}}\mathbf{B} &&\in \mathbb{R}^{r\times m},\\ \mathbf{C}_r &= \mathbf{C}\mathbf{V}_r &&\in \mathbb{R}^{p\times r},\\ \text{and}\quad \mathbf{M}_{k,r} &= \mathbf{V}_r^{\mathsf{T}}\mathbf{M}_k\mathbf{V}_r &&\in \mathbb{R}^{r\times r}. \end{aligned} \tag{4}$$

The approximation quality of the reduced model hinges upon the underlying projection subspaces $\operatorname{span}(\mathbf{W}_r)$ and $\operatorname{span}(\mathbf{V}_r)$, not the particular bases $\mathbf{W}_r$ and $\mathbf{V}_r$ used in (4). So, $\mathbf{W}_r$ and $\mathbf{V}_r$ are typically replaced by orthonormal matrices to avoid ill-conditioning and singularities in the reduced model. Following Remark 2.1, we also assume that $\mathbf{M}_{k,r}$ is symmetric for each k going forward.

2.2 The $\mathcal{H}_2$ system norm

The nonlinearity in (1) is entirely captured by the quadratic term in the output equation; the state equations evolve linearly in $\mathbf{x}$. So, the time-domain input-to-output map modeled by the LQO system in (1) is entirely characterized by *two* Volterra kernels. For any input function $\mathbf{u}$ and $t \geq 0$, the corresponding output $\mathbf{y}$ of (1) can be written as

$$\begin{aligned} \mathbf{y}(t) &= \int_0^t \mathbf{h}_1(\tau)\mathbf{u}(t-\tau)\,d\tau\\ &+ \int_0^t\int_0^t \mathbf{h}_2(\tau_1,\tau_2)\left(\mathbf{u}(t-\tau_1)\otimes\mathbf{u}(t-\tau_2)\right)d\tau_1 d\tau_2. \end{aligned} \tag{5}$$

The Volterra kernels $\mathbf{h}_1\colon [0,\infty) \to \mathbb{R}^{p\times m}$ and $\mathbf{h}_2\colon [0,\infty)\times[0,\infty) \to \mathbb{R}^{p\times m^2}$ are defined as

$$\begin{aligned} \mathbf{h}_1(t) &= \mathbf{C}e^{\mathbf{A}t}\mathbf{B},\\ \mathbf{h}_2(t_1,t_2) &= \mathbf{M}\left(e^{\mathbf{A}t_1}\mathbf{B}\otimes e^{\mathbf{A}t_2}\mathbf{B}\right), \end{aligned} \tag{6}$$

where $\mathbf{M} \in \mathbb{R}^{p\times n^2}$ is given according to (2). Here, the univariate kernel $\mathbf{h}_1(t)$ and convolution in (5) describe the purely *linear* term in the output; that is, $\mathbf{y}_1(t) = \mathbf{C}\mathbf{x}(t)$. The bivariate kernel $\mathbf{h}_2(t_1,t_2)$ and convolution in (5) describe the purely *quadratic* term in the output; i.e., $\mathbf{y}_2(t) = \mathbf{M}\left(\mathbf{x}(t)\otimes\mathbf{x}(t)\right)$. This description of $\mathcal{S}$ can be directly obtained from the governing equations (1); see [7, Section 4] for a derivation. From the Volterra kernels in (5), an $\mathcal{H}_2$ inner product and the corresponding norm for LQO systems can be defined [7, Definifiton 3.1].

Definition 2.1. *Let $\mathcal{S}$ and $\mathcal{S}_r$ be asymptotically stable LQO systems as in (1) and (3) having the kernels $\mathbf{h}_1(t)$, $\mathbf{h}_{1,r}(t)$ and $\mathbf{h}_2(t_1,t_2)$, $\mathbf{h}_{2,r}(t_1,t_2)$ defined according to (6), respectively. The $\mathcal{H}_2$* inner product *of $\mathcal{S}$ and $\mathcal{S}_r$ is*

$$\begin{aligned} \langle\mathcal{S},\mathcal{S}_r\rangle_{\mathcal{H}_2} &:= \int_0^\infty \operatorname{tr}\left(\mathbf{h}_1(\tau)\mathbf{h}_{1,r}(\tau)^{\mathsf{T}}\right)d\tau\\ &+ \int_0^\infty\int_0^\infty \operatorname{tr}\left(\mathbf{h}_2(\tau_1,\tau_2)\mathbf{h}_{2,r}(\tau_1,\tau_2)^{\mathsf{T}}\right)d\tau_1 d\tau_2, \end{aligned} \tag{7}$$

where $\operatorname{tr}(\cdot)$ denotes the trace of a matrix. Moreover, if $\mathcal{S} = \mathcal{S}_r$, the expression in (7) defines the $\mathcal{H}_2$ norm *of $\mathcal{S}$:*

$$\begin{aligned} \|\mathcal{S}\|_{\mathcal{H}_2}^2 &:= \int_0^\infty \|\mathbf{h}_1(\tau)\|_{\mathsf{F}}^2 d\tau\\ &+ \int_0^\infty\int_0^\infty \|\mathbf{h}_2(\tau_1,\tau_2)\|_{\mathsf{F}}^2 d\tau_1 d\tau_2, \end{aligned} \tag{8}$$

where $\|\cdot\|_{\mathsf{F}}$ denotes the matrix Frobenius norm.

If $\mathbf{M}_k$, for $k = 1,\ldots,p$, in (2) are identically zero, then $\mathcal{S} = \mathcal{S}_1$ is purely an LTI system with $\mathbf{h}_2 = \mathbf{0}_{p\times m^2}$, and so (7) and (8) agree with the usual $\mathcal{H}_2$ inner product and norm defined for linear dynamical systems; cf. [1, Section 5.1]. Lastly, it follows from Definition 2.1 that the $\mathcal{H}_2$ norm of an asymptotically stable LQO system is finite.

Next, we consider an alternative and more computationally tractable characterization of the $\mathcal{H}_2$ system norm in Definition 2.1 based on the *Gramians* of an LQO system. These formulations will be paramount in deriving gradients of the squared $\mathcal{H}_2$ system error and associated $\mathcal{H}_2$ optimality conditions in Section 3.

2.2.1 The quadratic output observability Gramian

Because the nonlinearity in (1) is limited to the output equation, the input-to-state map is identical to that of the related LTI system $\mathcal{S}_1$ having the same state, input, and linear output matrices, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^n$, and $\mathbf{C} \in \mathbb{R}^{p\times n}$,

with $\mathbf{M}_k = \mathbf{0}_{n\times n}$ for all k. Thus, the *reachability Gramian* $\mathbf{P} \in \mathbb{R}^{n\times n}$ of the LQO system $\mathcal{S}$ in (1) is the same as the classical reachability Gramian defined for purely LTI systems [1, Section 4.3], i.e.

$$\mathbf{P} = \int_0^\infty e^{\mathbf{A}\tau}\mathbf{B}\left(e^{\mathbf{A}\tau}\mathbf{B}\right)^\mathsf{T} d\tau. \tag{9}$$

Under the assumption that $\mathcal{S}$ is asymptotically stable, $\mathbf{P}$ is unique [1, Prop. 6.2], and can be computed as the solution of the Lyapunov equation

$$\mathbf{A}\mathbf{P} + \mathbf{P}\mathbf{A}^\mathsf{T} + \mathbf{B}\mathbf{B}^\mathsf{T} = \mathbf{0}. \tag{10}$$

A *quadratic output observability Gramian* based on the nonlinear state-to-output map of an LQO system was introduced in [7]. In a sense, the QO observability Gramian generalizes the classical observability Gramian of an LTI system [1, Section 4.3]. Consider an LQO system $\mathcal{S}$ as in (1) and define the intermediate matrices $\mathbf{Q}_1 \in \mathbb{R}^{n\times n}$ and $\mathbf{Q}_2^{(k)} \in \mathbb{R}^{n\times n}$ by

$$\mathbf{Q}_1 \coloneqq \int_0^\infty e^{\mathbf{A}^\mathsf{T}\tau}\mathbf{C}^\mathsf{T}\left(e^{\mathbf{A}^\mathsf{T}\tau}\mathbf{C}^\mathsf{T}\right)^\mathsf{T} d\tau, \tag{11}$$

$$\begin{aligned}\mathbf{Q}_2^{(k)} \coloneqq \int_0^\infty \int_0^\infty e^{\mathbf{A}^\mathsf{T}\tau_1}\mathbf{M}_k e^{\mathbf{A}\tau_2}\mathbf{B}\times \\ \left(e^{\mathbf{A}^\mathsf{T}\tau_1}\mathbf{M}_k e^{\mathbf{A}\tau_2}\mathbf{B}\right)^\mathsf{T} d\tau_1 d\tau_2,\end{aligned} \tag{12}$$

for each $k = 1, \dots p$. (Note that $\mathbf{Q}_1 \in \mathbb{R}^{n\times n}$ is just the classical observability Gramian of the linear system $\mathcal{S}_1$.) Then the QO observability Gramian $\mathbf{Q} \in \mathbb{R}^{n\times n}$ of $\mathcal{S}$ is defined as

$$\mathbf{Q} \coloneqq \mathbf{Q}_1 + \sum_{k=1}^p \mathbf{Q}_2^{(k)}. \tag{13}$$

Akin to the linear system Gramians, if $\mathcal{S}$ is asymptotically stable, then $\mathbf{Q}$ is the unique solution to the following Lyapunov equation [7, Section 4.1]

$$\mathbf{A}^\mathsf{T}\mathbf{Q} + \mathbf{Q}\mathbf{A} + \mathbf{C}^\mathsf{T}\mathbf{C} + \sum_{k=1}^p \mathbf{M}_k\mathbf{P}\mathbf{M}_k = \mathbf{0}, \tag{14}$$

where $\mathbf{P}$ is the reachability Gramian of $\mathcal{S}$ according to (9). See [7, Section 4.1] for a more detailed derivation of $\mathbf{Q}$. A balanced truncation model reduction algorithm based upon the balancing of energy functionals defined by the Gramians in (9) and (13) was proposed in [7].

It was proven in [7, Proposition 3.3] that the QO observability Gramian $\mathbf{Q}$ in (13) can be used to compute the $\mathcal{H}_2$ norm (8) of an LQO system. The $\mathcal{H}_2$ inner product of two LQO systems (7) can similarly be computed using the solution to a Sylvester equation. Next, we provide a proof of this result for LQO systems with multiple quadratic outputs and also show that the $\mathcal{H}_2$ inner product and norm for linear quadratic output systems in Definition 2.1 can be calculated from the reachability Gramian $\mathbf{P}$ in (9) and quadratic output matrices $\mathbf{M}_k$. The result [7, Prop. 3.3] proves the formulae (17) and (19) in Theorem 2.1 for the special case of a single quadratic output with $\mathbf{C} = \mathbf{0}_{p\times n}$. A related formula was independently proven in the recent work [20, Lemma 5.2] for systems of differential-algebraic equations with quadratic output functions.

Theorem 2.1. *Let $\mathcal{S}$ and $\mathcal{S}_r$ be asymptotically stable LQO systems as in (1) and (3), respectively. Let $\mathbf{X} \in \mathbb{R}^{n\times r}$ and $\mathbf{Z} \in \mathbb{R}^{n\times r}$ denote the unique solutions to the following Sylvester equations*

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}_r^\mathsf{T} + \mathbf{B}\mathbf{B}_r^\mathsf{T} = \mathbf{0}, \tag{15}$$

$$\text{and}\quad \mathbf{A}^\mathsf{T}\mathbf{Z} + \mathbf{Z}\mathbf{A}_r - \sum_{k=1}^p \mathbf{M}_k\mathbf{X}\mathbf{M}_{k,r} - \mathbf{C}^\mathsf{T}\mathbf{C}_r = \mathbf{0}. \tag{16}$$

Then, the $\mathcal{H}_2$ inner product of $\mathcal{S}$ and $\mathcal{S}_r$ is given by

$$\langle \mathcal{S}, \mathcal{S}_r\rangle_{\mathcal{H}_2} = -\operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{Z}\mathbf{B}_r\right) \tag{17}$$

$$= \operatorname{tr}\left(\mathbf{C}\mathbf{X}\mathbf{C}_r^\mathsf{T}\right) + \sum_{k=1}^p \operatorname{tr}\left(\mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X}\mathbf{M}_{k,r}\right). \tag{18}$$

Moreover, if instead $\mathcal{S} = \mathcal{S}_r$, then $\mathbf{X} = \mathbf{P} \in \mathbb{R}^{n\times n}$ and $\mathbf{Z} = \mathbf{Q} \in \mathbb{R}^{n\times n}$ according to (9) and (13), and the $\mathcal{H}_2$ norm of $\mathcal{S}$ is given by

$$\begin{aligned}\|\mathcal{S}\|_{\mathcal{H}_2}^2 &= \operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{Q}_1\mathbf{B}\right) + \sum_{k=1}^p \operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{Q}_2^{(k)}\mathbf{B}\right) \\ &= \operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{Q}\mathbf{B}\right) \qquad (19)\\ &= \operatorname{tr}\left(\mathbf{C}\mathbf{P}\mathbf{C}^\mathsf{T}\right) + \sum_{k=1}^p \operatorname{tr}\left(\mathbf{P}\mathbf{M}_k\mathbf{P}\mathbf{M}_k\right). \qquad (20)\end{aligned}$$

Proof. Throughout this proof, take $\mathbf{h}_1(t)$, $\mathbf{h}_2(t_1, t_2)$ and $\mathbf{h}_{1,r}(t)$, $\mathbf{h}_{2,r}(t_1, t_2)$ to denote the Volterra kernels of $\mathcal{S}$ and $\mathcal{S}_r$ respectively according to (6). Consider the Sylvester equation

$$\mathbf{A}^\mathsf{T}\mathbf{Z}_1 + \mathbf{Z}_1\mathbf{A}_r - \mathbf{C}^\mathsf{T}\mathbf{C}_r = \mathbf{0}. \tag{21}$$

Because $\mathcal{S}$ and $\mathcal{S}_r$ are asymptotically stable, the spectra of $\mathbf{A}$ and $-\mathbf{A}_r$ are disjoint. Thus, the Sylvester equations (15)

and (21) have unique solutions [1, Prop. 6.2]. These solutions can be explicitly written as

$$\mathbf{X} = \int_0^\infty e^{\mathbf{A}\tau}\mathbf{B}\left(e^{\mathbf{A}_r\tau}\mathbf{B}_r\right)^\mathsf{T} d\tau,$$
$$\mathbf{Z}_1 = -\int_0^\infty e^{\mathbf{A}^\mathsf{T}\tau}\mathbf{C}^\mathsf{T}\left(e^{\mathbf{A}_r^\mathsf{T}\tau}\mathbf{C}_r^\mathsf{T}\right)^\mathsf{T} d\tau.$$

By the same argument, for each $k = 1, \ldots, p$, the Sylvester equation

$$\mathbf{A}^\mathsf{T}\mathbf{Z}_k + \mathbf{Z}_k\mathbf{A}_r - \mathbf{M}_k\mathbf{X}\mathbf{M}_{k,r} = \mathbf{0} \tag{22}$$

has a unique solution of the form

$$\mathbf{Z}_2^{(k)} = -\int_0^\infty e^{\mathbf{A}^\mathsf{T}\tau}\mathbf{M}_k\mathbf{X}\left(e^{\mathbf{A}_r^\mathsf{T}\tau}\mathbf{M}_{k,r}\right)^\mathsf{T} d\tau.$$

Summing up the equations (21) and (22) over all k yields (16). By uniqueness, the solution $\mathbf{Z} \in \mathbb{R}^{n\times r}$ to (16) is such that

$$\mathbf{Z} = \mathbf{Z}_1 + \sum_{k=1}^p \mathbf{Z}_2^{(k)}.$$

Plugging the integral form of $\mathbf{X}$ into $\mathbf{Z}_2^{(k)}$ for each k and moving terms inside the integrand, $\mathbf{Z}$ can be expressed as

$$\begin{aligned}\mathbf{Z} = &-\int_0^\infty e^{\mathbf{A}^\mathsf{T}\tau}\mathbf{C}^\mathsf{T}\left(e^{\mathbf{A}_r^\mathsf{T}\tau}\mathbf{C}_r^\mathsf{T}\right)^\mathsf{T} d\tau \\ &-\sum_{k=1}^p \int_0^\infty\int_0^\infty \left(e^{\mathbf{A}^\mathsf{T}\tau_1}\mathbf{M}_k e^{\mathbf{A}\tau_2}\mathbf{B}\right)\times \\ &\qquad \left(e^{\mathbf{A}_r^\mathsf{T}\tau_1}\mathbf{M}_{k,r}e^{\mathbf{A}_r\tau_2}\mathbf{B}_r\right)^\mathsf{T} d\tau_2 d\tau_1.\end{aligned}$$

Then, by the invariance of the trace under cyclic permutation we have

$$\begin{aligned}\operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{Z}\mathbf{B}_r\right) = &-\int_0^\infty \operatorname{tr}\left(\mathbf{C}e^{\mathbf{A}\tau}\mathbf{B}\left(\mathbf{C}_r e^{\mathbf{A}_r\tau}\mathbf{B}_r\right)^\mathsf{T}\right) d\tau \\ &-\sum_{k=1}^p \int_0^\infty\int_0^\infty \operatorname{tr}\Big(\mathbf{B}^\mathsf{T}e^{\mathbf{A}^\mathsf{T}\tau_1}\mathbf{M}_k e^{\mathbf{A}\tau_2}\mathbf{B}\times \\ &\qquad \left(\mathbf{B}_r^\mathsf{T}e^{\mathbf{A}_r^\mathsf{T}\tau_1}\mathbf{M}_{k,r}e^{\mathbf{A}_r\tau_2}\mathbf{B}_r\right)^\mathsf{T}\Big) d\tau_2 d\tau_1.\end{aligned}$$

Note that the first term in the expression above is precisely the first term in the $\mathcal{H}_2$ inner product (7). It remains to be shown that the remaining terms in the summand over k equate to the second term in the expression for the inner product (7). To show this, we observe that by construction of $\mathbf{M}$ in (2) and properties of the Kronecker product [11], the bivariate kernel $\mathbf{h}_2(t_1, t_2) = \mathbf{M}\left(e^{\mathbf{A}t_1}\mathbf{B} \otimes e^{\mathbf{A}t_2}\mathbf{B}\right)$ can be expressed as

$$\mathbf{h}_2(t_1, t_2) = \begin{bmatrix}\operatorname{vec}(\mathbf{B}^\mathsf{T}e^{\mathbf{A}^\mathsf{T}t_1}\mathbf{M}_1 e^{\mathbf{A}t_2}\mathbf{B})^\mathsf{T} \\ \vdots \\ \operatorname{vec}(\mathbf{B}^\mathsf{T}e^{\mathbf{A}^\mathsf{T}t_1}\mathbf{M}_p e^{\mathbf{A}t_2}\mathbf{B})^\mathsf{T}\end{bmatrix},$$

and likewise for $\mathbf{h}_{2,r}(t_1, t_2)$. As a consequence

$$\begin{aligned}&\operatorname{tr}\left(\mathbf{h}_2(t_1,t_2)\mathbf{h}_{2,r}(t_1,t_2)^\mathsf{T}\right) \\ &\quad = \sum_{k=1}^p \operatorname{tr}\left(\mathbf{B}^\mathsf{T}e^{\mathbf{A}^\mathsf{T}t_1}\mathbf{M}_k e^{\mathbf{A}t_2}\mathbf{B}\left(\mathbf{B}_r^\mathsf{T}e^{\mathbf{A}_r^\mathsf{T}t_1}\mathbf{M}_{k,r}e^{\mathbf{A}_r t_2}\mathbf{B}_r\right)^\mathsf{T}\right),\end{aligned}$$

because the trace inner product of two matrices equates to the $p = 2$ vector inner product of their vectorized forms. Integrating both sides of the above equality yields

$$\begin{aligned}&\int_0^\infty\int_0^\infty \operatorname{tr}\left(\mathbf{h}_2(\tau_1,\tau_2)\mathbf{h}_{2,r}(\tau_1,\tau_2)^\mathsf{T}\right) d\tau_1 d\tau_2 \\ &= -\sum_{k=1}^p \int_0^\infty\int_0^\infty \operatorname{tr}\Big(\mathbf{B}^\mathsf{T}e^{\mathbf{A}^\mathsf{T}\tau_1}\mathbf{M}_k e^{\mathbf{A}\tau_2}\mathbf{B}\times \\ &\qquad \left(\mathbf{B}_r e^{\mathbf{A}_r^\mathsf{T}\tau_1}\mathbf{M}_{k,r}e^{\mathbf{A}_r\tau_2}\mathbf{B}_r\right)^\mathsf{T}\Big) d\tau_2 d\tau_1,\end{aligned}$$

proving that $\langle \mathcal{S}, \mathcal{S}_r\rangle_{\mathcal{H}_2} = -\operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{Z}\mathbf{B}_r\right)$, as claimed. The formula for the $\mathcal{H}_2$ norm in (19) follows directly by replacing $\mathcal{S}$ with $\mathcal{S}_r$.

The proof of (18) now follows straightforwardly from (17). First note that the matrix equations (15) and (16) can be reformulated as equivalent linear systems

$$\begin{aligned}\left(\mathbf{I}_n \otimes \mathbf{A} + \mathbf{A}_r \otimes \mathbf{I}_r\right)\operatorname{vec}(\mathbf{X}) &= -\operatorname{vec}(\mathbf{B}\mathbf{B}_r^\mathsf{T}), \\ \left(\mathbf{I}_n \otimes \mathbf{A}^\mathsf{T} + \mathbf{A}_r^\mathsf{T} \otimes \mathbf{I}_r\right)\operatorname{vec}(\mathbf{Z}) &= \operatorname{vec}(\mathbf{C}^\mathsf{T}\mathbf{C}_r) \\ &\quad + \sum_{k=1}^p (\mathbf{M} \otimes \mathbf{M}_r)\operatorname{vec}(\mathbf{X}).\end{aligned}$$

Using the characterization of $\langle \mathcal{S}, \mathcal{S}_r\rangle_{\mathcal{H}_2}$ in (17) as well as properties of the trace and Kronecker product, it follows that

$$\begin{aligned}\langle \mathcal{S}, \mathcal{S}_r\rangle_{\mathcal{H}_2} &= -\operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{Z}\mathbf{B}_r\right) = -\operatorname{vec}\left(\mathbf{B}\mathbf{B}_r^\mathsf{T}\right)^\mathsf{T}\operatorname{vec}\left(\mathbf{Z}\right) \\ &= \operatorname{vec}\left(\mathbf{X}\right)^\mathsf{T}\left(\mathbf{I}_n \otimes \mathbf{A}^\mathsf{T} + \mathbf{A}_r^\mathsf{T} \otimes \mathbf{I}_r\right)\operatorname{vec}\left(\mathbf{Z}\right) \\ &= \operatorname{vec}\left(\mathbf{X}\right)^\mathsf{T}\left(\operatorname{vec}(\mathbf{C}^\mathsf{T}\mathbf{C}_r) + \sum_{k=1}^p (\mathbf{M} \otimes \mathbf{M}_r)\operatorname{vec}\left(\mathbf{X}\right)\right) \\ &= \operatorname{tr}\left(\mathbf{C}\mathbf{X}\mathbf{C}_r^\mathsf{T}\right) + \sum_{k=1}^p \operatorname{tr}\left(\mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X}\mathbf{M}_{k,r}\right).\end{aligned}$$

This proves (18); the analogous formula for the norm (20) follows from replacing $\mathcal{S}_r$ with $\mathcal{S}$. □

2.2.2 Relating the $\mathcal{H}_2$ system error and $\mathcal{L}_\infty$ output error

We take a moment here to motivate the choice of the $\mathcal{H}_2$ norm as a performance metric in the model reduction of linear quadratic output systems. Consider the full and reduced-order systems $\mathcal{S}$ and $\mathcal{S}_r$ in (1) and (3), respectively. Recall that an effective surrogate model should replicate the full quantity of interest $\mathbf{y}_r \approx \mathbf{y}$ for a variety of external inputs $\mathbf{u}$. Suppose that one desires a reduced model so that the error due to the approximate output is uniformly small over all values $t > 0$. In other words, the $\mathcal{L}_\infty$ error

$$\|\mathbf{y} - \mathbf{y}_r\|_{\mathcal{L}_\infty^p} = \sup_{t \geq 0} \|\mathbf{y}(t) - \mathbf{y}_r(t)\|_\infty$$

should be small for admissible $\mathbf{u}$, e.g. $\mathbf{u} \in \mathcal{L}_2[0, \infty)$. Following (5), the error in the output at any time $t > 0$ may be expressed as

$$\begin{aligned}\mathbf{y}(t) - \mathbf{y}_r(t) = &\int_0^t (\mathbf{h}_1(\tau) - \mathbf{h}_{1,r}(\tau))\, \mathbf{u}(t-\tau)\, d\tau \\ &+ \int_0^t \int_0^t (\mathbf{h}_2(\tau_1, \tau_2) - \mathbf{h}_{2,r}(\tau_1, \tau_2)) \times \\ &\qquad (\mathbf{u}(t-\tau_1) \otimes \mathbf{u}(t-\tau_2))\, d\tau_1 d\tau_2.\end{aligned}$$

Applying the vector norm $\|\cdot\|_\infty \colon \mathbb{R}^p \to \mathbb{R}$ to both sides of this equality, and using the equivalence of $\|\cdot\|_\infty$ and $\|\cdot\|_2$ as well as the Cauchy-Schwarz inequality, the output error at any time $t > 0$ satisfies the inequality

$$\begin{aligned}\|\mathbf{y}(t) - \mathbf{y}_r(t)\|_\infty^2 \leq &\Bigg(\underbrace{\|\mathbf{h}_1 - \mathbf{h}_{1,r}\|_{\mathcal{L}_2^{p\times m}}^2 + \|\mathbf{h}_2 - \mathbf{h}_{2,r}\|_{\mathcal{L}_2^{p\times m^2}}^2}_{= \|\mathcal{S} - \mathcal{S}_r\|_{\mathcal{H}_2}^2} \Bigg) \\ &\times \left(\|\mathbf{u}\|_{\mathcal{L}_2^m}^2 + \|\mathbf{u} \otimes \mathbf{u}\|_{\mathcal{L}_2^{m^2}}^2 \right).\end{aligned}$$

Because $t > 0$ is arbitrarily specified, taking the supremum over all time reveals that the $\mathcal{L}_\infty$ *output error is bounded above by the* $\mathcal{H}_2$ *linear quadratic output system error*

$$\|\mathbf{y} - \mathbf{y}_r\|_{\mathcal{L}_\infty^p}^2 \leq \|\mathcal{S} - \mathcal{S}_r\|_{\mathcal{H}_2}^2 \left(\|\mathbf{u}\|_{\mathcal{L}_2^m}^2 + \|\mathbf{u} \otimes \mathbf{u}\|_{\mathcal{L}_2^{m^2}}^2 \right). \quad (23)$$

This upper bound (23) from [7, Theorem 3.4] shows that for an admissible input $\mathbf{u}$, a small $\mathcal{H}_2$ error guarantees that the reduced model output $\mathbf{y}_r$ is a high-fidelity $\mathcal{L}_\infty$ approximation to the original output $\mathbf{y}$.

The bound in (23) is a powerful motivator for using the $\mathcal{H}_2$ error as a performance measure in LQO-MOR. Indeed, if one wants the 'worst case' error in the output, $\|\mathbf{y} - \mathbf{y}_r\|_{\mathcal{L}_\infty}$ to be small, then one should seek a reduced model $\mathcal{S}_r$ so that the $\mathcal{H}_2$ error $\|\mathcal{S} - \mathcal{S}_r\|_{\mathcal{H}_2}$ appearing in (23) is as small as possible. Motivated by this, we consider the following $\mathcal{H}_2$ *optimal model reduction problem for linear quadratic output systems*: Given the full order asymptotically stable LQO system $\mathcal{S}$ in (1), we seek a reduced-order, and also asymptotically stable, LQO reduced model $\mathcal{S}_r$ represented in (3) such that the $\mathcal{H}_2$ norm of the error system is *minimized*:

$$\mathcal{S}_r = \underset{\substack{\widehat{\mathcal{S}}_r \text{ stable} \\ \dim(\widehat{\mathcal{S}}_r) = r}}{\arg\min}\ \mathcal{J}(\widehat{\mathcal{S}}_r), \quad \mathcal{J}(\widehat{\mathcal{S}}_r) := \|\mathcal{S} - \widehat{\mathcal{S}}_r\|_{\mathcal{H}_2}^2. \quad (24)$$

(The squared $\mathcal{H}_2$ error in (24) is used solely for the sake of ease in calculating gradients of $\mathcal{J}$ in Section 3.) By the asymptotic stability assumption imposed on the full and reduced-order models, the corresponding $\mathcal{H}_2$ error is guaranteed to be finite. Nonetheless, the minimization problem in (24) is in general non-convex, and the characterization of global minimizers is elusive. Instead, we wish to identify *local minimizers* of the $\mathcal{H}_2$ error in (24) which satisfy first-order necessary conditions (FONCs) for $\mathcal{H}_2$ optimality.

Before presenting our major theoretical results in the next section, we establish notation pertaining to gradients of functions defined over normed vector spaces; our presentation follows that of [12]. Consider a Fréchet differentiable function $f \colon U \to \mathbb{R}$ defined on an open subset U of a Hilbert space X endowed with the inner product $\langle \cdot, \cdot \rangle_X \colon X \times X \to \mathbb{R}$. For any $x_0 \in U$, the *gradient* of f at x_0 is the unique element $\nabla f(x_0) \in X$ so that

$$f(x_0 + h) = f(x_0) + \langle \nabla f(x_0), h \rangle_X + O(\|h\|_X^2), \quad (25)$$

for all h in a neighborhood of zero. We write $g(x) = O(\|h\|_X^2)$ if $\lim_{h\to 0} \frac{g(h)}{\|h\|_X} = 0$. If $\nabla f(x_0) = 0$, we call $x_0 \in X$ a *critical* point of f. If f has a local extremum at a point x_0, then necessarily x_0 is a critical point [12, Cor. 2.5]. For a multivariate function $f \colon X_1 \times \ldots \times X_\ell \to \mathbb{R}$, partial gradients $\nabla_{x_i} f(x_1, \ldots, x_\ell)$ are defined analogously.

3 Optimal $\mathcal{H}_2$ model reduction

This section contains the main theoretical results of the paper. Theorem 3.1 presents gradients of the squared $\mathcal{H}_2$ LQO system error in (24) with respect to the reduced-order matrices of the LQO-ROM in (3) as parameters. The stationary points of these gradients automatically yield Gramian-based FONCs for the $\mathcal{H}_2$ optimal model reduction of LQO systems, which we present in Theorem 3.2. To set the stage for these results and make comparisons later on, we first review the Gramian-based $\mathcal{H}_2$ optimality conditions for linear model reduction due to Wilson [29].

3.1 Optimal $\mathcal{H}_2$ model reduction of purely linear systems

For the discussion in this subsection, we restrict our attention to *purely* LTI dynamical systems. Consider

$$\mathcal{S}_1 : \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad \mathbf{x}(0) = \mathbf{0}, \\ \mathbf{y}_1(t) = \mathbf{C}\mathbf{x}(t). \end{cases} \tag{26}$$

The state, input, and output dimensions are the same as in (1). Analogous to the LQO model reduction problem, we seek a linear reduced model of the form

$$\mathcal{S}_{1,r} : \begin{cases} \dot{\mathbf{x}}_r(t) = \mathbf{A}_r\mathbf{x}_r(t) + \mathbf{B}_r\mathbf{u}(t), \quad \mathbf{x}_r(0) = \mathbf{0}, \\ \mathbf{y}_{1,r}(t) = \mathbf{C}_r\mathbf{x}_r(t), \end{cases} \tag{27}$$

It will be fruitful to view the LTI system in (26) as a 'special case' of the LQO system class (26) having the realization $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{0}_{n\times n}, \ldots, \mathbf{0}_{n\times n})$. With this perspective, much of the systems theory introduced in Section 2 neatly reduces to the analogous linear systems theory. The input-to-output map of the LTI system $\mathcal{S}_1$ in (26) is fully realized by the one-dimensional kernel $\mathbf{h}_1(t) = \mathbf{C}e^{\mathbf{A}t}\mathbf{B}$; the two-dimensional kernel $\mathbf{h}_2(t_1, t_2)$ is identically zero in this context. As a consequence, the $\mathcal{H}_2$ norm of $\mathcal{S}_1$ as given in Definition 2.1 agrees with the linear $\mathcal{H}_2$ system norm; cf. [1, Section 5.1]. The QO observability Gramian of $\mathcal{S}_1$ becomes the classical observability Gramian of a linear system, which is $\mathbf{Q}_1 \in \mathbb{R}^{n\times n}$ in (11), since $\mathbf{Q}_{2,k} = \mathbf{0}_{n\times n}$ for all k. The Lyapunov equation (14) reduces to

$$\mathbf{A}^\mathsf{T}\mathbf{Q}_1 + \mathbf{Q}_1\mathbf{A} + \mathbf{C}^\mathsf{T}\mathbf{C} = \mathbf{0}. \tag{28}$$

(When discussing the observability Gramian of a purely linear system, we keep the subscript to clearly differentiate from the general QO observability Gramian in (13).) The reachability Gramian $\mathbf{P}$ in (9) is unchanged. A similar bound to (23) can be derived, i.e.,

$$\|\mathbf{y}_1 - \mathbf{y}_{1,r}\|_{\mathcal{L}^p_\infty} \leq \|\mathcal{S}_1 - \mathcal{S}_{1,r}\|_{\mathcal{H}_2}\|\mathbf{u}\|_{\mathcal{L}^m_2}.$$

The Gramian-based framework for linear $\mathcal{H}_2$ optimal model reduction is attributed to Wilson [29] for multi-input, multi-output systems. (These results were later shown to be equivalent to interpolation-based optimality conditions [17, 19, 28].) The starting point for deriving the Gramian-based optimality conditions in [28, 29] is expressing the $\mathcal{H}_2$ norm as a function of the *error system* $\mathcal{S}_1 - \mathcal{S}_{1,r}$, which is itself an order-$(n + r)$ linear system of the form (26). A state-space realization of $\mathcal{S}_1 - \mathcal{S}_{1,r}$ is given by $(\mathbf{A}_e, \mathbf{B}_e, \mathbf{C}_e)$, where $\mathbf{A}_e \in \mathbb{R}^{(n+r)\times(n+r)}$, $\mathbf{B}_e \in \mathbb{R}^{(n+r)\times m}$, and $\mathbf{C}_e \in \mathbb{R}^{p\times(n+r)}$ are given by

$$\mathbf{A}_e = \begin{bmatrix} \mathbf{A} & \\ & \mathbf{A}_r \end{bmatrix}, \ \mathbf{B}_e = \begin{bmatrix} \mathbf{B} \\ \mathbf{B}_r \end{bmatrix}, \ \mathbf{C}_e = \begin{bmatrix} \mathbf{C} & -\mathbf{C}_r \end{bmatrix}. \tag{29}$$

The internal state and output of the error system are given by $\mathbf{x}_e = \begin{bmatrix} \mathbf{x} \\ \mathbf{x}_r \end{bmatrix}$ and $\mathbf{y}_1 - \mathbf{y}_{1,r}$. Take $\mathbf{P}_e, \mathbf{Q}_{1,e} \in \mathbb{R}^{(n+r)\times(n+r)}$ to denote the reachability and observability Gramians of (29) which solve the Lyapunov equations (10) and (28). These are expressed in 2×2 block form as

$$\mathbf{P}_e = \begin{bmatrix} \mathbf{P} & \mathbf{X} \\ \mathbf{X}^\mathsf{T} & \mathbf{P}_r \end{bmatrix} \quad \text{and} \quad \mathbf{Q}_{1,e} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Z}_1 \\ \mathbf{Z}_1^\mathsf{T} & \mathbf{Q}_{1,r} \end{bmatrix}, \tag{30}$$

where $\mathbf{P}, \mathbf{Q}_1 \in \mathbb{R}^{n\times n}$ and $\mathbf{P}_r, \mathbf{Q}_{1,r} \in \mathbb{R}^{r\times r}$ are the reachability, observability Gramians of the full-order and reduced-order linear models in (26) and (27), respectively, while $\mathbf{X} \in \mathbb{R}^{n\times r}$ and $\mathbf{Z}_1 \in \mathbb{R}^{n\times r}$ solve the matrix equations

$$\begin{aligned} \mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}_r^\mathsf{T} + \mathbf{B}\mathbf{B}_r^\mathsf{T} &= \mathbf{0}, \\ \text{and} \quad \mathbf{A}^\mathsf{T}\mathbf{Z}_1 + \mathbf{Z}_1\mathbf{A}_r - \mathbf{C}^\mathsf{T}\mathbf{C}_r &= \mathbf{0}. \end{aligned} \tag{31}$$

These equalities (31) can be deduced by solving for $\mathbf{X}$ and $\mathbf{Z}_1$ directly from equations (10) and (28) applied to the error system. Evidently, the $\mathcal{H}_2$ norm of the error system (29) is precisely the approximation error induced by $\mathcal{S}_{1,r}$. Applying Theorem 2.1 to $\mathcal{S}_1 - \mathcal{S}_{1,r}$, the squared $\mathcal{H}_2$ system error can be written as

$$\mathcal{J}(\mathcal{S}_{1,r}) = \|\mathcal{S}_1 - \mathcal{S}_{1,r}\|^2_{\mathcal{H}_2} = \operatorname{tr}\left(\mathbf{B}_e^\mathsf{T}\mathbf{Q}_{1,e}\mathbf{B}_e\right) = \operatorname{tr}\left(\mathbf{C}_e\mathbf{P}_e\mathbf{C}_e^\mathsf{T}\right).$$

From this expression, [28, Theorem 3.3], [29] show that gradients of $\mathcal{J}$ with respect to $\mathbf{A}_r$, $\mathbf{B}_r$, and $\mathbf{C}_r$ are given by

$$\begin{aligned} \nabla_{\mathbf{A}_r}\mathcal{J} &= 2(\mathbf{P}_r\mathbf{Q}_r + \mathbf{X}^\mathsf{T}\mathbf{Z}_1), \\ \nabla_{\mathbf{B}_r}\mathcal{J} &= 2(\mathbf{Q}_r\mathbf{B}_r + \mathbf{Z}_1^\mathsf{T}\mathbf{B}), \\ \nabla_{\mathbf{C}_r}\mathcal{J} &= 2(\mathbf{C}_r\mathbf{P}_r - \mathbf{C}\mathbf{X}). \end{aligned} \tag{32}$$

(We drop the dependence of $\mathcal{J}$ on $\mathcal{S}_{1,r}$ for convenience of notation.) The matrices $\mathbf{X} \in \mathbb{R}^{n\times r}$ and $\mathbf{Z}_1 \in \mathbb{R}^{n\times r}$ appearing in (32) are those that satisfy (31). If the reduced LTI system $\mathcal{S}_{1,r}$ in (27) is a local minimizer of the $\mathcal{H}_2$ error, then the gradients in (32) are identically zero, and the reduced order matrices $(\mathbf{A}_r, \mathbf{B}_r, \mathbf{C}_r)$ satisfy the following identities

$$\begin{aligned} \mathbf{0} &= \mathbf{P}_r\mathbf{Q}_r + \mathbf{X}^\mathsf{T}\mathbf{Z}_1, \\ \mathbf{0} &= \mathbf{Q}_{1,r}\mathbf{B}_r + \mathbf{Z}_1^\mathsf{T}\mathbf{B}, \\ \mathbf{0} &= \mathbf{C}_r\mathbf{P}_r - \mathbf{C}\mathbf{X}. \end{aligned} \tag{33}$$

These are the *Gramian-based* or, *Wilson* conditions for the $\mathcal{H}_2$ optimal model reduction of LTI systems. Under the assumption that $\mathbf{P}_r$ and $\mathbf{Q}_{1,r}$ are nonsingular, if $\mathcal{S}_{1,r}$ is $\mathcal{H}_2$ optimal then it is realized in a Petrov-Galerkin framework [29], [28, Theorem 3.4], where the projection matrices in (4) are given by $\mathbf{V}_r = \mathbf{X}\mathbf{P}_r^{-1}$ and $\mathbf{W}_r = -\mathbf{Z}_1\mathbf{Q}_{1,r}^{-1}$. Evidently, $\mathbf{V}_r$ and $\mathbf{W}_r$ depend explicitly upon the $\mathcal{H}_2$ optimal

reduced model $\mathcal{S}_{1,r}$, which of course is not known *a priori*. To handle this issue, the *two-sided iteration algorithm* (TSIA) based on iterative projection using the solutions to the pair of Sylvester equations in (31) was proposed in [30], yielding a ROM that satisfies (33).

3.2 Optimal $\mathcal{H}_2$ model reduction of linear quadratic output systems

We now return our attention to the $\mathcal{H}_2$ optimal model reduction problem for LQO systems described in Section 2.2. Recall the minimization problem in (24) as well as the associated cost function

$$\mathcal{J}(\mathcal{S}_r) = \mathcal{J}(\mathbf{A}_r, \mathbf{B}_r, \mathbf{C}_r, \mathbf{M}_{1,r}, \ldots, \mathbf{M}_{k,r}) = \|\mathcal{S} - \mathcal{S}_r\|_{\mathcal{H}_2}^2.$$

We view the objective function $\mathcal{J}\colon (\mathbb{R}^{r\times r} \times \mathbb{R}^{r\times m} \times \mathbb{R}^{p\times n} \times \cdots \times \mathbb{R}^{r\times r}) \to \mathbb{R}$ as taking the reduced-order matrices that determine $\mathcal{S}_r$ in (3) as arguments. That is, $\mathcal{J}$ is a multivariate function defined over a real-valued Hilbert space of matrices in each variable. Per [12, Cor. 2.5], if the reduced LQO system $\mathcal{S}_r$ in (3) having the realization $(\mathbf{A}_r, \mathbf{B}_r, \mathbf{C}_r, \mathbf{M}_{1,r}, \ldots, \mathbf{M}_{k,r})$ is a local minimizer of the $\mathcal{H}_2$ error, then the partial gradients of $\mathcal{J}$ with respect to the reduced-order matrices $\mathbf{A}_r$, $\mathbf{B}_r$, $\mathbf{C}_r$, and $\mathbf{M}_{k,r}$ are necessarily zero. So, computing gradients of $\mathcal{J}$ in (24) and setting them to zero paves the way for developing the Wilson optimality conditions for the LQO setting.

Following suit with the linear problem, our starting point for computing gradients of $\mathcal{J}$ is expressing the objective function in terms of the error system $\mathcal{S} - \mathcal{S}_r$. The matrices $(\mathbf{A}_e, \mathbf{B}_e, \mathbf{C}_e, \mathbf{M}_{1,e}, \ldots, \mathbf{M}_{p,e})$, where $\mathbf{A}_e \in \mathbb{R}^{(n+r)\times(n+r)}$, $\mathbf{B}_e \in \mathbb{R}^{(n+r)\times m}$, $\mathbf{C}_e \in \mathbb{R}^{p\times(n+r)}$ and $\mathbf{M}_{k,e} \in \mathbb{R}^{(n+r)\times(n+r)}$ are defined as

$$\begin{aligned} \mathbf{A}_e &= \begin{bmatrix} \mathbf{A} & \\ & \mathbf{A}_r \end{bmatrix}, \ \mathbf{B}_e = \begin{bmatrix} \mathbf{B} \\ \mathbf{B}_r \end{bmatrix}, \ \mathbf{C}_e = \begin{bmatrix} \mathbf{C} & -\mathbf{C}_r \end{bmatrix} \\ \mathbf{M}_{k,e} &= \begin{bmatrix} \mathbf{M}_k & \\ & -\mathbf{M}_{k,r} \end{bmatrix}, \text{ for each } k = 1, \ldots p, \end{aligned} \tag{34}$$

constitute a state-space realization as in (1) of the LQO error system. The reachability and quadratic output observability Gramians $\mathbf{P}_e, \mathbf{Q}_e \in \mathbb{R}^{(n+r)\times(n+r)}$ of the error system uniquely satisfy the Lyapunov equations (10) and (14) for the representation in (34):

$$\begin{aligned} \mathbf{A}^\mathsf{T}\mathbf{Q}_e + \mathbf{Q}_e\mathbf{A}_e + \mathbf{C}_e^\mathsf{T}\mathbf{C}_e + \sum_{k=1}^{p} \mathbf{M}_e\mathbf{P}_e\mathbf{M}_e &= \mathbf{0}, \\ \mathbf{A}_e\mathbf{P}_e + \mathbf{P}_e\mathbf{A}_e^\mathsf{T} + \mathbf{B}_e\mathbf{B}_e^\mathsf{T} &= \mathbf{0}. \end{aligned} \tag{35}$$

The Gramians $\mathbf{P}_e$ and $\mathbf{Q}_e$ have the particular structure

$$\mathbf{P}_e = \begin{bmatrix} \mathbf{P} & \mathbf{X} \\ \mathbf{X}^\mathsf{T} & \mathbf{P}_r \end{bmatrix} \quad \text{and} \quad \mathbf{Q}_e = \begin{bmatrix} \mathbf{Q} & \mathbf{Z} \\ \mathbf{Z}^\mathsf{T} & \mathbf{Q}_r \end{bmatrix}, \tag{36}$$

where $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{n\times n}$ are the full-order system Gramians, and the submatrices $\mathbf{P}_r, \mathbf{Q}_r \in \mathbb{R}^{r\times r}$ and $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^{n\times r}$ are the unique solutions to the matrix equations

$$\mathbf{A}_r\mathbf{P}_r + \mathbf{P}_r\mathbf{A}_r^\mathsf{T} + \mathbf{B}_r\mathbf{B}_r^\mathsf{T} = \mathbf{0}, \tag{37a}$$

$$\mathbf{A}_r^\mathsf{T}\mathbf{Q}_r + \mathbf{Q}_r\mathbf{A}_r + \sum_{k=1}^{p} \mathbf{M}_{k,r}\mathbf{P}_r\mathbf{M}_{k,r} + \mathbf{C}_r^\mathsf{T}\mathbf{C}_r = \mathbf{0}, \tag{37b}$$

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}_r^\mathsf{T} + \mathbf{B}\mathbf{B}_r^\mathsf{T} = \mathbf{0}, \tag{37c}$$

$$\mathbf{A}^\mathsf{T}\mathbf{Z} + \mathbf{Z}\mathbf{A}_r - \sum_{k=1}^{p} \mathbf{M}_k\mathbf{X}\mathbf{M}_{k,r} - \mathbf{C}^\mathsf{T}\mathbf{C}_r = \mathbf{0}. \tag{37d}$$

Note that the matrices $\mathbf{P}_r$ and $\mathbf{Q}_r$ are symmetric since they are the reachability and quadratic output observability Gramians of the reduced model $\mathcal{S}_r$. In addition to the constraints in (37a) – (37d), it will be useful to recall that $\mathbf{Z}_1 \in \mathbb{R}^{n\times r}$ and $\mathbf{Q}_{1,r} \in \mathbb{R}^{r\times r}$ in (30) satisfy

$$\mathbf{A}_r^\mathsf{T}\mathbf{Q}_{1,r} + \mathbf{Q}_{1,r}\mathbf{A}_r + \mathbf{C}_r^\mathsf{T}\mathbf{C}_r = \mathbf{0}, \tag{38a}$$

$$\mathbf{A}^\mathsf{T}\mathbf{Z}_1 + \mathbf{Z}_1\mathbf{A}_r - \mathbf{C}^\mathsf{T}\mathbf{C}_r = \mathbf{0}. \tag{38b}$$

Equations (38a) and (38b) simplify to (37b) and (37d) when $\mathbf{M}_k = \mathbf{0}_{n\times n}$ for all k, as in the LTI setting. Applying the result of Theorem 2.1 to the realization in (34), the squared $\mathcal{H}_2$ error in LQO-MOR can be expressed as

$$\begin{aligned} \mathcal{J}(\mathcal{S}_r) &= \operatorname{tr}\left(\mathbf{B}_e^\mathsf{T}\mathbf{Q}_e\mathbf{B}_e\right), \\ &= \operatorname{tr}\left(\mathbf{C}_e\mathbf{P}_e\mathbf{C}_e^\mathsf{T}\right) + \sum_{k=1}^{p} \operatorname{tr}\left(\mathbf{P}_e\mathbf{M}_{k,e}\mathbf{P}_e\mathbf{M}_{k,e}\right). \end{aligned} \tag{39}$$

Basic algebraic manipulations of the trace in (39) reveal that the squared $\mathcal{H}_2$ error can be re-written as

$$\mathcal{J}(\mathcal{S}_r) = \operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{Q}\mathbf{B} + 2\mathbf{B}^\mathsf{T}\mathbf{Z}\mathbf{B}_r + \mathbf{B}_r^\mathsf{T}\mathbf{Q}_r\mathbf{B}_r\right), \tag{40}$$

or equivalently

$$\begin{aligned} \mathcal{J}(\mathcal{S}_r) = \operatorname{tr}\left(\mathbf{C}\mathbf{P}\mathbf{C}^\mathsf{T} - 2\mathbf{C}\mathbf{X}\mathbf{C}_r^\mathsf{T} + \mathbf{C}_r\mathbf{P}_r\mathbf{C}_r^\mathsf{T}\right) \\ + \sum_{k=1}^{p} \operatorname{tr}\left(\mathbf{P}\mathbf{M}_k\mathbf{P}\mathbf{M}_k - 2\mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X}\mathbf{M}_{k,r} + \mathbf{P}_r\mathbf{M}_{k,r}\mathbf{P}_r\mathbf{M}_{k,r}\right), \end{aligned} \tag{41}$$

where $\mathbf{P}_r, \mathbf{Q}_r, \mathbf{Q}_{1,r} \in \mathbb{R}^{r\times r}$ and $\mathbf{X}, \mathbf{Z}, \mathbf{Z}_1 \in \mathbb{R}^{n\times r}$ satisfy the constraints in (37a) – (38b).

This brings us to our first major result, which we present in Theorem 3.1. We leverage the formulations in (40) and (41)

in order to establish partial gradients of the squared $\mathcal{H}_2$ error $\mathcal{J}(\mathcal{S}_r) = \|\mathcal{S} - \mathcal{S}_r\|_{\mathcal{H}_2}^2$ in LQO-MOR with respect to the reduced-order matrices $\mathbf{A}_r$, $\mathbf{B}_r$, $\mathbf{C}_r$, and $\mathbf{M}_{k,r}$ for all $k = 1, \dots p$. The stationary points of the gradients in turn yield FONCs for the $\mathcal{H}_2$ optimal model reduction of LQO systems. To simplify the proof of Theorem 3.1, we recall the following result that we will invoke repeatedly.

Lemma 3.1 ([31, Lemma A.1]). *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{A}_r \in \mathbb{R}^{r\times r}$ and $\mathbf{D}, \mathbf{F} \in \mathbb{R}^{n\times r}$ be such that the matrices $\mathbf{Y}, \mathbf{W} \in \mathbb{R}^{n\times r}$ solve the Sylvester equations*

$$\mathbf{A}\mathbf{Y} + \mathbf{Y}\mathbf{A}_r^\mathsf{T} + \mathbf{D} = \mathbf{0}$$
$$\text{and} \quad \mathbf{A}^\mathsf{T}\mathbf{W} + \mathbf{W}\mathbf{A}_r + \mathbf{F} = \mathbf{0},$$

then, $\operatorname{tr}\left(\mathbf{D}^\mathsf{T}\mathbf{W}\right) = \operatorname{tr}\left(\mathbf{F}^\mathsf{T}\mathbf{Y}\right)$.

From Lemma 3.1, we also have, e.g., $\operatorname{tr}\left(\mathbf{W}^\mathsf{T}\mathbf{D}\right) = \operatorname{tr}\left(\mathbf{Y}^\mathsf{T}\mathbf{F}\right)$. In the subsequent result, we take for granted any identities that arise from cyclic permutations or transposes applied to the result of Lemma 3.1. When it is helpful for simplifying notation, we drop the dependence of $\mathcal{J}$ on $\mathcal{S}_r$.

Theorem 3.1. *Let $\mathcal{S}$ and $\mathcal{S}_r$ be asymptotically stable LQO systems as in (1) and (3), respectively. Then the gradients $\nabla_{\mathbf{A}_r}\mathcal{J}$, $\nabla_{\mathbf{B}_r}\mathcal{J}$, $\nabla_{\mathbf{C}_r}\mathcal{J}$, $\nabla_{\mathbf{M}_{k,r}}\mathcal{J}$, $k = 1, \dots, p$ of the squared $\mathcal{H}_2$ error $\mathcal{J}(\mathcal{S}_r) = \|\mathcal{S} - \mathcal{S}_r\|_{\mathcal{H}_2}^2$ are given explicitly as*

$$\nabla_{\mathbf{A}_r}\mathcal{J} = 2\left((2\mathbf{Q}_r - \mathbf{Q}_{1,r})\mathbf{P}_r + (2\mathbf{Z}^\mathsf{T} - \mathbf{Z}_1^\mathsf{T})\mathbf{X}\right) \tag{42a}$$
$$\nabla_{\mathbf{B}_r}\mathcal{J} = 2\left((2\mathbf{Q}_r - \mathbf{Q}_{1,r})\mathbf{B}_r + (2\mathbf{Z}^\mathsf{T} - \mathbf{Z}_1^\mathsf{T})\mathbf{B}\right) \tag{42b}$$
$$\nabla_{\mathbf{C}_r}\mathcal{J} = 2\left(\mathbf{C}_r\mathbf{P}_r - \mathbf{C}\mathbf{X}\right) \tag{42c}$$
$$\nabla_{\mathbf{M}_{k,r}}\mathcal{J} = 2\left(\mathbf{P}_r\mathbf{M}_{k,r}\mathbf{P}_r - \mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X}\right), \ k = 1, \dots, p, \tag{42d}$$

where $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^{n\times r}$ and $\mathbf{P}_r, \mathbf{Q}_r \in \mathbb{R}^{r\times r}$ satisfy the constraint equations (37a) – (37d), and $\mathbf{Z}_1 \in \mathbb{R}^{n\times r}$, $\mathbf{Q}_1 \in \mathbb{R}^{r\times r}$ satisfy (38a) and (38b), respectively.

Proof. Note that $\mathcal{S}$ is fixed; we begin with the gradients with respect to the quadratic output matrices (42d). For any $k = 1, \dots, p$, consider an arbitrary perturbation $\Delta_{\mathbf{M}_{k,r}} \in \mathbb{R}^{r\times r}$ about zero. By (25), it suffices to show the claimed form of the gradient $\nabla_{\mathbf{M}_{k,r}}\mathcal{J}$ in (42d) satisfies

$$\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r}) = \mathcal{J}(\mathcal{S}_r) + \langle \nabla_{\mathbf{M}_{k,r}}\mathcal{J}, \Delta_{\mathbf{M}_{k,r}}\rangle_\mathsf{F} + O(\|\Delta_{\mathbf{M}_{k,r}}\|_\mathsf{F}^2)$$

where $\Delta_{\mathcal{S}_r}$ denotes the perturbation in the reduced model $\mathcal{S}_r$ due to $\Delta_{\mathbf{M}_{k,r}}$. The first-order perturbation in $\mathbf{M}_{k,r}$ in turn induces perturbations $\Delta_\mathbf{Z} \in \mathbb{R}^{n\times r}$ and $\Delta_{\mathbf{Q}_r} \in \mathbb{R}^{r\times r}$ in the solutions to (37d) and (37b). Expanding upon the perturbed equations (37d) and (37b) reveal that $\Delta_\mathbf{Z}$ and $\Delta_{\mathbf{Q}_r}$ satisfy

$$\mathbf{A}^\mathsf{T}\Delta_\mathbf{Z} + \Delta_\mathbf{Z}\mathbf{A}_r - \mathbf{M}_k\mathbf{X}\Delta_{\mathbf{M}_{k,r}}^\mathsf{T} = \mathbf{0} \tag{43a}$$
$$\text{and} \quad \mathbf{A}_r^\mathsf{T}\Delta_{\mathbf{Q}_r} + \Delta_{\mathbf{Q}_r}\mathbf{A}_r + \mathbf{M}_{k,r}\mathbf{P}_r\Delta_{\mathbf{M}_{k,r}}^\mathsf{T} + \Delta_{\mathbf{M}_{k,r}}\mathbf{P}_r\mathbf{M}_{k,r} + \Delta_{\mathbf{M}_{k,r}}\mathbf{P}_r\Delta_{\mathbf{M}_{k,r}}^\mathsf{T} = \mathbf{0}. \tag{43b}$$

Note that the term $\Delta_{\mathbf{M}_{k,r}}\mathbf{P}_r\Delta_{\mathbf{M}_{k,r}}^\mathsf{T}$ is $O(\|\Delta_{\mathbf{M}_{k,r}}\|_\mathsf{F}^2)$ in the sense of (25). From (40), $\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r})$ can be expressed as

$$\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r}) = \mathcal{J}(\mathcal{S}_r) + 2\operatorname{tr}(\mathbf{B}^\mathsf{T}\Delta_\mathbf{Z}\mathbf{B}_r) + \operatorname{tr}(\mathbf{B}_r^\mathsf{T}\Delta_{\mathbf{Q}_r}\mathbf{B}_r).$$

By applying Lemma 3.1 to Sylvester equations (37c) and (43a), as well as using the properties of the trace, the terms in the first-order expansion above can be re-written as

$$\begin{aligned}\operatorname{tr}\left(\mathbf{B}^\mathsf{T}\Delta_\mathbf{Z}\mathbf{B}_r\right) = \operatorname{tr}\left(\mathbf{B}_r\mathbf{B}^\mathsf{T}\Delta_\mathbf{Z}\right) &= \operatorname{tr}\left(-\mathbf{M}_k\mathbf{X}\Delta_{\mathbf{M}_{k,r}}^\mathsf{T}\mathbf{X}^\mathsf{T}\right)\\ &= \operatorname{tr}\left(-\mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X}\Delta_{\mathbf{M}_{k,r}}\right).\end{aligned}$$

Lemma 3.1 can be applied to equations (37a) and (43b) in a similar way to show that

$$\operatorname{tr}\left(\mathbf{B}_r^\mathsf{T}\Delta_{\mathbf{Q}_r}\mathbf{B}_r\right) = 2\operatorname{tr}\left(\mathbf{P}_r\mathbf{M}_{k,r}\mathbf{P}_r\Delta_{\mathbf{M}_{k,r}}\right) + O(\|\Delta_{\mathbf{M}_{k,r}}\|_\mathsf{F}^2).$$

Plugging these into the expression for $\mathcal{J}(\mathcal{S} + \Delta_{\mathcal{S}_r})$, we get

$$\begin{aligned}\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r}) &= \mathcal{J}(\mathcal{S}_r) + 2\operatorname{tr}\left(-\mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X}\Delta_{\mathbf{M}_{k,r}}\right)\\ &\quad + 2\operatorname{tr}\left(\mathbf{P}_r\mathbf{M}_{k,r}\mathbf{P}_r\Delta_{\mathbf{M}_{k,r}}\right) + O(\|\Delta_{\mathbf{M}_{k,r}}\|_\mathsf{F}^2)\\ &= \mathcal{J}(\mathcal{S}_r) + \langle 2(\mathbf{P}_r\mathbf{M}_{k,r}\mathbf{P}_r - \mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X}), \Delta_{\mathbf{M}_{k,r}}\rangle_\mathsf{F}\\ &\quad + O(\|\Delta_{\mathbf{M}_{k,r}}\|_\mathsf{F}^2).\end{aligned}$$

So, the gradients with respect to $\mathbf{M}_{k,r}$ satisfy $\nabla_{\mathbf{M}_{k,r}}\mathcal{J} = 2(\mathbf{P}_r\mathbf{M}_{k,r}\mathbf{P}_r - \mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X})$ for each k as claimed in (42d).

Next we compute the gradient $\nabla_{\mathbf{C}_r}\mathcal{J}$ in (42c). Consider an arbitrary perturbation $\Delta_{\mathbf{C}_r} \in \mathbb{R}^{p\times r}$ about zero to $\mathbf{C}_r$; let $\Delta_{\mathcal{S}_r}$ be the perturbation to $\mathcal{S}_r$ corresponding to $\Delta_{\mathbf{C}_r}$. From (41), the first-order expansion of $\mathcal{J}(\mathcal{S} + \Delta_{\mathcal{S}_r})$ can be expressed as

$$\begin{aligned}\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r}) &= \mathcal{J}(\mathcal{S}_r) - \operatorname{tr}\left(2\mathbf{C}\mathbf{X}\Delta_{\mathbf{C}_r}^\mathsf{T}\right)\\ &\quad + \operatorname{tr}\left(2\mathbf{C}_r\mathbf{P}_r\Delta_{\mathbf{C}_r}^\mathsf{T}\right) + O(\|\Delta_{\mathbf{C}_r}\|_\mathsf{F}^2)\\ &= \mathcal{J}(\mathcal{S}_r) + \langle 2(\mathbf{C}_r\mathbf{P}_r - \mathbf{C}\mathbf{X}), \Delta_{\mathbf{C}_r}\rangle_\mathsf{F} + O(\|\Delta_{\mathbf{C}_r}\|_\mathsf{F}^2).\end{aligned}$$

So $\nabla_{\mathbf{C}_r}\mathcal{J} = 2(\mathbf{C}_r\mathbf{P}_r - \mathbf{C}\mathbf{X})$ as claimed in (42c).

We show the formula for $\nabla_{\mathbf{B}_r}\mathcal{J}$ in (42b). Consider an arbitrary perturbation $\Delta_{\mathbf{B}_r} \in \mathbb{R}^r$ about zero to $\mathbf{B}_r$, which induces perturbations $\Delta_\mathbf{X} \in \mathbb{R}^{n\times r}$ and $\Delta_{\mathbf{P}_r} \in \mathbb{R}^{n\times r}$ in the

solutions of (37c) and (37a). The resulting perturbations satisfy

$$\mathbf{A}\Delta_{\mathbf{X}} + \Delta_{\mathbf{X}}\mathbf{A}_r^{\mathsf{T}} + \mathbf{B}\Delta_{\mathbf{B}_r}^{\mathsf{T}} = \mathbf{0}, \tag{44a}$$

$$\text{and} \quad \mathbf{A}_r\Delta_{\mathbf{P}_r} + \Delta_{\mathbf{P}_r}\mathbf{A}_r^{\mathsf{T}} + \Delta_{\mathbf{B}_r}\mathbf{B}_r^{\mathsf{T}} + \mathbf{B}_r\Delta_{\mathbf{B}_r}^{\mathsf{T}} + O(\|\Delta_{\mathbf{B}_r}\|_{\mathsf{F}}^2) = \mathbf{0}. \tag{44b}$$

The solutions to (37b) and (37d) depend linearly upon $\mathbf{X}$ and $\mathbf{P}_r$, so the perturbations $\Delta_{\mathbf{X}}$ and $\Delta_{\mathbf{P}_r}$ induce further perturbations $\Delta_{\mathbf{Z}} \in \mathbb{R}^{n\times r}$ and $\Delta_{\mathbf{Q}_r} \in \mathbb{R}^{n\times r}$ in the solutions to (37d) and (37b). These perturbations satisfy

$$\mathbf{A}^{\mathsf{T}}\Delta_{\mathbf{Z}} + \Delta_{\mathbf{Z}}\mathbf{A}_r - \sum_{k=1}^{p}\mathbf{M}_k\Delta_{\mathbf{X}}\mathbf{M}_{k,r} = \mathbf{0}, \tag{45a}$$

$$\text{and} \quad \mathbf{A}_r^{\mathsf{T}}\Delta_{\mathbf{Q}_r} + \Delta_{\mathbf{Q}_r}\mathbf{A}_r + \sum_{k=1}^{p}\mathbf{M}_{k,r}\Delta_{\mathbf{P}_r}\mathbf{M}_{k,r} = \mathbf{0}. \tag{45b}$$

From (40) we may expand $\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r})$ as

$$\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r}) = \mathcal{J}(\mathcal{S}_r) + 2\operatorname{tr}\left(\mathbf{B}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{B}_r} + \mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right) + \operatorname{tr}\left(2\mathbf{B}_r^{\mathsf{T}}\,\mathbf{Q}_r\Delta_{\mathbf{B}_r} + \mathbf{B}_r\mathbf{B}_r^{\mathsf{T}}\Delta_{\mathbf{Q}_r}\right) + O(\|\Delta_{\mathbf{B}_r}\|_{\mathsf{F}}^2).$$

We handle the terms in this expansion individually. First, $\operatorname{tr}\left(\mathbf{B}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{B}_r} + \mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right)$ splits into the individual terms $\operatorname{tr}\left(\mathbf{B}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{B}_r}\right)$ and $\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right)$. Using properties of the trace, and applying Lemma 3.1 to equations (37c) and (45a), we see that $\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right)$ can be written

$$\begin{aligned}
\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right) &= \operatorname{tr}\left(-\left(\sum_{k=1}^{p}\mathbf{M}_{k,r}\Delta_{\mathbf{X}}^{\mathsf{T}}\mathbf{M}_k\right)\mathbf{X}\right) \\
&= \operatorname{tr}\left(-\Delta_{\mathbf{X}}\sum_{k=1}^{p}\mathbf{M}_{k,r}\mathbf{X}^{\mathsf{T}}\mathbf{M}_k\right) \\
&= \operatorname{tr}\left(-(\mathbf{Z}^{\mathsf{T}}\mathbf{A} + \mathbf{Z}^{\mathsf{T}}\mathbf{A}_r^{\mathsf{T}})\Delta_{\mathbf{X}}\right) + \operatorname{tr}\left(\Delta_{\mathbf{X}}\mathbf{C}_r^{\mathsf{T}}\mathbf{C}\right) \\
&= \operatorname{tr}\left(-\mathbf{Z}^{\mathsf{T}}(\mathbf{A}\Delta_{\mathbf{X}} + \Delta_{\mathbf{X}}\mathbf{A}_r^{\mathsf{T}})\right) + \operatorname{tr}\left(\Delta_{\mathbf{X}}\mathbf{C}_r^{\mathsf{T}}\mathbf{C}\right) \\
&= \operatorname{tr}\left(\mathbf{Z}^{\mathsf{T}}\mathbf{B}\Delta_{\mathbf{B}_r}^{\mathsf{T}}\right) + \operatorname{tr}\left(\Delta_{\mathbf{X}}\mathbf{C}_r^{\mathsf{T}}\mathbf{C}\right) \\
&= \operatorname{tr}\left(\mathbf{B}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{B}_r}\right) + \operatorname{tr}\left(\Delta_{\mathbf{X}}\mathbf{C}_r^{\mathsf{T}}\mathbf{C}\right).
\end{aligned}$$

Applying Lemma 3.1 to equations (38b) and (44a), it follows that

$$\begin{aligned}
\operatorname{tr}\left(\Delta_{\mathbf{X}}\mathbf{C}_r\mathbf{C}^{\mathsf{T}}\right) = \operatorname{tr}\left(\mathbf{C}_r\mathbf{C}^{\mathsf{T}}\Delta_{\mathbf{X}}\right) &= \operatorname{tr}\left(-\Delta_{\mathbf{B}_r}\mathbf{B}^{\mathsf{T}}\mathbf{Z}_1\right) \\
&= \operatorname{tr}\left(-\mathbf{B}^{\mathsf{T}}\mathbf{Z}_1\Delta_{\mathbf{B}_r}\right).
\end{aligned}$$

So the term $2\operatorname{tr}(\mathbf{B}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{B}_r} + \mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}})$ in the expression of $\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r})$ becomes

$$2\operatorname{tr}\left(\mathbf{B}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{B}_r} + \mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right) = \operatorname{tr}\left(4\mathbf{B}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{B}_r}\right) - \operatorname{tr}\left(2\mathbf{B}^{\mathsf{T}}\mathbf{Z}_1\Delta_{\mathbf{B}_r}\right).$$

The term $\operatorname{tr}\left(2\mathbf{B}_r^{\mathsf{T}}\mathbf{Q}_r\Delta_{\mathbf{B}_r} + \mathbf{B}_r\mathbf{B}_r^{\mathsf{T}}\Delta_{\mathbf{Q}_r}\right)$ in $\mathcal{J}(\mathcal{S} + \Delta_{\mathcal{S}_r})$ can be dealt with using similar ideas to be written as

$$\operatorname{tr}\left(2\mathbf{B}_r^{\mathsf{T}}\mathbf{Q}_r\Delta_{\mathbf{B}_r} + \mathbf{B}_r\mathbf{B}_r^{\mathsf{T}}\Delta_{\mathbf{Q}_r}\right) = \operatorname{tr}\left(4\mathbf{B}_r^{\mathsf{T}}\mathbf{Q}_r\Delta_{\mathbf{B}_r}\right) - \operatorname{tr}\left(2\mathbf{B}_r^{\mathsf{T}}\,\mathbf{Q}_{1,r}\Delta_{\mathbf{B}_r}\right) + O(\|\Delta_{\mathbf{B}_r}\|_{\mathsf{F}}^2).$$

The expression for $\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r})$ in this case becomes

$$\begin{aligned}
\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r}) &= \mathcal{J}(\mathcal{S}_r) + \operatorname{tr}\left(4\mathbf{B}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{B}_r}\right) - \operatorname{tr}\left(2\mathbf{B}^{\mathsf{T}}\mathbf{Z}_1\Delta_{\mathbf{B}_r}\right) + \\
&\quad \operatorname{tr}\left(4\mathbf{B}_r^{\mathsf{T}}\,\mathbf{Q}_r\Delta_{\mathbf{B}_r}\right) - \operatorname{tr}\left(2\mathbf{B}_r^{\mathsf{T}}\mathbf{Q}_{1,r}\Delta_{\mathbf{B}_r}\right) + O(\|\Delta_{\mathbf{B}_r}\|_{\mathsf{F}}^2) \\
&= \mathcal{J}(\mathcal{S}_r) + \langle 2\left((2\mathbf{Q}_r - \mathbf{Q}_{1,r})\mathbf{B}_r + (2\mathbf{Z}^{\mathsf{T}} - \mathbf{Z}_1^{\mathsf{T}})\mathbf{B}\right), \Delta_{\mathbf{B}_r}\rangle_{\mathsf{F}} \\
&\quad + O(\|\Delta_{\mathbf{B}_r}\|_{\mathsf{F}}^2).
\end{aligned}$$

So, $\nabla_{\mathbf{B}_r}\mathcal{J} = 2\left((2\mathbf{Q}_r - \mathbf{Q}_{1,r})\mathbf{B}_r + (2\mathbf{Z}^{\mathsf{T}} - \mathbf{Z}_1^{\mathsf{T}})\mathbf{B}\right)$ as claimed.

Lastly, we compute $\nabla_{\mathbf{A}_r}\mathcal{J}$ in (42a). Consider an arbitrary perturbation $\Delta_{\mathbf{A}_r} \in \mathbb{R}^{r\times r}$ about zero to $\mathbf{A}_r$. As was the case for the gradient with respect to $\mathbf{B}_r$, this first-order perturbation induces perturbations $\Delta_{\mathbf{X}}$, $\Delta_{\mathbf{Z}} \in \mathbb{R}^{n\times r}$, and $\Delta_{\mathbf{P}_r}$, $\Delta_{\mathbf{Q}_r} \in \mathbb{R}^{r\times r}$ in the solutions to (37a) – (37d). These satisfy

$$\mathbf{A}\Delta_{\mathbf{X}} + \Delta_{\mathbf{X}}\mathbf{A}_r^{\mathsf{T}} + \mathbf{X}\Delta_{\mathbf{A}_r}^{\mathsf{T}} + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2) = \mathbf{0}, \tag{46a}$$

$$\mathbf{A}_r\Delta_{\mathbf{P}_r} + \Delta_{\mathbf{P}_r}\mathbf{A}_r^{\mathsf{T}} + \Delta_{\mathbf{A}_r}\mathbf{P}_r + \mathbf{P}_r\Delta_{\mathbf{A}_r}^{\mathsf{T}} + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2) = \mathbf{0}, \tag{46b}$$

$$\mathbf{A}^{\mathsf{T}}\Delta_{\mathbf{Z}} + \Delta_{\mathbf{Z}}\mathbf{A}_r + \mathbf{Z}\Delta_{\mathbf{A}_r} - \sum_{k=1}^{p}\mathbf{M}_k\Delta_{\mathbf{X}}\mathbf{M}_{k,r} + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2) = \mathbf{0}, \tag{46c}$$

$$\text{and} \quad \mathbf{A}_r^{\mathsf{T}}\Delta_{\mathbf{Q}_r} + \Delta_{\mathbf{Q}_r}\mathbf{A}_r + \Delta_{\mathbf{A}_r}^{\mathsf{T}}\mathbf{Q}_r + \mathbf{Q}_r\Delta_{\mathbf{A}_r} + \sum_{k=1}^{p}\mathbf{M}_{k,r}\Delta_{\mathbf{P}_r}\mathbf{M}_{k,r} + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2) = \mathbf{0}. \tag{46d}$$

By (40), the error $\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r})$ may be expanded as

$$\mathcal{J}(\mathcal{S}_r + \Delta_{\mathcal{S}_r}) = \mathcal{J}(\mathcal{S}_r) + 2\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right) + \operatorname{tr}\left(\mathbf{B}_r\mathbf{B}_r^{\mathsf{T}}\Delta_{\mathbf{Q}_r}\right).$$

We deal with the terms in the above expansion individually as follows. Applying Lemma 3.1 to equations (37c) and (46c), $\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right)$ can be re-written as

$$\begin{aligned}
\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right) &= \operatorname{tr}\left(-\left(\sum_{k=1}^{p}\mathbf{M}_{k,r}\Delta_{\mathbf{X}}^{\mathsf{T}}\mathbf{M}_k - \Delta_{\mathbf{A}_r}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\right)\mathbf{X}\right) \\
&\quad + o(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}) \\
&= \operatorname{tr}\left(\mathbf{X}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{A}_r}\right) - \operatorname{tr}\left(\sum_{k=1}^{p}\mathbf{M}_{k,r}\mathbf{X}^{\mathsf{T}}\mathbf{M}_k\Delta_{\mathbf{X}}\right) \\
&\quad + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2).
\end{aligned}$$

From (37d) and (46a), observe that

$$\begin{aligned}\operatorname{tr}\left(\sum_{k=1}^{p}\mathbf{M}_{k,r}\mathbf{X}^{\mathsf{T}}\mathbf{M}_k\Delta_{\mathbf{X}}\right) &= \operatorname{tr}\left((\mathbf{Z}^{\mathsf{T}}\mathbf{A}+\mathbf{A}_r^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}})\Delta_{\mathbf{X}}\right)\\ &\quad - \operatorname{tr}\left(\mathbf{C}_r^{\mathsf{T}}\mathbf{C}\Delta_{\mathbf{X}}\right)\\ &= \operatorname{tr}\left(\mathbf{Z}^{\mathsf{T}}(\mathbf{A}\Delta_{\mathbf{X}}+\Delta_{\mathbf{X}}\mathbf{A}_r^{\mathsf{T}})\right) - \operatorname{tr}\left(\mathbf{C}_r^{\mathsf{T}}\mathbf{C}\Delta_{\mathbf{X}}\right)\\ &= -\operatorname{tr}\left(\mathbf{Z}^{\mathsf{T}}\mathbf{X}\Delta_{\mathbf{A}_r}^{\mathsf{T}}\right) - \operatorname{tr}\left(\mathbf{C}_r^{\mathsf{T}}\mathbf{C}\Delta_{\mathbf{X}}\right) + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2).\end{aligned}$$

Applying Lemma 3.1 to (38b) and (46a) reveals

$$\begin{aligned}\operatorname{tr}\left(\sum_{k=1}^{p}\mathbf{M}_{k,r}\mathbf{X}^{\mathsf{T}}\mathbf{M}_k\Delta_{\mathbf{X}}\right) &= -\operatorname{tr}\left(\Delta_{\mathbf{A}_r}\mathbf{X}^{\mathsf{T}}\mathbf{Z}\right)\\ &\quad - \operatorname{tr}\left(\mathbf{X}^{\mathsf{T}}\mathbf{Z}_1\Delta_{\mathbf{A}_r}\right) + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2).\end{aligned}$$

In aggregate, the term $\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right)$ in the expansion of $\mathcal{J}(\mathcal{S}_r+\Delta_{\mathcal{S}_r})$ becomes

$$\begin{aligned}\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}^{\mathsf{T}}\Delta_{\mathbf{Z}}\right) &= \operatorname{tr}\left(2\mathbf{X}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{A}_r}\right) - \operatorname{tr}\left(\mathbf{X}^{\mathsf{T}}\mathbf{Z}_1\Delta_{\mathbf{A}_r}\right)\\ &\quad + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2).\end{aligned}$$

Using a similar line of reasoning, the term $\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}_r^{\mathsf{T}}\Delta_{\mathbf{Q}_r}\right)$ can be re-written as

$$\begin{aligned}\operatorname{tr}\left(\mathbf{B}_r\mathbf{B}_r^{\mathsf{T}}\Delta_{\mathbf{Q}_r}\right) &= \operatorname{tr}\left(4\mathbf{P}_r\mathbf{Q}_r\Delta_{\mathbf{A}_r}\right) - \operatorname{tr}\left(2\mathbf{P}_r\mathbf{Q}_{1,r}\Delta_{\mathbf{A}_r}\right)\\ &\quad + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2).\end{aligned}$$

So, the expansion $\mathcal{J}(\mathcal{S}_r+\Delta_{\mathcal{S}_r})$ simplifies to

$$\begin{aligned}\mathcal{J}(\mathcal{S}_r+\Delta_{\mathcal{S}_r}) &= \mathcal{J}(\mathcal{S}_r) + \operatorname{tr}\left(4\mathbf{X}^{\mathsf{T}}\mathbf{Z}\Delta_{\mathbf{A}_r}\right) - \operatorname{tr}\left(2\mathbf{X}^{\mathsf{T}}\mathbf{Z}_1\Delta_{\mathbf{A}_r}\right)\\ &\quad + \operatorname{tr}\left((4\mathbf{P}_r\mathbf{Q}_r\Delta_{\mathbf{A}_r}\right) - \operatorname{tr}\left(2\mathbf{P}_r\mathbf{Q}_{1,r}\Delta_{\mathbf{A}_r}\right) + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2)\\ &= \mathcal{J}(\mathcal{S}_r) + \langle 2\left((2\mathbf{Z}^{\mathsf{T}}-\mathbf{Z}_1^{\mathsf{T}})\mathbf{X}+(2\mathbf{Q}_r-\mathbf{Q}_{1,r})\mathbf{P}_r\right),\Delta_{\mathbf{A}_r}\rangle_{\mathsf{F}}\\ &\quad + O(\|\Delta_{\mathbf{A}_r}\|_{\mathsf{F}}^2).\end{aligned}$$

And $\nabla_{\mathbf{A}_r}\mathcal{J} = 2\left((2\mathbf{Z}^{\mathsf{T}}-\mathbf{Z}_1^{\mathsf{T}})\mathbf{X}+(2\mathbf{Q}_r-\mathbf{Q}_{1,r})\mathbf{P}_r\right)$. This completes the proof. □

Theorem 3.1 now yields Theorem 3.2, which contains our second significant theoretical contribution: The stationary points of the gradients of $\mathcal{J}$ in Theorem 3.1 characterize FONCs for the $\mathcal{H}_2$ optimal approximation of LQO systems, and a $\mathcal{H}_2$ optimal LQO-ROM is necessarily determined by Petrov-Galerkin projection (4). This result generalizes that of [28, Theorem 3.4] to the LQO problem setting.

Theorem 3.2. *Let $\mathcal{S}$ and $\mathcal{S}_r$ be asymptotically stable LQO systems as in (1) and (3), respectively. Let $\mathcal{S}_r$ be a local minimizer of the squared $\mathcal{H}_2$ error $\mathcal{J}(\mathcal{S}_r) = \|\mathcal{S}-\mathcal{S}_r\|_{\mathcal{H}_2}^2$. Then, $\mathcal{S}_r$ satisfies the first-order optimality conditions*

$$\mathbf{0} = \left((2\mathbf{Q}_r-\mathbf{Q}_{1,r})\mathbf{P}_r + (2\mathbf{Z}^{\mathsf{T}}-\mathbf{Z}_1^{\mathsf{T}})\mathbf{X}\right) \tag{47a}$$

$$\mathbf{0} = \left((2\mathbf{Q}_r-\mathbf{Q}_{1,r})\mathbf{B}_r + (2\mathbf{Z}^{\mathsf{T}}-\mathbf{Z}_1^{\mathsf{T}})\mathbf{B}\right) \tag{47b}$$

$$\mathbf{0} = \mathbf{C}_r\mathbf{P}_r - \mathbf{C}\mathbf{X}, \text{ and} \tag{47c}$$

$$\mathbf{0} = \mathbf{P}_r\mathbf{M}_{k,r}\mathbf{P}_r - \mathbf{X}^{\mathsf{T}}\mathbf{M}_k\mathbf{X}, \ k=1,\ldots,p, \tag{47d}$$

where $\mathbf{X},\mathbf{Z}\in\mathbb{R}^{n\times r}$, $\mathbf{P}_r,\mathbf{Q}_r\in\mathbb{R}^{r\times r}$ satisfy (37a) – (37d), and $\mathbf{Z}_1\in\mathbb{R}^{n\times r}$, $\mathbf{Q}_{1,r}\in\mathbb{R}^{r\times r}$ satisfy (38a) and (38b). Further, if $\mathbf{P}_r$ and $2\mathbf{Q}_r-\mathbf{Q}_{1,r}$ are nonsingular, then the $\mathcal{H}_2$ optimal reduced model can be obtained via a Petrov-Galerkin projection as in (4) where the model reduction matrices $\mathbf{V}_r,\mathbf{W}_r\in\mathbb{R}^{n\times r}$ are given by

$$\mathbf{V}_r = \mathbf{X}\mathbf{P}_r^{-1} \quad \text{and} \quad \mathbf{W}_r = -(2\mathbf{Z}-\mathbf{Z}_1)(2\mathbf{Q}_r-\mathbf{Q}_{1,r})^{-1},$$

and satisfy $\mathbf{W}_r^{\mathsf{T}}\mathbf{V}_r = \mathbf{I}_r$.

Proof. The FONCs in (47a) – (47d) follow as a direct result of Theorem 3.1 along with the assumption that $\mathcal{S}_r$ locally minimizes the $\mathcal{H}_2$ error. Indeed, if $\mathcal{S}_r$ is a local minimum of the function $\mathcal{J}(\mathcal{S}_r) = \|\mathcal{S}-\mathcal{S}_r\|_{\mathcal{H}_2}^2$, then the gradients of $\mathcal{J}$ are identically zero at $\mathcal{S}_r$.

What is left to prove that the optimal reduced model is obtained via a Petrov-Galerkin projection. Since $\mathbf{P}_r\in\mathbb{R}^{r\times r}$ and $2\mathbf{Q}_r-\mathbf{Q}_{1,r}\in\mathbb{R}^{r\times r}$ are assumed to be nonsingular, define the matrices $\mathbf{V}_r = \mathbf{X}\mathbf{P}_r^{-1}\in\mathbb{R}^{n\times r}$ and $\mathbf{W}_r = -(2\mathbf{Z}-\mathbf{Z}_1)(2\mathbf{Q}_r-\mathbf{Q}_{1,r})^{-1}\in\mathbb{R}^{n\times r}$. Then, the condition (47a) implies

$$\begin{aligned}-(2\mathbf{Q}_r-\mathbf{Q}_{1,r})\mathbf{P}_r &= (2\mathbf{Z}^{\mathsf{T}}-\mathbf{Z}_1^{\mathsf{T}})\mathbf{X}\\ \implies \mathbf{I}_r &= -(2\mathbf{Q}_r-\mathbf{Q}_{1,r})^{-1}(2\mathbf{Z}^{\mathsf{T}}-\mathbf{Z}_1^{\mathsf{T}})\mathbf{X}\mathbf{P}_r^{-1}\\ &= \mathbf{W}_r^{\mathsf{T}}\mathbf{V}_r.\end{aligned}$$

Then, $\mathbf{V}_r = \mathbf{X}\mathbf{P}_r^{-1}$ and the FONCs (47c) and (47d) imply

$$\begin{aligned}\mathbf{0} = \mathbf{P}_r\mathbf{C}_r^{\mathsf{T}} - \mathbf{X}^{\mathsf{T}}\mathbf{C}^{\mathsf{T}} &\implies \mathbf{C}_r = \mathbf{C}\mathbf{X}\mathbf{P}_r^{-1} = \mathbf{C}\mathbf{V}_r,\\ \mathbf{0} = \mathbf{P}_r\mathbf{M}_{k,r}\mathbf{P}_r - \mathbf{X}^{\mathsf{T}}\mathbf{M}_k\mathbf{X} &\implies \mathbf{M}_{k,r} = (\mathbf{P}_r^{-1}\mathbf{X}^{\mathsf{T}})\mathbf{M}_k(\mathbf{X}\mathbf{P}_r^{-1})\\ &\qquad\qquad = \mathbf{V}_r^{\mathsf{T}}\mathbf{M}_k\mathbf{V}_r,\end{aligned}$$

for each $k=1,\ldots,p$. $\mathbf{W}_r = -(2\mathbf{Z}-\mathbf{Z}_1)(2\mathbf{Q}_r-\mathbf{Q}_{1,r})^{-1}$. And the condition (47b) jointly imply

$$\begin{aligned}\mathbf{0} &= \left((2\mathbf{Q}_r-\mathbf{Q}_{1,r})\mathbf{B}_r + (2\mathbf{Z}^{\mathsf{T}}-\mathbf{Z}_1^{\mathsf{T}})\mathbf{B}\right)\\ \implies \mathbf{B}_r &= -(2\mathbf{Q}_r-\mathbf{Q}_{1,r})^{-1}(2\mathbf{Z}^{\mathsf{T}}-\mathbf{Z}_1^{\mathsf{T}})\mathbf{B}\\ &= \mathbf{W}_r^{\mathsf{T}}\mathbf{B}.\end{aligned}$$

Lastly, $\mathbf{V}_r = \mathbf{X}\mathbf{P}_r^{-1}$ implies $\mathbf{X} = \mathbf{V}_r\mathbf{P}_r$. This identity, and multiplying (37c) by $\mathbf{W}_r^{\mathsf{T}}$ yields

$$\begin{aligned}&\mathbf{W}_r^{\mathsf{T}}\mathbf{A}(\mathbf{V}_r\mathbf{P}_r) + \underbrace{\mathbf{W}_r^{\mathsf{T}}(\mathbf{V}_r}_{=\mathbf{I}_r}\mathbf{P}_r)\mathbf{A}_r^{\mathsf{T}} + \underbrace{(\mathbf{W}_r^{\mathsf{T}}\mathbf{B})}_{=\mathbf{B}_r}\mathbf{B}_r^{\mathsf{T}} = \mathbf{0}\\ \implies\ &(\mathbf{W}_r^{\mathsf{T}}\mathbf{A}\mathbf{V}_r)\mathbf{P}_r + \mathbf{P}_r\mathbf{A}_r^{\mathsf{T}} + \mathbf{B}_r\mathbf{B}_r^{\mathsf{T}} = \mathbf{0},\end{aligned}$$

which, by comparison with (37a), yields the remaining identity $\mathbf{A}_r = \mathbf{W}_r^\mathsf{T}\mathbf{A}\mathbf{V}_r$ since $\mathbf{P}_r$ is nonsingular. This completes the proof. □

Two remarks are in order.

Remark 3.1. *Both the gradients of $\mathcal{J}(\mathcal{S}_r) = \|\mathcal{S} - \mathcal{S}_r\|_{\mathcal{H}_2}^2$ in Theorem 3.1 and the Gramian-based FONCs for $\mathcal{H}_2$ optimality in Theorem 3.2 generalize the analogous results in the LTI setting to LQO systems, i.e., they establish the Wilson framework for $\mathcal{H}_2$ optimal model reduction for LQO systems. Indeed, in the particular instance where $\mathbf{M}_k = \mathbf{0}_{n\times n}$, and so, $\mathbf{M}_{k,r} = \mathbf{0}_{r\times r}$ as well, for each output k and $\mathcal{S} = \mathcal{S}_1$ is an LTI system as in (26), the reduced-order QO observability Gramian $\mathbf{Q}_r \in \mathbb{R}^{r\times r}$ and solution $\mathbf{Z} \in \mathbb{R}^{n\times r}$ to (37d) reduce to $\mathbf{Q}_r = \mathbf{Q}_{1,r} \in \mathbb{R}^{r\times r}$ in (11) and $\mathbf{Z} = \mathbf{Z}_1 \in \mathbb{R}^{n\times r}$ solving (38b), respectively. If we apply Theorem 3.1 in this case, the gradients with respect to $\mathbf{A}_r$ and $\mathbf{B}_r$ then become*

$$\nabla_{\mathbf{A}_r}\mathcal{J} = 2\left(\mathbf{Q}_{1,r}\mathbf{P}_r + \mathbf{Z}_1^\mathsf{T}\mathbf{X}\right)$$
$$\text{and}\quad \nabla_{\mathbf{B}_r}\mathcal{J} = 2\left(\mathbf{Q}_{1,r}\mathbf{B}_r + \mathbf{Z}_1^\mathsf{T}\mathbf{B}\right)$$

which are precisely those in (32); the gradient with respect to $\mathbf{C}_r$ is unchanged. The Gramian-based FONCs in Theorem 3.2 reduce in an obviously similar way. Thus Theorem 3.1 and Theorem 3.2 contain (32) and (33) as a special case.

Remark 3.2. *In some applications, there is no linear component in the output, i.e., $\mathbf{C} = \mathbf{0}_{p\times n}$, so the observation term $\mathbf{y} = \mathbf{y}_2$ of (1) is purely quadratic. In such instances, $\mathbf{Q}_1 \in \mathbb{R}^{n\times n}$, $\mathbf{Q}_{1,r} \in \mathbb{R}^{r\times r}$ and $\mathbf{Z}_1 \in \mathbb{R}^{r\times r}$ are all zero. The gradients of $\mathcal{J}$ in Theorem 3.1 then become*

$$\nabla_{\mathbf{A}_r}\mathcal{J} = 4\left(\mathbf{Q}_r\mathbf{P}_r + \mathbf{Z}^\mathsf{T}\mathbf{X}\right)$$
$$\nabla_{\mathbf{B}_r}\mathcal{J} = 4\left(\mathbf{Q}_r\mathbf{B}_r + \mathbf{Z}^\mathsf{T}\mathbf{B}\right)$$
$$\nabla_{\mathbf{M}_{k,r}}\mathcal{J} = 2\left(\mathbf{P}_r\mathbf{M}_{k,r}\mathbf{P}_r - \mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X}\right),\ k = 1\ldots,p,$$

where $\mathbf{Q}_r \in \mathbb{R}^{r\times r}$, $\mathbf{Z} \in \mathbb{R}^{n\times r}$ solve (37b) and (37d) respectively, with $\mathbf{C}$ and $\mathbf{C}_r$ equal to zero.

4 A two-sided iteration algorithm for LQO model reduction

Theorem 3.2 states that any local minimizer (3) of the $\mathcal{H}_2$ error in (24) is necessarily defined by a Petrov-Galerkin framework (4) where the optimal reduction matrices are given as $\mathbf{V}_r = \mathbf{X}\mathbf{P}_r^{-1} \in \mathbb{R}^{n\times r}$ and $\mathbf{W}_r = -(2\mathbf{Z}-\mathbf{Z}_1)(2\mathbf{Q}_r - \mathbf{Q}_{1,r})^{-1} \in \mathbb{R}^{n\times r}$. It follows directly that the $\mathcal{H}_2$ optimal LQO-ROM in question has an equivalent state-space realization given by

$$\begin{aligned}
\mathbf{A}_r &= \left((2\mathbf{Z}^\mathsf{T} - \mathbf{Z}_1^\mathsf{T})\mathbf{X}\right)^{-1}\left(2\mathbf{Z}^\mathsf{T} - \mathbf{Z}_1^\mathsf{T}\right)\mathbf{A}\mathbf{X}\\
\mathbf{B}_r &= \left((2\mathbf{Z}^\mathsf{T} - \mathbf{Z}_1^\mathsf{T})\mathbf{X}\right)^{-1}\left(2\mathbf{Z}^\mathsf{T} - \mathbf{Z}_1^\mathsf{T}\right)\mathbf{B}\\
\mathbf{C}_r &= \mathbf{C}\mathbf{X}\\
\mathbf{M}_{k,r} &= \mathbf{X}^\mathsf{T}\mathbf{M}_k\mathbf{X},\quad k = 1\ldots,p,
\end{aligned}$$

under the change of coordinate transformation $\mathbf{T} = \mathbf{P}_r \in \mathbb{R}^{r\times r}$ (and with $\mathbf{T}^{-1} = -(2\mathbf{Q}_r - \mathbf{Q}_{1,r}) \in \mathbb{R}^{r\times r}$). In other words, a $\mathcal{H}_2$ optimal LQO-ROM is defined by the projection framework (4) along with the matrices $\mathbf{V}_r = \mathbf{X} \in \mathbb{R}^{n\times r}$ and $\mathbf{W}_r = 2\mathbf{Z} - \mathbf{Z}_1 \in \mathbb{R}^{n\times r}$, where $\mathbf{W}_r^\mathsf{T}\mathbf{V}_r = (2\mathbf{Z}^\mathsf{T} - \mathbf{Z}_1^\mathsf{T})\mathbf{X}$ is invertible due to the identity from (47a)

$$(2\mathbf{Z}^\mathsf{T} - \mathbf{Z}_1^\mathsf{T})\mathbf{X} = -(2\mathbf{Q}_r - \mathbf{Q}_{1,r})\mathbf{P}_r,$$

and the assumption that $2\mathbf{Q}_r - \mathbf{Q}_{1,r}$ and $\mathbf{P}_r$ are nonsingular. The right projection matrix $\mathbf{V}_r = \mathbf{X}$ satisfies the Sylvester equation (37c) corresponding to the $\mathcal{H}_2$ optimal LQO-ROM. Because $\mathbf{Z} \in \mathbb{R}^{n\times r}$ and $\mathbf{Z}_1 \in \mathbb{R}^{n\times r}$ are the solutions to (37d) and (38b), the left projection matrix $\mathbf{W}_r = \widehat{\mathbf{Z}} \coloneqq 2\mathbf{Z} - \mathbf{Z}_1$ satisfies a linear combination of (37d) and (38b), i.e.

$$\begin{aligned}
&\mathbf{A}^\mathsf{T}\widehat{\mathbf{Z}} + \widehat{\mathbf{Z}}\mathbf{A}_r - 2\sum_{k=1}^{p}\mathbf{M}_k\mathbf{X}\mathbf{M}_{k,r} - \mathbf{C}^\mathsf{T}\mathbf{C}_r = \mathbf{0}\\
&\text{where } \mathbf{X} \in \mathbb{R}^{n\times r} \text{ satisfies } \ \mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}_r^\mathsf{T} + \mathbf{B}\mathbf{B}_r^\mathsf{T} = \mathbf{0}.
\end{aligned} \tag{48}$$

Note that the Sylvester equations in (48) depend explicitly upon the $\mathcal{H}_2$ optimal reduced model. So, if we wanted to project down the full-order matrices using the optimal choice of $\mathbf{V}_r = \mathbf{X}$ and $\mathbf{W}_r = \widehat{\mathbf{Z}}$, this would require *a priori* knowledge of the optimal reduced-order quantities $\mathbf{A}_r$, $\mathbf{B}_r$, $\mathbf{C}_r$, and $\mathbf{M}_{k,r}$, which is of course unavailable.

Based on these observations, we propose a computationally efficient algorithm that uses iterated projection with the solutions to the Sylvester equations (48) in Algorithm 1. This leads to a fixed point iteration where at every step, the equations (48) are solved to project the full-order LQO system to obtain a new reduced model. This procedure is repeated until some stopping criterion is satisfied. The idea of using fixed-point algorithms in $\mathcal{H}_2$ optimal model reduction is not new; indeed, ours is inspired by a similar approach for linear $\mathcal{H}_2$ optimal model reduction introduced in [30]. The computational procedure proposed in [30] is called the *two-sided iteration algorithm*, which performs iterative projection on the linear full-order model using the solutions to (37c) and (38b). Practical improvements that make TSIA

more computationally efficient were considered in [8]. Because of this, we call Algorithm 1 the *linear quadratic output two-sided iteration algorithm* (LQO-TSIA).

Algorithm 1 Linear quadratic output two-sided iteration algorithm (LQO-TSIA) for $\mathcal{H}_2$ optimal LQO-MOR

Input: LQO order-n model $\mathcal{S} = (\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{M}_1, \ldots, \mathbf{M}_k)$, $1 \leq r < n$, and initial order-r LQO-ROM $\mathcal{S}_r^{(0)} = (\mathbf{A}_r^{(0)}, \mathbf{B}_r^{(0)}, \mathbf{C}_r^{(0)}, \mathbf{M}_{1,r}^{(0)}, \ldots, \mathbf{M}_{p,r}^{(0)})$ with $\lambda(\mathbf{A})$, $\lambda(\mathbf{A}_r^{(0)}) \subset \mathbb{C}_-$, $\mathbf{M}_k, \mathbf{M}_{k,r}^{(0)}$ symmetric for all k, and convergence tolerance $\epsilon > 0$.

1: **while error** $> \epsilon$ **do**

2: Solve the Sylvester equations (48) for $\mathbf{X}, \widehat{\mathbf{Z}} \in \mathbb{R}^{n \times r}$:

$$\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{A}_r^{(j)^\mathsf{T}} + \mathbf{B}\mathbf{B}_r^{(j)^\mathsf{T}} = \mathbf{0},$$
$$\mathbf{A}^\mathsf{T}\widehat{\mathbf{Z}} + \widehat{\mathbf{Z}}\mathbf{A}_r^{(j)} - 2\sum_{k=1}^{p} \mathbf{M}_k \mathbf{X} \mathbf{M}_{k,r}^{(j)} - \mathbf{C}^\mathsf{T}\mathbf{C}_r^{(j)} = \mathbf{0}.$$

3: Perform orthogonalization on the solution matrices:

$$\mathbf{V}_r = \operatorname{orth}(\mathbf{X}), \quad \mathbf{W}_r = \operatorname{orth}(\widehat{\mathbf{Z}}).$$

4: Compute reduced-order matrices:

$$\begin{aligned}
\mathbf{A}_r^{(j+1)} &= (\mathbf{W}_r^\mathsf{T}\mathbf{V}_r)^{-1}\mathbf{W}_r^\mathsf{T}\mathbf{A}\mathbf{V}_r, \\
\mathbf{B}_r^{(j+1)} &= (\mathbf{W}_r^\mathsf{T}\mathbf{V}_r)^{-1}\mathbf{W}_r^\mathsf{T}\mathbf{B}, \\
\mathbf{C}_r^{(j+1)} &= \mathbf{C}\mathbf{V}_r, \\
\mathbf{M}_{k,r}^{(j+1)} &= \mathbf{V}_r^\mathsf{T}\mathbf{M}_k\mathbf{V}_r, \quad k = 1, \ldots, p.
\end{aligned}$$

5: **end while**.

Output: Order-r LQO-ROM specified by the reduced order matrices $\mathcal{S}_r = (\mathbf{A}_r, \mathbf{B}_r, \mathbf{C}_r, \mathbf{M}_{1,r}, \ldots, \mathbf{M}_{p,r})$.

The main computational cost at each step lies in solving the Sylvester equations in Step 2 of Algorithm 1. However, because the solution matrices $\mathbf{X}, \widehat{\mathbf{Z}} \in \mathbb{R}^{n \times r}$ are tall and skinny, they can be obtained efficiently, e.g., by computing a Schur decomposition of *the reduced matrix* $\mathbf{A}_r$, and solving for their columns via shifted linear solves. Often, the full-order coefficient matrix $\mathbf{A}$ has some structured sparsity which can also be exploited [8, Sec. 3]. The iteration in Algorithm 1 repeats until either some preset number of steps is reached, or the algorithm converges within the inputted tolerance based on some pre-determined stopping criterion. As is the case with any optimization problem, there exist a variety of possible choices for measuring convergence. Because we are seeking to minimize the squared $\mathcal{H}_2$ error $\mathcal{J}$ in (24), for simplicity we use the change in the (relative) squared $\mathcal{H}_2$ error between consecutive iterates to monitor convergence. By (40), the square of the relative error due to $\mathcal{S}_r^{(j)}$ at step j of the iteration is given by

$$\eta^{(j)} := \frac{\|\mathcal{S} - \mathcal{S}_r^{(j)}\|_{\mathcal{H}_2}^2}{\|\mathcal{S}\|_{\mathcal{H}_2}^2} = \frac{\|\mathcal{S}\|_{\mathcal{H}_2}^2 + \|\mathcal{S}_r^{(j)}\|_{\mathcal{H}_2}^2 + 2\operatorname{tr}(\mathbf{B}\mathbf{Z}\mathbf{B}_r^{(j)^\mathsf{T}})}{\|\mathcal{S}\|_{\mathcal{H}_2}^2}, \tag{49}$$

where $\mathbf{Z} \in \mathbb{R}^{n \times r}$ satisfies (37d) for $\mathcal{S}_r^{(j)}$. Then, Algorithm 1 is deemed to have converged if $|\eta^{(j)} - \eta^{(j-1)}|/\eta^{(1)} \leq \epsilon$ for $\epsilon > 0$. The $\mathcal{H}_2$ norm of the FOM in (49) can be pre-computed at the start of the iteration. Another natural option is to monitor changes in the gradients $\nabla_{\mathbf{A}_r}\mathcal{J}$, $\nabla_{\mathbf{B}_r}\mathcal{J}$, $\nabla_{\mathbf{C}_r}\mathcal{J}$ and $\nabla_{\mathbf{M}_{k,r}}\mathcal{J}$ of the error function $\mathcal{J}$, and terminate when they are sufficiently small. The information required to compute these quantities is readily available from the iteration itself. However, for a relative metric that uses scaled gradients (as is usually done in practice) this would require computing the Hessian of $\mathcal{J}$, which is not directly available from already computed quantities. So, we do not consider this convergence criterion further. Upon convergence of Algorithm 1, the optimality conditions in Theorem 3.2 are satisfied. As is the case for the linear TSIA [30] (and IRKA [17] for the interpolatory formulation of the linear $\mathcal{H}_2$ optimality framework) convergence of Algorithm 1 is *not* guaranteed in general since it is a fixed-point iteration. However, similar to TSIA and IRKA, in practice the algorithm performs well. We include a study of the convergence of Algorithm 1 in practice in Section 5. For guaranteed convergence, one may consider developing a descent-based algorithm based on the explicit gradient formulae (42) we derived. We leave those computational considerations to a future work.

Remark 4.1. *For large-scale problems, computing the $\mathcal{H}_2$ norm of the FOM entails the solution of the large-scale Lyapunov equation in (14), which may not be feasible in practice. The examples we consider in Section 5 have a modest state-space dimension, and so the $\mathcal{H}_2$ norm of the full-order system $\mathcal{S}$ can be computed without issue. However, note that in the error formula (49), the only part that varies in each iteration is the 'tail', that is*

$$\tau^{(j)} := \|\mathcal{S}_r^{(j)}\|_{\mathcal{H}_2}^2 + 2\operatorname{tr}(\mathbf{B}\mathbf{Z}\mathbf{B}_r^{(j)^\mathsf{T}}) \tag{50}$$

Therefore, if computing the true (or an approximate) $\mathcal{H}_2$ norm of the FOM is not feasible, one may monitor the tail $\tau^{(j)}$ and terminate the algorithm when the change in the tails fall below the inputted convergence tolerance. We investigate this choice in Section 5.

We conclude this section with a comment regarding the recent work [15, Algorithm 1] where a two-sided iteration, similar to Algorithm 1, was proposed. The algorithm in [15] was proposed without making explicit reference to $\mathcal{H}_2$ optimality conditions; it was developed heuristically based on the corresponding two-sided iteration for linear systems in [30]. As a result, despite the structural similarities, the method of [15] solves a different Sylvester equation to compute $\widehat{\mathbf{Z}}$ and thus, unlike Algorithm 1, the resulting ROM does not satisfy the $\mathcal{H}_2$ optimality conditions in Theorem 3.2 upon convergence.

5 Numerical Results

We test the effectiveness of the approach in Algorithm 1 on a numerical example. We compare our method to the structure-preserving balanced truncation algorithm for LQO systems from [7]. We refer to this approach by LQO-BT. All experiments were performed on a MacBook Air with 8 gigabytes of RAM and an Apple M2 processor running macOS Ventura version 13.4 with MATLAB 23.2.0.2515942 (R2023b) Update 7. The source codes for recreating the numerical experiments and the computed results are available at [23].

We consider the example of a 1D Advection-diffusion equation from [13, Section 4.1]. We briefly overview the relevant details of the example. The governing equations are

$$\begin{aligned} \frac{\partial}{\partial t}v(t,x) - \alpha\frac{\partial^2}{\partial x^2}v(t,x) + \beta\frac{\partial}{\partial x}v(t,x) = 0, \\ v(t,0) = u_0(t), \quad \alpha\frac{\partial}{\partial x}v(t,1) = u_1(t), \\ v(0,x) = 0, \end{aligned} \tag{51}$$

for $x \in (0,1)$ and $t \in (0,T)$ and inputs $u_0, u_1 \in \mathcal{L}_2(0,T)$; the diffusion and advection coefficients are $\alpha > 0$ and $\beta \geq 0$, respectively. The quantity of interest that we consider is

$$\frac{1}{2}\int_0^1 |v(t,x) - 1|^2 dx, \tag{52}$$

Such an observable may arise from, e.g., the objective function in a quadratic optimal control problem. Discretizing the equations in (51) using $n+1$ equidistant spatial points yields an order-n state-space model of the form (1) with $m = 2$ inputs and $p = 1$ output y. Let $\mathbf{x}(t) \in \mathbb{R}^n$ denote the spatial discretization of $v(t,x)$, $h := 1/n$, and $\mathbf{1} \in \mathbb{R}^n$ the vector consisting of all ones. Then, the discretization provides an approximation to the quadratic cost function (52)

$$\begin{aligned} \frac{h}{2}\|\mathbf{x}(t) - \mathbf{1}\|_2^2 &= \underbrace{-h\mathbf{1}^\mathsf{T}\mathbf{x}(t)}_{=y_1(t)} + \underbrace{\frac{h}{2}\mathbf{x}(t)^\mathsf{T}\mathbf{x}(t)}_{=y_2(t)} + \frac{h}{2}\|\mathbf{1}\|_2^2 \\ &= y(t) + \frac{h}{2}\|\mathbf{1}\|_2^2. \end{aligned}$$

To fit (1), we consider the single output of the discretized system to be given by $y(t) = \mathbf{c}\mathbf{x}(t) + \mathbf{x}(t)^\mathsf{T}\mathbf{M}\mathbf{x}(t)$ where $\mathbf{c} := -h\mathbf{1}^\mathsf{T} \in \mathbb{R}^{1\times n}$ and $\mathbf{M} := \frac{h}{2}\mathbf{I}_n \in \mathbb{R}^{n\times n}$ is the $n \times n$ identity matrix. The approximation to the cost is recovered from the output $y(t)$ via $\frac{h}{2}\|\mathbf{x}(t) - \mathbf{1}\|_2^2 = y(t) + \frac{h}{2}\|\mathbf{1}\|_2^2$.

We perform a discretization of (51) using $n = 300$ spatial grid points to obtain an LQO system in state-space form (1); the advection and diffusion parameters are selected as $\alpha = 0.01$ and $\beta = 1$. Reduced-order models of order $r = 30$ are computed using the LQO-TSIA and LQO-BT approaches. The LQO-TSIA is run with an overtly strict tolerance of $\epsilon = 10^{-14}$ to investigate the two convergence criteria discussed in Section 4 (in practice one would choose a lower stopping criterion). The change in the relative squared $\mathcal{H}_2$ error functional at each step (49) is used to determine convergence. As an initialization, $\mathbf{A}_r^{(0)} \in \mathbb{R}^{r\times r}$ is taken as a diagonal matrix with $r = 30$ logarithmically spaced points chosen in the interval from -10^0 and -10^4, while $\mathbf{B}_r^{(0)} \in \mathbb{R}^{r\times m}$ is the leading m columns of the $r \times r$ identity matrix, $\mathbf{C}_r^{(0)} \in \mathbb{R}^{p\times r}$ is the leading p rows of the $r \times r$ identity matrix, and $\mathbf{M}_r^{(0)} \in \mathbb{R}^{r\times r}$ is the $r \times r$ identity matrix. We test the performance of the two reduced-order models in reproducing the output of the full-order system using two choices for the input u_1; in either case, we enforce the Neumann boundary condition of $u_0(t) = 0$. We first simulate the full and reduced-order dynamics influenced by the sinusoidal input $u_1(t) = 0.5\cos(\pi t) + 1$ and then by the exponentially damped quadratic input $u_1(t) = t^2 e^{-t/5}$. To capture the full spread of the output dynamics in each experiment, the first simulation occurs over $T = 10$ seconds, and the second over $T = 30$ seconds. We plot the output response of the full and reduced-order systems due to the sinusoidal and exponentially damped inputs and the associated relative pointwise error in Figure 1 and Figure 2, respectively. As evidenced by the error plots, both LQO-TSIA and LQO-BT are successful in replicating the full-order output for both inputs, while the approximation due to LQO-TSIA performs slightly better on average. (After we numerically investigate different convergence criteria of LQO-TSIA next, we will provide a more detailed comparison to LQO-BT with respect to the $\mathcal{H}_2$ error.)

To numerically investigate the behavior of the two different convergence criteria for LQO-TSIA discussed in Section 4,

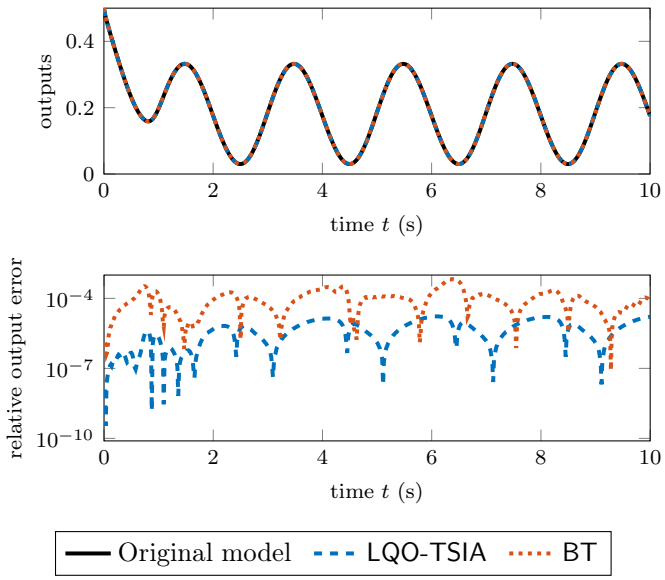

Figure 1: Output response of the FOM and order $r = 30$ ROMs computed by LQO-TSIA and LQO-BT driven by the inputs $u_0(t) = 0$ and $u_1(t) = .5\cos(\pi t) + 1$.

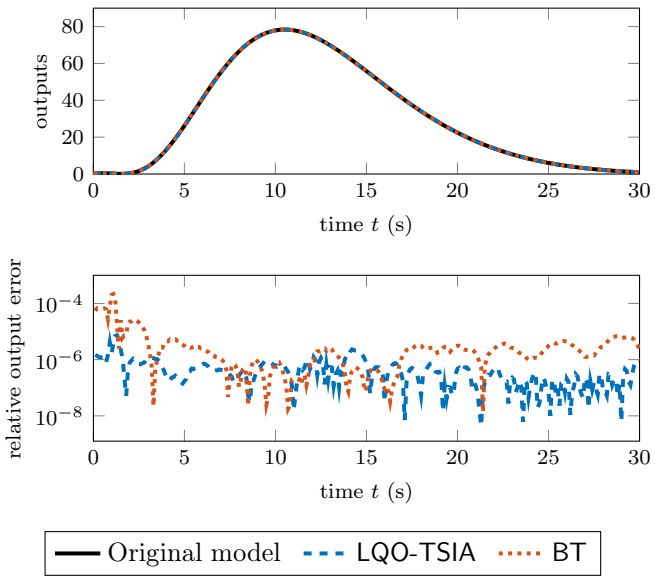

Figure 2: Output response of the FOM and order $r = 30$ ROMs computed by LQO-TSIA and LQO-BT driven by the inputs $u_0(t) = 0$ and $u_1(t) = t^2e^{-t/5}$.

we track the following two quantities during the iteration for the previously discussed $r = 30$ case: (1) The relative change in the (relative) squared $\mathcal{H}_2$ error, $|\eta^{(j)} - \eta^{(j-1)}|/|\eta^{(1)}|$ for $\eta^{(j)}$ in (49), and (2) The relative change in the variable tails of the error, $|\tau^{(j)} - \tau^{(j-1)}|/|\tau^{(1)}|$ for $\tau^{(j)}$ in (50) mentioned in Remark 4.1. The magnitude of these quantities throughout the iteration is plotted in Figure 3. The relative changes in the $\mathcal{H}_2$ errors $|\eta^{(j)} - \eta^{(j-1)}|/\eta^{(1)}$ between consecutive iterates level off around the iterates $j = 75$ to $j = 80$. At this point, a local minima of the squared $\mathcal{H}_2$ error has been found, and the algorithm converges shortly after. Evidently, $|\tau^{(j)} - \tau^{(j-1)}|/|\tau^{(1)}$ tracks $|\eta^{(j)} - \eta^{(j-1)}|/\eta^{(1)}$ very well until convergence. At the point at which the $\mathcal{H}_2$ error begins to stagnate, the change in tails continues to decay. So, this convergence criterion may cause Algorithm 1 to continue to iterate past the point at which relative the $\mathcal{H}_2$ error has stagnated. However, these two measures start to stagnate around a relative measure of 10^{-10}, a tolerance much tighter than one would usually employ in practice. Therefore, even though we conclude that monitoring changes in the relative squared $\mathcal{H}_2$ error (49) to be a more reliable convergence metric, monitoring changes in the tails according to (50) prove to be a robust alternative if computing the $\mathcal{H}_2$ norm of $\mathcal{S}$ is intractable.

Finally, to better compare the performance of LQO-TSIA and LQO-BT, we compute a hierarchy of reduced models for orders $r = 2, 4, \ldots, 30$ using both LQO-TSIA and LQO-

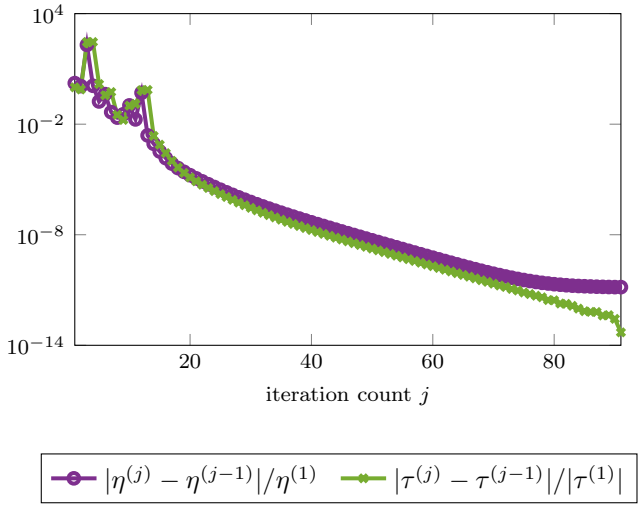

Figure 3: Convergence of the LQO-TSIA algorithm for $r = 30$. The relative changes in $\eta^{(j)}$ and $\tau^{(j)}$ according to (49) and (50) are measured throughout the iteration.

BT. The same initialization and convergence criterion from before are used for LQO-TSIA. For each order, the squared relative $\mathcal{H}_2$ error due to the order-r LQO-TSIA and LQO-BT reduced models are computed using formula (19). The results are plotted in Figure 4. As the order-r increases, the

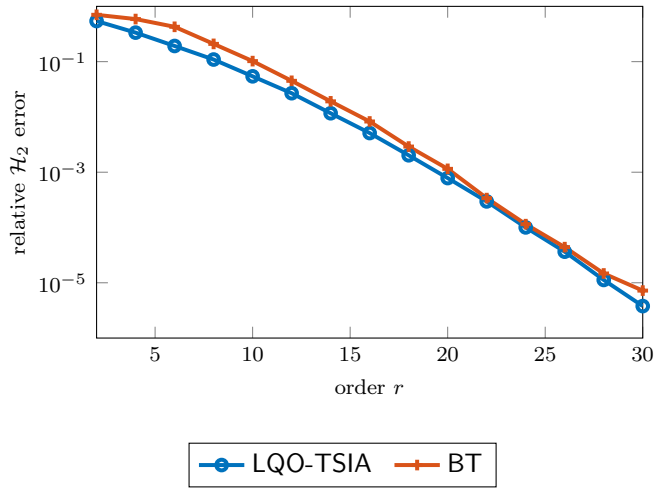

Figure 4: Relative $\mathcal{H}_2$ errors for hierarchy of LQO-TSIA and LQO-BT reduced models for orders of reduction $r = 2, 4, \ldots, 30$.

relative $\mathcal{H}_2$ error due to the LQO-TSIA and LQO-BT ROMs decrease as well; the LQO-TSIA models exhibit a slightly smaller error for all orders of reduction.

6 Conclusions

In this work, we have presented a novel $\mathcal{H}_2$ optimality framework for the model reduction of LQO dynamical systems. The novel contributions are the computation of gradients of the squared $\mathcal{H}_2$ error for such systems in Theorem 3.1, and the generalization the well-known Gramian-based $\mathcal{H}_2$ optimality framework for linear dynamical systems to the model reduction of LQO systems in Theorem 3.2. Finally, a linear quadratic-output two-sided iteration algorithm, LQO-TSIA, is proposed in Algorithm 1 for the efficient $\mathcal{H}_2$ model reduction of LQO systems. We illustrate the effectiveness of LQO-TSIA on a numerical example resulting from an optimal control problem. In the future, we will consider alternative $\mathcal{H}_2$ optimality frameworks for LQO system using the concept on rational function interpolation.

Acknowledgments

The work of Gugercin and Reiter is based upon work supported by the National Science Foundation under Grant No. AMPS-1923221. We thank Alejandro Diaz and Matthias Heinkenschloss for providing the code for creating the test examples in Section 5.

References

[1] Antoulas, A.C.: Approximation of large-scale dynamical systems. SIAM, Philadelphia, PA (2005)

[2] Antoulas, A.C., Beattie, C.A., Gugercin, S.: Interpolatory Methods for Model Reduction. Computational Science & Engineering. SIAM, Philadelphia, PA (2020)

[3] Antoulas, A.C., Sorensen, D.C., Zhou, Y.: On the decay rate of Hankel singular values and related issues. Systems & Control Letters **46**(5), 323–342 (2002)

[4] Aumann, Q., Werner, S.W.R.: Structured model order reduction for vibro-acoustic problems using interpolation and balancing methods. J. Sound Vib. **543**, 117363 (2023)

[5] Benner, P., Breiten, T.: Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems. SIAM Journal on Matrix Analysis and Applications **33**(3), 859–885 (2012)

[6] Benner, P., Goyal, P., Gugercin, S.: $\mathcal{H}_2$-quasi-optimal model order reduction for quadratic-bilinear control systems. SIAM Journal on Matrix Analysis and Applications **39**(2), 983–1032 (2018)

[7] Benner, P., Goyal, P., Pontes Duff, I.: Gramians, energy functionals, and balanced truncation for linear dynamical systems with quadratic outputs. IEEE Transactions on Automatic Control **67**(2), 886–893 (2021)

[8] Benner, P., Köhler, M., Saak, J.: Sparse-dense Sylvester equations in $\mathcal{H}_2$-model order reduction. Preprint MPIMD/11-11, Max Planck Institute Magdeburg (2011)

[9] Benner, P., Mehrmann, V., Sorensen, D.C.: Dimension reduction of large-scale systems, vol. 45. Springer (2005)

[10] Benner, P., Ohlberger, M., Cohen, A., Willcox, K.: Model reduction and approximation: theory and algorithms. SIAM, Philadelphia, PA (2017)

[11] Brewer, J.: Kronecker products and matrix calculus in system theory. IEEE Transactions on circuits and

systems **25**(9), 772–781 (1978)

[12] Coleman, R.: Calculus on normed vector spaces. Springer Science & Business Media (2012)

[13] Diaz, A.N., Heinkenschloss, M., Gosea, I.V., Antoulas, A.C.: Interpolatory model reduction of quadratic-bilinear dynamical systems with quadratic-bilinear outputs. Advances in Computational Mathematics **49**(6), 1–28 (2023)

[14] Flagg, G., Gugercin, S.: Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems. SIAM Journal on Matrix Analysis and Applications **36**(2), 549–579 (2015)

[15] Gosea, I.V., Antoulas, A.C.: A two-sided iterative framework for model reduction of linear systems with quadratic output. In: 2019 IEEE 58th Conference on Decision and Control (CDC), pp. 7812–7817. IEEE (2019)

[16] Gosea, I.V., Gugercin, S.: Data-driven modeling of linear dynamical systems with quadratic output in the AAA framework. Journal of Scientific Computing **91**(1), 16 (2022)

[17] Gugercin, S., Antoulas, A.C., Beattie, C.: $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems. SIAM Journal on Matrix Analysis and Applications **30**(2), 609–638 (2008)

[18] Holicki, T., Nicodemus, J., Schwerdtner, P., Unger, B.: Energy matching in reduced passive and port-Hamiltonian systems. e-prints 2309.05778, arXiv (2023). URL https://arxiv.org/abs/2309.05778

[19] Meier, L., Luenberger, D.: Approximation of linear constant systems. IEEE Transactions on Automatic Control **12**(5), 585–588 (1967)

[20] Przybilla, J., Pontes Duff, I., Goyal, P., Benner, P.: Balanced truncation of descriptor systems with a quadratic output. e-prints 2402.14716, arXiv (2024). URL https://arxiv.org/abs/2402.14716

[21] Pulch, R.: Energy-based model order reduction for linear stochastic Galerkin systems of second order. PAMM **23**(3), e202300038 (2023)

[22] Pulch, R., Narayan, A.: Balanced truncation for model order reduction of linear dynamical systems with quadratic outputs. SIAM Journal on Scientific Computing **41**(4), A2270–A2295 (2019)

[23] Reiter, S.: Code, data, and results for numerical experiments in "$\mathcal{H}_2$ optimal model reduction of linear systems with multiple quadratic outputs" (version 1.1) (2024). doi:10.5281/zenodo.11104814

[24] Reiter, S., Werner, S.W.R.: Interpolatory model order reduction of large-scale dynamical systems with root mean squared error measures. e-prints 2403.08894, arXiv (2024). URL https://arxiv.org/abs/2403.08894

[25] Song, Q.Y., Zulfiqar, U., Xiao, Z.H., Uddin, M.M., Sreeram, V.: Balanced truncation of linear systems with quadratic outputs in limited time and frequency intervals. e-prints 2402.11445, arXiv (2024). URL https://arxiv.org/abs/2402.11445

[26] Van Beeumen, R., Meerbergen, K.: Model reduction by balanced truncation of linear systems with a quadratic output. In: AIP Conference Proceedings, vol. 1281, pp. 2033–2036. American Institute of Physics (2010)

[27] Van Beeumen, R., Van Nimmen, K., Lombaert, G., Meerbergen, K.: Model reduction for dynamical systems with quadratic output. International Journal for Numerical Methods in Engineering **91**(3), 229–248 (2012)

[28] Van Dooren, P., Gallivan, K.A., Absil, P.A.: H_2-optimal model reduction of MIMO systems. Applied Mathematics Letters **21**(12), 1267–1273 (2008)

[29] Wilson, D.: Optimum solution of model-reduction problem. In: Proceedings of the Institution of Electrical Engineers, vol. 117, pp. 1161–1165. IET (1970)

[30] Xu, Y., Zeng, T.: Optimal $\mathcal{H}_2$ model reduction for large scale MIMO systems via tangential interpolation. Int. J. Numer. Anal. Model. **8**(1), 174–188 (2011)

[31] Yan, W.Y., Lam, J.: An approximate approach to H_2 optimal model reduction. IEEE Transactions on Automatic Control **44**(7), 1341–1358 (1999)

[32] Zhang, L., Lam, J.: On H_2 model reduction of bilinear systems. Automatica **38**(2), 205–216 (2002)