

\textbf{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.

\textbf{Key words.} Model Reduction, Rational Krylov, $\mathcal{H}_2$ Approximation

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

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 is 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 (Micro Electro Mechanical System) devices. For an overview of model reduction for large-scale dynamical systems we refer to the book [4]. 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:

\begin{equation}
G: \begin{cases}
\dot{x}(t) = Ax(t) + bu(t) \\
y(t) = c^T x(t)
\end{cases}
\text{ or } G(s) = c^T (sI - A)^{-1} b,
\end{equation}

where $A \in \mathbb{R}^{n \times n}$, $b, c \in \mathbb{R}^n$; $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, $y(t) \in \mathbb{R}$, are the state, input, and output, respectively, of the system. $G(s)$ is the transfer function of the system: if $\hat{u}(s)$ and $\hat{y}(s)$ denote the Laplace transforms of the input and output $u(t)$ and $y(t)$,
The model reduction process will yield another system

\[
G_r: \begin{cases}
\dot{x}_r(t) = A_r x_r(t) + b_r u(t) \\
y_r(t) = c_r^T x_r(t)
\end{cases} \quad \text{or} \quad G_r(s) = c_r^T (sI - A_r)^{-1} b_r \quad (1.2)
\]

having (much) smaller dimension \( r \ll n \), with \( A_r \in \mathbb{R}^{r \times r} \) and \( b_r, c_r \in \mathbb{R}^r \).

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, \( \int_0^\infty |u(t)|^2 \, dt \leq 1 \). Observe first that \( \tilde{y}(s) - \tilde{y}_r(s) = [G(s) - G_r(s)] \tilde{u}(s) \) and then:

\[
\max_{t>0} |y(t) - y_r(t)| = \max_{t>0} \left| \int_0^\infty \frac{1}{2\pi} \int_{-\infty}^\infty \left( \tilde{y}( \omega ) - \tilde{y}_r( \omega ) \right) e^{i \omega t} \, d\omega \right|
\]

\[
\leq \frac{1}{2\pi} \int_{-\infty}^\infty \left| \tilde{y}( \omega ) - \tilde{y}_r( \omega ) \right| \, d\omega = \frac{1}{2\pi} \int_{-\infty}^\infty |G( \omega ) - G_r( \omega )| \left| \tilde{u}( \omega ) \right| \, d\omega
\]

\[
\leq \left( \frac{1}{2\pi} \int_{-\infty}^\infty |G( \omega ) - G_r( \omega )|^2 \, d\omega \right)^{1/2} \left( \frac{1}{2\pi} \int_{-\infty}^\infty |\tilde{u}( \omega )|^2 \, d\omega \right)^{1/2}
\]

\[
\leq \left( \frac{1}{2\pi} \int_{-\infty}^\infty |G( \omega ) - G_r( \omega )|^2 \, d\omega \right)^{1/2} \left( \int_0^\infty |u(t)|^2 \, dt \right)^{1/2}
\]

\[
\leq \left( \frac{1}{2\pi} \int_{-\infty}^\infty |G( \omega ) - G_r( \omega )|^2 \, d\omega \right)^{1/2} \overset{\text{def}}{=} \|G - G_r\|_{\mathcal{H}_2}
\]

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 that exist in \( G \) (such as stability) exist also in \( G_r \); and

iii) the computation of \( G_r \) (i.e., the computation of \( A_r, b_r, \) and \( c_r \)) is 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, 33, 31, 9, 20, 25, 22, 32, 24, 12] and 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 optimality. Most methods that identify reduced order models that satisfy such first-order necessary conditions require dense matrix operations, e.g., solution of a sequence of Lyapunov equations, which can rapidly become intractable as the dimension increases. Such methods are unsuitable even for medium scale problems. Here, we propose an iterative algorithm which is based on computationally effective use of rational Krylov subspaces. The proposed method is suitable for systems whose dimension \( n \) is of the order of many thousands of state variables.

The rest of the paper is organized as follows. In Section 2, we review the moment matching problem for model reduction, its connection with rational Krylov methods,
and basic features of the H₂ norm and inner product. Section 3 describes the main results of the paper which comprise a new structured orthogonality condition necessary for optimality and its equivalence to two other frameworks used in evaluating H₂ optimality, thus showing these frameworks to be equivalent as well. The development of an algorithmic approach to solve the H₂ model reduction problem is described in Section 4 and 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 points σ_k in the complex plane. For our purposes, simple Hermite interpolation suffices so our problem is to find A_r, b_r, and 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

\[ c^T(\sigma_k I - A)^{-1} b = c^T_r(\sigma_k I - A_r)^{-1} b_r \quad \text{and} \quad c^T(\sigma_k I - A)^{-2} b = c^T_r(\sigma_k I - A_r)^{-2} b_r \]

for \( k = 1, \ldots, r \). The quantity \( c^T(\sigma_k I - A)^{-1(j+1)} b \) is called the \( j \)th moment of \( G(s) \) at \( \sigma_k \). Moment matching for finite \( \sigma \in C \), becomes rational interpolation, see for example [2]. 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 \( \mathcal{V}_r \cap \mathcal{W}_r = \{0\} \). Then for any input \( u(t) \) the reduced-order output \( y_r(t) \) is defined by

\[
\text{Find } v(t) \in \mathcal{V}_r \text{ such that } \dot{v}(t) - Av(t) - bu(t) \perp \mathcal{W}_r \text{ for all } t \quad (2.1)
\]

and then \( y_r(t) \triangleq c^T v(t) \)

Denote by \( \text{Ran}(M) \), the range of a matrix \( M \). If \( \mathcal{V}_r \in \mathbb{R}^{n \times r} \) and \( \mathcal{W}_r \in \mathbb{R}^{n \times r} \) are matrices defined so that \( \mathcal{V}_r = \text{Ran}(\mathcal{V}_r) \) and \( \mathcal{W}_r = \text{Ran}(\mathcal{W}_r) \) then the assumption that \( \mathcal{V}_r \cap \mathcal{W}_r = \{0\} \) is equivalent to \( \mathcal{W}_r^T \mathcal{V}_r \) being nonsingular and the Galerkin approximation (2.1) can be interpreted as \( v(t) = \mathcal{V}_r x_r(t) \) with \( x_r(t) \in \mathbb{R}^r \) for each \( t \) and

\[ \mathcal{W}_r^T (\mathcal{V}_r \dot{x}_r(t) - A \mathcal{V}_r x_r(t) - b u(t)) = 0 \]

leading then to the reduced order model (1.2) with

\[
A_r = (\mathcal{W}_r^T \mathcal{V}_r)^{-1} \mathcal{W}_r^T A \mathcal{V}_r, \quad b_r = (\mathcal{W}_r^T \mathcal{V}_r)^{-1} \mathcal{W}_r^T b, \quad \text{and} \quad c_r^T = c^T \mathcal{V}_r. \quad (2.2)
\]

Evidently the choice of \( \mathcal{V}_r \) and \( \mathcal{W}_r \) will determine the quality of the reduced order model.

Rational interpolation by projection was first proposed by Skelton et. al. in [10, 34, 35]. Grimme [16] showed how one can obtain the required projection using the rational Krylov method of Ruhe [30]. 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 [11].
LEMMA 2.1. Suppose $\sigma \in \mathbb{C}$ is not an eigenvalue, neither of $A$ nor of $A_r$.

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

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

If both $(\sigma I - A)^{-1}b \in \mathcal{V}_r$ and $(\sigma I - A^T)^{-1}c \in \mathcal{W}_r$

then $G_r(\sigma) = G(\sigma)$ and $G'_r(\sigma) = G'(\sigma)$. 
(2.5)

Proof. Define $N_r(z) = V_r(zI - A_r)^{-1}(W_r^TV_r)^{-1}W_r^T(zI - A)$ and $\tilde{N}_r(z) = (zI - A)N_r(z)(zI - A)^{-1}$. Both $N_r(z)$ and $\tilde{N}_r(z)$ are analytic matrix-valued functions in a neighborhood of $z = \sigma$. One may directly verify that $N_r^2(z) = N_r(z)$ and $\tilde{N}_r^2(z) = \tilde{N}_r(z)$ and that $\mathcal{V}_r = \text{Ran } N_r(z) = \text{Ker } (I - N_r(z))$ and $\mathcal{W}_r^\perp = \text{Ker } \tilde{N}_r(z) = \text{Ran } (I - \tilde{N}_r(z))$ for all $z$ in a neighborhood of $\sigma$. Then

$$G(z) - G_r(z) = \left[(zI - A^T)^{-1}c\right]^T(I - \tilde{N}_r(z)) (zI - A) (I - N_r(z)) (zI - A)^{-1}b$$

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

$$G(\sigma + \varepsilon) - G_r(\sigma + \varepsilon) = O(\varepsilon^2)$$

which gives (2.5) as a consequence. \(\blacksquare\)

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

$$\text{Ran } (\mathcal{V}_r) = \text{span } \{(\sigma_1 I - A)^{-1}b, \cdots, (\sigma_r I - A)^{-1}b\}$$

and

$$\text{Ran } (\mathcal{W}_r) = \text{span } \{(\sigma_1 I - A^T)^{-1}c, \cdots, (\sigma_r I - A^T)^{-1}c\}.$$ 

Then $\mathcal{V}_r$ and $\mathcal{W}_r$ can be chosen to be real matrices and the reduced order system $G_r$ defined by $A_r = (W_r^TV_r)^{-1}W_r^TAV_r$, $b_r = (W_r^TV_r)^{-1}W_r^Tb$, $c_r^T = c^T V_r$, is itself real and matches the first two moments of $G(s)$ at each of the interpolation points $\sigma_k$, i.e., $G(\sigma_k) = G_r(\sigma_k)$ and $G'(\sigma_k) = G'_r(\sigma_k)$, for $k = 1, \cdots, r$.

For Krylov-based model reduction, one chooses interpolation points and then constructs $\mathcal{V}_r$ and $\mathcal{W}_r$ as above. 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 [4]), Krylov-based model reduction requires only matrix-vector multiplications, some sparse linear solvers, and can be iteratively implemented; hence it is computationally effective; for details, see also [14, 15].

2.2. 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, $Re(z) > 0$, and such that for each fixed $Re(z) = x > 0$, $g(x + iy)$ is square integrable as a function of $y \in (-\infty, \infty)$ in such a way that

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

$\mathcal{H}_2$ 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

$$\langle G, H \rangle_{\mathcal{H}_2} \overset{\text{def}}{=} \frac{1}{2\pi} \int_{-\infty}^{\infty} G(\omega) H(\omega) d\omega = \int_{-\infty}^{\infty} G(-i\omega) H(\omega) d\omega \tag{2.6}$$

with a norm defined as

$$\|G\|_{\mathcal{H}_2} \overset{\text{def}}{=} \left( \frac{1}{2\pi} \int_{-\infty}^{\infty} |G(\omega)|^2 d\omega \right)^{1/2} \tag{2.7}$$

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 $(G, H)_{\mathcal{H}_2}$ must be real.

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

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

$$G(s) = c^T (sI - A)^{-1} b \quad \text{and} \quad H(s) = \tilde{c}^T (sI - B)^{-1} \tilde{b}.$$

Then the inner product $\langle G, H \rangle_{\mathcal{H}_2}$ is associated with solutions to Sylvester equations as

If $P$ solves $AP + PB^T + bb^T = 0$ then $\langle G, H \rangle_{\mathcal{H}_2} = c^T P \tilde{c}$ \tag{2.8}

If $Q$ solves $QA + B^TQ + \tilde{c} \tilde{c}^T = 0$ then $\langle G, H \rangle_{\mathcal{H}_2} = \tilde{b}^T Qb$ \tag{2.9}

If $R$ solves $AR + RB + b \tilde{c}^T = 0$ then $\langle G, H \rangle_{\mathcal{H}_2} = c^T R\tilde{b}$ \tag{2.10}

Note that if $A = B, b = \tilde{b}$, and $c = \tilde{c}$ then $P$ is the “reachability gramian” of $G(s)$, $Q$ is the “observability gramian” of $G(s)$, and $R$ is the “cross gramian” of $G(s)$; and

$$\|G\|^2_{\mathcal{H}_2} = c^T P c = b^T Q b = c^T R \tilde{b} \tag{2.11}$$

Gramians play a prominent role in the analysis of linear dynamical systems; refer to [4] for more information.

**Proof.** We detail the proof of (2.8); proofs of (2.9) and (2.10) are similar. Since $A$ and $B$ are stable, the solution, $P$, to the Sylvester equation of (2.8) exists and is unique. For any $\omega \in \mathbb{R}$, rearrange this equation to obtain in sequence:

$$(\omega I - A) P + P (\omega I - B^T)^{-1} - b \tilde{b}^T = 0$$

$$(\omega I - A)^{-1} P + P (\omega I - B^T)^{-1} = (\omega I - A)^{-1} b \tilde{b}^T (\omega I - B^T)^{-1}$$

$$c^T (\omega I - A)^{-1} P \tilde{c} + c^T P (\omega I - B^T)^{-1} \tilde{c} = G(-\omega) H(\omega)$$

and finally,

$$c^T \left( \int_{-L}^{L} (\omega I - A)^{-1} d\omega \right) P \tilde{c} + c^T P \left( \int_{-L}^{L} (\omega I - B^T)^{-1} d\omega \right) \tilde{c}$$

$$= \int_{-L}^{L} G(-\omega) H(\omega) d\omega$$
Taking $L \to \infty$ and using Lemma A.1 in the Appendix, leads to
\[
\int_{-\infty}^{\infty} G(-\omega)H(\omega) \, d\omega = c^T \left( \text{P.V.} \int_{-\infty}^{\infty} (-\omega I - A)^{-1} \, d\omega \right) P\tilde{c} \\
+ c^T P \left( \text{P.V.} \int_{-\infty}^{\infty} (\omega I - B^T)^{-1} \, d\omega \right) \tilde{c} \\
= 2\pi \, c^T P\tilde{c}
\]

Recently, Antoulas [4] 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 (2.11). 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 $\lambda$, denote the residue of $f(s)$ at $\lambda$ by $\text{res}[f(s), \lambda]$. Thus, if $\lambda$ is a simple pole of $f(s)$ then $\text{res}[f(s), \lambda] = \lim_{s \to \lambda} (s-\lambda)f(s)$ and if $\lambda$ is a double pole of $f(s)$ then $\text{res}[f(s), \lambda] = \lim_{s \to \lambda} \frac{d}{ds} [(s-\lambda)^2 f(s)]$.

**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 halfplane. Then
\[
\langle G, H \rangle_{\mathcal{H}_2} = \sum_{k=1}^{m} \text{res}[G(-s)H(s), \mu_k] = \sum_{k=1}^{n} \text{res}[H(-s)G(s), \lambda_k] \quad (2.12)
\]

In particular,

- if $\mu_k$ is a simple pole of $H(s)$ then
  \[\text{res}[G(-s)H(s), \mu_k] = G(-\mu_k)\text{res}[H(s), \mu_k];\]

- if $\mu_k$ is a double pole of $H(s)$ then
  \[\text{res}[G(-s)H(s), \mu_k] = G(-\mu_k) \text{res}[H(s), \mu_k] - G'(-\mu_k) \cdot h_0(\mu_k)\]
  where $h_0(\mu_k) = \lim_{s \to \mu_k} ((s-\mu_k)^2 H(s))$.

**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 halfplane:
\[
\Gamma_R = \{ z = j\omega \text{ with } \omega \in [-R, R] \} \cup \left\{ z = Re^{j\theta} \text{ with } \theta \in \left[ \frac{\pi}{2}, \frac{3\pi}{2} \right] \right\}.
\]

$\Gamma_R$ bounds a region that for sufficiently large $R$ contains all the system poles of $H(s)$ and so, by the Residue Theorem
\[
\langle G, H \rangle_{\mathcal{H}_2} = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(-\omega)H(\omega) \, d\omega \\
= \lim_{R \to \infty} \frac{1}{2\pi j} \int_{\Gamma_R} G(-s)H(s) \, ds \\
= \sum_{k=1}^{m} \text{res}[G(-s)H(s), \mu_k]
\]
Evidently, if $\mu_k$ is a simple pole for $H(s)$, it is also a simple pole for $G(-s)H(s)$ and
\[
\text{res}[G(-s)H(s), \mu_k] = \lim_{s \to \mu_k} (s - \mu_k)G(-s)H(s) = G(-\mu_k) \lim_{s \to \mu_k} (s - \mu_k)H(s).
\]
If $\mu_k$ is a double pole for $H(s)$, then it is also a double pole for $G(-s)H(s)$ and
\[
\text{res}[G(-s)H(s), \mu_k] = \lim_{s \to \mu_k} \frac{d}{ds} (s - \mu_k)^2 G(-s)H(s) = \lim_{s \to \mu_k} \frac{d}{ds} (s - \mu_k)^2 H(s) - G'(-s)(s - \mu_k)H(s) = G(-\mu_k) \lim_{s \to \mu_k} \frac{d}{ds} (s - \mu_k)^2 H(s) - G'(-\mu_k) \lim_{s \to \mu_k} (s - \mu_k)^2 H(s).
\]

**Lemma 2.4** immediately yields the expression for $\|G\|_{\mathcal{H}_2}$ given by Antoulas [4] based on poles and residues of the transfer function $G(s)$.

**Corollary 2.5** ([4, p. 145]). If $G(s)$ has simple poles at $\lambda_1, \lambda_2, \ldots, \lambda_n$ then
\[
\|G\|_{\mathcal{H}_2} = \left( \sum_{k=1}^{n} \text{res}[G(s), \lambda_k]G(-\lambda_k) \right)^{1/2}
\]

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 [25] and Lyapunov-based [32, 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 $r^{th}$ order dynamical system approximating $G$ with respect to the $\mathcal{H}_2$ norm:
\[
\|G - G_r\|_{\mathcal{H}_2} = \min_{\dim(G_r) = r} \|G - G_r\|_{\mathcal{H}_2}.
\]

Many researchers have worked on the problem (3.1), the optimal $\mathcal{H}_2$ model reduction problem. See [33, 31, 9, 20, 25, 22, 32, 24] and references therein.

3.1. **Structured orthogonality optimality conditions.** The set of all stable $r^{th}$ order dynamical systems do not constitute a subspace of $\mathcal{H}_2$, so the best $r^{th}$ order $\mathcal{H}_2$ approximation is not so easy to characterize, the Hilbert space structure of $\mathcal{H}_2$ notwithstanding. This observation does suggest a 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 $\mathcal{M}(\mu)$ to be the set of all proper rational functions that have simple poles exactly at $\mu_1, \mu_2, \ldots, \mu_r$. Then
\( H \in \mathcal{M}(\mu) \) implies that \( H \) is the transfer function of a stable dynamical system with \( \dim(H) = r \).

\( \mathcal{M}(\mu) \) is an \( r - 1 \) dimensional subspace of \( \mathcal{H}_2 \).

\( G_r \in \mathcal{M}(\mu) \) solves

\[
\|G - G_r\|_{\mathcal{H}_2} = \min_{\tilde{G}_r \in \mathcal{M}(\mu)} \|G - \tilde{G}_r\|_{\mathcal{H}_2}. \tag{3.2}
\]

if and only if

\[
(G - G_r, H)_{\mathcal{H}_2} = 0 \quad \text{for all } H \in \mathcal{M}(\mu). \tag{3.3}
\]

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

Proof. The key observation is that \( \mathcal{M}(\mu) \) 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. [29]).

One consequence of Theorem 3.1 is that if \( G_r(s) \) interpolates \( G(s) \) at the mirror images of the poles of \( G_r(s) \), 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 \( \{a_i\} \). The set of stable \( r^{th} \) order dynamical systems is not convex and so, the 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,

\[
\|G - G_r\|_{\mathcal{H}_2} \leq \|G - \tilde{G}_r^{(c)}\|_{\mathcal{H}_2}, \tag{3.4}
\]

for all stable dynamical systems, \( \tilde{G}_r^{(c)} \) with \( \dim(\tilde{G}_r^{(c)}) = r \) and \( \|G_r - \tilde{G}_r^{(c)}\|_{\mathcal{H}_2} \leq C \varepsilon \), \( C \) is a constant that may depend on the particular family \( \tilde{G}_r^{(c)} \) 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.

We derive necessary conditions that must occur for the reduced order system \( G_r \), to be a local minimizer that amount to structured orthogonality conditions:

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

\[
(G - G_r, G_r \cdot H_1 + H_2)_{\mathcal{H}_2} = 0 \tag{3.5}
\]

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

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

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

\[
\|G - G_r\|_{\mathcal{H}_2}^2 \leq \|G - \tilde{G}_r^{(c)}\|_{\mathcal{H}_2}^2
\]

\[
\leq \|(G - G_r) + (G_r - \tilde{G}_r^{(c)})\|_{\mathcal{H}_2}^2
\]

\[
\leq \|G - G_r\|_{\mathcal{H}_2}^2 + 2 \langle G - G_r, G_r - \tilde{G}_r^{(c)} \rangle_{\mathcal{H}_2} + \|G_r - \tilde{G}_r^{(c)}\|_{\mathcal{H}_2}^2
\]
This in turn implies for all $\varepsilon > 0$ sufficiently small,

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

We will consider a few different choices for $\tilde{G}_r^{(\varepsilon)}$. 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 $\mu_i = \alpha_i + i\beta_i$. Any real rational function having the same poles as $G_r$ 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 $\gamma_i$, $\rho_i$, and $\tau_i$. Now suppose that $\mu$ is a real pole for $G_r$ and that

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

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

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

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 that

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

which leads to

$$G_r(s) - \tilde{G}_r^{(\varepsilon)}(s) = \mp \varepsilon \frac{G_r(s)}{s - \mu} + \mathcal{O}(\varepsilon^2)$$

and

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

This implies as before as $\varepsilon \to 0$, $0 < \left| \left\langle G - G_r, \frac{G_r(s)}{s - \mu} \right\rangle_{\mathcal{H}_2} \right| \leq C\varepsilon$ for some constant $C$, leading again to a contradiction.

Now suppose that $\mu = \alpha + i\beta$ 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

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

\[
\langle G - G_r, \frac{G_r(s)}{(s - \alpha)^2 + \beta^2} \rangle_{\mathcal{H}_2} = 0,
\]

and consider

\[
\tilde{G}_r^{(\varepsilon)}(s) = \frac{p_{r-1}(s)}{[(s - \alpha)^2 + \beta^2 (\pm \varepsilon)] q_{r-2}(s)};
\]

\[
\langle G - G_r, \frac{(s - \alpha) G_r(s)}{(s - \alpha)^2 + \beta^2} \rangle_{\mathcal{H}_2} = 0,
\]

and consider

\[
\tilde{G}_r^{(\varepsilon)}(s) = \frac{p_{r-1}(s)}{[(s - \alpha - (\pm \varepsilon))^2 + \beta^2] q_{r-2}(s)}.
\]

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

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

\[\blacksquare\]

Theorem 3.2 describes a new framework for the necessary conditions of 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 \[25\] and Lyapunov-based conditions \[32, 22\] 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 (2.12); the Lyapunov framework uses the Sylvester equation formulation of the \( \mathcal{H}_2 \) norm as in (2.8).

### 3.2. Interpolation-based optimality conditions.

Corollary 2.5 immediately yields the following 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 \( \lambda_i \) and \( \bar{\lambda}_i \) be the poles of \( G(s) \) and \( G_r(s) \), respectively, and suppose that the poles of \( G_r(s) \) are distinct. Let \( \phi_i \) and \( \bar{\phi}_i \) denote the residues of the transfer functions \( G(s) \) and \( G_r(s) \) at their poles \( \lambda_i \) and \( \bar{\lambda}_i \), respectively: \( \phi_i = \text{res}[G(s), \lambda_i] \) \( i = 1, \ldots, n \)
and \( \phi_j = \text{res}[G_r(s), \tilde{\lambda}_j] \) for \( j = 1, \ldots, r \). The \( H_2 \) norm of the error system is given by

\[
\|G - G_r\|_{H_2}^2 = \sum_{i=1}^{n} \text{res}\{(G(-s) - G_r(-s)) (G(s) - G_r(s)) , \lambda_i\}
\]

\[
+ \sum_{j=1}^{r} \text{res}\{(G(-s) - G_r(-s)) (G(s) - G_r(s)) , \tilde{\lambda}_j\}
\]

\[
= \sum_{i=1}^{n} \phi_i \left( G(-\lambda_i) - G_r(-\lambda_i) \right) - \sum_{j=1}^{r} \phi_j \left( G(-\tilde{\lambda}_j) - G_r(-\tilde{\lambda}_j) \right) \tag{3.10}
\]

The \( H_2 \) error expression (3.10) generalizes a result of [19, 18], proven only for model reduction by the Lanczos procedure, to the most general setting, valid for any reduced order model regardless of the underlying reduction technique.

Proposition 3.3 has the system-theoretic interpretation that the \( H_2 \) error is due to mismatch of the transfer functions \( G(s) \) and \( G_r(s) \) at the mirror images of the full-order poles \( \lambda_i \) and the reduced order poles \( \tilde{\lambda}_i \). Hence, this expression reveals that for a good \( H_2 \) performance, \( G_r(s) \) should approximate \( G(s) \) well at \( -\lambda_i \) and \( -\tilde{\lambda}_j \). Note that \( \lambda_i \) is not known \textit{a priori}. Therefore, to minimize the \( H_2 \) error, Gugercin and Antoulas [19] proposed choosing \( \sigma_i = -\lambda_i(A) \) where \( \lambda_i(A) \) are the poles with big residuals \( \phi_i \). They have illustrated that this selection of interpolation points works quite well, see [18, 19]. However, as (3.10) illustrates, there is a second part of the \( H_2 \) error due to the mismatch at \( -\tilde{\lambda}_j \). Indeed, as we will show below, interpolation at \( -\tilde{\lambda}_i \) is more important for model reduction and is a necessary condition for optimal \( H_2 \) model reduction, i.e. \( \sigma_i = -\tilde{\lambda}_i \) is the optimal shift selection for the \( H_2 \) model reduction.

**Theorem 3.4.** Given a stable SISO system \( G(s) = c^T(sI - A)^{-1}b \), let \( G_r(s) = c^T_r(sI - A_r)^{-1}b \) be a local minimizer of dimension \( r \) for the optimal \( H_2 \) model reduction problem (3.1) and suppose that \( G_r(s) \) has simple poles at \( \tilde{\lambda}_i \), \( i = 1, \ldots, r \). Then \( G_r(s) \) interpolates both \( G(s) \) and its first derivative at \( -\tilde{\lambda}_i \), \( i = 1, \ldots, r \):

\[
G_r(-\tilde{\lambda}_i) = G(-\tilde{\lambda}_i) \quad \text{and} \quad G'_r(-\tilde{\lambda}_i) = G'(-\tilde{\lambda}_i), \quad \text{for} \; i = 1, \ldots, r. \tag{3.11}
\]

**Proof.** From (3.5), consider the first the case \( H_1 = 0 \) and \( H_2 \) is an arbitrary transfer function with simple poles at \( \tilde{\lambda}_i \), \( i = 1, \ldots, r \). Denote \( \phi_i = \text{res}[H_2(s), \tilde{\lambda}_i] \). Then ( ) leads to

\[
\langle G - G_r, H_2 \rangle_{H_2} = \sum_{i=1}^{r} \text{res}\{(G(-s) - G_r(-s)) H_2(s) , \tilde{\lambda}_i\}
\]

\[
= \sum_{i=1}^{r} \phi_i \left( G(-\tilde{\lambda}_i) - G_r(-\tilde{\lambda}_i) \right) = 0.
\]

Since this is true for arbitrary choices of \( \phi_i \), we have \( G(-\tilde{\lambda}_i) = G_r(-\tilde{\lambda}_i) \). Now consider the case \( H_2 = 0 \) and \( H_1 \) is an arbitrary transfer function with simple poles at \( \tilde{\lambda}_i \), \( i = 1, \ldots, r \). Then \( G_r(s) H_1(s) \) has double poles at \( \tilde{\lambda}_i \), \( i = 1, \ldots, r \), and since
\[ G(-\tilde{\lambda}_i) = G_r(-\tilde{\lambda}_i) \] we have
\[
\langle G - G_r, G_r \cdot H_1 \rangle_{\mathcal{H}_2} = \sum_{i=1}^{r} \text{res}[G(-s) - G_r(-s)] G_r(s) H_1(s), \tilde{\lambda}_i \in \mathbb{C}
\]
\[
= - \sum_{i=1}^{r} \hat{\phi}_i \text{res}[G_r, \tilde{\lambda}_i] \left( G'(-\tilde{\lambda}_i) - G_r'(-\tilde{\lambda}_i) \right) = 0.
\]

where we have calculated
\[
\lim_{s \to \tilde{\lambda}_i} \left( (s - \tilde{\lambda}_i)^2 G_r(s) \cdot H_1(s) \right) = \text{res}[H_1(s), \tilde{\lambda}_i] \cdot \text{res}[G_r(s), \tilde{\lambda}_i] = \hat{\phi}_i \text{res}[G_r, \tilde{\lambda}_i].
\]

The first order conditions (3.11), we refer to as Meier-Luenberger conditions, recognizing the work of [25]. As stated earlier, these conditions are directly obtained from the newly derived structured orthogonality conditions (3.5).

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

**Remark 3.1.** In the multiple input/multiple output (MIMO) case, the transfer function becomes matrix valued: \( G(s) = C(sI - A)^{-1}B \) with \( A \in \mathbb{R}^{n \times n} \) as before and \( B \in \mathbb{R}^{n \times m}, C \in \mathbb{R}^{p \times n} \) for \( m, p \geq 1 \). We seek a reduced order system with the same number of inputs \( m \) and outputs \( p \), but of lower order \( r < n \), and (matrix-valued) transfer function \( G_r(s) = C_r(sI - A_r)^{-1}B_r \) with \( V_r \in \mathbb{R}^{n \times r} \) and \( W_r \in \mathbb{R}^{n \times r} \) as before and
\[
A_r = (W_r^T V_r)^{-1}W_1^T AV_r, B_r = (W_r^T V_r)^{-1}W_r^T B_r, \text{ and } C_r = CV_r.
\]

Assume that \( G_r \) minimizes \( \|G - G_r\|_{\mathcal{H}_2} \) over all (stable) transfer functions of that form and that \( A_r \) has distinct eigenvalues, \( \tilde{\lambda}_i \), and is diagonalized as \( Y^* A X = \text{diag}(\tilde{\lambda}_i) \). Denote the columns of \( C_\lambda A X = \hat{c}_i \) and the rows of \( Y^* B_r \) as \( \hat{b}_i^* \). Then we can write \( G_r(s) = \sum_i \frac{1}{s - \tilde{\lambda}_i} \hat{c}_i \hat{b}_i^* \) and the residue of \( G_r(s) \) at \( \tilde{\lambda}_i \) is matrix valued: \( \text{res}[G_r(s), \tilde{\lambda}_i] = \hat{c}_i \hat{b}_i^* \). A similar analysis to what we have carried out above yields analogous error expressions (see also [4]), and recent investigations by the authors have yielded the following first order necessary conditions for the MIMO optimal \( \mathcal{H}_2 \) reduction problem:
\[
G(-\tilde{\lambda}_i) \hat{b}_i = G_r(-\tilde{\lambda}_i) \hat{b}_i, \quad \hat{c}_i^* G(-\tilde{\lambda}_i) = \hat{c}_i^* G_r(-\tilde{\lambda}_i), \quad \text{and} \quad \hat{c}_i^* G'(-\tilde{\lambda}_i) \hat{b}_i = \hat{c}_i^* G_r'(-\tilde{\lambda}_i) \hat{b}_i, \quad \text{for } i = 1, \ldots, r
\] (3.12)

as before, a prime (′) denotes first derivative with respect to \( s \). Thus the single-input single-output (\( m = p = 1 \)) conditions are replaced in the multiple-input multiple-output case by left tangential, right tangential as well as bi-tangential interpolation conditions. Following the considerations of section 2.1, we are able to conclude that if the 1\(^{st} \) column of \( V_r \) is \((\lambda_i I + A)^{-1}B_b \) and the i\(^{th} \) row of \( W_r^T \) is \( \hat{c}_i^* C(\lambda_i I + A)^{-1} \), for each \( i = 1, 2, ..., r \) then the \( \mathcal{H}_2 \) optimality conditions given above hold. The MIMO case will be presented in detail in a separate work by the authors.
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 $G(s) = c^T(sI-A)^{-1}b$, let $G_r(s) = c_r^T(sI-A_r)^{-1}b_r$ 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 $\lambda_i$, $i = 1, \ldots, r$. Let $G_r(s)$ defined by $A_r$, $b_r$, and $c_r^T$ be a local minimizer for the optimal $\mathcal{H}_2$ problem (3.1).

It is convenient to define the error system

$$G_{err}(s) \triangleq G(s) - G_r(s) = c_{err}^T(sI - A_{err})^{-1}b_{err}$$

with $A_{err} = \begin{bmatrix} A & 0 \\ 0 & A_r \end{bmatrix}$, $b_{err} = \begin{bmatrix} b \\ b_r \end{bmatrix}$, and $c_{err}^T = [c_r - c_{err}]$. (3.14)

Let $P_{err}$ and $Q_{err}$ be the gramians for the error system $G_{err}(s)$, i.e. $P_{err}$ and $Q_{err}$ solve

$$A_{err}P_{err} + P_{err}A_{err}^T + b_{err}b_{err}^T = 0 \quad (3.15)$$

$$Q_{err}A_{err} + A_{err}^TQ_{err} + c_{err}c_{err}^T = 0 \quad (3.16)$$

Partition $P_{err}$ and $Q_{err}$:

$$P_{err} = \begin{bmatrix} P_{11} & P_{12} \\ P_{12} & P_{22} \end{bmatrix} \quad Q_{err} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12} & Q_{22} \end{bmatrix}$$

Wilson [32] showed that the reduced order model $G_r(s) = c_r^T(sI - A_r)^{-1}b_r$ can be defined in terms of a Galerkin framework as well by taking

$$V_r = P_{12}P_{22}^{-1} \quad \text{and} \quad W_r = -Q_{12}Q_{22}^{-1}$$

obtained by a projection satisfies the first-order conditions of the optimal $\mathcal{H}_2$ problem. It was also shown in [32] that $W_r^T V_r = I$. The next result states the Lyapunov-based Wilson conditions for $\mathcal{H}_2$ optimality and shows their equivalence to our structured orthogonality framework:

**Theorem 3.5. The Wilson conditions**

$$P_{12}^T Q_{12} + P_{22}Q_{22} = 0, \quad (3.19)$$

$$Q_{12}^T b + Q_{22}b_r = 0 \quad (3.20)$$

$$c_r^T P_{22} - c_r^T P_{12} = 0. \quad (3.21)$$

for $\mathcal{H}_2$ optimality 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 $\lambda_i$, $i = 1, \ldots, r$. Write $H_2(s) = \tilde{c}^T(sI - A_r)^{-1}\tilde{b}$ where $\tilde{b}$ and $\tilde{c}$ can vary arbitrarily. Then from (2.8), if for any $\tilde{b} \neq 0$, $[\tilde{P}_1^T, \tilde{P}_2^T]^T$ solves

$$\begin{bmatrix} A & 0 \\ 0 & A_r \end{bmatrix} \begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix} + \begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix} A_r^T + \begin{bmatrix} b \\ b_r \end{bmatrix} \tilde{b}^T = 0, \quad (3.22)$$

we have for arbitrary $\tilde{c}$,

$$\langle G - G_r, H_2 \rangle_{\mathcal{H}_2} = [c_r^T - c_{err}] \begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix} \tilde{c} = 0.$$
Notice that $\tilde{P}_1$ and $\tilde{P}_2$ are independent of $\tilde{c}$ so for each choice of $\tilde{b}$, we must have
\[
c^T \tilde{P}_1 - c_r^T \tilde{P}_2 = 0.
\]
For $\tilde{b} = b_r$, one may check directly that $\tilde{P}_1 = P_{12}$ and $\tilde{P}_2 = P_{22}$ in $P_{err}$ that solves (3.15) in Wilson’s conditions.

Likewise, from (2.9) for each choice of $\tilde{c}$, if $[\tilde{Q}_1, \tilde{Q}_2]$ solves
\[
[\tilde{Q}_1, \tilde{Q}_2] \begin{bmatrix} A & 0 \\ 0 & A_r \end{bmatrix} + A_r^T [\tilde{Q}_1, \tilde{Q}_2] + \tilde{c} [c^T, -c_r^T] = 0
\]
then we have for every $\tilde{b}$,
\[
\langle G - G_r, H_2 \rangle_{\mathcal{H}_2} = \tilde{b}^T [\tilde{Q}_1, \tilde{Q}_2] \begin{bmatrix} b \\ b_r \end{bmatrix} = 0.
\]
Similarly to the first case, $[\tilde{Q}_1, \tilde{Q}_2]$ is independent of $\tilde{b}$ so for each choice of $\tilde{c}$, we must have
\[
\tilde{Q}_1 b + \tilde{Q}_2 b_r = 0,
\]
and for the particular case $\tilde{c} = -c_r$, one may check directly that $\tilde{Q}_1 = Q_{12}^T$ and $\tilde{Q}_2 = Q_{22}$ in $Q_{err}$ that solves (3.16) 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 (3.20) and (3.21).

The additional orthogonality condition $\langle G - G_r, G_r \cdot H \rangle_{\mathcal{H}_2} = 0$ taken over all systems $H(s)$ with the same poles as $G_r$ will yield the remaining Wilson condition (3.19)

Observe that
\[
G_r(s)H(s) = c_r^T(sI - A_r)^{-1} b_r \tilde{c}^T(sI - A_r)^{-1} \tilde{b}
\]
\[
= [c_r^T, 0] \left[ sI_{2r} - \begin{bmatrix} A_r & b_r \tilde{c}^T \\ 0 & A_r \end{bmatrix} \right]^{-1} \begin{bmatrix} 0 \\ b_r \end{bmatrix}
\]
Referring to (2.8), the condition $\langle G - G_r, G_r \cdot H \rangle_{\mathcal{H}_2} = 0$ leads to a Sylvester equation
\[
\begin{bmatrix} A & 0 \\ 0 & A_r \end{bmatrix} \begin{bmatrix} \tilde{W}_1 & \tilde{P}_1 \\ \tilde{W}_2 & \tilde{P}_2 \end{bmatrix} + \begin{bmatrix} \tilde{W}_1 & \tilde{P}_1 \\ \tilde{W}_2 & \tilde{P}_2 \end{bmatrix} \begin{bmatrix} A_r^T & 0 \\ \tilde{c} b_r^T & A_r^T \end{bmatrix} + \begin{bmatrix} b \\ b_r \end{bmatrix} [0, b_r] = 0,
\]
where the use of $\tilde{P}_1$ and $\tilde{P}_2$ reflects that they solve (3.22) as well. Then
\[
\langle G - G_r, G_r \cdot H_2 \rangle_{\mathcal{H}_2} = [c_r^T, -c_r^T] \begin{bmatrix} \tilde{W}_1 & \tilde{P}_1 \\ \tilde{W}_2 & \tilde{P}_2 \end{bmatrix} \begin{bmatrix} c_r \\ 0 \end{bmatrix} = 0.
\]
Alternatively, from (2.9),
\[
[\tilde{Q}_1, \tilde{Q}_2] \begin{bmatrix} A & 0 \\ 0 & A_r \end{bmatrix} + \begin{bmatrix} A_r^T & 0 \\ \tilde{c} b_r^T & A_r^T \end{bmatrix} [\tilde{Q}_1, \tilde{Q}_2] + \begin{bmatrix} c_r \\ 0 \end{bmatrix} [c_r^T, -c_r^T] = 0
\]
(3.24)
Define \( \tilde{e} \), the eigenvectors of \( A \), by eigenvalues, \( \tilde{e} \) and is omitted. The first equation asserts that

\[
(G - G_r, G_r \cdot H_2)_{\mathcal{H}_2} = [0, \tilde{b}]^T \begin{bmatrix} \tilde{Q}_1 & \tilde{Q}_2 \\ \tilde{Y}_1 & \tilde{Y}_2 \end{bmatrix} \begin{bmatrix} b_r \\ \tilde{b} \end{bmatrix} = 0.
\]

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

\[
\begin{bmatrix} \tilde{Q}_1 & \tilde{Q}_2 \\ \tilde{Y}_1 & \tilde{Y}_2 \end{bmatrix} \begin{bmatrix} b_r \\ \tilde{b} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.
\]

Define \( \begin{bmatrix} \tilde{Q}_1 & \tilde{Q}_2 \\ \tilde{Y}_1 & \tilde{Y}_2 \end{bmatrix} \begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix} = \begin{bmatrix} \tilde{R}_1 \\ \tilde{R}_2 \end{bmatrix} \). We will show that \( \tilde{R}_1 = 0 \). Premultiply (3.22) by \( \begin{bmatrix} \tilde{Q}_1 & \tilde{Q}_2 \\ \tilde{Y}_1 & \tilde{Y}_2 \end{bmatrix} \), postmultiply (3.24) by \( \begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix} \), and subtract the resulting equations to get

\[
\begin{align*}
\tilde{R}_1 A_r^T - A_r^T \tilde{R}_1 &= 0, \\
\tilde{R}_2 A_r^T - A_r^T \tilde{R}_2 &= c \bar{b}^T \tilde{R}_1,
\end{align*}
\]

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

\[
\begin{align*}
\tilde{x}_i^T \left( \tilde{R}_2 A_r^T - A_r^T \tilde{R}_2 \right) \tilde{y}_i &= \tilde{x}_i^T c \bar{b}^T \tilde{R}_1 \tilde{y}_i, \\
\tilde{x}_i^T R_2 \tilde{y}_i \lambda_i - \bar{\lambda}_i \tilde{x}_i^T R_2 \tilde{y}_i &= \tilde{x}_i^T c \bar{b}^T \tilde{R}_1 \tilde{y}_i, \\
0 &= (\tilde{x}_i^T \bar{c}) (b_i^T \tilde{y}_i) \ d_i
\end{align*}
\]

Either \( d_i = 0 \) or one of \( \tilde{x}_i^T \bar{c} \) and \( b_i^T \tilde{y}_i \) must vanish, which would then imply then that either \( \dim H < r \) or \( \dim G_r < r \). Thus \( d_i = 0 \) for all \( i = 1, \ldots, r \) and \( \tilde{R}_1 = 0 \) which proves the final Wilson condition (3.19).

The converse follows in a straightforward way by reversing the preceding arguments and is omitted.

Hyland and Bernstein [22] offered conditions that are equivalent to the Wilson conditions. Suppose \( G_r(s) \) defined by \( A_r, b_r, \) and \( c_r^T \) solves the optimal \( \mathcal{H}_2 \) problem. Then there exist positive nonnegative matrices \( P, Q \in \mathbb{R}^{n \times n} \) such that

\[
PQ = V_r W_r^T, \quad W_r^T V_r = I,
\]

where \( M \) is similar to a positive definite matrix. Then \( G_r(s) \) is given by \( A_r, b_r, \) and \( c_r^T \) with \( A_r = W_r^T A V_r, \quad b_r = W_r^T b, \) and \( c_r^T = c V_r \) such that, with the skew projection \( \Pi = V_r W_r^T \), the following conditions are satisfied:

\[
\begin{align*}
\text{rank}(P) &= \text{rank}(Q) = \text{rank}(PQ) \quad \text{(3.26)} \\
\Pi [AP + PA^T + bb^T] &= 0 \quad \text{(3.27)} \\
[A^T Q + QA + cc^T] \Pi &= 0 \quad \text{(3.28)}
\end{align*}
\]
Note that in both [32] and [22], the first-order necessary conditions are given in terms of (coupled) Lyapunov equations. Both [32] 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. [36] discusses computational issues related to solving associated linearized problems within each step.

Theorems 3.4 and 3.5 shows the equivalence between the structured orthogonality conditions and Lyapunov and interpolation based conditions for $H_2$ optimality, respectively. To complete the discussion, we formally state the equivalence between the Lyapunov and interpolation frameworks.

**Lemma 3.6. (Equivalence of Lyapunov and Interpolation Frameworks)**

The first-order necessary conditions of both [22] as given in (3.26)-(3.28) and [32] as given in (3.18) are equivalent to those of [25] as given in (3.11). That is, the Lyapunov-based first-order conditions [32, 22] for optimal $H_2$ 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 $H_2$ 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 which produces a reduced order model $G_r(s)$ that satisfies 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 $G_r(s)$ 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 a reduced order model which satisfies some form of a first-order necessary conditions; see [33, 31, 9, 20, 25, 22, 32, 24]. However, these approaches do not seem to be suitable for large-scale problems. The ones based on Lyapunov-based conditions, e.g. [32, 22, 31, 33], require solving a couple of Lyapunov equations at each step of the iteration. To our knowledge, the only methods which depend on interpolation-based necessary conditions have been proposed in [24] and [25]. The authors work with the transfer functions of $G(s)$ and $G_r(s)$; make an iteration on the denominator [24] or poles and residues [25] 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 [27, 28] 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 $\sigma$ 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 $\lambda(\sigma) = \{\tilde{\lambda}_1, \ldots, \tilde{\lambda}_r\}$ denote the resulting reduced order poles of $G_r(s)$; hence $\lambda(\sigma)$ is a function from $\mathbb{C}^r \mapsto \mathbb{C}^r$. Define the function $g(\sigma) = \lambda(\sigma) + \sigma$. Note that $g(\sigma) : \mathbb{C}^r \mapsto \mathbb{C}^r$. Aside from issues related to the ordering of the reduced order poles, $g(\sigma) = 0$ is equivalent to (3.11) and, hence, is a necessary condition for
H2-optimality of the reduced order model, G_r(s). Thus one can formulate a search for optimal H2 reduced order systems by considering the root finding problem g(σ) = 0. Many plausible approaches to this problem originate with Newton’s method, which appears as

\[ \sigma^{(k+1)} = \sigma^{(k)} - (I + J)^{-1} \left( \sigma^{(k)} + \lambda \left( \sigma^{(k)} \right) \right) \]  

(4.1)

In (4.1), J is the usual r × r Jacobian of λ(σ) with respect to σ: for J = [J_{i,j}],

\[ J_{i,j} = \frac{\partial \lambda_i}{\partial \sigma_j} \]

for i, j = 1, …, r. How to compute 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 H2-optimal shift set, the entries of the Jacobian matrix become small and simply setting J = 0 might serve as a relaxed iteration strategy. This leads to a successive substitution framework: \( \sigma_i \leftarrow -\lambda_i(A_r) \); 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 \(i\)th 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 \( \sigma_i \), for \( i = 1, \ldots, r \) that is closed under conjugation and fix a convergence tolerance tol.
2. Choose \( V_r \) and \( W_r \) so that \( \text{Ran}(V_r) = \text{span}\{ (\sigma_1 - A)^{-1}b, \ldots, (\sigma_r - A)^{-1}b \} \)
   \( \text{Ran}(W_r) = \text{span}\{ (\sigma_1 - A^T)^{-1}c, \ldots, (\sigma_r - A^T)^{-1}c \} \) and \( W_r^T V_r = I \).
3. while (relative change in \{ \sigma_i \} > tol )
   (a) \( A_r = W_r^T A V_r \),
   (b) Assign \( \sigma_i \leftarrow -\lambda_i(A_r) \) for \( i = 1, \ldots, r \)
   (c) Update \( V_r \) and \( W_r \) so \( \text{Ran}(V_r) = \text{span}\{ (\sigma_1 - A)^{-1}b, \ldots, (\sigma_r - A)^{-1}b \} \)
   \( \text{Ran}(W_r) = \text{span}\{ (\sigma_1 - A^T)^{-1}c, \ldots, (\sigma_r - A^T)^{-1}c \} \) and \( W_r^T V_r = I \).
4. \( A_r = W_r^T A V_r \), \( b_r = W_r^T b \), \( c_r^T = c^T V_r \)

Upon convergence, the first-order necessary conditions (3.11) for H2 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 for 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 H2 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 H2 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, 24, 33, 31], 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 which can be expected to improve the results. Recall that at convergence, interpolation points are mirror images of the eigenvalues of $A_r$. The eigenvalues of $A_r$ might be expected to approximate the eigenvalues of $A$. Hence, at convergence, interpolation points will lie in the mirror spectrum of $A$. Therefore, one could choose initial shifts randomly distributed within a region containing the mirror image of the numerical range of $A$. The boundary of the numerical range can be estimated by computing the eigenvalues of $A$ with the smallest and largest real and imaginary parts using numerically effective tools such as 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 $\sigma_i = -\lambda_i(A)$ where $\lambda_i(A)$ are the poles with big residues $\phi_i$ 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 $\phi_i$ might be directly read from the entries of the matrices $b$ and $c^T$.

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 $A$, 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 $A$ were between $-1.5668 \times 10^{-1}$ and $-2.0621 \times 10^{-3}$ 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 (3.12) introduced in Remark 3.1 for MIMO systems $G(s) = C(sI - A)^{-1}B$, one can easily extend IRKA to the MIMO case by replacing $(\sigma_i I - A)^{-1}b$ with $(\sigma_i I - A)^{-1}Bb_i$ and $(\sigma_i I - A^T)^{-1}c$ with $(\sigma_i I - A)^{-1}C^Tc_i$ in Algorithm 4.1 where $b_i$ and $c_i$ are as defined in Remark 3.1. As stated before, the MIMO case will be presented in a separate work.

Remark 4.2. Similar to its continuous-time analog (1.1), an $n^{th}$ order SISO discrete-time dynamical system is defined by a set of difference equations

$$G: \begin{cases} x(t+1) = Ax(t) + b \ u(t) \\ y(t) = c^T x(t) \end{cases} \quad \text{or} \quad G(z) = c^T (zI - A)^{-1}b, \quad (4.2)$$
where $t \in \mathbb{Z}$ and $A \in \mathbb{R}^{n \times n}$, $b, c \in \mathbb{R}^n$. $G(z)$ is the transfer function of the system, in other words, if $u(z)$ and $y(z)$ 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 $|\lambda_i(A)| < 1$ for $i = 1, \ldots, n$. Also, the $h_2$ norm is defined as $\|G\|_{h_2}^2 = \frac{1}{2\pi} \int_0^{2\pi} |G(e^{j\theta})|^2 d\theta$. The model reduction for discrete-time systems are defined similarly. In this setting, the interpolatory necessary conditions for the $r^{th}$ order reduced model $G_r(z) = c_r^T(zI - A_r)^{-1}b_r$ to be $h_2$ optimal become $G \left(1/\tilde{\lambda}_i\right) = G_r \left(1/\tilde{\lambda}_i\right)$ and $G' \left(1/\tilde{\lambda}_i\right) = G_r' \left(1/\tilde{\lambda}_i\right)$ for $i = 1, \ldots, r$ where $\tilde{\lambda}_i$ denotes the $i^{th}$ eigenvalue of $A_r$. Hence, in the discrete time case, the root-finding problem becomes $g(\sigma) = \Sigma\lambda(\sigma) - e$, where $e^T = [1, 1, \ldots, 1]$ and $\Sigma = \text{diag}(\sigma)$. Therefore, for discrete-time systems, Step 3b of Algorithm 4.1 becomes $\sigma_i \leftarrow 1/\tilde{\lambda}_i(A_r)$ for $i = 1, \ldots, r$. Moreover, the associated Newton step is

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

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

### 4.3. A Newton Framework for IRKA

As discussed above, Algorithm 4.1 uses the successive substitution framework by simply setting $J = 0$ 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 $A \in \mathbb{R}^{n \times n}$ and $b, c \in \mathbb{R}^n$, suppose that $\sigma_i$, $i = 1, \ldots, r$, are $r$ distinct points in $\mathbb{C}$ none of which are eigenvalues of $A$, and define the complex $r$-tuple $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_r]^T \in \mathbb{C}^r$ together with related matrices:

$$V_r(\sigma) = \begin{bmatrix} (\sigma_1 I - A)^{-1}b & (\sigma_2 I - A)^{-1}b & \cdots & (\sigma_r I - A)^{-1}b \end{bmatrix} \in \mathbb{C}^{n \times r} \quad (4.3)$$

and

$$W_r^T(\sigma) = \begin{bmatrix} c^T(\sigma_1 I - A)^{-1} \\ c^T(\sigma_2 I - A)^{-1} \\ \vdots \\ c^T(\sigma_r I - A)^{-1} \end{bmatrix} \in \mathbb{C}^{r \times n}. \quad (4.4)$$

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

**Lemma 4.2.** Let $\bar{x}_i$ be an eigenvector of $A_r = (W_r^T V_r)^{-1} W_r^T A V_r$ associated with $\tilde{\lambda}_i$, normalized so that $|\bar{x}_i^T W_r^T V_r \bar{x}_i| = 1$. Then $W_r^T A V_r \bar{x}_i = \tilde{\lambda}_i W_r^T V_r \bar{x}_i$. Then,

$$\frac{\partial \tilde{\lambda}_i}{\partial \sigma_j} = \bar{x}_i^T \frac{\partial}{\partial \sigma_j} W_r^T \left( A V_r \bar{x}_i - \tilde{\lambda}_i V_r \bar{x}_i \right) + \left( \bar{x}_i^T W_r^T A - \bar{x}_i^T W_r^T \tilde{\lambda}_i \right) \frac{\partial}{\partial \sigma_j} V_r \bar{x}_i \quad (4.5)$$

where $\frac{\partial}{\partial \sigma_j} W_r^T = -e_j c(\sigma_j I - A)^{-2}$ and $\frac{\partial}{\partial \sigma_j} V_r = -(\sigma_j I - A)^{-2} b e_j^T$. 


Proof. With $V_r(\sigma) = V_r$ and $W_r(\sigma) = W_r$ defined as in (4.3) and (4.4), both $W_r^TAV_r$ and $W_r^T V_r$ are complex symmetric matrices. Write $\tilde{\lambda}$ for $\tilde{\lambda}_i$ and $\tilde{x}$ for $\tilde{x}_i$; so

\[(a) \ W_r^TAV_r, \tilde{x} = \tilde{\lambda}W_r^T V_r, \tilde{x} \quad \text{and} \quad \text{(b) } \tilde{x}^T W_r^T A V_r = \tilde{\lambda} \tilde{x}^T W_r^T V_r. \tag{4.6}\]

(4.6b) is obtained by transposition of (4.6a). $\tilde{x}^T W_r^T V_r$ is a left eigenvector for $A_r$ associated with $\tilde{\lambda}_i$. Differentiate (4.6a) with respect to $\sigma_j$, premultiply with $\tilde{x}^T$, and simplify using (4.6b):

\[
\tilde{x}^T \frac{\partial}{\partial \sigma_j} W_r^T \left( A V_r - \tilde{\lambda} V_r \right) + \tilde{x}^T W_r^T A - \tilde{\lambda} \tilde{x}^T W_r^T \frac{\partial}{\partial \sigma_j} V_r \tilde{x} = \left( \frac{\partial \tilde{\lambda}}{\partial \sigma_j} \right) \tilde{x}^T W_r^T V_r \tilde{x}
\]

where $\frac{\partial}{\partial \sigma_j} W_r = \frac{\partial}{\partial \sigma_j} W_r^T = e_j e^T (\sigma_j I - A)^{-2}$ and $\frac{\partial}{\partial \sigma_j} V_r = (\sigma_j I - A)^{-2} b e_j^T$. This completes the proof. \[\Box\]

5. Numerical Examples. In this section, we first compare our approach with the existing ones [22, 24, 33] for a number of low order benchmark examples presented in these papers, and show that in each case, we attain the minimum; the main difference, however, is that we achieve this minimum in a numerically efficient way. Then we test our method in large-scale settings.

5.1. Low-order Models and Comparisons. We consider the following 4 models:

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

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

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 [33], denoted by GFM; the orthogonal projection method of [22], denoted by OPM; and the balanced truncation method, denoted by BTM.

- FOM-2: Example in [24]. 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 [24], denoted by LMPV.

- FOM-3: Example 1 in [31]. 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 [31], denoted by SMM.
• FOM-4: Example 2 in [31]. 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 \( H_2 \) errors \( \frac{\|G(s) - G_r(s)\|_{H_2}}{\|G(s)\|_{H_2}} \) are tabulated in Table 5.1 below.

<table>
<thead>
<tr>
<th>Model</th>
<th>( r )</th>
<th>IRKA</th>
<th>GFM</th>
<th>OPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOM-1</td>
<td>1</td>
<td>4.2683 \times 10^{-2}</td>
<td>4.2709 \times 10^{-1}</td>
<td>4.2683 \times 10^{-1}</td>
</tr>
<tr>
<td>FOM-1</td>
<td>2</td>
<td>3.9290 \times 10^{-2}</td>
<td>3.9299 \times 10^{-2}</td>
<td>3.9290 \times 10^{-2}</td>
</tr>
<tr>
<td>FOM-1</td>
<td>3</td>
<td>1.347 \times 10^{-3}</td>
<td>1.3107 \times 10^{-3}</td>
<td>1.3047 \times 10^{-3}</td>
</tr>
<tr>
<td>FOM-2</td>
<td>3</td>
<td>1.171 \times 10^{-3}</td>
<td>1.171 \times 10^{-3}</td>
<td>Divergent</td>
</tr>
<tr>
<td>FOM-2</td>
<td>4</td>
<td>8.199 \times 10^{-3}</td>
<td>8.199 \times 10^{-3}</td>
<td>8.199 \times 10^{-3}</td>
</tr>
<tr>
<td>FOM-2</td>
<td>5</td>
<td>2.132 \times 10^{-3}</td>
<td>2.132 \times 10^{-3}</td>
<td>Divergent</td>
</tr>
<tr>
<td>FOM-2</td>
<td>6</td>
<td>5.817 \times 10^{-4}</td>
<td>5.817 \times 10^{-4}</td>
<td>5.817 \times 10^{-4}</td>
</tr>
<tr>
<td>FOM-3</td>
<td>1</td>
<td>4.818 \times 10^{-4}</td>
<td>4.818 \times 10^{-4}</td>
<td>4.818 \times 10^{-4}</td>
</tr>
<tr>
<td>FOM-3</td>
<td>2</td>
<td>2.443 \times 10^{-4}</td>
<td>2.443 \times 10^{-4}</td>
<td>Divergent</td>
</tr>
<tr>
<td>FOM-3</td>
<td>3</td>
<td>5.74 \times 10^{-3}</td>
<td>5.98 \times 10^{-3}</td>
<td>5.74 \times 10^{-3}</td>
</tr>
<tr>
<td>FOM-4</td>
<td>1</td>
<td>9.85 \times 10^{-3}</td>
<td>9.85 \times 10^{-3}</td>
<td>9.85 \times 10^{-3}</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>( r )</th>
<th>BTM</th>
<th>LMPV</th>
<th>SMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOM-1</td>
<td>1</td>
<td>4.3212 \times 10^{-4}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOM-1</td>
<td>2</td>
<td>3.9378 \times 10^{-2}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOM-1</td>
<td>3</td>
<td>1.3107 \times 10^{-3}</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FOM-2</td>
<td>3</td>
<td>2.384 \times 10^{-4}</td>
<td>1.171 \times 10^{-4}</td>
<td></td>
</tr>
<tr>
<td>FOM-2</td>
<td>4</td>
<td>8.226 \times 10^{-4}</td>
<td>8.199 \times 10^{-4}</td>
<td></td>
</tr>
<tr>
<td>FOM-2</td>
<td>5</td>
<td>2.452 \times 10^{-4}</td>
<td>2.132 \times 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>FOM-2</td>
<td>6</td>
<td>5.822 \times 10^{-4}</td>
<td>2.864 \times 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>FOM-3</td>
<td>1</td>
<td>4.848 \times 10^{-4}</td>
<td>4.818 \times 10^{-4}</td>
<td></td>
</tr>
<tr>
<td>FOM-3</td>
<td>2</td>
<td>3.332 \times 10^{-4}</td>
<td>2.443 \times 10^{-4}</td>
<td></td>
</tr>
<tr>
<td>FOM-3</td>
<td>3</td>
<td>5.99 \times 10^{-3}</td>
<td>5.74 \times 10^{-3}</td>
<td></td>
</tr>
<tr>
<td>FOM-4</td>
<td>1</td>
<td>9.949 \times 10^{-4}</td>
<td>9.985 \times 10^{-4}</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1 clearly illustrates that the proposed method is the only one which 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 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 [33] 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 \( H_2 \) 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 \( \tilde{\lambda}_1 = -6.2217 \) and \( \tilde{\lambda}_{2,3} = -6.1774 \times 10^{-1} \pm 1.5628 \); and it can be shown that \( G_3(s) \) interpolates the first two moments of \( G(s) \) at \( -\tilde{\lambda}_i \), 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:

\[
\begin{align*}
S_1 &= \{-1.01, -2.01, -30000\} \\
S_2 &= \{0, 10, 3\} \\
S_3 &= \{1, 10, 3\} \\
S_4 &= \{0.01, 20, 10000\}
\end{align*}
\]

With selection \( S_1 \), we have initiated the algorithm with some negative shifts close to system poles, and consequently with a relative \( H_2 \) error bigger than 1. However, in all four cases including \( S_1 \), the algorithm converged in 5 steps to the same reduced model. The results are depicted in Figure 5.1.

**Fig. 5.1.** \( H_2 \) 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 [31], since \( r = 1 \), the optimal \( H_2 \) problem can be formulated as only a function of the reduced system pole. It was shown in [31] that
there are two local minima: one corresponding to a reduced pole at $-0.0052$ and consequently a reduced order model $G^l_1(s) = \frac{1.0313}{s+0.0052}$ and a relative error of 0.0949; and one to a reduced pole at $-4998$ and consequently a reduced model $G^g_1 = \frac{9999}{s+4998}$ with a relative error of 0.0985. It follows that the latter, i.e. $G^g_1(s)$ is the global minimum. The first order balanced truncation for FOM-4 can be easily computed as $G^b_1(s) = \frac{1.0308}{s+0.0052}$. Therefore, it is highly likely that if one starts from a balanced realization, the algorithm would converge to the local minimum $G^l_1(s)$. This was indeed the case as reported in [31]. 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 converges 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 to avoid the local minimum if we start from a bad shift. These observation perfectly agree with the discussion of Section 4.2. Note that, transfer function of FOM-4 can be written as

$$G(s) = \frac{100000}{s^2 + 5000s + 25} = \frac{0.99}{s + 0.0050} + \frac{9999}{s + 5000}.$$ 

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 [5], 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 $H_\infty$ and $H_2$ 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 as leading to small $H_\infty$ and $H_2$ error norms, see [5, 17], mainly due to global information available through the usage of 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 $H_2$ error 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 $[10^{-1}, 10^1]$ 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 $H_2$ 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 $H_2$ 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 of our algorithm with the existing methods and this makes the proposed algorithm numerically effective in large-scale settings.

![Graph](image.png)

Fig. 5.2. Relative \( H_2 \) norm of the error system vs \( r \)

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 \( H_2 \) error due to the current estimate and plot this error vs 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 \( H_2 \) norm of the error is reduced. (2) The algorithm converges after 3 steps. The resulting reduced models are stable for both cases.

5.3. A Semi-discretized 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 \( 1.382 \times 10^{-2} \) results in a system of the form

\[
\dot{\mathbf{x}}(t) = A\mathbf{x}(t) + B\mathbf{u}(t), \quad y(t) = c^T\mathbf{x}(t)
\]

with state-dimension \( n = 20209 \), i.e., \( A, E \in \mathbb{R}^{20209 \times 20209} \), \( b \in \mathbb{R}^{20209 \times 7} \), \( c^T \in \mathbb{R}^{6 \times 20209} \). Note that in this case \( E \neq I \), but the algorithm works with the obvious modifications. For details regarding the modeling, discretization, optimal control
design, and model reduction for this example, see [26, 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 $7.85 \times 10^{-3}$. Note that in order to apply balanced truncation in this example, one would need to solve two generalized Lyapunov equations (since $E \neq I$) of order 20209.

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 $H_2$ reduced model for $r = 1$ as

$$G_r(s) = \frac{0.97197}{s + 0.272727}$$

One can easily show that this reduced model interpolates $G(s)$ and its derivative at $\sigma = 0.272727$. 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 below. The figure shows that for this example, the successive substitution framework is divergent and indeed $\frac{d\lambda}{d\sigma} \approx 1.3728$. 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 $H_2$-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 $H_2$-optimality.

REFERENCES

Fig. 5.5. Comparison for FOM-1
Successive substitution framework with $\sigma_0 = 0.2$

Newton framework with $\sigma_0 = 2000$

Fig. 5.6. Comparison for the random third order model

of the 42nd IEEE Conference on Decision and Control, December 2003.


[34] A. Yousouff and R.E. Skelton, Covariance equivalent realizations with applications to model reduction of large-scale systems, in Control and Dynamic Systems, C.T. Leondes ed., Aca-
A. Appendix. Lemma A.1. For any stable matrix $M$,

\[ \text{P.V.} \int_{-\infty}^{+\infty} (\omega - M)^{-1} d\omega \overset{\text{def}}{=} \lim_{L \to \infty} \int_{-L}^{L} (\omega - M)^{-1} d\omega = \pi I \]

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

\[ \int_{-L}^{L} (\omega - M)^{-1} d\omega = \int_{-L}^{L} (-\omega - M)(\omega^2 + M^2)^{-1} d\omega \\
= \int_{-L}^{L} (-M)(\omega^2 I + M^2)^{-1} d\omega \\
\]

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

\[ -M(\omega^2 I + M^2)^{-1} = \frac{1}{2\pi i} \int_{\Gamma} \frac{-z}{\omega^2 + z^2} (zI - M)^{-1} dz \]

For any fixed value $z$ in the left half plane,

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

Thus,

\[ \lim_{L \to \infty} \int_{-L}^{L} (-M)(\omega^2 I + M^2)^{-1} d\omega = \frac{1}{2\pi i} \int_{\Gamma} \lim_{L \to \infty} \left( \int_{-L}^{L} \frac{-z}{\omega^2 + z^2} d\omega \right) (zI - M)^{-1} dz \\
= \frac{1}{2\pi i} \int_{\Gamma} \pi (zI - M)^{-1} dz = \pi I \]