SIAM J. MATRIX ANAL. APPL.
Vol. 30, No. 2, pp. 609–638

$\mathcal{H}_2$ MODEL REDUCTION FOR LARGE-SCALE LINEAR DYNAMICAL SYSTEMS*

S. GUGERCIN†, A. C. ANTOULAS‡, AND C. BEATTIE†

Abstract. The optimal $\mathcal{H}_2$ model reduction problem is of great importance in the area of dynamical systems and simulation. In the literature, two independent frameworks have evolved focusing either on solution of Lyapunov equations on the one hand or interpolation of transfer functions on the other, without any apparent connection between the two approaches. In this paper, we develop a new unifying framework for the optimal $\mathcal{H}_2$ approximation problem using best approximation properties in the underlying Hilbert space. This new framework leads to a new set of local optimality conditions taking the form of a structured orthogonality condition. We show that the existing Lyapunov- and interpolation-based conditions are each equivalent to our conditions and so are equivalent to each other. Also, we provide a new elementary proof of the interpolation-based condition that clarifies the importance of the mirror images of the reduced system poles. Based on the interpolation framework, we describe an iteratively corrected rational Krylov algorithm for $\mathcal{H}_2$ model reduction. The formulation is based on finding a reduced order model that satisfies interpolation-based first-order necessary conditions for $\mathcal{H}_2$ optimality and results in a method that is numerically effective and suited for large-scale problems. We illustrate the performance of the method with a variety of numerical experiments and comparisons with existing methods.

Key words. model reduction, rational Krylov, $\mathcal{H}_2$ approximation

AMS subject classifications. 34C20, 41A05, 49K15, 49M05, 93A15, 93C05, 93C15

DOI. 10.1137/060666123

1. Introduction.

Given a dynamical system described by a set of first-order differential equations, the model reduction problem seeks to replace this original set of equations with a (much) smaller set of such equations so that the behavior of both systems is similar in an appropriately defined sense. Such situations arise frequently when physical systems need to be simulated or controlled; the greater the level of detail that is required, the greater the number of resulting equations. In large-scale settings, computations become infeasible due to limitations on computational resources as well as growing inaccuracy due to numerical ill-conditioning. In all these cases the number of equations involved may range from a few hundred to a few million. Examples of large-scale systems abound, ranging from the design of VLSI (very large scale integration) chips to the simulation and control of MEMS (microelectromechanical system) devices. For an overview of model reduction for large-scale dynamical systems we refer to the book [2]. See also [23] for a recent collection of large-scale benchmark problems.

In this paper, we consider single input/single output (SISO) linear dynamical systems represented as

$$G : \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}\,u(t) \\ y(t) = \mathbf{c}^T\mathbf{x}(t) \end{cases} \qquad \text{or} \qquad G(s) = \mathbf{c}^T(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \tag{1.1}$$

*Received by the editors July 26, 2006; accepted for publication (in revised form) by P. Benner February 25, 2008; published electronically June 6, 2008.
http://www.siam.org/journals/simax/30-2/66612.html

†Department of Mathematics, Virginia Tech, Blacksburg, VA (gugercin@vt.edu, beattie@vt.edu). The work of these authors was supported in part by the NSF through grants DMS-050597 and DMS-0513542, and by the AFOSR through grant FA9550-05-1-0449.

‡Department of Electrical and Computer Engineering, Rice University, Houston, TX (aca@ece.rice.edu). The work of this author was supported in part by the NSF through grants CCR-0306503 and ACI-0325081.

where $\mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$; we define $\mathbf{x}(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$ as the *state*, *input*, and *output*, respectively, of the system. (We comment on extensions to the multiple input/multiple output (MIMO) case in section 3.2.1, but will confine our analysis and examples to the SISO case.)

$G(s)$ is the transfer function of the system: if $\widehat{u}(s)$ and $\widehat{y}(s)$ denote the Laplace transforms of the input and output $u(t)$ and $y(t)$, respectively, then $\widehat{y}(s) = G(s)\widehat{u}(s)$. With a standard abuse of notation, we will denote both the system and its transfer function by G. The "dimension of G" is taken to be the dimension of the underlying state space, $\dim G = n$ in this case. It will always be assumed that the system, G, is *stable*, that is, that the eigenvalues of $\mathbf{A}$ have strictly negative real parts.

The model reduction process will yield another system,

$$G_r : \begin{cases} \dot{\mathbf{x}}_r(t) = \mathbf{A}_r\mathbf{x}_r(t) + \mathbf{b}_r u(t) \\ y_r(t) = \mathbf{c}_r^T\mathbf{x}_r(t) \end{cases} \qquad \text{or} \qquad G_r(s) = \mathbf{c}_r^T(s\mathbf{I} - \mathbf{A}_r)^{-1}\mathbf{b}_r, \tag{1.2}$$

having (much) smaller dimension $r \ll n$, with $\mathbf{A}_r \in \mathbb{R}^{r\times r}$ and $\mathbf{b}_r$, $\mathbf{c}_r \in \mathbb{R}^r$.

We want $y_r(t) \approx y(t)$ over a large class of inputs $u(t)$. Different measures of approximation and different choices of input classes will lead to different model reduction goals. Suppose one wants to ensure that $\max_{t>0} |y(t) - y_r(t)|$ is small uniformly over all inputs, $u(t)$, having bounded "energy," that is, $\int_0^\infty |u(t)|^2\, dt \le 1$. Observe first that $\widehat{y}(s) - \widehat{y}_r(s) = [G(s) - G_r(s)]\, \widehat{u}(s)$ and then

$$\begin{aligned}
\max_{t>0} |y(t) - y_r(t)| &= \max_{t>0} \left| \frac{1}{2\pi} \int_{-\infty}^{\infty} (\widehat{y}(\imath\omega) - \widehat{y}_r(\imath\omega))\, e^{\imath\omega t}\, d\omega \right| \\
&\le \frac{1}{2\pi} \int_{-\infty}^{\infty} |\widehat{y}(\imath\omega) - \widehat{y}_r(\imath\omega)|\ d\omega = \frac{1}{2\pi} \int_{-\infty}^{\infty} |G(\imath\omega) - G_r(\imath\omega)|\ |\widehat{u}(\imath\omega)|\ d\omega \\
&\le \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} |G(\imath\omega) - G_r(\imath\omega)|^2\ d\omega \right)^{1/2} \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} |\widehat{u}(\imath\omega)|^2\ d\omega \right)^{1/2} \\
&\le \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} |G(\imath\omega) - G_r(\imath\omega)|^2\ d\omega \right)^{1/2} \left(\int_0^{\infty} |u(t)|^2\ dt \right)^{1/2} \\
&\le \left(\frac{1}{2\pi} \int_{-\infty}^{+\infty} |G(\imath\omega) - G_r(\imath\omega)|^2\, d\omega \right)^{1/2} \stackrel{\text{def}}{=} \|G - G_r\|_{\mathcal{H}_2}.
\end{aligned}$$

We seek a reduced order dynamical system, G_r, such that

(i) $\|G - G_r\|_{\mathcal{H}_2}$, the "$\mathcal{H}_2$ error," is as small as possible;
(ii) critical system properties for G (such as stability) exist also in G_r; and
(iii) the computation of G_r (i.e., the computation of $\mathbf{A}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r$) is both efficient and numerically stable.

The problem of finding reduced order models that yield a small $\mathcal{H}_2$ error has been the object of many investigations; see, for instance, [6, 37, 34, 9, 21, 26, 22, 36, 25, 13] and the references therein. Finding a global minimizer of $\|G - G_r\|_{\mathcal{H}_2}$ is a hard task, so the goal in making $\|G - G_r\|_{\mathcal{H}_2}$ "as small as possible" becomes, as for many optimization problems, identification of reduced order models, G_r, that satisfy first-order necessary conditions for local optimality. There is a wide variety of such conditions that may be derived, yet their interconnections are generally unclear. Most methods that can identify reduced order models satisfying such first-order necessary conditions will require dense matrix operations, typically the solution of a sequence of matrix Lyapunov equations, a task which becomes computationally intractable

rapidly as the dimension increases. Such methods are unsuitable even for medium-scale problems. In section 2, we review the moment matching problem for model reduction, its connection with rational Krylov methods (which are very useful for large-scale problems), and basic features of the $\mathcal{H}_2$ norm and inner product.

We offer in section 3 what appears to be a new set of first-order necessary conditions for local optimality of a reduced order model comprising in effect a structured orthogonality condition. We also show its equivalence with two other $\mathcal{H}_2$ optimality conditions that have been previously known (thus showing them all to be equivalent).

An iterative algorithm that is designed to force optimality with respect to a set of conditions that is computationally tractable is described in section 4. The proposed method also forces optimality with respect to the other equivalent conditions as well. It is based on computationally effective use of rational Krylov subspaces and so is suitable for systems whose dimension n is of the order of many thousands of state variables. Numerical examples are presented in section 5.

2. Background.

2.1. Model reduction by moment matching. Given the system (1.1), reduction by *moment matching* consists in finding a system (1.2) so that $G_r(s)$ interpolates the values of $G(s)$, and perhaps also derivative values as well, at selected interpolation points (also called *shifts*) σ_k in the complex plane. For our purposes, simple Hermite interpolation suffices, so our problem is to find $\mathbf{A}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r$ so that

$$G_r(\sigma_k) = G(\sigma_k) \quad \text{and} \quad G_r'(\sigma_k) = G'(\sigma_k) \quad \text{for} \quad k = 1, \ldots, r$$

or, equivalently,

$$\mathbf{c}^T(\sigma_k\mathbf{I} - \mathbf{A})^{-1}\mathbf{b} = \mathbf{c}_r^T(\sigma_k\mathbf{I} - \mathbf{A}_r)^{-1}\mathbf{b}_r \ \text{ and } \ \mathbf{c}^T(\sigma_k\mathbf{I} - \mathbf{A})^{-2}\mathbf{b} = \mathbf{c}_r^T(\sigma_k\mathbf{I} - \mathbf{A}_r)^{-2}\mathbf{b}_r$$

for $k = 1, \ldots, r$. The quantity $\mathbf{c}^T(\sigma_k\mathbf{I} - \mathbf{A})^{-(j+1)}\mathbf{b}$ is called the jth moment of $G(s)$ at σ_k. Moment matching for finite $\sigma \in \mathbb{C}$ becomes *rational interpolation*; see, for example, [3]. Importantly, these problems can be solved in a recursive and numerically effective way by means of *rational Lanczos/Arnoldi* procedures.

To see this we first consider reduced order models that are constructed by Galerkin approximation: Let $\mathcal{V}_r$ and $\mathcal{W}_r$ be given r-dimensional subspaces of $\mathbb{R}^n$ that are generic in the sense that $\mathcal{V}_r \cap \mathcal{W}_r^{\perp} = \{0\}$. Then for any input $u(t)$ the reduced order output $y_r(t)$ is defined by

$$\text{Find} \quad \mathbf{v}(t) \in \mathcal{V}_r \quad \text{such that} \quad \dot{\mathbf{v}}(t) - \mathbf{A}\mathbf{v}(t) - \mathbf{b}u(t) \perp \mathcal{W}_r \quad \text{for all } t; \tag{2.1}$$
$$\text{then} \quad y_r(t) \stackrel{\text{def}}{=} \mathbf{c}^T\mathbf{v}(t).$$

Denote by $\mathrm{Ran}(\mathbf{M})$ the range of a matrix $\mathbf{M}$. Let $\mathbf{V}_r \in \mathbb{R}^{n\times r}$ and $\mathbf{W}_r \in \mathbb{R}^{n\times r}$ be matrices defined so that $\mathcal{V}_r = \mathrm{Ran}(\mathbf{V}_r)$ and $\mathcal{W}_r = \mathrm{Ran}(\mathbf{W}_r)$. Then the assumption $\mathcal{V}_r \cap \mathcal{W}_r^{\perp} = \{0\}$ is equivalent to $\mathbf{W}_r^T\mathbf{V}_r$ being nonsingular. The Galerkin approximation (2.1) can be interpreted as $\mathbf{v}(t) = \mathbf{V}_r\mathbf{x}_r(t)$ with $\mathbf{x}_r(t) \in \mathbb{R}^r$ for each t and

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

leading then to the reduced order model (1.2) with

$$\mathbf{A}_r = (\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T\mathbf{A}\mathbf{V}_r, \quad \mathbf{b}_r = (\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T\mathbf{b}, \quad \text{and} \quad \mathbf{c}_r^T = \mathbf{c}^T\mathbf{V}_r. \tag{2.2}$$

Evidently the choice of $\mathbf{V}_r$ and $\mathbf{W}_r$ determines the quality of the reduced order model.

Rational interpolation by projection was first proposed by Skelton et al. in [11, 38, 39]. Grimme [17] showed how one can obtain the required projection using the rational Krylov method of Ruhe [33]. Krylov-based methods are able to match moments without ever computing them explicitly. This is important since the computation of moments is in general ill-conditioned. This is a fundamental motivation behind the Krylov-based methods [12].

In Lemma 2.1 and Corollary 2.2 below, we present new short proofs of rational interpolation by Krylov projection that are substantially simpler than those found in the original works [17, 11, 38, 39].

LEMMA 2.1. *Suppose $\sigma \in \mathbb{C}$ is not an eigenvalue of either $\mathbf{A}$ or $\mathbf{A}_r$.*

$$\text{If} \quad (\sigma \mathbf{I} - \mathbf{A})^{-1}\mathbf{b} \in \mathcal{V}_r, \quad \text{then} \quad G_r(\sigma) = G(\sigma). \tag{2.3}$$

$$\text{If} \quad (\overline{\sigma}\, \mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c} \in \mathcal{W}_r, \quad \text{then} \quad G_r(\sigma) = G(\sigma). \tag{2.4}$$

$$\text{If both} \quad (\sigma \mathbf{I} - \mathbf{A})^{-1}\mathbf{b} \in \mathcal{V}_r \quad \text{and} \quad (\overline{\sigma}\, \mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c} \in \mathcal{W}_r,$$
$$\text{then} \quad G_r(\sigma) = G(\sigma) \quad \text{and} \quad G_r'(\sigma) = G'(\sigma). \tag{2.5}$$

Proof. Define $\mathbf{N}_r(z) = \mathbf{V}_r(z\mathbf{I} - \mathbf{A}_r)^{-1}(\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T(z\mathbf{I} - \mathbf{A})$ and $\widetilde{\mathbf{N}}_r(z) = (z\mathbf{I}-\mathbf{A})\mathbf{N}_r(z)(z\mathbf{I}-\mathbf{A})^{-1}$. Both $\mathbf{N}_r(z)$ and $\widetilde{\mathbf{N}}_r(z)$ are analytic matrix-valued functions in a neighborhood of $z = \sigma$. One may directly verify that $\mathbf{N}_r^2(z) = \mathbf{N}_r(z)$ and $\widetilde{\mathbf{N}}_r^2(z) = \widetilde{\mathbf{N}}_r(z)$ and that $\mathcal{V}_r = \text{Ran}\, \mathbf{N}_r(z) = \text{Ker}\, (\mathbf{I} - \mathbf{N}_r(z))$ and $\mathcal{W}_r^{\perp} = \text{Ker}\, \widetilde{\mathbf{N}}_r(z) = \text{Ran}\big(\mathbf{I} - \widetilde{\mathbf{N}}_r(z)\big)$ for all z in a neighborhood of σ. Then

$$G(z) - G_r(z) = \left[(z\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}\right]^T \left(\mathbf{I} - \widetilde{\mathbf{N}}_r(z)\right)(z\mathbf{I} - \mathbf{A})\Big(\mathbf{I} - \mathbf{N}_r(z)\Big)(z\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}.$$

Evaluating at $z = \sigma$ leads to (2.3) and (2.4). Evaluating at $z = \sigma + \varepsilon$ and observing that $(\sigma\mathbf{I} + \varepsilon\mathbf{I} - \mathbf{A})^{-1} = (\sigma\mathbf{I} - \mathbf{A})^{-1} - \varepsilon(\sigma\mathbf{I} - \mathbf{A})^{-2} + \mathcal{O}(\varepsilon^2)$ yields

$$G(\sigma + \varepsilon) - G_r(\sigma + \varepsilon) = \mathcal{O}(\varepsilon^2),$$

which gives (2.5) as a consequence. ∎

COROLLARY 2.2. *Consider the system G defined by $\mathbf{A}, \mathbf{b}, \mathbf{c}$, a set of distinct shifts given by $\{\sigma_k\}_{k=1}^r$, that is closed under conjugation (i.e., shifts are either real or occur in conjugate pairs), and subspaces spanned by the columns of $\mathbf{V}_r$ and $\mathbf{W}_r$ with*

$$\text{Ran}(\mathbf{V}_r) = \text{span}\left\{(\sigma_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \quad \ldots, \quad (\sigma_r\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\right\} \quad \text{and} \tag{2.6}$$

$$\text{Ran}(\mathbf{W}_r) = \text{span}\left\{(\sigma_1\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}, \quad \ldots, \quad (\sigma_r\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}\right\}. \tag{2.7}$$

Then $\mathbf{V}_r$ and $\mathbf{W}_r$ can be chosen to be real matrices and the reduced order system G_r defined by $\mathbf{A}_r = (\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T\mathbf{A}\mathbf{V}_r$, $\mathbf{b}_r = (\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T\mathbf{b}$, $\mathbf{c}_r^T = \mathbf{c}^T\mathbf{V}_r$ is itself real and matches the first two moments of $G(s)$ at each of the interpolation points σ_k, i.e., $G(\sigma_k) = G_r(\sigma_k)$ and $G'(\sigma_k) = G_r'(\sigma_k)$ for $k = 1, \ldots, r$.

For Krylov-based model reduction, one chooses interpolation points and then constructs $\mathbf{V}_r$ and $\mathbf{W}_r$ satisfying (2.6) and (2.7), respectively. Note that, in a numerical implementation, one does not actually compute $(\sigma_i\mathbf{I} - \mathbf{A})^{-1}$, but instead computes a (potentially sparse) factorization (one for each interpolation point σ_i), uses it to solve a system of equations having $\mathbf{b}$ as a right-hand side, and uses its transpose to solve a system of equations having $\mathbf{c}$ as a right-hand side. The interpolation points are chosen so as to minimize the deviation of G_r from G in a sense that is detailed in the next section. Unlike Gramian-based model reduction methods such as balanced truncation (see section 2.2 below and [2]), Krylov-based model reduction requires only

matrix-vector multiplications and some sparse linear solvers, and can be iteratively implemented; hence it is computationally effective; for details, see also [15, 16].

2.2. Model reduction by balanced truncation. One of the most common model reduction techniques is *balanced truncation* [28, 27]. In this case, the modeling subspaces $\mathbf{V}_r$ and $\mathbf{W}_r$ depend on the solutions to the two Lyapunov equations

$$\mathbf{AP} + \mathbf{PA}^T + \mathbf{bb}^T = \mathbf{0}, \qquad \mathbf{A}^T\mathbf{Q} + \mathbf{QA} + \mathbf{c}^T\mathbf{c} = \mathbf{0}. \tag{2.8}$$

$\mathbf{P}$ and $\mathbf{Q}$ are called the reachability and observability Gramians, respectively. Under the assumption that $\mathbf{A}$ is stable, both $\mathbf{P}$ and $\mathbf{Q}$ are positive semidefinite matrices. Square roots of the eigenvalues of the product $\mathbf{PQ}$ are the singular values of the Hankel operator associated with $G(s)$ and are called the Hankel singular values of $G(s)$, denoted by $\eta_i(G)$.

Let $\mathbf{P} = \mathbf{UU}^T$ and $\mathbf{Q} = \mathbf{LL}^T$. Let $\mathbf{U}^T\mathbf{L} = \mathbf{ZSY}^T$ be the singular value decomposition with $\mathbf{S} = \operatorname{diag}(\eta_1, \eta_2, \ldots, \eta_n)$. Let $\mathbf{S}_r = \operatorname{diag}(\eta_1, \eta_2, \ldots, \eta_r)$, $r < n$. Construct

$$\mathbf{W}_r = \mathbf{LY}_r\mathbf{S}_r^{-1/2} \qquad \text{and} \qquad \mathbf{V}_r = \mathbf{UZ}_r\mathbf{S}_r^{-1/2}, \tag{2.9}$$

where $\mathbf{Z}_r$ and $\mathbf{Y}_r$ denote the leading r columns of left singular vectors, $\mathbf{Z}$, and right singular vectors, $\mathbf{Y}$, respectively. The rth-order reduced order model via *balanced truncation*, $G_r(s)$, is obtained by reducing $G(s)$ using $\mathbf{W}_r$ and $\mathbf{V}_r$ from (2.9).

Another important dynamical systems norm (besides the $\mathcal{H}_2$ norm) is the $\mathcal{H}_\infty$ norm defined as $\|G\|_{\mathcal{H}_\infty} := \sup_{\omega\in\mathbb{R}} |G(\imath\omega)|$. The reduced order system $G_r(s)$ obtained by balanced truncation is asymptotically stable and the $\mathcal{H}_\infty$ norm of the error system satisfies $\|G - G_r\|_{\mathcal{H}_\infty} \le 2(\eta_{r+1} + \cdots + \eta_n)$.

The value of having, for reduced order models, guaranteed stability and an explicit error bound is widely recognized, though it is achieved at potentially considerable cost. As described above, balanced truncation requires the solution of two Lyapunov equations of order n, which is a formidable task in large-scale settings. For more details and background on balanced truncation, see section III.7 of [2].

2.3. The $\mathcal{H}_2$ norm. $\mathcal{H}_2$ will denote the set of functions, $g(z)$, that are analytic for z in the open right half plane, $\operatorname{Re}(z) > 0$, and such that for each fixed $\operatorname{Re}(z) = x > 0$, $g(x + \imath y)$ is square integrable as a function of $y \in (-\infty, \infty)$ in such a way that

$$\sup_{x>0} \int_{-\infty}^{\infty} |g(x + \imath y)|^2 \, dy < \infty.$$

$\mathcal{H}_2$ is a Hilbert space and holds our interest because transfer functions associated with stable SISO finite-dimensional dynamical systems are elements of $\mathcal{H}_2$. Indeed, if $G(s)$ and $H(s)$ are transfer functions associated with real stable SISO dynamical systems, then the $\mathcal{H}_2$ inner product can be defined as

$$\langle G,\ H\rangle_{\mathcal{H}_2} \stackrel{\text{def}}{=} \frac{1}{2\pi}\int_{-\infty}^{\infty} \overline{G(\imath\omega)}H(\imath\omega)\, d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} G(-\imath\omega)H(\imath\omega)\, d\omega, \tag{2.10}$$

with a norm defined as

$$\|G\|_{\mathcal{H}_2} \stackrel{\text{def}}{=} \left(\frac{1}{2\pi}\int_{-\infty}^{+\infty} |G(\imath\omega)|^2\, d\omega\right)^{1/2}. \tag{2.11}$$

Notice in particular that if $G(s)$ and $H(s)$ represent real dynamical systems, then $\langle G,\ H\rangle_{\mathcal{H}_2} = \langle H,\ G\rangle_{\mathcal{H}_2}$ and $\langle G,\ H\rangle_{\mathcal{H}_2}$ must be real.

There are two alternate characterizations of this inner product that make it far more computationally accessible.

LEMMA 2.3. *Suppose* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *and* $\mathbf{B} \in \mathbb{R}^{m\times m}$ *are stable and, given* $\mathbf{b}$, $\mathbf{c} \in \mathbb{R}^n$ *and* $\widetilde{\mathbf{b}}$, $\widetilde{\mathbf{c}} \in \mathbb{R}^m$, *define associated transfer functions,*

$$G(s) = \mathbf{c}^T (s\mathbf{I} - \mathbf{A})^{-1} \mathbf{b} \quad \textit{and} \quad H(s) = \widetilde{\mathbf{c}}^T (s\mathbf{I} - \mathbf{B})^{-1} \widetilde{\mathbf{b}}.$$

The inner product $\langle G,\ H\rangle_{\mathcal{H}_2}$ *is associated with solutions to Sylvester equations as follows:*

$$\textit{If}\quad \mathbf{P}\quad \textit{solves}\quad \mathbf{AP} + \mathbf{PB}^T + \mathbf{b}\widetilde{\mathbf{b}}^T = 0, \quad \textit{then}\quad \langle G,\ H\rangle_{\mathcal{H}_2} = \mathbf{c}^T\mathbf{P}\widetilde{\mathbf{c}}. \tag{2.12}$$

$$\textit{If}\quad \mathbf{Q}\quad \textit{solves}\quad \mathbf{QA} + \mathbf{B}^T\mathbf{Q} + \widetilde{\mathbf{c}}\mathbf{c}^T = 0, \quad \textit{then}\quad \langle G,\ H\rangle_{\mathcal{H}_2} = \widetilde{\mathbf{b}}^T\mathbf{Q}\mathbf{b}. \tag{2.13}$$

$$\textit{If}\quad \mathbf{R}\quad \textit{solves}\quad \mathbf{AR} + \mathbf{RB} + \mathbf{b}\widetilde{\mathbf{c}}^T = 0, \quad \textit{then}\quad \langle G,\ H\rangle_{\mathcal{H}_2} = \mathbf{c}^T\mathbf{R}\widetilde{\mathbf{b}}. \tag{2.14}$$

Note that if $\mathbf{A} = \mathbf{B}$, $\mathbf{b} = \widetilde{\mathbf{b}}$, *and* $\mathbf{c} = \widetilde{\mathbf{c}}$, *then* $\mathbf{P}$ *is the "reachability Gramian" of* $G(s)$, $\mathbf{Q}$ *is the "observability Gramian" of* $G(s)$, *and* $\mathbf{R}$ *is the "cross Gramian" of* $G(s)$; *and*

$$\|G\|_{\mathcal{H}_2}^2 = \mathbf{c}^T\mathbf{P}\mathbf{c} = \mathbf{b}^T\mathbf{Q}\mathbf{b} = \mathbf{c}^T\mathbf{R}\mathbf{b}. \tag{2.15}$$

Gramians play a prominent role in the analysis of linear dynamical systems; refer to [2] for more information.

Proof. We detail the proof of (2.12); proofs of (2.13) and (2.14) are similar. Since $\mathbf{A}$ and $\mathbf{B}$ are stable, the solution, $\mathbf{P}$, to the Sylvester equation of (2.12) exists and is unique. For any $\omega \in \mathbb{R}$, rearrange this equation to obtain in sequence

$$\begin{aligned}
(-\imath\omega\mathbf{I} - \mathbf{A})\,\mathbf{P} + \mathbf{P}\left(\imath\omega\mathbf{I} - \mathbf{B}^T\right) - \mathbf{b}\widetilde{\mathbf{b}}^T &= 0,\\
(-\imath\omega\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{P} + \mathbf{P}\left(\imath\omega\mathbf{I} - \mathbf{B}^T\right)^{-1} &= (-\imath\omega\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{b}\widetilde{\mathbf{b}}^T\left(\imath\omega\mathbf{I} - \mathbf{B}^T\right)^{-1},\\
\mathbf{c}^T(-\imath\omega\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{P}\widetilde{\mathbf{c}} + \mathbf{c}^T\mathbf{P}\left(\imath\omega\mathbf{I} - \mathbf{B}^T\right)^{-1}\widetilde{\mathbf{c}} &= G(-\imath\omega)H(\imath\omega),
\end{aligned}$$

and finally

$$\begin{aligned}
\mathbf{c}^T\left(\int_{-L}^{L}(-\imath\omega\mathbf{I} - \mathbf{A})^{-1}\,d\omega\right)\mathbf{P}\widetilde{\mathbf{c}} + \mathbf{c}^T\mathbf{P}\left(\int_{-L}^{L}\left(\imath\omega\mathbf{I} - \mathbf{B}^T\right)^{-1}\,d\omega\right)\widetilde{\mathbf{c}}&\\
= \int_{-L}^{L} G(-\imath\omega)H(\imath\omega)\,d\omega.&
\end{aligned}$$

Taking $L \to \infty$ and using Lemma A.1 in the appendix leads to

$$\begin{aligned}
\int_{-\infty}^{\infty} G(-\imath\omega)H(\imath\omega)\,d\omega &= \mathbf{c}^T\left(\text{P.V.}\int_{-\infty}^{\infty}(-\imath\omega\mathbf{I} - \mathbf{A})^{-1}\,d\omega\right)\mathbf{P}\widetilde{\mathbf{c}}\\
&\quad + \mathbf{c}^T\mathbf{P}\left(\text{P.V.}\int_{-\infty}^{\infty}\left(\imath\omega\mathbf{I} - \mathbf{B}^T\right)^{-1}\,d\omega\right)\widetilde{\mathbf{c}}\\
&= 2\pi\ \mathbf{c}^T\mathbf{P}\widetilde{\mathbf{c}}. \qquad \square
\end{aligned}$$

Recently, Antoulas [2] obtained a new expression for $\|G\|_{\mathcal{H}_2}$ based on the poles and residues of the transfer function $G(s)$ that complements the widely known alternative expression (2.15). We provide a compact derivation of this expression and the associated $\mathcal{H}_2$ inner product.

If $f(s)$ is a meromorphic function with a pole at λ, denote the residue of $f(s)$ at λ by $\mathsf{res}[f(s), \lambda]$. Thus, if λ is a simple pole of $f(s)$, then $\mathsf{res}[f(s), \lambda] = \lim_{s\to\lambda}(s-\lambda)f(s)$, and if λ is a double pole of $f(s)$, then $\mathsf{res}[f(s), \lambda] = \lim_{s\to\lambda} \frac{d}{ds}\left[(s-\lambda)^2 f(s)\right]$.

LEMMA 2.4. *Suppose that $G(s)$ has poles at $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $H(s)$ has poles at $\mu_1, \mu_2, \ldots, \mu_m$, both sets contained in the open left half plane. Then*

$$\langle G,\ H\rangle_{\mathcal{H}_2} = \sum_{k=1}^{m} \mathsf{res}[G(-s)H(s), \mu_k] = \sum_{k=1}^{n} \mathsf{res}[H(-s)G(s), \lambda_k]. \tag{2.16}$$

In particular,

- *if μ_k is a simple pole of $H(s)$, then*

$$\mathsf{res}[G(-s)H(s), \mu_k] = G(-\mu_k)\mathsf{res}[H(s), \mu_k];$$

- *if μ_k is a double pole of $H(s)$, then*

$$\mathsf{res}[G(-s)H(s), \mu_k] = G(-\mu_k)\,\mathsf{res}[H(s), \mu_k] - G'(-\mu_k)\cdot h_0(\mu_k),$$

where $h_0(\mu_k) = \lim_{s\to\mu_k}\left((s-\mu_k)^2 H(s)\right)$.

Proof. Notice that the function $G(-s)H(s)$ has singularities at $\mu_1, \mu_2, \ldots, \mu_m$ and $-\lambda_1, -\lambda_2, \ldots, -\lambda_n$. For any $R > 0$, define the semicircular contour in the left half plane:

$$\Gamma_R = \{z\,|z = \imath\omega \text{ with } \omega \in [-R, R]\,\} \cup \left\{ z \,\middle|\, z = R\,e^{\imath\theta} \text{ with } \theta \in \left[\frac{\pi}{2}, \frac{3\pi}{2}\right] \right\}.$$

Γ_R bounds a region that for sufficiently large R contains all the system poles of $H(s)$ and so, by the residue theorem,

$$\begin{aligned}\langle G,\ H\rangle_{\mathcal{H}_2} &= \frac{1}{2\pi}\int_{-\infty}^{+\infty} G(-\imath\omega)H(\imath\omega)\,d\omega \\ &= \lim_{R\to\infty}\frac{1}{2\pi\imath}\int_{\Gamma_R} G(-s)H(s)\ ds = \sum_{k=1}^{m}\mathsf{res}[G(-s)H(s), \mu_k].\end{aligned}$$

Evidently, if μ_k is a simple pole for $H(s)$, it is also a simple pole for $G(-s)H(s)$ and

$$\mathsf{res}[G(-s)H(s),\ \mu_k] = \lim_{s\to\mu_k}(s-\mu_k)G(-s)H(s) = G(-\mu_k)\lim_{s\to\mu_k}(s-\mu_k)H(s).$$

If μ_k is a double pole for $H(s)$, then it is also a double pole for $G(-s)H(s)$ and

$$\begin{aligned}\mathsf{res}[G(-s)H(s),\ \mu_k] &= \lim_{s\to\mu_k}\frac{d}{ds}(s-\mu_k)^2 G(-s)H(s) \\ &= \lim_{s\to\mu_k} G(-s)\frac{d}{ds}(s-\mu_k)^2 H(s) - G'(-s)(s-\mu_k)^2 H(s) \\ &= G(-\mu_k)\lim_{s\to\mu_k}\frac{d}{ds}(s-\mu_k)^2 H(s) - G'(-\mu_k)\lim_{s\to\mu_k}(s-\mu_k)^2 H(s). \qquad \square\end{aligned}$$

Lemma 2.4 immediately yields the expression for $\|G\|_{\mathcal{H}_2}$ given by Antoulas [2, p. 145] based on poles and residues of the transfer function $G(s)$.

COROLLARY 2.5. *If $G(s)$ has simple poles at $\lambda_1, \lambda_2, \ldots, \lambda_n$, then*

$$\|G\|_{\mathcal{H}_2} = \left(\sum_{k=1}^{n}\mathsf{res}[G(s), \lambda_k]G(-\lambda_k)\right)^{1/2}.$$

3. Optimal $\mathcal{H}_2$ model reduction. In this section, we investigate three frameworks of necessary conditions for $\mathcal{H}_2$ optimality. The first utilizes the inner product structure $\mathcal{H}_2$ and leads to what could be thought of as a geometric condition for optimality. This appears to be a new characterization of $\mathcal{H}_2$ optimality for reduced order models. The remaining two frameworks, interpolation-based [26] and Lyapunov-based [36, 22], are easily derived from the first framework and in this way can be seen to be equivalent to one another—a fact that is not a priori evident. This equivalence proves that solving the optimal $\mathcal{H}_2$ problem in the Krylov framework is equivalent to solving it in the Lyapunov framework, which leads to the proposed Krylov-based method for $\mathcal{H}_2$ model reduction in section 4.

Given G, a stable SISO finite-dimensional dynamical system as described in (1.1), we seek a stable reduced order system G_r of order r as described in (1.2), which is the best stable rth-order dynamical system approximating G with respect to the $\mathcal{H}_2$ norm:

$$\|G - G_r\|_{\mathcal{H}_2} = \min_{\substack{\dim(\widetilde{G}_r)=r \\ \widetilde{G}_r \text{ : stable}}} \|G - \widetilde{G}_r\|_{\mathcal{H}_2}. \tag{3.1}$$

Many researchers have worked on problem (3.1), the *optimal $\mathcal{H}_2$ model reduction problem*. See [37, 34, 9, 21, 26, 22, 36, 25] and the references therein.

3.1. Structured orthogonality optimality conditions. The set of all stable rth-order dynamical systems do not constitute a subspace of $\mathcal{H}_2$, so the best rth-order $\mathcal{H}_2$ approximation is not so easy to characterize, the Hilbert space structure of $\mathcal{H}_2$ notwithstanding. This observation does suggest the following narrower though simpler result.

THEOREM 3.1. *Let $\mu_1, \mu_2, \ldots, \mu_r \subset \mathbb{C}$ be distinct points in the open left half plane and define $\mathcal{M}(\boldsymbol{\mu})$ to be the set of all proper rational functions that have simple poles exactly at $\mu_1, \mu_2, \ldots, \mu_r$. Then*

- *$H \in \mathcal{M}(\boldsymbol{\mu})$ implies that H is the transfer function of a stable dynamical system with $\dim(H) = r$;*
- *$\mathcal{M}(\boldsymbol{\mu})$ is an $(r-1)$-dimensional subspace of $\mathcal{H}_2$;*
- *$G_r \in \mathcal{M}(\boldsymbol{\mu})$ solves*

$$\|G - G_r\|_{\mathcal{H}_2} = \min_{\widetilde{G}_r \in \mathcal{M}(\boldsymbol{\mu})} \|G - \widetilde{G}_r\|_{\mathcal{H}_2} \tag{3.2}$$

if and only if

$$\langle G - G_r,\, H \rangle_{\mathcal{H}_2} = 0 \qquad \text{for all } H \in \mathcal{M}(\boldsymbol{\mu}). \tag{3.3}$$

Furthermore the solution, G_r, to (3.2) exists and is unique.

Proof. The key observation is that $\mathcal{M}(\boldsymbol{\mu})$ is a closed subspace of $\mathcal{H}_2$. Then the equivalence of (3.2) and (3.3) follows from the classic projection theorem in Hilbert space (cf. [32]). ∎

One consequence of Theorem 3.1 is that if $G_r(s)$ interpolates a real system $G(s)$ at the *mirror images* of its own poles (i.e., at the poles of $G_r(s)$ reflected across the imaginary axis), then $G_r(s)$ is guaranteed to be an *optimal* approximation of $G(s)$ relative to the $\mathcal{H}_2$ norm among all reduced order systems having the same reduced system poles $\{\mu_i\}_{i=1}^r$. An analogous result for optimal rational approximants to analytic functions on the unit disk can be found in [14]. The set of stable rth-order dynamical systems is not convex, and so the original problem (3.1) allows for

multiple minimizers. Indeed there may be "local minimizers" that do not solve (3.1). A reduced order system, G_r, is a *local minimizer* for (3.1) if, for all $\varepsilon > 0$ sufficiently small,

$$\|G - G_r\|_{\mathcal{H}_2} \leq \|G - \widetilde{G}_r^{(\varepsilon)}\|_{\mathcal{H}_2} \tag{3.4}$$

for all stable dynamical systems $\widetilde{G}_r^{(\varepsilon)}$ with $\dim(\widetilde{G}_r^{(\varepsilon)}) = r$ and $\|G_r - \widetilde{G}_r^{(\varepsilon)}\|_{\mathcal{H}_2} \leq C\,\varepsilon$, with C being a constant that may depend on the particular family $\widetilde{G}_r^{(\varepsilon)}$ considered. As a practical matter, the global minimizers that solve (3.1) are difficult to obtain with certainty; current approaches favor seeking reduced order models that satisfy a local (first-order) necessary condition for optimality. Even though such strategies do not guarantee global minimizers, they often produce effective reduced order models nonetheless. In this spirit, we give necessary conditions for optimality for the reduced order system, G_r, that appear as structured orthogonality conditions similar to (3.3).

THEOREM 3.2. *If G_r is a local minimizer to G as described in* (3.4) *and G_r has simple poles, then*

$$\langle G - G_r,\ G_r \cdot H_1 + H_2 \rangle_{\mathcal{H}_2} = 0 \tag{3.5}$$

for all real dynamical systems H_1 and H_2 having the same poles with the same multiplicities as G_r.

($G_r \cdot H_1$ here denotes pointwise multiplication of scalar functions.)

Proof. Theorem 3.1 implies (3.5) with $H_1 = 0$, so it suffices to show that the hypotheses imply that $\langle G - G_r,\ G_r \cdot H \rangle_{\mathcal{H}_2} = 0$ for all real dynamical systems H having the same poles with the same multiplicities as G_r.

Suppose that $\{\widetilde{G}_r^{(\varepsilon)}\}_{\varepsilon>0}$ is a family of real stable dynamical systems with $\dim(\widetilde{G}_r^{(\varepsilon)}) = r$ and $\|G_r - \widetilde{G}_r^{(\varepsilon)}\|_{\mathcal{H}_2} < C\varepsilon$ for some constant $C > 0$. Then for all $\varepsilon > 0$ sufficiently small,

$$\begin{aligned}
\|G - G_r\|_{\mathcal{H}_2}^2 &\leq \|G - \widetilde{G}_r^{(\varepsilon)}\|_{\mathcal{H}_2}^2 \\
&\leq \|(G - G_r) + (G_r - \widetilde{G}_r^{(\varepsilon)})\|_{\mathcal{H}_2}^2 \\
&\leq \|G - G_r\|_{\mathcal{H}_2}^2 + 2\left\langle G - G_r,\ G_r - \widetilde{G}_r^{(\varepsilon)} \right\rangle_{\mathcal{H}_2} + \|G_r - \widetilde{G}_r^{(\varepsilon)})\|_{\mathcal{H}_2}^2.
\end{aligned}$$

This in turn implies for all $\varepsilon > 0$ sufficiently small that

$$0 \leq 2\left\langle G - G_r,\ G_r - \widetilde{G}_r^{(\varepsilon)} \right\rangle_{\mathcal{H}_2} + \|G_r - \widetilde{G}_r^{(\varepsilon)}\|_{\mathcal{H}_2}^2. \tag{3.6}$$

By considering a few different "directions of approach" of $\widetilde{G}_r^{(\varepsilon)}$ to G_r as $\varepsilon \to 0$, (3.6) will lead to a few different necessary conditions for G_r to be a locally optimal reduced order model. Denote the poles of G_r as $\mu_1, \mu_2, \ldots, \mu_r$ and suppose they are ordered so that the first m_R are real and the next m_C are in the upper half plane. Write $\mu_i = \alpha_i + \imath\beta_i$. Any real rational function having the same poles as $G_r(s)$ can be written as

$$H(s) = \sum_{i=1}^{m_R} \frac{\gamma_i}{s - \mu_i} + \sum_{i=m_R+1}^{m_R+m_C} \frac{\rho_i(s - \alpha_i) + \tau_i}{(s - \alpha_i)^2 + \beta_i^2},$$

with arbitrary real-valued choices for γ_i, ρ_i, and τ_i. Now suppose that μ is a real pole for G_r and that

$$\left\langle G - G_r,\ \frac{G_r(s)}{s - \mu} \right\rangle_{\mathcal{H}_2} \neq 0. \tag{3.7}$$

Write $G_r(s) = \frac{p_{r-1}(s)}{(s-\mu)\,q_{r-1}(s)}$ for real polynomials $p_{r-1},\ q_{r-1} \in \mathcal{P}_{r-1}$ and define

$$\widetilde{G}_r^{(\varepsilon)}(s) = \frac{p_{r-1}(s)}{[s-\mu-(\pm\varepsilon)]\ q_{r-1}(s)},$$

where the sign of $\pm\varepsilon$ is chosen to match that of $\left\langle G - G_r,\ \frac{G_r(s)}{s-\mu}\right\rangle_{\mathcal{H}_2}$. Then we have

$$\widetilde{G}_r^{(\varepsilon)}(s) = G_r(s) \pm \varepsilon \frac{p_{r-1}(s)}{(s-\mu)^2\, q_{r-1}(s)} + \mathcal{O}(\varepsilon^2),$$

which leads to $G_r(s) - \widetilde{G}_r^{(\varepsilon)}(s) = \mp\varepsilon \frac{G_r(s)}{s-\mu} + \mathcal{O}(\varepsilon^2)$ and

$$\left\langle G - G_r,\ G_r - \widetilde{G}_r^{(\varepsilon)}\right\rangle_{\mathcal{H}_2} = -\varepsilon \left| \left\langle G - G_r,\ \frac{G_r(s)}{s-\mu}\right\rangle_{\mathcal{H}_2}\right| + \mathcal{O}(\varepsilon^2). \tag{3.8}$$

Then (3.6) implies that as $\varepsilon \to 0$, $0 < \left|\left\langle G - G_r,\ \frac{G_r(s)}{s-\mu}\right\rangle_{\mathcal{H}_2}\right| \le C\varepsilon$ for some constant C, which then contradicts (3.7).

Now suppose that $\mu = \alpha + \imath\beta$ is a pole for G_r with a nontrivial imaginary part, $\beta \neq 0$, and so is one of a conjugate pair of poles for G_r. Suppose further that

$$\left\langle G - G_r,\ \frac{G_r(s)}{(s-\alpha)^2+\beta^2}\right\rangle_{\mathcal{H}_2} \neq 0 \quad \text{and} \quad \left\langle G - G_r,\ \frac{(s-\alpha)\,G_r(s)}{(s-\alpha)^2+\beta^2}\right\rangle_{\mathcal{H}_2} \neq 0. \tag{3.9}$$

Write $G_r(s) = \frac{p_{r-1}(s)}{[(s-\alpha)^2+\beta^2]\,q_{r-2}(s)}$ for some choice of real polynomials $p_{r-1} \in \mathcal{P}_{r-1}$ and $q_{r-2} \in \mathcal{P}_{r-2}$. Arguments exactly analogous to the previous case lead to the remaining assertions. In particular,

$$\begin{aligned}
\text{to show} \quad & \left\langle G - G_r,\ \frac{G_r(s)}{(s-\alpha)^2+\beta^2}\right\rangle_{\mathcal{H}_2} = 0, \\
& \text{consider} \quad \widetilde{G}_r^{(\varepsilon)}(s) = \frac{p_{r-1}(s)}{[(s-\alpha)^2+\beta^2-(\pm\varepsilon)]\,q_{r-2}(s)}; \\
\text{to show} \quad & \left\langle G - G_r,\ \frac{(s-\alpha)\,G_r(s)}{(s-\alpha)^2+\beta^2}\right\rangle_{\mathcal{H}_2} = 0, \\
& \text{consider} \quad \widetilde{G}_r^{(\varepsilon)}(s) = \frac{p_{r-1}(s)}{[(s-\alpha-(\pm\varepsilon))^2+\beta^2]\,q_{r-2}(s)}.
\end{aligned}$$

The conclusion follows then by observing that if G_r is a locally optimal $\mathcal{H}_2$ reduced order model, then

$$\begin{aligned}
\langle G - G_r,\ G_r \cdot H_1 + H_2\rangle_{\mathcal{H}_2} = & \sum_{i=1}^{m_R} \gamma_i \left\langle G - G_r,\ \frac{G_r(s)}{s-\mu_i}\right\rangle_{\mathcal{H}_2} \\
& + \sum_{i=m_R+1}^{m_R+m_C} \rho_i \left\langle G - G_r,\ \frac{(s-\alpha_i)\,G_r(s)}{(s-\alpha_i)^2+\beta_i^2}\right\rangle_{\mathcal{H}_2} \\
& + \sum_{i=m_R+1}^{m_R+m_C} \tau_i \left\langle G - G_r,\ \frac{G_r(s)}{(s-\alpha_i)^2+\beta_i^2}\right\rangle_{\mathcal{H}_2} \\
& + \langle G - G_r,\ H_2(s)\rangle_{\mathcal{H}_2} = 0.
\end{aligned}$$

Theorem 3.2 describes new necessary conditions for the $\mathcal{H}_2$ approximation problem as structured orthogonality conditions. This new formulation amounts to a unifying framework for the optimal $\mathcal{H}_2$ problem. Indeed, as we show in sections 3.2 and 3.3, two other known optimality frameworks, namely, interpolatory- [26] and Lyapunov-based conditions [36, 22], can be directly obtained from our new conditions by using an appropriate form for the $\mathcal{H}_2$ inner product. The interpolatory framework uses the residue formulation of the $\mathcal{H}_2$ inner product as in (2.16); the Lyapunov framework uses the Sylvester equation formulation of the $\mathcal{H}_2$ norm as in (2.12).

3.2. Interpolation-based optimality conditions. Corollary 2.5 immediately yields an observation regarding the $\mathcal{H}_2$ norm of the error system, which serves as a main motivation for the interpolation framework of the optimal $\mathcal{H}_2$ problem.

PROPOSITION 3.3. *Given the full-order model $G(s)$ and a reduced order model $G_r(s)$, let λ_i and $\widetilde{\lambda}_i$ be the poles of $G(s)$ and $G_r(s)$, respectively, and suppose that the poles of $G_r(s)$ are distinct. Let ϕ_i and $\widetilde{\phi}_j$ denote the residues of the transfer functions $G(s)$ and $G_r(s)$ at their poles λ_i and $\widetilde{\lambda}_i$, respectively: $\phi_i = \mathsf{res}[G(s), \lambda_i]$ for $i = 1, \ldots, n$ and $\widetilde{\phi}_j = \mathsf{res}[G_r(s), \widetilde{\lambda}_j]$ for $j = 1, \ldots, r$. The $\mathcal{H}_2$ norm of the error system is given by*

$$\begin{aligned}\|G - G_r\|_{\mathcal{H}_2}^2 &= \sum_{i=1}^{n} \mathsf{res}[(G(-s) - G_r(-s))\,(G(s) - G_r(s))\,, \lambda_i] \\ &\quad + \sum_{j=1}^{r} \mathsf{res}[(G(-s) - G_r(-s))\,(G(s) - G_r(s))\,, \widetilde{\lambda}_j] \\ &= \sum_{i=1}^{n} \phi_i \Big(G(-\lambda_i) - G_r(-\lambda_i)\Big) - \sum_{j=1}^{r} \widetilde{\phi}_j \Big(G(-\widetilde{\lambda}_j) - G_r(-\widetilde{\lambda}_j)\Big). \end{aligned} \tag{3.10}$$

The $\mathcal{H}_2$ error expression (3.10) is valid for any reduced order model regardless of the underlying reduction technique and generalizes a result of [20, 18] to the most general setting.

Proposition 3.3 has the system-theoretic interpretation that the $\mathcal{H}_2$ error is due to mismatch of the transfer functions $G(s)$ and $G_r(s)$ at mirror images of the full-order poles λ_i and reduced order poles $\widetilde{\lambda}_i$. This expression reveals that for good $\mathcal{H}_2$ performance, $G_r(s)$ should approximate $G(s)$ well at $-\lambda_i$ and $-\widetilde{\lambda}_j$. Note that $\widetilde{\lambda}_i$ is not known a priori. Therefore, to minimize the $\mathcal{H}_2$ error, Gugercin and Antoulas [20] proposed choosing $\sigma_i = -\lambda_i(\mathbf{A})$, where $\lambda_i(\mathbf{A})$ are those system poles having big residuals ϕ_i. They have illustrated that this selection of interpolation points works quite well; see [18, 20]. However, as (3.10) illustrates, there is a second part of the $\mathcal{H}_2$ error due to the mismatch at $-\widetilde{\lambda}_j$. Indeed, as we will show below, interpolation at $-\widetilde{\lambda}_i$ is more important for model reduction and is a necessary condition for optimal $\mathcal{H}_2$ model reduction; i.e., $\sigma_i = -\widetilde{\lambda}_i$ is the optimal shift selection.

THEOREM 3.4. *Given a stable SISO system $G(s) = \mathbf{c}^T(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}$, let $G_r(s) = \mathbf{c}_r^T(s\mathbf{I} - \mathbf{A}_r)^{-1}\mathbf{b}_r$ be a local minimizer of dimension r for the optimal $\mathcal{H}_2$ model reduction problem* (3.1) *and suppose that $G_r(s)$ has simple poles at $\widetilde{\lambda}_i$, $i = 1, \ldots, r$. Then $G_r(s)$ interpolates both $G(s)$ and its first derivative at $-\widetilde{\lambda}_i$, $i = 1, \ldots, r$:*

$$G_r(-\widetilde{\lambda}_i) = G(-\widetilde{\lambda}_i) \quad \text{and} \quad G_r'(-\widetilde{\lambda}_i) = G'(-\widetilde{\lambda}_i) \quad \text{for } i = 1, \ldots, r. \tag{3.11}$$

Proof. From (3.5), consider first the case $H_1 = 0$ and H_2 is an arbitrary transfer function with simple poles at $\widetilde{\lambda}_i$, $i = 1, \ldots, r$. Denote $\widetilde{\phi}_i = \mathsf{res}[H_2(s), \widetilde{\lambda}_i]$. Then (2.16)

leads to

$$\langle G - G_r,\ H_2\rangle_{\mathcal{H}_2} = \sum_{i=1}^{r} \mathsf{res}[(G(-s) - G_r(-s))\, H_2(s), \widetilde{\lambda}_i] = \sum_{i=1}^{r} \widetilde{\phi}_i \left(G(-\widetilde{\lambda}_i) - G_r(-\widetilde{\lambda}_i)\right) = 0.$$

Since this is true for arbitrary choices of $\widetilde{\phi}_i$, we have $G(-\widetilde{\lambda}_i) = G_r(-\widetilde{\lambda}_i)$. Now consider the case $H_2 = 0$ and H_1 is an arbitrary transfer function with simple poles at $\widetilde{\lambda}_i$, $i = 1, \ldots, r$. Then $G_r(s)H_1(s)$ has double poles at $\widetilde{\lambda}_i$, $i = 1, \ldots, r$, and since $G(-\widetilde{\lambda}_i) = G_r(-\widetilde{\lambda}_i)$ we have

$$\langle G - G_r,\ G_r \cdot H_1\rangle_{\mathcal{H}_2} = \sum_{i=1}^{r} \mathsf{res}[(G(-s) - G_r(-s))\, G_r(s)H_1(s), \widetilde{\lambda}_i] = -\sum_{i=1}^{r} \widetilde{\phi}_i\, \mathsf{res}[G_r, \widetilde{\lambda}_i] \left(G'(-\widetilde{\lambda}_i) - G_r'(-\widetilde{\lambda}_i)\right) = 0,$$

where we have calculated

$$\lim_{s\to\widetilde{\lambda}_i} \left((s - \widetilde{\lambda}_i)^2 G_r(s) \cdot H_1(s)\right) = \mathsf{res}[H_1(s), \widetilde{\lambda}_i] \cdot \mathsf{res}[G_r(s), \widetilde{\lambda}_i] = \widetilde{\phi}_i\, \mathsf{res}[G_r, \widetilde{\lambda}_i]. \qquad \square$$

We refer to the first-order conditions (3.11) as Meier–Luenberger conditions, recognizing the work of [26], although we have here directly obtained them from the newly derived structured orthogonality conditions (3.5).

In Theorem 3.4, we assume that the reduced order poles (eigenvalues of $\mathbf{A}_r$) are simple; analogous results for the case that G_r has a higher order pole are straightforward and correspond to interpolation conditions of higher derivatives at the mirror images of reduced order poles.

3.2.1. Multiple input/multiple output systems. Many of these considerations extend naturally to the multiple input/multiple output (MIMO) setting:

$$\mathbf{G} : \begin{cases} \dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\,\mathbf{u}(t) \\ \mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \end{cases} \qquad \text{or} \qquad \mathbf{G}(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}, \tag{3.12}$$

where the state vector $\mathbf{x}(t) \in \mathbb{R}^n$ as before, but now the system has an *input vector* $\mathbf{u}(t) \in \mathbb{R}^m$ and *output vector* $\mathbf{y}(t) \in \mathbb{R}^p$, so that $\mathbf{B} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{p\times n}$ for some $m, p \geq 1$. The transfer function, $\mathbf{G}(s)$, in (3.12) becomes matrix valued. A reduced order system analogous to (1.2) is sought with the same number of inputs m and outputs p, but with lower state space dimension $r \ll n$. If $\mathbf{V}_r \in \mathbb{R}^{n\times r}$ and $\mathbf{W}_r \in \mathbb{R}^{n\times r}$ such that $\mathbf{W}_r^T\mathbf{V}_r$ is nonsingular, we can define a (matrix-valued) reduced order transfer function $\mathbf{G}_r(s) = \mathbf{C}_r(s\mathbf{I} - \mathbf{A}_r)^{-1}\mathbf{B}_r$ with

$$\mathbf{A}_r = (\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T\mathbf{A}\mathbf{V}_r, \quad \mathbf{B}_r = (\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T\mathbf{B}, \quad \text{and} \quad \mathbf{C}_r = \mathbf{C}\mathbf{V}_r.$$

In order to assess "closeness" of MIMO systems, there is a natural extension of the Hilbert space, $\mathcal{H}_2$, to $p \times m$ matrix-valued functions. In particular, if $\mathbf{G}(s)$ and $\mathbf{H}(s)$ are $p \times m$ matrix-valued transfer functions associated with real stable MIMO dynamical systems, then the associated $\mathcal{H}_2$ inner product is

$$\langle \mathbf{G},\ \mathbf{H}\rangle_{\mathcal{H}_2} \stackrel{\text{def}}{=} \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathsf{tr}\left(\overline{\mathbf{G}(\imath\omega)}\,\mathbf{H}^T(\imath\omega)\right) d\omega = \frac{1}{2\pi}\int_{-\infty}^{\infty} \mathsf{tr}\left(\mathbf{G}(-\imath\omega)\mathbf{H}^T(\imath\omega)\right) d\omega, \tag{3.13}$$

where "$\mathsf{tr}(\mathbf{M})$" denotes the trace of the matrix $\mathbf{M}$. The $\mathcal{H}_2$ norm is then

$$\|\mathbf{G}\|_{\mathcal{H}_2} \stackrel{\text{def}}{=} \left(\frac{1}{2\pi}\int_{-\infty}^{+\infty} \|\mathbf{G}(\imath\omega)\|_F^2\, d\omega\right)^{1/2}, \tag{3.14}$$

where $\|\mathbf{F}\|_F \stackrel{\text{def}}{=} \left(\sum_{ij} |F_{ij}|^2\right)^{1/2}$ denotes the usual Frobenius matrix norm. As before, if $\mathbf{G}(s)$ and $\mathbf{H}(s)$ represent real dynamical systems, then $\langle \mathbf{G},\ \mathbf{H}\rangle_{\mathcal{H}_2} = \langle \mathbf{H},\ \mathbf{G}\rangle_{\mathcal{H}_2}$ and $\langle \mathbf{G},\ \mathbf{H}\rangle_{\mathcal{H}_2}$ is real.

Necessary conditions for $\mathcal{H}_2$ optimality built on structured orthogonality paralleling the results of section 3.1 can be derived in this setting as well. In particular, the residue form for the inner product is a straightforward analogue of Lemma 2.4 and leads naturally to interpolation conditions. If $\mathbf{F}(s)$ is a matrix-valued meromorphic function with a pole at λ, then $\mathbf{F}(s)$ has a Laurent expansion (with matrix coefficients), and its residue, $\mathsf{res}[\mathbf{F}(s), \lambda]$, will be the coefficient matrix associated with the expansion term $(s-\lambda)^{-1}$. For example, suppose that $\mathbf{F}(s)$ has the realization $\mathbf{F}(s) = \widetilde{\mathbf{C}}\big(s\mathbf{I} - \widetilde{\mathbf{A}}\big)^{-1}\widetilde{\mathbf{B}}$. If λ is a simple pole of $\mathbf{F}(s)$, then we can assume that λ is a simple eigenvalue of $\widetilde{\mathbf{A}}$ associated with a rank-1 spectral projector $\mathbf{E}_\lambda$ and then $\mathbf{F}(s) = \frac{1}{s-\lambda}\mathbf{E}_\lambda + \mathbf{D}(s)$, where $\mathbf{D}(s)$ is analytic at $s=\lambda$, and $\mathsf{res}[\mathbf{F}(s),\lambda] = \lim_{s\to\lambda}(s-\lambda)\mathbf{F}(s) = \widetilde{\mathbf{C}}\mathbf{E}_\lambda\widetilde{\mathbf{B}}$. If λ is a double pole, then we can assume that λ is a double eigenvalue of $\widetilde{\mathbf{A}}$ associated with a rank-2 spectral projector $\mathbf{E}_\lambda$ and a rank-1 nilpotent matrix $\mathbf{N}_\lambda$ such that $\widetilde{\mathbf{A}}\mathbf{E}_\lambda = \lambda\mathbf{E}_\lambda + \mathbf{N}_\lambda$. Then $\mathbf{F}(s) = \frac{1}{(s-\lambda)^2}\mathbf{N}_\lambda + \frac{1}{(s-\lambda)}\mathbf{E}_\lambda + \mathbf{D}(s)$, where $\mathbf{D}(s)$ is analytic at $s=\lambda$, and so $\mathsf{res}[\mathbf{F}(s),\lambda] = \lim_{s\to\lambda}\frac{d}{ds}\left[(s-\lambda)^2\mathbf{F}(s)\right] = \widetilde{\mathbf{C}}\mathbf{E}_\lambda\widetilde{\mathbf{B}}$.

LEMMA 3.5. *Suppose that $\mathbf{G}(s)$ has poles at $\lambda_1, \lambda_2, \ldots, \lambda_n$ and $\mathbf{H}(s)$ has poles at $\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots, \widetilde{\lambda}_{\tilde{n}}$, with both sets contained in the open left half plane. Then*

$$\langle \mathbf{G},\ \mathbf{H}\rangle_{\mathcal{H}_2} = \sum_{k=1}^{\tilde{n}} \mathsf{tr}\left(\mathsf{res}[\mathbf{G}(-s)\mathbf{H}^T(s), \widetilde{\lambda}_k]\right). \tag{3.15}$$

In particular, suppose $\mathbf{H}(s)$ has a realization $\mathbf{H}(s) = \widetilde{\mathbf{C}}(s\mathbf{I} - \widetilde{\mathbf{A}})^{-1}\widetilde{\mathbf{B}}$:

- *If $\widetilde{\lambda}_k$ is a simple pole of $\mathbf{H}(s)$, and $\widetilde{\lambda}_k$ is associated with left and right eigenvectors of $\widetilde{\mathbf{A}}$, $\widetilde{\mathbf{y}}_k$, and $\widetilde{\mathbf{x}}_k$, respectively,*

  $$\widetilde{\mathbf{A}}\widetilde{\mathbf{x}}_k = \widetilde{\lambda}_k\,\widetilde{\mathbf{x}}_k, \qquad \widetilde{\mathbf{y}}_k^*\widetilde{\mathbf{A}} = \widetilde{\lambda}_k\,\widetilde{\mathbf{y}}_k^*, \qquad \text{and} \quad \widetilde{\mathbf{y}}_k^*\widetilde{\mathbf{x}}_k = 1,$$

  then

  $$\mathsf{tr}\left(\mathsf{res}[\mathbf{G}(-s)\mathbf{H}^T(s), \widetilde{\lambda}_k]\right) = \widetilde{\mathbf{c}}_k^T\mathbf{G}(-\widetilde{\lambda}_k)\widetilde{\mathbf{b}}_k,$$

  where $\widetilde{\mathbf{b}}_k^T = \widetilde{\mathbf{y}}_k^*\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{c}}_k = \widetilde{\mathbf{C}}\widetilde{\mathbf{x}}_k$.
- *If $\widetilde{\lambda}_k$ is a double pole of $\mathbf{H}(s)$, and $\widetilde{\lambda}_k$ is associated with left and right eigenvectors $\widetilde{\mathbf{y}}_k$ and $\widetilde{\mathbf{x}}_k$ of $\widetilde{\mathbf{A}}$, and generalized eigenvectors, $\widetilde{\mathbf{z}}_k$ and $\widetilde{\mathbf{w}}_k$, respectively,*

  $$\widetilde{\mathbf{A}}\widetilde{\mathbf{x}}_k = \widetilde{\lambda}_k\,\widetilde{\mathbf{x}}_k, \quad \widetilde{\mathbf{A}}\widetilde{\mathbf{w}}_k = \widetilde{\lambda}_k\,\widetilde{\mathbf{w}}_k + \widetilde{\mathbf{x}}_k, \quad \widetilde{\mathbf{y}}_k^*\widetilde{\mathbf{A}} = \widetilde{\lambda}_k\,\widetilde{\mathbf{y}}_k^*, \quad \widetilde{\mathbf{z}}_k^*\widetilde{\mathbf{A}} = \widetilde{\lambda}_k\,\widetilde{\mathbf{z}}_k^* + \widetilde{\mathbf{y}}_k^*,$$
  $$\text{and} \quad \widetilde{\mathbf{y}}_k^*\widetilde{\mathbf{x}}_k = 0, \quad \widetilde{\mathbf{z}}_k^*\widetilde{\mathbf{w}}_k = 0, \quad \text{and} \quad \widetilde{\mathbf{z}}_k^*\widetilde{\mathbf{x}}_k = \widetilde{\mathbf{y}}_k^*\widetilde{\mathbf{w}}_k = 1,$$

  then

  $$\mathsf{tr}\left(\mathsf{res}[\mathbf{G}(-s)\mathbf{H}^T(s), \widetilde{\lambda}_k]\right) = \widetilde{\mathbf{d}}_k^T\mathbf{G}(-\widetilde{\lambda}_k)\widetilde{\mathbf{b}}_k + \widetilde{\mathbf{c}}_k^T\mathbf{G}(-\widetilde{\lambda}_k)\widetilde{\mathbf{e}}_k - \widetilde{\mathbf{c}}_k^T\mathbf{G}'(-\widetilde{\lambda}_k)\widetilde{\mathbf{b}}_k,$$

  where $\widetilde{\mathbf{b}}_k$ and $\widetilde{\mathbf{c}}_k$ are as above and $\widetilde{\mathbf{e}}_k^T = \widetilde{\mathbf{z}}_k^*\widetilde{\mathbf{B}}$ and $\widetilde{\mathbf{d}}_k = \widetilde{\mathbf{C}}\widetilde{\mathbf{w}}_k$.

Now assume that $\mathbf{G}_r$ is an optimal reduced order model minimizing $\|\mathbf{G}-\mathbf{G}_r\|_{\mathcal{H}_2}$ in the sense described in (3.1) and suppose further that $\mathbf{G}_r$ has simple poles $\widetilde{\lambda}_i$. Take $\mathbf{H}(s)=\mathbf{G}_r$ in (3.5) so that $\mathbf{G}_r(s)=\sum_k \frac{1}{s-\widetilde{\lambda}_k}\widetilde{\mathbf{c}}_k\widetilde{\mathbf{b}}_k^T$ and the residue of $\mathbf{G}_r(s)$ at $\widetilde{\lambda}_k$ is matrix valued and rank one: $\mathsf{res}[\mathbf{G}_r(s),\widetilde{\lambda}_k]=\widetilde{\mathbf{c}}_k\widetilde{\mathbf{b}}_k^T$. An analysis paralleling what we have carried out above yields analogous error expressions (see also [2]) and first-order necessary conditions for the MIMO optimal $\mathcal{H}_2$ reduction problem:

$$\begin{aligned}\mathbf{G}(-\widetilde{\lambda}_k)\widetilde{\mathbf{b}}_k &= \mathbf{G}_r(-\widetilde{\lambda}_k)\widetilde{\mathbf{b}}_k,\\ \widetilde{\mathbf{c}}_k^T\mathbf{G}(-\widetilde{\lambda}_k) &= \widetilde{\mathbf{c}}_k^T\mathbf{G}_r(-\widetilde{\lambda}_k), \quad \text{and}\\ \widetilde{\mathbf{c}}_k^T\mathbf{G}'(-\widetilde{\lambda}_k)\widetilde{\mathbf{b}}_k &= \widetilde{\mathbf{c}}_k^T\mathbf{G}_r'(-\widetilde{\lambda}_k)\widetilde{\mathbf{b}}_k, \quad \text{for } k=1,\ldots,r.\end{aligned} \tag{3.16}$$

The SISO ($m=p=1$) conditions are replaced in the MIMO case by left tangential, right tangential, as well as bi-tangential interpolation conditions. From the discussion of section 2.1, if $\mathrm{Ran}(\mathbf{V}_r)$ contains $(\widetilde{\lambda}_k\mathbf{I}+\mathbf{A})^{-1}\mathbf{B}\widetilde{\mathbf{b}}_k$ and $\mathrm{Ran}(\mathbf{W}_r)$ contains $(\widetilde{\lambda}_k\mathbf{I}+\mathbf{A})^{-T}\mathbf{C}^T\widetilde{\mathbf{c}}_k$ for each $k=1,2,\ldots,r$, then the $\mathcal{H}_2$ optimality conditions given above hold. First-order interpolatory MIMO conditions have been obtained recently in other independent works as well; see [24, 35].

3.2.2. The discrete time case. An nth-order SISO discrete-time dynamical system is defined by a set of difference equations

$$G:\begin{cases}\mathbf{x}(t+1)=\mathbf{A}\mathbf{x}(t)+\mathbf{b}\,u(t)\\ y(t)=\mathbf{c}^T\mathbf{x}(t)\end{cases} \qquad \text{or} \qquad G(z)=\mathbf{c}^T(z\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}, \tag{3.17}$$

where $t\in\mathbb{Z}$ and $\mathbf{A}\in\mathbb{R}^{n\times n}$, $\mathbf{b},\mathbf{c}\in\mathbb{R}^n$. $G(z)$ is the transfer function of the system, so that if $\hat{u}(z)$ and $\hat{y}(z)$ denote the z-transforms of $u(t)$ and $y(t)$, respectively, then $\hat{y}(z)=G(z)\hat{u}(z)$. In this case, stability of G means that $\mid\lambda_i(\mathbf{A})\mid<1$ for $i=1,\ldots,n$. Also, the h_2 norm is defined as $\|G\|_{h_2}^2=\frac{1}{2\pi}\int_0^{2\pi}\mid G(e^{\imath\theta})\mid^2 d\theta$. Model reduction for discrete-time systems is defined similarly. In this setting, interpolatory (necessary) conditions for h_2 optimality of the rth-order reduced model $G_r(z)=\mathbf{c}_r^T(z\mathbf{I}-\mathbf{A}_r)^{-1}\mathbf{b}_r$ become $G\big(1/\widetilde{\lambda}_i\big)=G_r\big(1/\widetilde{\lambda}_i\big)$ and $G'\big(1/\widetilde{\lambda}_i\big)=G_r'\big(1/\widetilde{\lambda}_i\big)$ for $i=1,\ldots,r$, where $\widetilde{\lambda}_i$ denotes the ith eigenvalue of $\mathbf{A}_r$. This is a special case of results for discrete-time MIMO systems formulated previously in [10].

3.3. Lyapunov-based $\mathcal{H}_2$ optimality conditions. In this section we briefly review the Lyapunov framework for the first-order $\mathcal{H}_2$ optimality conditions and present its connection to our structured orthogonality framework.

Given a stable SISO system $G(s)=\mathbf{c}^T(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}$, let $G_r(s)=\mathbf{c}_r^T(s\mathbf{I}-\mathbf{A}_r)^{-1}\mathbf{b}_r$ be a local minimizer of dimension r for the optimal $\mathcal{H}_2$ model reduction problem (3.1) and suppose that $G_r(s)$ has simple poles at $\widetilde{\lambda}_i$, $i=1,\ldots,r$.

It is convenient to define the error system

$$G_{err}(s)\stackrel{\text{def}}{=}G(s)-G_r(s)=\mathbf{c}_{err}^T\left(s\mathbf{I}-\mathbf{A}_{err}\right)^{-1}\mathbf{b}_{err} \tag{3.18}$$

$$\text{with}\quad \mathbf{A}_{err}=\begin{bmatrix}\mathbf{A} & 0\\ 0 & \mathbf{A}_r\end{bmatrix},\quad \mathbf{b}_{err}=\begin{bmatrix}\mathbf{b}\\ \mathbf{b}_r\end{bmatrix},\quad \text{and } \mathbf{c}_{err}^T=[\mathbf{c}^T\ \ -\mathbf{c}_r^T]. \tag{3.19}$$

Let $\mathbf{P}_{err}$ and $\mathbf{Q}_{err}$ be the Gramians for the error system $G_{err}(s)$; i.e., $\mathbf{P}_{err}$ and $\mathbf{Q}_{err}$ solve

$$\mathbf{A}_{err}\mathbf{P}_{err}+\mathbf{P}_{err}\mathbf{A}_{err}^T+\mathbf{b}_{err}\mathbf{b}_{err}^T=0, \tag{3.20}$$

$$\mathbf{Q}_{err}\mathbf{A}_{err}+\mathbf{A}_{err}^T\mathbf{Q}_{err}+\mathbf{c}_{err}\mathbf{c}_{err}^T=0. \tag{3.21}$$

Partition $\mathbf{P}_{err}$ and $\mathbf{Q}_{err}$:

$$\mathbf{P}_{err} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{12}^T & \mathbf{P}_{22} \end{bmatrix}, \qquad \mathbf{Q}_{err} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{12}^T & \mathbf{Q}_{22} \end{bmatrix}, \tag{3.22}$$

where $\mathbf{P}_{11}, \mathbf{Q}_{11} \in \mathbb{R}^{n\times n}$ and $\mathbf{P}_{22}, \mathbf{Q}_{22} \in \mathbb{R}^{r\times r}$. Wilson [36] showed that the reduced order model $G_r(s) = \mathbf{c}_r^T(s\mathbf{I}-\mathbf{A}_r)^{-1}\mathbf{b}_r$ can be defined in terms of a Galerkin framework as well by taking

$$\mathbf{V}_r = \mathbf{P}_{12}\mathbf{P}_{22}^{-1} \text{ and } \mathbf{W}_r = -\mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}, \tag{3.23}$$

and the resulting reduced order model satisfies the first-order conditions of the optimal $\mathcal{H}_2$ problem. It was also shown in [36] that $\mathbf{W}_r^T\mathbf{V}_r = \mathbf{I}$. The next result states the Lyapunov-based Wilson conditions for $\mathcal{H}_2$ optimality and shows their equivalence to our structured orthogonality framework.

THEOREM 3.6. *The Wilson conditions for $\mathcal{H}_2$ optimality,*

$$\mathbf{P}_{12}^T\mathbf{Q}_{12} + \mathbf{P}_{22}\mathbf{Q}_{22} = 0, \tag{3.24}$$

$$\mathbf{Q}_{12}^T\mathbf{b} + \mathbf{Q}_{22}\mathbf{b}_r = 0, \tag{3.25}$$

$$\mathbf{c}_r^T\mathbf{P}_{22} - \mathbf{c}^T\mathbf{P}_{12} = 0, \tag{3.26}$$

are equivalent to the structured orthogonality conditions of Theorem 3.2.

Proof. From (3.5), consider first the case $H_1 = 0$ and H_2 is an arbitrary transfer function with simple poles at $\widetilde{\lambda}_i$, $i = 1, \ldots, r$. Write $H_2(s) = \widetilde{\mathbf{c}}^T(s\mathbf{I} - \mathbf{A}_r)^{-1}\widetilde{\mathbf{b}}$, where $\widetilde{\mathbf{b}}$ and $\widetilde{\mathbf{c}}$ can vary arbitrarily. Then from (2.12), if, for any $\widetilde{\mathbf{b}} \neq 0$, $[\widetilde{\mathbf{P}}_1^T, \widetilde{\mathbf{P}}_2^T]^T$ solves

$$\begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{A}_r \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{P}}_1 \\ \widetilde{\mathbf{P}}_2 \end{bmatrix} + \begin{bmatrix} \widetilde{\mathbf{P}}_1 \\ \widetilde{\mathbf{P}}_2 \end{bmatrix} \mathbf{A}_r^T + \begin{bmatrix} \mathbf{b} \\ \mathbf{b}_r \end{bmatrix} \widetilde{\mathbf{b}}^T = 0, \tag{3.27}$$

we have for arbitrary $\widetilde{\mathbf{c}}$

$$\langle G - G_r,\ H_2\rangle_{\mathcal{H}_2} = [\mathbf{c}^T\ -\mathbf{c}_r^T] \begin{bmatrix} \widetilde{\mathbf{P}}_1 \\ \widetilde{\mathbf{P}}_2 \end{bmatrix} \widetilde{\mathbf{c}} = 0.$$

Notice that $\widetilde{\mathbf{P}}_1$ and $\widetilde{\mathbf{P}}_2$ are independent of $\widetilde{\mathbf{c}}$, so for each choice of $\widetilde{\mathbf{b}}$ we must have

$$\mathbf{c}^T\widetilde{\mathbf{P}}_1 - \mathbf{c}_r^T\widetilde{\mathbf{P}}_2 = 0.$$

For $\widetilde{\mathbf{b}} = \mathbf{b}_r$, one may check directly that $\widetilde{\mathbf{P}}_1 = \mathbf{P}_{12}$ and $\widetilde{\mathbf{P}}_2 = \mathbf{P}_{22}$ in $\mathbf{P}_{err}$ that solves (3.20) in Wilson's conditions.

Likewise, from (2.13) for each choice of $\widetilde{\mathbf{c}}$, if $[\widetilde{\mathbf{Q}}_1, \widetilde{\mathbf{Q}}_2]$ solves

$$[\widetilde{\mathbf{Q}}_1, \widetilde{\mathbf{Q}}_2] \begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{A}_r \end{bmatrix} + \mathbf{A}_r^T[\widetilde{\mathbf{Q}}_1, \widetilde{\mathbf{Q}}_2] + \widetilde{\mathbf{c}}[\mathbf{c}^T,\ -\mathbf{c}_r^T] = 0, \tag{3.28}$$

then we have for every $\widetilde{\mathbf{b}}$

$$\langle G - G_r,\ H_2\rangle_{\mathcal{H}_2} = \widetilde{\mathbf{b}}^T[\widetilde{\mathbf{Q}}_1, \widetilde{\mathbf{Q}}_2] \begin{bmatrix} \mathbf{b} \\ \mathbf{b}_r \end{bmatrix} = 0.$$

Similarly to the first case, $[\widetilde{\mathbf{Q}}_1, \widetilde{\mathbf{Q}}_2]$ is independent of $\widetilde{\mathbf{b}}$, so for each choice of $\widetilde{\mathbf{c}}$ we must have

$$\widetilde{\mathbf{Q}}_1\mathbf{b} + \widetilde{\mathbf{Q}}_2\mathbf{b}_r = 0,$$

and for the particular case $\widetilde{\mathbf{c}} = -\mathbf{c}_r$, one may check directly that $\widetilde{\mathbf{Q}}_1 = \mathbf{Q}_{12}^T$ and $\widetilde{\mathbf{Q}}_2 = \mathbf{Q}_{22}$ in $\mathbf{Q}_{err}$ that solves (3.21) in Wilson's conditions. The structured orthogonality condition $\langle G - G_r,\ H\rangle_{\mathcal{H}_2} = 0$ taken over all systems $H(s)$ with the same poles as G_r leads directly to the Wilson conditions (3.25) and (3.26).

The additional orthogonality condition $\langle G - G_r,\ G_r \cdot H\rangle_{\mathcal{H}_2} = 0$ taken over all $H(s)$ with the same poles as G_r will yield the remaining Wilson condition (3.24).

Observe that

$$\begin{aligned} G_r(s)H(s) &= \mathbf{c}_r^T(s\mathbf{I} - \mathbf{A}_r)^{-1}\mathbf{b}_r\,\widetilde{\mathbf{c}}^T(s\mathbf{I} - \mathbf{A}_r)^{-1}\widetilde{\mathbf{b}} \\ &= [\mathbf{c}_r^T, 0]\left(s\mathbf{I}_{2r} - \begin{bmatrix} \mathbf{A}_r & \mathbf{b}_r\widetilde{\mathbf{c}}^T \\ 0 & \mathbf{A}_r \end{bmatrix}\right)^{-1}\begin{bmatrix} 0 \\ \widetilde{\mathbf{b}} \end{bmatrix}. \end{aligned}$$

Referring to (2.12), the condition $\langle G - G_r,\ G_r \cdot H\rangle_{\mathcal{H}_2} = 0$ leads to a Sylvester equation,

$$\begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{A}_r \end{bmatrix}\begin{bmatrix} \widetilde{\mathbf{W}}_1 & \widetilde{\mathbf{P}}_1 \\ \widetilde{\mathbf{W}}_2 & \widetilde{\mathbf{P}}_2 \end{bmatrix} + \begin{bmatrix} \widetilde{\mathbf{W}}_1 & \widetilde{\mathbf{P}}_1 \\ \widetilde{\mathbf{W}}_2 & \widetilde{\mathbf{P}}_2 \end{bmatrix}\begin{bmatrix} \mathbf{A}_r^T & 0 \\ \widetilde{\mathbf{c}}\mathbf{b}_r^T & \mathbf{A}_r^T \end{bmatrix} + \begin{bmatrix} \mathbf{b} \\ \mathbf{b}_r \end{bmatrix}[0,\ \widetilde{\mathbf{b}}^T] = 0,$$

where the use of $\widetilde{\mathbf{P}}_1$ and $\widetilde{\mathbf{P}}_2$ is intended to indicate that they solve (3.27) as well. Then

$$\langle G - G_r,\ G_r \cdot H_2\rangle_{\mathcal{H}_2} = [\mathbf{c}^T,\ -\mathbf{c}_r^T]\begin{bmatrix} \widetilde{\mathbf{W}}_1 & \widetilde{\mathbf{P}}_1 \\ \widetilde{\mathbf{W}}_2 & \widetilde{\mathbf{P}}_2 \end{bmatrix}\begin{bmatrix} \mathbf{c}_r \\ 0 \end{bmatrix} = 0.$$

Alternatively, from (2.13),

(3.29)
$$\begin{bmatrix} \widetilde{\mathbf{Q}}_1 & \widetilde{\mathbf{Q}}_2 \\ \widetilde{\mathbf{Y}}_1 & \widetilde{\mathbf{Y}}_2 \end{bmatrix}\begin{bmatrix} \mathbf{A} & 0 \\ 0 & \mathbf{A}_r \end{bmatrix} + \begin{bmatrix} \mathbf{A}_r^T & 0 \\ \widetilde{\mathbf{c}}\mathbf{b}_r^T & \mathbf{A}_r^T \end{bmatrix}\begin{bmatrix} \widetilde{\mathbf{Q}}_1 & \widetilde{\mathbf{Q}}_2 \\ \widetilde{\mathbf{Y}}_1 & \widetilde{\mathbf{Y}}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{c}_r \\ 0 \end{bmatrix}[\mathbf{c}^T,\ -\mathbf{c}_r^T] = 0$$

($\widetilde{\mathbf{Q}}_1$ and $\widetilde{\mathbf{Q}}_2$ here also solve (3.28)) and

$$\langle G - G_r,\ G_r \cdot H_2\rangle_{\mathcal{H}_2} = [0,\ \widetilde{\mathbf{b}}^T]\begin{bmatrix} \widetilde{\mathbf{Q}}_1 & \widetilde{\mathbf{Q}}_2 \\ \widetilde{\mathbf{Y}}_1 & \widetilde{\mathbf{Y}}_2 \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{b}_r \end{bmatrix} = 0.$$

Since this last equality is true for all $\widetilde{\mathbf{b}}$, and since $\widetilde{\mathbf{Y}}_1$ and $\widetilde{\mathbf{Y}}_2$ are independent of $\widetilde{\mathbf{b}}$, we see that $\widetilde{\mathbf{Y}}_1\mathbf{b} + \widetilde{\mathbf{Y}}_2\mathbf{b}_r = 0$. We know already that $\widetilde{\mathbf{Q}}_1\mathbf{b} + \widetilde{\mathbf{Q}}_2\mathbf{b}_r = 0$, so

$$\begin{bmatrix} \widetilde{\mathbf{Q}}_1 & \widetilde{\mathbf{Q}}_2 \\ \widetilde{\mathbf{Y}}_1 & \widetilde{\mathbf{Y}}_2 \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{b}_r \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Define $\left[\begin{smallmatrix} \tilde{\mathbf{Q}}_1 & \tilde{\mathbf{Q}}_2 \\ \tilde{\mathbf{Y}}_1 & \tilde{\mathbf{Y}}_2 \end{smallmatrix}\right]\left[\begin{smallmatrix} \tilde{\mathbf{P}}_1 \\ \tilde{\mathbf{P}}_2 \end{smallmatrix}\right] = \left[\begin{smallmatrix} \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_2 \end{smallmatrix}\right]$. We will show that $\widetilde{\mathbf{R}}_1 = 0$. Premultiply (3.27) by $\left[\begin{smallmatrix} \tilde{\mathbf{Q}}_1 & \tilde{\mathbf{Q}}_2 \\ \tilde{\mathbf{Y}}_1 & \tilde{\mathbf{Y}}_2 \end{smallmatrix}\right]$, postmultiply (3.29) by $\left[\begin{smallmatrix} \tilde{\mathbf{P}}_1 \\ \tilde{\mathbf{P}}_2 \end{smallmatrix}\right]$, and subtract the resulting equations to get

$$\widetilde{\mathbf{R}}_1\mathbf{A}_r^T - \mathbf{A}_r^T\widetilde{\mathbf{R}}_1 = 0 \qquad \text{and} \qquad \widetilde{\mathbf{R}}_2\mathbf{A}_r^T - \mathbf{A}_r^T\widetilde{\mathbf{R}}_2 = \widetilde{\mathbf{c}}\mathbf{b}_r^T\widetilde{\mathbf{R}}_1.$$

The first equation asserts that $\widetilde{\mathbf{R}}_1$ commutes with $\mathbf{A}_r^T$, and since $\mathbf{A}_r^T$ has distinct eigenvalues, $\widetilde{\mathbf{R}}_1$ must have the same eigenvectors as $\mathbf{A}_r^T$. Let $\widetilde{\mathbf{y}}_i$, $\widetilde{\mathbf{x}}_i$ be left and right eigenvectors of $\mathbf{A}_r$ associated with $\widetilde{\lambda}_i$ (respectively, right and left eigenvectors of $\mathbf{A}_r^T$): $\mathbf{A}_r\widetilde{\mathbf{x}}_i = \widetilde{\lambda}_i\widetilde{\mathbf{x}}_i$ and $\widetilde{\mathbf{y}}_i^T\mathbf{A}_r = \widetilde{\lambda}_i\widetilde{\mathbf{y}}_i^T$. Then $\widetilde{\mathbf{R}}_1\widetilde{\mathbf{y}}_i = d_i\widetilde{\mathbf{y}}_i$. Now premultiply the second equation by $\widetilde{\mathbf{x}}_i^T$ and postmultiply by $\widetilde{\mathbf{y}}_i$ to find

$$\begin{aligned}\widetilde{\mathbf{x}}_i^T\left(\widetilde{\mathbf{R}}_2\mathbf{A}_r^T - \mathbf{A}_r^T\widetilde{\mathbf{R}}_2\right)\widetilde{\mathbf{y}}_i &= \widetilde{\mathbf{x}}_i^T\widetilde{\mathbf{c}}\mathbf{b}_r^T\widetilde{\mathbf{R}}_1\widetilde{\mathbf{y}}_i,\\ \widetilde{\mathbf{x}}_i^T\widetilde{\mathbf{R}}_2\widetilde{\mathbf{y}}_i\widetilde{\lambda}_i - \widetilde{\lambda}_i\widetilde{\mathbf{x}}_i^T\widetilde{\mathbf{R}}_2\widetilde{\mathbf{y}}_i &= \widetilde{\mathbf{x}}_i^T\widetilde{\mathbf{c}}\mathbf{b}_r^T\widetilde{\mathbf{R}}_1\widetilde{\mathbf{y}}_i,\\ 0 &= \left(\widetilde{\mathbf{x}}_i^T\widetilde{\mathbf{c}}\right)\left(\mathbf{b}_r^T\widetilde{\mathbf{y}}_i\right)\ d_i.\end{aligned}$$

Either $d_i = 0$ or one of $\widetilde{\mathbf{x}}_i^T\widetilde{\mathbf{c}}$ and $\mathbf{b}_r^T\widetilde{\mathbf{y}}_i$ must vanish, which would then imply that either $\dim H < r$ or $\dim G_r < r$. Thus $d_i = 0$ for all $i = 1, \ldots, r$ and $\widetilde{\mathbf{R}}_1 = 0$, which proves the final Wilson condition (3.24).

The converse is omitted here since it follows in a straightforward way by reversing the preceding arguments. ∎

Hyland and Bernstein [22] offered conditions that are equivalent to the Wilson conditions. Suppose $G_r(s)$ defined by $\mathbf{A}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r^T$ solves the optimal $\mathcal{H}_2$ problem. Then there exist positive nonnegative matrices $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{n\times n}$ and two $n \times r$ matrices $\mathbf{F}_r$ and $\mathbf{Y}_r$ such that

$$\mathbf{PQ} = \mathbf{F}_r\mathbf{M}\mathbf{Y}_r^T, \quad \mathbf{Y}_r^T\mathbf{F}_r = \mathbf{I}_r, \tag{3.30}$$

where $\mathbf{M}$ is similar to a positive definite matrix. Then $G_r(s)$ is given by $\mathbf{A}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r^T$ with $\mathbf{A}_r = \mathbf{Y}_r^T\mathbf{A}\mathbf{F}_r$, $\mathbf{b}_r = \mathbf{Y}_r^T\mathbf{b}$, and $\mathbf{c}_r^T = \mathbf{c}^T\mathbf{Y}_r$ such that, with the skew projection $\mathbf{\Pi} = \mathbf{F}_r\mathbf{Y}_r^T$, the following conditions are satisfied:

$$\operatorname{rank}(\mathbf{P}) = \operatorname{rank}(\mathbf{Q}) = \operatorname{rank}(\mathbf{PQ}), \tag{3.31}$$

$$\mathbf{\Pi}\left[\mathbf{AP} + \mathbf{P}\mathbf{A}^T + \mathbf{b}\mathbf{b}^T\right] = 0, \tag{3.32}$$

$$\left[\mathbf{A}^T\mathbf{Q} + \mathbf{Q}\mathbf{A} + \mathbf{c}\mathbf{c}^T\right]\mathbf{\Pi} = 0. \tag{3.33}$$

Note that in both [36] and [22], the first-order necessary conditions are given in terms of (coupled) Lyapunov equations. Both [36] and [22] proposed iterative algorithms to obtain a reduced order model satisfying these Lyapunov-based first-order conditions. However, the main drawback in each case is that both approaches require solving two large-scale Lyapunov equations at each step of the algorithm. [40] discusses computational issues related to solving associated linearized problems within each step.

Theorems 3.4 and 3.6 show the equivalence between the structured orthogonality conditions and Lyapunov- and interpolation-based conditions for $\mathcal{H}_2$ optimality, respectively. To complete the discussion, we formally state the equivalence between the Lyapunov and interpolation frameworks.

LEMMA 3.7 (equivalence of Lyapunov and interpolation frameworks). *The first-order necessary conditions of both* [22] *as given in* (3.31)–(3.33) *and* [36] *as given in* (3.23) *are equivalent to those of* [26] *as given in* (3.11). *That is, the Lyapunov-based first-order conditions* [36, 22] *for the optimal $\mathcal{H}_2$ problem are equivalent to the interpolation-based Meier–Luenberger conditions.*

We note that the connection between the Lyapunov and interpolation frameworks has not been observed in the literature before. This result shows that solving the optimal $\mathcal{H}_2$ problem in the Krylov framework is equivalent to solving it in the Lyapunov framework. This leads to the Krylov-based method proposed in the next section.

4. Iterated interpolation. We propose an effective numerical algorithm that produces a reduced order model $G_r(s)$ satisfying the interpolation-based first-order necessary conditions (3.11). Effectiveness of the proposed algorithm results from the fact that we use rational Krylov steps to construct a $G_r(s)$ that meets the first-order conditions (3.11). No Lyapunov solvers or dense matrix decompositions are needed. Therefore, the method is suited for large-scale systems where $n \gg 1000$.

Several approaches have been proposed in the literature to compute reduced order models that satisfy *some form* of first-order necessary conditions; see [37, 34, 9, 21, 26, 22, 36, 25]. However, these approaches do not seem to be suitable for large-scale problems. The ones based on Lyapunov-based conditions, e.g., [36, 22, 34, 37], require solving a couple of Lyapunov equations at each step of the iteration. To our knowledge, the only methods that depend on interpolation-based necessary conditions have been proposed in [25] and [26]. The authors work directly with the transfer functions of $G(s)$ and $G_r(s)$; make an iteration on the denominator [25] or poles and residues [26] of $G_r(s)$; and explicitly compute $G(s)$, $G_r(s)$, and their derivatives at certain points in the complex plane. However, working with the transfer function, its values, and its derivative values explicitly is not desirable in large-scale settings. Indeed, one will most likely be given a state space representation of $G(s)$ rather than the transfer function. And trying to compute the coefficients of the transfer function can be highly ill-conditioned. These approaches are similar to [30, 31], where interpolation is done by explicit usage of transfer functions. On the other hand, our approach, which is detailed below, is based on the connection between interpolation and effective rational Krylov iteration, and is therefore numerically effective and stable.

Let $\boldsymbol{\sigma}$ denote the set of interpolation points $\{\sigma_1, \ldots, \sigma_r\}$; use these interpolation points to construct a reduced order model, $G_r(s)$, that interpolates both $G(s)$ and $G'(s)$ at $\{\sigma_1, \ldots, \sigma_r\}$; let $\boldsymbol{\lambda}(\boldsymbol{\sigma}) = \{\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_r\}$ denote the resulting reduced order poles of $G_r(s)$; hence $\boldsymbol{\lambda}(\boldsymbol{\sigma})$ is a function from $\mathbb{C}^r \mapsto \mathbb{C}^r$. Define the function $\mathbf{g}(\boldsymbol{\sigma}) = \boldsymbol{\lambda}(\boldsymbol{\sigma}) + \boldsymbol{\sigma}$. Note that $\mathbf{g}(\boldsymbol{\sigma}) : \mathbb{C}^r \mapsto \mathbb{C}^r$. Aside from issues related to the ordering of the reduced order poles, $\mathbf{g}(\boldsymbol{\sigma}) = \mathbf{0}$ yields $\boldsymbol{\lambda}(\boldsymbol{\sigma}) = -\boldsymbol{\sigma}$; i.e., the reduced order poles $\boldsymbol{\lambda}(\boldsymbol{\sigma})$ are mirror images of the interpolation points $\boldsymbol{\sigma}$. Hence, $\mathbf{g}(\boldsymbol{\sigma}) = \mathbf{0}$ is equivalent to (3.11) and is a necessary condition for $\mathcal{H}_2$ optimality of the reduced order model, $G_r(s)$. Thus one can formulate a search for optimal $\mathcal{H}_2$ reduced order systems by considering the root-finding problem $\mathbf{g}(\boldsymbol{\sigma}) = \mathbf{0}$. Many plausible approaches to this problem originate with Newton's method, which appears as

$$\boldsymbol{\sigma}^{(k+1)} = \boldsymbol{\sigma}^{(k)} - (\mathbf{I} + \mathbf{J})^{-1} \left(\boldsymbol{\sigma}^{(\mathbf{k})} + \boldsymbol{\lambda} \left(\boldsymbol{\sigma}^{(\mathbf{k})} \right) \right). \tag{4.1}$$

In (4.1), $\mathbf{J}$ is the usual $r \times r$ Jacobian of $\boldsymbol{\lambda}(\boldsymbol{\sigma})$ with respect to $\boldsymbol{\sigma}$: for $\mathbf{J} = [J_{i,j}]$, $J_{i,j} = \frac{\partial \widetilde{\lambda}_i}{\partial \sigma_j}$ for $i, j = 1, \ldots, r$. How to compute $\mathbf{J}$ will be clarified in section 4.3.

4.1. Proposed algorithm. We seek a reduced order transfer function $G_r(s)$ that interpolates $G(s)$ at the mirror images of the poles of $G_r(s)$ by solving the equivalent root-finding problem, say by a variant of (4.1). It is often the case that in the neighborhood of an $\mathcal{H}_2$ optimal shift set, the entries of the Jacobian matrix become small and simply setting $\mathbf{J} = 0$ might serve as a relaxed iteration strategy. This leads to a successive substitution framework: $\sigma_i \leftarrow -\lambda_i(\mathbf{A}_r)$; successive interpolation steps using a rational Krylov method are used so that at the $(i+1)$st step interpolation points are chosen as the mirror images of the Ritz values from the ith step. Despite its simplicity, this appears to be a very effective strategy in many circumstances.

Here is a sketch of the proposed algorithm.

ALGORITHM 4.1. An iterative rational Krylov algorithm (IRKA).

1. *Make an initial selection of* σ_i *for* $i = 1, \ldots, r$ *that is closed under conjugation and fix a convergence tolerance* tol.
2. *Choose* $\mathbf{V}_r$ *and* $\mathbf{W}_r$ *so that* $\mathrm{Ran}(\mathbf{V}_r) = \mathrm{span}\left\{(\sigma_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \ldots, (\sigma_r\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\right\}$, $\mathrm{Ran}(\mathbf{W}_r) = \mathrm{span}\left\{(\sigma_1\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}, \ldots, (\sigma_r\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}\right\}$, *and* $\mathbf{W}_r^T\mathbf{V}_r = \mathbf{I}$.
3. *while (relative change in* $\{\sigma_i\} >$ tol*)*
   (a) $\mathbf{A_r} = \mathbf{W_r^T A V_r}$,
   (b) *Assign* $\sigma_i \longleftarrow -\lambda_i(\mathbf{A_r})$ *for* $i = 1, \ldots, r$
   (c) *Update* $\mathbf{V}_r$ *and* $\mathbf{W}_r$ *so* $\mathrm{Ran}(\mathbf{V}_r) = \mathrm{span}\left\{(\sigma_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \ldots, (\sigma_r\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\right\}$, $\mathrm{Ran}(\mathbf{W}_r) = \mathrm{span}\left\{(\sigma_1\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}, \ldots, (\sigma_r\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}\right\}$, *and* $\mathbf{W}_r^T\mathbf{V}_r = \mathbf{I}$.
4. $\mathbf{A_r} = \mathbf{W_r^T A V_r}$, $\mathbf{b}_r = \mathbf{W}_r^T\mathbf{b}$, $\mathbf{c}_r^T = \mathbf{c}^T\mathbf{V}_r$

Upon convergence, the first-order necessary conditions (3.11) for $\mathcal{H}_2$ optimality will be satisfied. Notice that step 3(b) could be replaced with some variant of a Newton step (4.1).

We have implemented the above algorithm and applied it to many different large-scale systems. In each of our numerical examples, the algorithm worked very effectively: It has always converged after a small number of steps and resulted in stable reduced systems. For those standard test problems we tried where a global optimum is known, Algorithm 4.1 converged to this global optimum.

It should be noted that the solution is obtained via Krylov projection methods only and its computation is suitable for large-scale systems. To our knowledge, this is the first numerically effective approach for the optimal $\mathcal{H}_2$ reduction problem.

We know that the reduced model $G_r(s)$ resulting from the above algorithm will satisfy the first-order optimality conditions. Moreover, from Theorem 3.1 this reduced order model is globally optimal in the following sense.

COROLLARY 4.1. *Let* $G_r(s)$ *be the reduced model resulting from Algorithm* 4.1. *Then* $G_r(s)$ *is the optimal approximation of* $G(s)$ *with respect to the* $\mathcal{H}_2$ *norm among all reduced order systems having the same reduced system poles as* $G_r(s)$.

Therefore Algorithm 4.1 generates a reduced model, $G_r(s)$, which is the optimal solution for a restricted $\mathcal{H}_2$ problem.

4.2. Initial shift selection. For the proposed algorithm, the final reduced model can depend on the initial shift selection. Nonetheless for most of the cases, a random initial shift selection resulted in satisfactory reduced models. For small-order benchmark examples taken from [22, 25, 37, 34], the algorithm converged to the global minimizer. For larger problems, the results were as good as those obtained by balanced truncation. Therefore, while staying within a numerically effective Krylov projection framework, we have been able to produce results close to or better than those obtained by balanced truncation (which requires the solution of two large-scale Lyapunov equations).

We outline some initialization strategies that can be expected to improve the results. Recall that at convergence, interpolation points are mirror images of the eigenvalues of $\mathbf{A}_r$. The eigenvalues of $\mathbf{A}_r$ might be expected to approximate the eigenvalues of $\mathbf{A}$. Hence, at convergence, interpolation points will lie in the mirror spectrum of $\mathbf{A}$. Therefore, one could choose initial shifts randomly distributed within a region containing the mirror image of the numerical range of $\mathbf{A}$. The boundary of the numerical range can be estimated by computing the eigenvalues of $\mathbf{A}$ with the smallest and largest real and imaginary parts using numerically effective tools such as the implicitly restarted Arnoldi (IRA) algorithm.

The starting point for another initialization strategy is the $\mathcal{H}_2$ expression presented in Proposition 3.3. Based on this expression, it is appropriate to initiate the proposed algorithm with $\sigma_i = -\lambda_i(\mathbf{A})$, where $\lambda_i(\mathbf{A})$ are the poles with big residues, ϕ_i for $i = 1, \ldots, r$. The main disadvantage of this approach is that it requires a modal state space decomposition for $G(s)$, which will be numerically expensive for large-scale problems. However, there might be some applications where the original state space representation is in the modal form and ϕ_i might be directly read from the entries of the matrices $\mathbf{b}$ and $\mathbf{c}^T$.

Unstable reduced order models are not acceptable candidates for optimal $\mathcal{H}_2$ reduction. Nonetheless stability of a reduced model is not guaranteed a priori and might depend on the initial shift selection. We have observed that if one avoids making extremely unrealistic initial shift selections, stability will be preserved. In our simulations we have never generated an unstable system when the initial shift selection was not drastically different from the mirror spectrum of $\mathbf{A}$, but otherwise random. We were able to produce an unstable reduced order system; however, this occurred for a case where the real parts of the eigenvalues of $\mathbf{A}$ were between -1.5668×10^{-1} and -2.0621×10^{-3}, yet we chose initial shifts bigger than 50. We believe that with a good starting point, stability will not be an issue. These considerations are illustrated for many numerical examples in section 5.

Remark 4.1. Based on the first-order conditions (3.16) discussed in section 3.2.1 for MIMO systems $\mathbf{G}(s) = \mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}$, one can extend IRKA to the MIMO case by replacing $(\sigma_i\mathbf{I}-\mathbf{A})^{-1}\mathbf{b}$ with $(\sigma_i\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\widetilde{\mathbf{b}}_i$ and $(\sigma_i\mathbf{I}-\mathbf{A}^T)^{-1}\mathbf{c}$ with $(\sigma_i\mathbf{I}-\mathbf{A})^{-1}\mathbf{C}^T\widetilde{\mathbf{c}}_i$ in Algorithm 4.1, where $\widetilde{\mathbf{b}}_i$ and $\widetilde{\mathbf{c}}_i$ are as defined in section 3.2.1.

Remark 4.2. In the discrete-time case described in (3.17) above, the root-finding problem becomes $\mathbf{g}(\boldsymbol{\sigma}) = \boldsymbol{\Sigma}\boldsymbol{\lambda}(\boldsymbol{\sigma}) - \mathbf{e}$, where $\mathbf{e}^T = [1, 1, \ldots, 1]$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma})$. Therefore, for discrete-time systems, step 3(b) of Algorithm 4.1 becomes $\sigma_i \leftarrow 1/\lambda_i(\mathbf{A_r})$ for $i = 1, \ldots, r$. Moreover, the associated Newton step is

$$\boldsymbol{\sigma}^{(k+1)} = \boldsymbol{\sigma}^{(k)} - (\mathbf{I} + \boldsymbol{\Lambda}^{-1}\boldsymbol{\Sigma}\mathbf{J})^{-1}\left(\boldsymbol{\sigma}^{(\mathbf{k})} - \boldsymbol{\Lambda}^{-1}\mathbf{e}\right),$$

where $\Lambda = \mathrm{diag}(\boldsymbol{\lambda})$.

4.3. A Newton framework for IRKA. As discussed above, Algorithm 4.1 uses the successive substitution framework by simply setting $\mathbf{J} = \mathbf{0}$ in the Newton step (4.1). The Newton framework for IRKA can be easily obtained by replacing step 3(b) of Algorithm 4.1 with the Newton step (4.1). The only point to clarify for the Newton framework is the computation of the Jacobian, which measures the sensitivity of the reduced system poles with respect to shifts.

Given $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$, suppose that σ_i, $i = 1, \ldots, r$, are r distinct points in $\mathbb{C}$, none of which are eigenvalues of $\mathbf{A}$, and define the complex r-tuple $\boldsymbol{\sigma} = [\sigma_1, \sigma_2, \ldots, \sigma_r]^T \in \mathbb{C}^r$ together with related matrices:

$$\mathbf{V}_r(\boldsymbol{\sigma}) = \begin{bmatrix} (\sigma_1\mathbf{I}-\mathbf{A})^{-1}\mathbf{b} & (\sigma_2\mathbf{I}-\mathbf{A})^{-1}\mathbf{b} & \ldots & (\sigma_r\mathbf{I}-\mathbf{A})^{-1}\mathbf{b} \end{bmatrix} \in \mathbb{C}^{n\times r} \tag{4.2}$$

and

$$\mathbf{W}_r^T(\boldsymbol{\sigma}) = \begin{bmatrix} \mathbf{c}^T(\sigma_1\mathbf{I}-\mathbf{A})^{-1} \\ \mathbf{c}^T(\sigma_2\mathbf{I}-\mathbf{A})^{-1} \\ \vdots \\ \mathbf{c}^T(\sigma_r\mathbf{I}-\mathbf{A})^{-1} \end{bmatrix} \in \mathbb{C}^{r\times n}. \tag{4.3}$$

We normally suppress the dependence on $\boldsymbol{\sigma}$ and write $\mathbf{V}_r(\boldsymbol{\sigma}) = \mathbf{V}_r$ and $\mathbf{W}_r(\boldsymbol{\sigma}) = \mathbf{W}_r$. Hence, the reduced order system matrix $\mathbf{A}_r$ is given by $\mathbf{A}_r = (\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T\mathbf{A}\mathbf{V}_r$, where $(\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r$ plays the role of $\mathbf{W}_r$ in Algorithm 4.1. Let $\widetilde{\lambda}_i$, for $i = 1, \ldots, r$, denote the eigenvalues of $\mathbf{A}_r$. Hence, the Jacobian computation amounts to computing $\mathbf{J}(i,j) = \frac{\partial\widetilde{\lambda}_i}{\partial\sigma_j}$. The following result shows how to compute the Jacobian for the Newton formulation of the IRKA method proposed here.

LEMMA 4.2. *Let $\widetilde{\mathbf{x}}_i$ be an eigenvector of $\mathbf{A}_r = (\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T\mathbf{A}\mathbf{V}_r$ associated with $\widetilde{\lambda}_i$, normalized so that $|\widetilde{\mathbf{x}}_i^T\mathbf{W}_r^T\mathbf{V}_r\widetilde{\mathbf{x}}_i| = 1$. Then $\mathbf{W}_r^T\mathbf{A}\mathbf{V}_r\widetilde{\mathbf{x}}_i = \widetilde{\lambda}_i\mathbf{W}_r^T\mathbf{V}_r\widetilde{\mathbf{x}}_i$ and*

$$\frac{\partial\widetilde{\lambda}_i}{\partial\sigma_j} = \widetilde{\mathbf{x}}_i^T\partial_j\mathbf{W}_r^T\left(\mathbf{A}\mathbf{V}_r\widetilde{\mathbf{x}}_i - \widetilde{\lambda}_i\mathbf{V}_r\widetilde{\mathbf{x}}_i\right) + \left(\widetilde{\mathbf{x}}_i^T\mathbf{W}_r^T\mathbf{A} - \widetilde{\lambda}_i\widetilde{\mathbf{x}}_i^T\mathbf{W}_r^T\right)\partial_j\mathbf{V}_r\widetilde{\mathbf{x}}_i, \tag{4.4}$$

where $\partial_j\mathbf{W}_r^T = \frac{\partial}{\partial\sigma_j}\mathbf{W}_r^T = -\mathbf{e}_j\,\mathbf{c}(\sigma_j\mathbf{I}-\mathbf{A})^{-2}$ and $\partial_j\mathbf{V}_r = \frac{\partial}{\partial\sigma_j}\mathbf{V}_r = -(\sigma_j\mathbf{I}-\mathbf{A})^{-2}\mathbf{b}\,\mathbf{e}_j^T$.

Proof. With $\mathbf{V}_r(\boldsymbol{\sigma}) = \mathbf{V}_r$ and $\mathbf{W}_r(\boldsymbol{\sigma}) = \mathbf{W}_r$ defined as in (4.2) and (4.3), both $\mathbf{W}_r^T\mathbf{A}\mathbf{V}_r$ and $\mathbf{W}_r^T\mathbf{V}_r$ are complex symmetric matrices. Write $\widetilde{\lambda}$ for $\widetilde{\lambda}_i$ and $\widetilde{\mathbf{x}}$ for $\widetilde{\mathbf{x}}_i$, so

$$\text{(a) } \mathbf{W}_r^T\mathbf{A}\mathbf{V}_r\widetilde{\mathbf{x}} = \widetilde{\lambda}\mathbf{W}_r^T\mathbf{V}_r\widetilde{\mathbf{x}} \quad\text{and}\quad \text{(b) } \widetilde{\mathbf{x}}^T\mathbf{W}_r^T\mathbf{A}\mathbf{V}_r = \widetilde{\lambda}\,\widetilde{\mathbf{x}}^T\mathbf{W}_r^T\mathbf{V}_r. \tag{4.5}$$

Equation (4.5b) is obtained by transposition of (4.5a). $\widetilde{\mathbf{x}}^T\mathbf{W}_r^T\mathbf{V}_r$ is a left eigenvector for $\mathbf{A}_r$ associated with $\widetilde{\lambda}_i$. Differentiate (4.5a) with respect to σ_j, premultiply with $\widetilde{\mathbf{x}}^T$, and simplify using (4.5b):

$$\widetilde{\mathbf{x}}^T\partial_j\mathbf{W}_r^T\left(\mathbf{A}\mathbf{V}_r\widetilde{\mathbf{x}} - \widetilde{\lambda}\mathbf{V}_r\widetilde{\mathbf{x}}\right) + \left(\widetilde{\mathbf{x}}^T\mathbf{W}_r^T\mathbf{A} - \widetilde{\lambda}\widetilde{\mathbf{x}}^T\mathbf{W}_r^T\right)\partial_j\mathbf{V}_r\widetilde{\mathbf{x}} = \left(\frac{\partial\widetilde{\lambda}}{\partial\sigma_j}\right)\widetilde{\mathbf{x}}^T\mathbf{W}_r^T\mathbf{V}_r\widetilde{\mathbf{x}},$$

where $\partial_j\mathbf{W}_r^T = \frac{\partial}{\partial\sigma_j}\mathbf{W}_r^T = \mathbf{e}_j\,\mathbf{c}^T(\sigma_j\mathbf{I}-\mathbf{A})^{-2}$ and $\partial_j\mathbf{V}_r = \frac{\partial}{\partial\sigma_j}\mathbf{V} = (\sigma_j\mathbf{I}-\mathbf{A})^{-2}\mathbf{b}\,\mathbf{e}_j^T$. This completes the proof. ∎

5. Numerical examples. We first compare our approach with the earlier approaches [22, 25, 37] on *low-order* benchmark examples presented in those papers. We show that in each case we attain the minimum, the main difference being that we achieve this minimum in a numerically efficient way. For each low-order model, comparisons are made using data taken from the original sources [22, 25, 37]. We then test our method in large-scale settings.

5.1. Low-order models and comparisons. Consider the following 4 models:

- FOM-1: Example 6.1 in [22]. State space representation of FOM-1 is given by

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & -150 \\ 1 & 0 & 0 & -245 \\ 0 & 1 & 0 & -113 \\ 0 & 0 & 1 & -19 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 4 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

  We reduce the order to $r = 3, 2, 1$ using the proposed successive rational Krylov algorithm, denoted by IRKA, and compare our results with the gradient flow method of [37], denoted by GFM; the orthogonal projection method of [22], denoted by OPM; and the balanced truncation method, denoted by BTM.

- FOM-2: Example in [25]. Transfer function of FOM-2 is given by

$$G(s) = \frac{2s^6 + 11.5s^5 + 57.75s^4 + 178.625s^3 + 345.5s^2 + 323.625s + 94.5}{s^7 + 10s^6 + 46s^5 + 130s^4 + 239s^3 + 280s^2 + 194s + 60}.$$

We reduce the order to $r = 6, 5, 4, 3$ using IRKA and compare our results with GFM, OPM, BTM, and the method proposed in [25], denoted by LMPV.

- FOM-3: Example 1 in [34]. Transfer function of FOM-3 is given by

$$G(s) = \frac{s^2 + 15s + 50}{s^4 + 5s^3 + 33s^2 + 79s + 50}.$$

We reduce the order to $r = 3, 2, 1$ using IRKA and compare our results with GFM, OPM, BTM, and the method proposed in [34], denoted by SMM.

- FOM-4: Example 2 in [34]. Transfer function of FOM-4 is given by

$$G(s) = \frac{10000s + 5000}{s^2 + 5000s + 25}.$$

We reduce the order to $r = 1$ IRKA and compare our results with GFM, OPM, BTM, and SMM.

For all these cases, the resulting relative $\mathcal{H}_2$ errors $\frac{\|G(s)-G_r(s)\|_{\mathcal{H}_2}}{\|G(s)\|_{\mathcal{H}_2}}$ are tabulated in Table 5.1 below, which clearly illustrates that the proposed method is the only one that attains the minimum in each case. More importantly, the proposed method achieves this value in a numerically efficient way staying in the Krylov projection framework. No Lyapunov solvers or dense matrix decompositions are needed. The

TABLE 5.1
Comparison.

Model	r	IRKA	GFM	OPM
FOM-1	1	4.2683×10^{-1}	4.2709×10^{-1}	4.2683×10^{-1}
FOM-1	2	3.9290×10^{-2}	3.9299×19^{-2}	3.9290×10^{-2}
FOM-1	3	1.3047×10^{-3}	1.3107×19^{-3}	1.3047×10^{-3}
FOM-2	3	1.171×10^{-1}	1.171×10^{-1}	Divergent
FOM-2	4	8.199×10^{-3}	8.199×10^{-3}	8.199×10^{-3}
FOM-2	5	2.132×10^{-3}	2.132×10^{-3}	Divergent
FOM-2	6	5.817×10^{-5}	5.817×10^{-5}	5.817×10^{-5}
FOM-3	1	4.818×10^{-1}	4.818×10^{-1}	4.818×10^{-1}
FOM-3	2	2.443×10^{-1}	2.443×10^{-1}	Divergent
FOM-3	3	5.74×10^{-2}	5.98×10^{-2}	5.74×10^{-2}
FOM-4	1	9.85×10^{-2}	9.85×10^{-2}	9.85×10^{-2}

Model	r	BTM	LMPV	SMM
FOM-1	1	4.3212×10^{-1}		
FOM-1	2	3.9378×10^{-2}		
FOM-1	3	1.3107×10^{-3}		
FOM-2	3	2.384×10^{-1}	1.171×10^{-1}	
FOM-2	4	8.226×10^{-3}	8.199×10^{-3}	
FOM-2	5	2.452×10^{-3}	2.132×10^{-3}	
FOM-2	6	5.822×10^{-5}	2.864×10^{-4}	
FOM-3	1	4.848×10^{-1}		4.818×10^{-1}
FOM-3	2	3.332×10^{-1}		2.443×10^{-1}
FOM-3	3	5.99×10^{-2}		5.74×10^{-2}
FOM-4	1	9.949×10^{-1}	9.985×10^{-2}	

only arithmetic operations involved are LU decompositions and some linear solvers. Moreover, our method does not require starting from an initial balanced realization, as suggested in [37] and [22]. In all these simulations, we have chosen a random initial shift selection, and the algorithm converged in a small number of steps.

To illustrate the evolution of the $\mathcal{H}_2$ error throughout the iteration, consider the model FOM-2 with $r = 3$. The proposed method yields the following third-order optimal reduced model:

$$G_3(s) = \frac{2.155s^2 + 3.343s + 33.8}{s^3 + 7.457s^2 + 10.51s + 17.57}.$$

Poles of $G_3(s)$ are $\widetilde{\lambda}_1 = -6.2217$ and $\widetilde{\lambda}_{2,3} = -6.1774 \times 10^{-1} \pm \imath 1.5628$, and it can be shown that $G_3(s)$ interpolates the first two moments of $G(s)$ at $-\widetilde{\lambda}_i$ for $i = 1, 2, 3$. Hence, the first-order interpolation conditions are satisfied. This also means that if we start Algorithm 4.1 with the mirror images of these Ritz values, the algorithm converges at the first step. However, we will try four random, but *bad*, initial selections. In other words, we start away from the optimal solution. We test the following four selections: $\mathcal{S}_1 = \{-1.01,\ -2.01,\ -30000\}$, $\mathcal{S}_2 = \{0,\ 10,\ 3\}$, $\mathcal{S}_3 = \{1,\ 10,\ 3\}$, and $\mathcal{S}_4 = \{0.01,\ 20,\ 10000\}$. With selection $\mathcal{S}_1$, we have initiated the algorithm with some negative shifts close to system poles, and consequently with a relative $\mathcal{H}_2$ error bigger than 1. However, in all four cases including $\mathcal{S}_1$, the algorithm converged in 5 steps to the same reduced model. The results are depicted in Figure 5.1.

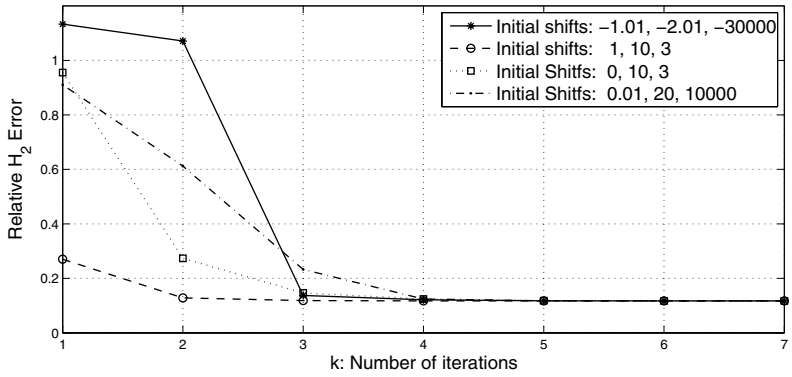

FIG. 5.1. *$\mathcal{H}_2$ norm of the error system vs. the number of iterations.*

Before testing the proposed method in large-scale settings, we investigate FOM-4 further. As pointed out in [34], since $r = 1$, the optimal $\mathcal{H}_2$ problem can be formulated as only a function of the reduced system pole. It was shown in [34] that there are two local minima: (i) one corresponding to a reduced pole at -0.0052 and consequently a reduced order model $G_1^l(s) = \frac{1.0313}{s+0.0052}$ and a relative error of 0.9949, and (ii) one to a reduced pole at -4998 and consequently a reduced model $G_1^g = \frac{9999}{s+4998}$ with a relative error of 0.0985. It follows that the latter, i.e., $G_1^g(s)$, is the global minimum. The first-order balanced truncation for FOM-4 can be easily computed as $G_1^b(s) = \frac{1.0308}{s+0.0052}$. Therefore, it is highly likely that if one starts from a balanced realization, the algorithm would converge to the local minimum $G_1^l(s)$. This was indeed the case as reported in [34]. SMM converged to the local minimum for all starting poles bigger than -0.47. On the other hand, SMM converged to the

global minimum when it was started with an initial pole smaller than -0.47. We have observed exactly the same situation in our simulations. When we start from an initial shift selection smaller than 0.48, IRKA converged to the local minimum. However, when we start with any initial shift bigger than 0.48, the algorithm converged to the global minimum in at most 3 steps. Therefore, for this example we were not able the avoid the local minimum if we started from a *bad* shift. These observations perfectly agree with the discussion of section 4.2. Note that the transfer function of FOM-4 can be written as

$$G(s) = \frac{10000s + 5000}{s^2 + 5000s + 25} = \frac{0.99}{s + 0.0050} + \frac{9999}{s + 5000}.$$

The pole at -5000 is the one corresponding to the large residue of 9999. Therefore, a good initial shift is 5000. And if we start the proposed algorithm with an initial shift at 5000, or close, the algorithm converges to the global minimum.

5.2. CD player example. The original model describes the dynamics between a lens actuator and the radial arm position in a portable CD player. The model has 120 states, i.e., n = 120, with a single input and a single output. As illustrated in [4], the Hankel singular values of this model do not decay rapidly and hence the model is relatively hard to reduce. Moreover, even though the Krylov-based methods resulted in good local behavior, they are observed to yield large $\mathcal{H}_\infty$ and $\mathcal{H}_2$ error compared to balanced truncation.

We compare the performance of the proposed method, Algorithm 4.1, with that of balanced truncation. Balanced truncation is well known to lead to small $\mathcal{H}_\infty$ and $\mathcal{H}_2$ error norms; see [4, 19]. This is due mainly to global information available through the two system Gramians, the reachability and observability Gramians, which are each solutions of a different Lyapunov equation. We reduce the order to r, with r varying from 2 to 40; and for each r value, we compare the $\mathcal{H}_2$ error norms due to balanced truncation and due to Algorithm 4.1. For the proposed algorithm, two different selections have been tried for the initial shifts. (1) Mirror images of the eigenvalues corresponding to large residuals, and (2) a random selection with real parts in the interval $[10^{-1}, 10^3]$ and the imaginary parts in the interval $[1, 10^5]$. To make this selection, we looked at the poles of $G(s)$ having the maximum/minimum real and imaginary parts. The results showing the relative $\mathcal{H}_2$ error for each r are depicted in Figure 5.2. The figure reveals that both selection strategies work quite well. Indeed, the random initial selection behaves better than the residual-based selection and outperforms balanced truncation for almost all the r values except $r = 2, 24, 36$. However, even for these r values, the resulting $\mathcal{H}_2$ error is not far away from the one due to balanced truncation. For the range $r = [12, 22]$, the random selection clearly outperforms the balanced truncation. We would like to emphasize that these results were obtained by a *random* shift selection and staying in the numerically effective Krylov projection framework *without* requiring any solutions to large-scale Lyapunov equations. This is the main difference our proposed algorithm has with existing methods and what makes it numerically effective in large-scale settings.

To examine convergence behavior, we reduce the order to $r = 8$ and $r = 10$ using Algorithm 4.1. At each step of the iteration, we compute the $\mathcal{H}_2$ error due to the current estimate and plot this error versus the iteration index. The results are shown in Figure 5.3. The figure illustrates two important properties for both cases $r = 8$ and $r = 10$: (1) At each step of the iteration, the $\mathcal{H}_2$ norm of the error is reduced. (2) The algorithm converges after 3 steps. The resulting reduced models are stable for both cases.

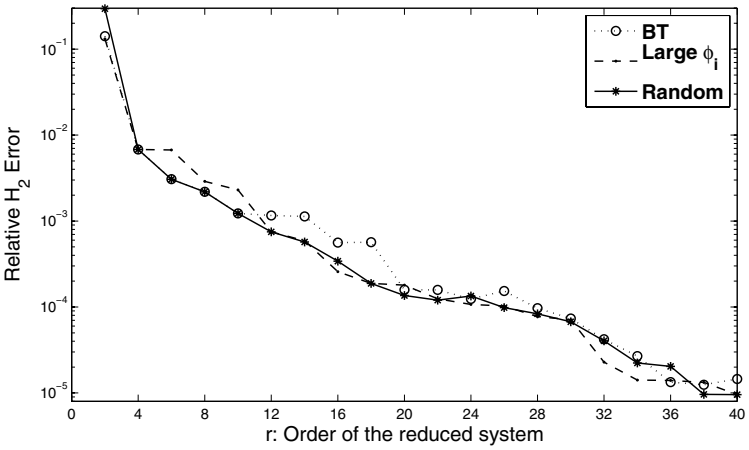

FIG. 5.2. *Relative $\mathcal{H}_2$ norm of the error system vs. r.*

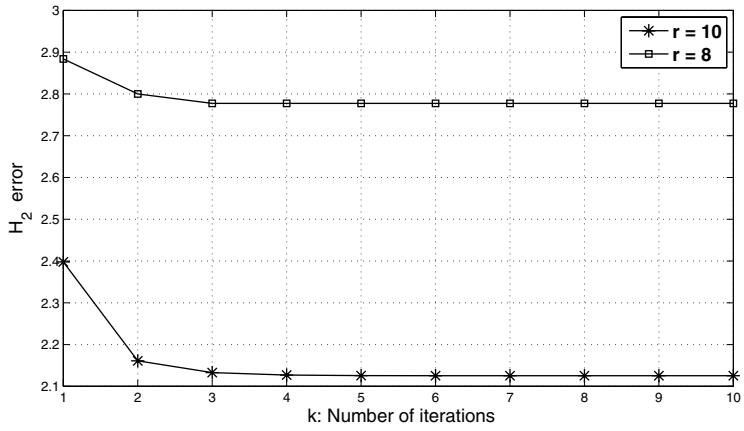

FIG. 5.3. *$\mathcal{H}_2$ norm of the error system vs. the number of iterations.*

5.3. A semidiscretized heat transfer problem for optimal cooling of steel profiles. This problem arises during a cooling process in a rolling mill when different steps in the production process require different temperatures of the raw material. To achieve high throughput, one seeks to reduce the temperature as fast as possible to the required level before entering the next production phase. This is realized by spraying cooling fluids on the surface and must be controlled so that material properties, such as durability or porosity, stay within given quality standards. The problem is modeled as boundary control of a two-dimensional heat equation. A finite element discretization using two steps of mesh refinement with maximum mesh width of 1.382×10^{-2} results in a system of the form

$$\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t), \qquad y(t) = \mathbf{c}^T\mathbf{x}(t),$$

with state dimension $n = 20209$, i.e., $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{20209\times 20209}$, $\mathbf{b} \in \mathbb{R}^{20209\times 7}$, $\mathbf{c}^T \in \mathbb{R}^{6\times 20209}$. Note that in this case $\mathbf{E} \neq \mathbf{I}$, but the algorithm works with the obvious

modifications. For details regarding the modeling, discretization, optimal control design, and model reduction for this example, see [29, 7, 8]. We consider the full-order SISO system that associates the sixth input of this system with the second output. We apply our algorithm and reduce the order to $r = 6$. Amplitude Bode plots of $G(s)$ and $G_r(s)$ are shown in Figure 5.4. The output response of $G_r(s)$ is virtually indistinguishable from $G(s)$ in the frequency range considered. IRKA converged in 7 iteration steps in this case, although some interpolation points converged in the first 2–3 steps. The relative $\mathcal{H}_\infty$ error obtained with our sixth order system was 7.85×10^{-3}. Note that in order to apply balanced truncation in this example, one would need to solve *two generalized* Lyapunov equations (since $\mathbf{E} \neq \mathbf{I}$) of order 20209. This presents a severe computational challenge, though there have been interesting approaches to addressing it (e.g., [5]).

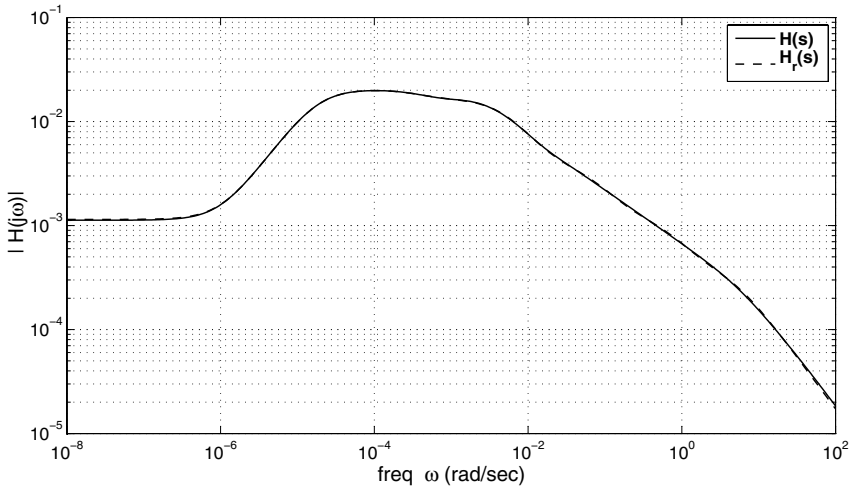

FIG. 5.4. *Amplitude Bode plots of $H(s)$ and $H_r(s)$.*

5.4. Successive substitution vs. Newton framework. In this section, we present two examples to show the effect of the Newton formulation for IRKA on two low-order examples.

The first example is FOM-1 from section 5.1. For this example, for reduction to $r = 1$, the optimal shift is $\sigma = 0.4952$. We initiate both iterations, successive substitution and Newton frameworks, away from this optimal value with an initial selection $\sigma_0 = 10^4$. Figure 5.5 illustrates how each process converges. As the figure shows, even though it takes almost 15 iterations with oscillations for the successive substitution framework to converge, the Newton formulation reaches the optimal shift in 4 steps.

The second example in this section is a third-order model with a transfer function

$$G = \frac{-s^2 + (7/4)s + 5/4}{s^3 + 2s^2 + (17/16)s + 15/32}.$$

One can exactly compute the optimal $\mathcal{H}_2$ reduced model for $r = 1$ as

$$G_r(s) = \frac{0.97197}{s + 0.2727272}$$

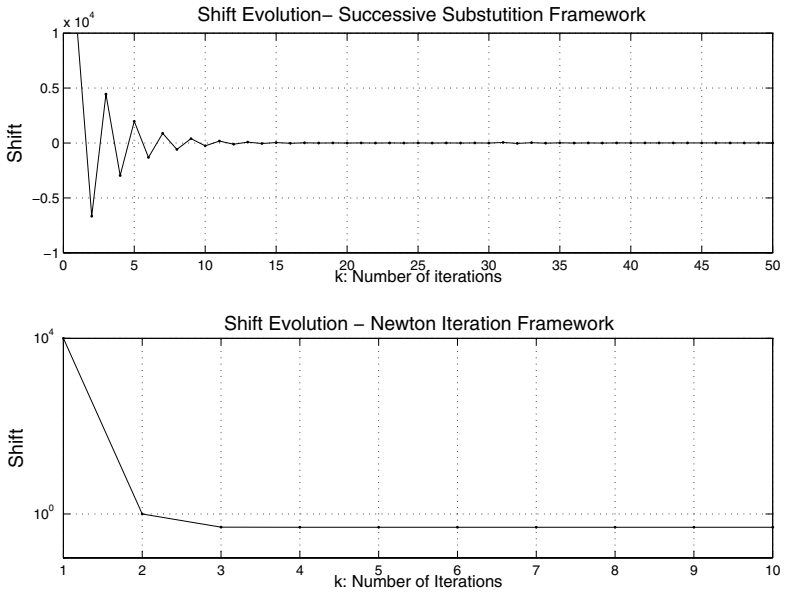

FIG. 5.5. *Comparison for* FOM-1.

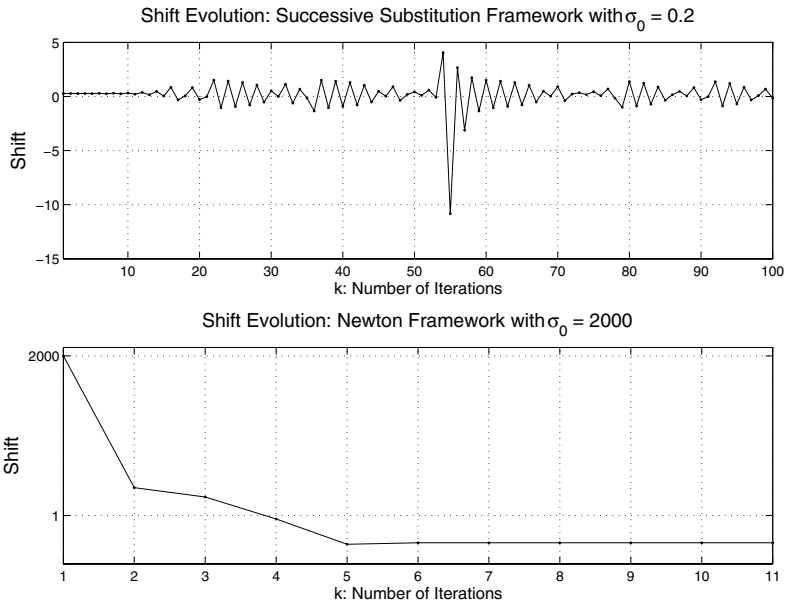

FIG. 5.6. *Comparison for the random third-order model.*

and easily show that this reduced model interpolates $G(s)$ and its derivative at $\sigma = 0.2727272$. We initiate Algorithm 4.1 with $\sigma_0 = 0.27$, very close to the optimal shift. We initiate the Newton framework at $\sigma_0 = 2000$, far away from the optimal solution. Convergence behavior of both models is depicted in Figure 5.6. The figure shows that for this example, the successive substitution framework is divergent and indeed

$\frac{\partial\tilde{\lambda}}{\partial\sigma} \approx 1.3728$. On the other hand, the Newton framework is able to converge to the optimal solution in a small number of steps.

6. Conclusions. We have developed an interpolation-based rational Krylov algorithm that iteratively corrects interpolation locations until first-order $\mathcal{H}_2$ optimality conditions are satisfied. The resulting method proves numerically effective and well suited for large-scale problems. A new derivation of the interpolation-based necessary conditions is presented and shown to be equivalent to two other common frameworks for $\mathcal{H}_2$ optimality.

Appendix.

LEMMA A.1. *For any stable matrix* $\mathbf{M}$*,*

$$P.V. \int_{-\infty}^{+\infty} (\imath\omega - \mathbf{M})^{-1}\, d\omega \stackrel{def}{=} \lim_{L\to\infty} \int_{-L}^{L} (\imath\omega - \mathbf{M})^{-1}\, d\omega = \pi\mathbf{I}.$$

Proof. Observe that for any $L > 0$,

$$\int_{-L}^{L} (\imath\omega - \mathbf{M})^{-1}\, d\omega = \int_{-L}^{L} (-\imath\omega - \mathbf{M})(\omega^2 + \mathbf{M}^2)^{-1}\, d\omega = \int_{-L}^{L} (-\mathbf{M})(\omega^2\mathbf{I} + \mathbf{M}^2)^{-1}\, d\omega.$$

Fix a contour Γ contained in the open left half plane so that the interior of Γ contains all eigenvalues of $\mathbf{M}$. Then

$$-\mathbf{M}(\omega^2\mathbf{I} + \mathbf{M}^2)^{-1} = \frac{1}{2\pi\imath} \int_{\Gamma} \frac{-z}{\omega^2 + z^2} (z\mathbf{I} - \mathbf{M})^{-1}\, dz.$$

For any fixed value z in the left half plane,

$$\text{P.V.} \int_{-\infty}^{+\infty} \frac{d\omega}{\imath\omega - z} = \lim_{L\to\infty} \int_{-L}^{L} \frac{-z}{\omega^2 + z^2}\, d\omega = \pi.$$

Thus,

$$\begin{aligned}\lim_{L\to\infty} \int_{-L}^{L} (-\mathbf{M})(\omega^2\mathbf{I} + \mathbf{M}^2)^{-1}\, d\omega &= \frac{1}{2\pi\imath} \int_{\Gamma} \lim_{L\to\infty} \left(\int_{-L}^{L} \frac{-z}{\omega^2 + z^2}\, d\omega \right) (z\mathbf{I} - \mathbf{M})^{-1}\, dz \\ &= \frac{1}{2\pi\imath} \int_{\Gamma} \pi\, (z\mathbf{I} - \mathbf{M})^{-1}\, dz = \pi\mathbf{I}. \qquad \square\end{aligned}$$

REFERENCES

[1] A.C. ANTOULAS, *Recursive modeling of discrete-time time series*, in Linear Algebra for Control Theory, P. Van Dooren and B. W. Wyman, eds., IMA Vol. Math. Appl. 62, Springer-Verlag, New York, 1993, pp. 1–20.

[2] A.C. ANTOULAS, *Approximation of Large-Scale Dynamical Systems*, Adv. Des. Control 6, SIAM, Philadelphia, 2005.

[3] A.C. ANTOULAS AND J.C. WILLEMS, *A behavioral approach to linear exact modeling*, IEEE Trans. Automat. Control, 38 (1993), pp. 1776–1802.

[4] A.C. ANTOULAS, D.C. SORENSEN, AND S. GUGERCIN, *A survey of model reduction methods for large scale systems*, in Structured Matrices in Mathematics, Computer Science, and Engineering, I (Boulder, CO, 1999), Contemp. Math. 280, AMS, Providence, RI, 2001, pp. 193–219.

[5] J. BADIA, P. BENNER, R. MAYO, E.S. QUINTANA-ORTÍ, G. QUINTANA-ORTÍ, AND J. SAAK, *Parallel order reduction via balanced truncation for optimal cooling of steel profiles*, in Proceedings of the 11th International European Conference on Parallel Processing, EuroPar 2005, Lisbon, J. C. Cunha and P. D. Medeiros, eds., Lecture Notes in Comput. Sci. 3648, Springer-Verlag, Berlin, 2005, pp. 857–866.

[6] L. Baratchart, M. Cardelli, and M. Olivi, *Identification and rational ℓ_2 approximation: A gradient algorithm*, Automat., 27 (1991), pp. 413–418.
[7] P. Benner, *Solving large-scale control problems*, IEEE Control Systems Mag., 24 (2004), pp. 44–59.
[8] P. Benner and J. Saak, *Efficient numerical solution of the LQR-problem for the heat equation*, Proc. Appl. Math. Mech., 4 (2004), pp. 648–649.
[9] A.E. Bryson and A. Carrier, *Second-order algorithm for optimal model order reduction*, J. Guidance Control Dynam., 13 (1990), pp. 887–892.
[10] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, and D. Wilczek, *H_2-optimal Model Reduction for Large Scale Discrete Dynamical MIMO Systems*, ZeTeM Technical Report 07-04, University of Bremen, 2007; available online from http://www.math.uni-bremen.de/zetem/reports/reports-liste.html#reports2007.
[11] C. De Villemagne and R. Skelton, *Model reduction using a projection formulation*, Internat. J. Control, 40 (1987), pp. 2141–2169.
[12] P. Feldman and R.W. Freund, *Efficient linear circuit analysis by Padé approximation via a Lanczos method*, IEEE Trans. Computer-Aided Design, 14 (1995), pp. 639–649.
[13] P. Fulcheri and M. Olivi, *Matrix rational H_2 approximation: A gradient algorithm based on Schur analysis*, SIAM J. Control Optim., 36 (1998), pp. 2103–2127.
[14] D. Gaier, *Lectures on Complex Approximation*, Birkhäuser, Boston, 1987.
[15] K. Gallivan, E. Grimme, and P. Van Dooren, *A rational Lanczos algorithm for model reduction*, Numer. Algorithms, 2 (1996), pp. 33–63.
[16] K. Gallivan, P. Van Dooren, and E. Grimme, *On some recent developments in projection-based model reduction*, in ENUMATH 97 (Heidelberg), World Scientific, River Edge, NJ, 1998, pp. 98–113.
[17] E.J. Grimme, *Krylov Projection Methods for Model Reduction*, Ph.D. thesis, University of Illinois, Urbana-Champaign, Urbana, IL, 1997.
[18] S. Gugercin, *Projection Methods for Model Reduction of Large-Scale Dynamical Systems*, Ph.D. thesis, Rice University, Houston, TX, 2002.
[19] S. Gugercin and A.C. Antoulas, *A comparative study of 7 model reduction algorithms*, in Proceedings of the 39th IEEE Conference on Decision and Control, Sydney, 2000, pp. 2367–2372.
[20] S. Gugercin and A.C. Antoulas, *An $\mathcal{H}_2$ error expression for the Lanczos procedure*, in Proceedings of the 42nd IEEE Conference on Decision and Control, 2003, pp. 1869–1872.
[21] Y. Halevi, *Frequency weighted model reduction via optimal projection*, in Proceedings of the 29th IEEE Conference on Decision and Control, 1990, pp. 2906–2911.
[22] D.C. Hyland and D.S. Bernstein, *The optimal projection equations for model reduction and the relationships among the methods of Wilson, Skelton, and Moore*, IEEE Trans. Automat. Control, 30 (1985), pp. 1201–1211.
[23] J.G. Korvink and E.B. Rudyni, *Oberwolfach benchmark collection*, in Dimension Reduction of Large-Scale Systems, P. Benner, G. Golub, V. Mehrmann, and D. Sorensen, eds., Lecture Notes in Comput. Sci. Engrg. 45, Springer-Verlag, New York, 2005.
[24] D. Kubalinska, A. Bunse-Gerstner, G. Vossen, and D. Wilczek, *H_2-optimal interpolation based model reduction for large-scale systems*, in Proceedings of the 16th International Conference on System Science, Wroclaw, Poland, 2007.
[25] A. Lepschy, G.A. Mian, G. Pinato, and U. Viaro, *Rational L_2 approximation: A non-gradient algorithm*, in Proceedings of the 30th IEEE Conference on Decision and Control, 1991, pp. 2321–2323.
[26] L. Meier and D.G. Luenberger, *Approximation of linear constant systems*, IEEE Trans. Automat. Control, 12 (1967), pp. 585–588.
[27] B.C. Moore, *Principal component analysis in linear system: Controllability, observability and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17–32.
[28] C.T. Mullis and R.A. Roberts, *Synthesis of minimum roundoff noise fixed point digital filters*, IEEE Trans. Circuits Systems, CAS-23 (1976), pp. 551–562.
[29] T. Penzl, *Algorithms for model reduction of large dynamical systems*, Linear Algebra Appl., 415 (2006), pp. 322–343.
[30] L.T. Pillage and R.A. Rohrer, *Asymptotic waveform evaluation for timing analysis*, IEEE Trans. Computer-Aided Design, 9 (1990), pp. 352–366.
[31] V. Raghavan, R.A. Rohrer, L.T. Pillage, J.Y. Lee, J.E. Bracken, and M.M. Alaybeyi, *AWE inspired*, in Proceedings of the IEEE Custom Integrated Circuits Conference, 1993, pp. 18.1.1–18.1.8.
[32] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Ungar, New York, 1955.
[33] A. Ruhe, *Rational Krylov algorithms for nonsymmetric eigenvalue problems* II: *Matrix pairs*, Linear Algebra Appl., 197 (1994), pp. 283–295.

[34] J.T. Spanos, M.H. Milman, and D.L. Mingori, *A new algorithm for $\mathcal{L}_2$ optimal model reduction*, Automat., 28 (1992), pp. 897–909.
[35] P. Van Dooren, K.A. Gallivan, and P.-A. Absil, *H_2 optimal model reduction of MIMO systems*, Appl. Math. Lett., to appear; doi:10.1016/j.aml.2007.09.015.
[36] D.A. Wilson, *Optimum solution of model reduction problem*, Proc. IEE-D, 117 (1970), pp. 1161–1165.
[37] W.-Y. Yan and J. Lam, *An approximate approach to $\mathcal{H}_2$ optimal model reduction*, IEEE Trans. Automat. Control, 44 (1999), pp. 1341–1358.
[38] A. Yousouff and R.E. Skelton, *Covariance equivalent realizations with applications to model reduction of large-scale systems*, in Control and Dynamic Systems, Vol. 22, C. T. Leondes, ed., Academic Press, New York, 1985, pp. 273–348.
[39] A. Yousouff, D.A. Wagie, and R.E. Skelton, *Linear system approximation via covariance equivalent realizations*, J. Math. Anal. Appl., 196 (1985), pp. 91–115.
[40] D. Zigic, L. Watson, and C. Beattie, *Contragredient transformations applied to the optimal projection equations*, Linear Algebra Appl., 188/189 (1993), pp. 665–676.