View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

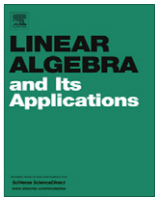
provided by Elsevier - Publisher Connector

Linear Algebra and its Applications 436 (2012) 2916–2943

Contents lists available at SciVerse ScienceDirect

Linear Algebra and its Applications

journal homepage: www.elsevier.com/locate/laa



Inexact solves in interpolatory model reduction ☆

Christopher Beattie *, Serkan Gugercin, Sarah Wyatt

Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123, United States

ARTICLE INFO

Article history:
Received 28 July 2010
Accepted 5 July 2011
Available online 3 August 2011

Submitted by V. Mehrmann

Dedicated to Danny Sorensen on the occasion of his 65th birthday

Keywords:
Model reduction
System order reduction
Tangential interpolation
Iterative solves
Petrov–Galerkin

ABSTRACT

We investigate the use of inexact solves for interpolatory model reduction and consider associated perturbation effects on the underlying model reduction problem. We give bounds on system perturbations induced by inexact solves and relate this to termination criteria for iterative solution methods. We show that when a Petrov–Galerkin framework is employed for the inexact solves, the associated reduced order model is an exact interpolatory model for a nearby full-order system; thus demonstrating backward stability. We also give evidence that for $\mathcal{H}_2$-optimal interpolation points, interpolatory model reduction is robust with respect to perturbations due to inexact solves. Finally, we demonstrate the effectiveness of direct use of inexact solves in optimal $\mathcal{H}_2$ approximation. The result is an effective model reduction strategy that is applicable in realistically large-scale settings.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

The simulation of dynamical systems constitutes a basic framework for the modeling and control of many complex phenomena of interest in science and industry. The need for ever greater model fidelity often leads to computational tasks that make unmanageably large demands on resources. Efficient model utilization becomes a critical consideration in such large-scale problem settings and motivates the development of strategies for model reduction.

We consider here linear time invariant multi-input/multi-output (MIMO) systems that have a state space form (in the Laplace transform domain) as

$$\text{Find } \widehat{\mathbf{v}}(s) \text{ such that } \mathcal{K}(s)\,\widehat{\mathbf{v}}(s) = \mathcal{B}(s)\widehat{\mathbf{u}}(s), \quad \text{then } \widehat{\mathbf{y}}(s) \stackrel{\text{def}}{=} \mathcal{C}(s)\,\widehat{\mathbf{v}}(s). \tag{1}$$

☆ This work was supported in part by the NSF through Grants DMS-0505971 and DMS-0645347.

* Corresponding author.
E-mail addresses: beattie@vt.edu (C. Beattie), gugercin@math.vt.edu (S. Gugercin), sawyatt@vt.edu (S. Wyatt).

0024-3795/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.laa.2011.07.015

Table 1
Examples of generalized coprime system realizations.

Descriptor systems	$\mathbf{C}(s\mathbf{E}-\mathbf{A})^{-1}\mathbf{B}$ ($\mathbf{E}$ possibly singular)
Delay systems	$(e^{-s\,\tau_{out}}\mathbf{C})(s\mathbf{I}-\mathbf{A}_0-e^{-s\,\tau_{sys}}\mathbf{A}_1)^{-1}(e^{-s\,\tau_{inp}}\mathbf{B})$
Second order systems	$(s\mathbf{C}_1+\mathbf{C}_0)(s^2\mathbf{M}+s\mathbf{G}+\mathbf{K})^{-1}\mathbf{B}$
Weighted systems	$\mathbf{W}_o(s)\mathbf{C}(s\mathbf{I}-\mathbf{A})^{-1}\mathbf{B}\mathbf{W}_i(s)$

Here, $\widehat{\mathbf{u}}(s)$ and $\widehat{\mathbf{y}}(s)$ denote Laplace-transformed system inputs and outputs, respectively; $\widehat{\mathbf{v}}(s)$ represents the internal system state. We assume that $\mathcal{C}(s) \in \mathbb{C}^{p\times n}$ and $\mathcal{B}(s) \in \mathbb{C}^{n\times m}$ are analytic in the right half plane; and that $\mathcal{K}(s) \in \mathbb{C}^{n\times n}$ is analytic and full rank throughout the right half plane. Solving for $\widehat{\mathbf{y}}(s)$ in terms of $\widehat{\mathbf{u}}(s)$, we obtain

$$\widehat{\mathbf{y}}(s) = \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)\widehat{\mathbf{u}}(s) = \mathcal{H}(s)\widehat{\mathbf{u}}(s). \tag{2}$$

This representation of the *transfer function*,

$$\mathcal{H}(s) = \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathcal{B}(s), \tag{3}$$

we refer to as a *generalized coprime realization*. Standard first-order descriptor system realizations, with $\mathcal{H}(s) = \mathbf{C}\,(s\mathbf{E}-\mathbf{A})^{-1}\,\mathbf{B}$ for constant matrices $\mathbf{E},\ \mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, and $\mathbf{C} \in \mathbb{R}^{p\times n}$ evidently fit this pattern with $\mathcal{C}(s) = \mathbf{C}$, $\mathcal{B}(s) = \mathbf{B}$, and $\mathcal{K}(s) = s\mathbf{E}-\mathbf{A}$. However, many dynamical systems can be described more naturally with generalized coprime realizations. For example, a system that includes internal system delays as well as transmission/propagation delays in its input and output could be described with a model

$$\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}_0\,\mathbf{x}(t) + \mathbf{A}_1\,\mathbf{x}(t-\tau_{sys}) + \mathbf{B}\,\mathbf{u}(t-\tau_{inp}), \quad \mathbf{y}(t) = \mathbf{C}\,\mathbf{x}(t-\tau_{out}) \tag{4}$$

for $\tau_{sys},\ \tau_{inp},\ \tau_{out} > 0$, and $\mathbf{E},\ \mathbf{A}_0,\ \mathbf{A}_1 \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{p\times n}$. Taking the Laplace transformation of (4) yields the transfer function

$$\mathcal{H}(s) = \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathcal{B}(s) = \left(e^{-s\,\tau_{out}}\,\mathbf{C}\right)\left(s\,\mathbf{E}-\mathbf{A}_0-e^{-s\,\tau_{sys}}\,\mathbf{A}_1\right)^{-1}\left(e^{-s\,\tau_{inp}}\,\mathbf{B}\right),$$

which has the form of (3). The form of (3) can accommodate greater generality than this, of course, including memory convolution involving higher derivatives, second and higher-order polynomial differential equations, systems described via integro-differential equations, and systems where state variables may be coupled through infinite dimensional subsystems (possibly modeling internal propagation or diffusion). See Table 1 for other examples and [6] for further discussion.

In many applications, the state space dimension, n, is too large for efficient system simulation and control computation, so the cases of interest for us here have state space dimension vastly larger than input and output dimensions: $n \gg m, p$. See [19] for a recent collection of such benchmark problems.

The goal is to produce a reduced system that will have approximately the same response (output) as the original system for any given input $\mathbf{u}(t)$. For a given reduced-order $r \ll n$, we construct reduced order models through a Petrov–Galerkin approximation of (1): Select full rank matrices $\mathbf{V}_r \in \mathbb{R}^{n\times r}$ and $\mathbf{W}_r \in \mathbb{R}^{n\times r}$. For any input, $\mathbf{u}(t)$, the reduced system output, $\mathbf{y}_r(t)$, is then defined (in the Laplace transform domain) as:

$$\text{Find } \widehat{\mathbf{v}}(s) \in \text{Ran}(\mathbf{V}_r) \text{ such that } \mathbf{W}_r^T\left(\mathcal{K}(s)\,\widehat{\mathbf{v}}(s) - \mathcal{B}(s)\widehat{\mathbf{u}}(s)\right) = 0 \tag{5}$$

$$\text{then } \widehat{\mathbf{y}}_r(s) \stackrel{\text{def}}{=} \mathcal{C}(s)\,\widehat{\mathbf{v}}(s), \tag{6}$$

which defines the reduced transfer function as,

$$\mathcal{H}_r(s) = \mathcal{C}_r(s)\mathcal{K}_r(s)^{-1}\mathcal{B}_r(s), \tag{7}$$

where

$$\mathcal{K}_r(s) = \mathbf{W}_r^T \mathcal{K}(s)\mathbf{V}_r \in \mathbb{C}^{r\times r}, \quad \mathcal{B}_r(s) = \mathbf{W}_r^T \mathcal{B}(s) \in \mathbb{C}^{r\times m}, \\ \text{and} \quad \mathcal{C}_r(s) = \mathcal{C}(s)\mathbf{V}_r \in \mathbb{C}^{p\times r}. \tag{8}$$

2. Interpolatory model reduction

Interpolatory reduced order models are designed to exactly reproduce certain system response components that result from inputs having specified frequency content and growth. The approach has been described for standard first-order system realizations in [13,2,11,3] and extended to generalized coprime realizations in [6]. We summarize the basic elements of this approach below.

A set of points $\{\mu_i\}_{i=1}^r \subset \mathbb{C}$ and (nontrivial) direction vectors $\{\mathbf{c}_i\}_{i=1}^r \subset \mathbb{C}^p$ constitute left tangential interpolation data for the reduced model, $\mathcal{H}_r(s)$, if

$$\mathbf{c}_i^T \mathcal{H}(\mu_i) = \mathbf{c}_i^T \mathcal{H}_r(\mu_i) \quad \text{for each } i = 1, \ \ldots, \ r. \tag{9}$$

Likewise, $\{\sigma_j\}_{j=1}^r$, and associated directions $\{\mathbf{b}_j\}_{j=1}^r \subset \mathbb{C}^m$, constitute right tangential interpolation data for the reduced model, $\mathcal{H}_r(s)$, if

$$\mathcal{H}(\sigma_j)\mathbf{b}_j = \mathcal{H}_r(\sigma_j)\mathbf{b}_j \quad \text{for each } j = 1, \ \ldots, \ r. \tag{10}$$

Given left and right tangential interpolating data, interpolatory model reduction may be implemented by first solving the linear systems:

$$\text{Find } \mathbf{w}_i \text{ such that } \mathbf{w}_i^T \mathcal{K}(\mu_i) = \mathbf{c}_i^T \mathcal{C}(\mu_i) \quad \text{for } i = 1, \ \ldots, \ r, \text{ and} \tag{11}$$

$$\text{find } \mathbf{v}_i \text{ such that } \mathcal{K}(\sigma_j)\mathbf{v}_j = \mathcal{B}(\sigma_j)\mathbf{b}_j \quad \text{for } j = 1, \ \ldots, \ r. \tag{12}$$

We assume that the two point sets $\{\mu_i\}_{i=1}^r$ and $\{\sigma_j\}_{j=1}^r$ each consist of r distinct points and that the vectors $\{\mathbf{v}_1, \ \ldots, \ \mathbf{v}_r\}$ and $\{\mathbf{w}_1, \ \ldots, \ \mathbf{w}_r\}$ are linearly independent sets. These vectors constitute "primitive bases" for the subspaces $\mathcal{V}_r = \text{span}\{\mathbf{v}_1, \ \ldots, \ \mathbf{v}_r\}$ and $\mathcal{W}_r = \text{span}\{\mathbf{w}_1, \ \ldots, \ \mathbf{w}_r\}$. Define the associated matrices:

$$\mathbf{V}_r = [\ \mathbf{v}_1, \ \ldots, \ \mathbf{v}_r\ ] = \left[\ \mathcal{K}(\sigma_1)^{-1}\mathcal{B}(\sigma_1)\mathbf{b}_1, \ \ldots, \ \mathcal{K}(\sigma_r)^{-1}\mathcal{B}(\sigma_r)\mathbf{b}_r\ \right], \tag{13}$$

$$\mathbf{W}_r^T = \begin{bmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_r^T \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1^T \mathcal{C}(\mu_1)\mathcal{K}(\mu_1)^{-1} \\ \vdots \\ \mathbf{c}_r^T \mathcal{C}(\mu_r)\mathcal{K}(\mu_r)^{-1} \end{bmatrix}. \tag{14}$$

The reduced model, $\mathcal{H}_r(s)$, as defined in (7) and (8) using $\mathbf{V}_r$ and $\mathbf{W}_r$ from (13) and (14), interpolates $\mathcal{H}(s)$ at the $2r$ points $\{\mu_i\}_{i=1}^r$ and $\{\sigma_j\}_{j=1}^r$, in respective output directions $\{\mathbf{c}_i\}_{i=1}^r$ and input directions $\{\mathbf{b}_j\}_{j=1}^r$; that is, conditions (9) and (10) are satisfied. If $\mu_k = \sigma_k$ for some k then first order bitangential moments match as well:

$$\mathbf{c}_k^T \mathcal{H}'(\mu_k)\, \mathbf{b}_k = \mathbf{c}_k^T \mathcal{H}_r'(\mu_k)\, \mathbf{b}_k.$$

Interpolation of higher order derivatives of $\mathcal{H}(s)$ can be accomplished with similar constructions as well; see [6,3] and references therein.

For large-scale settings with millions of degrees of freedom, interpolatory model reduction has become the method of choice since it does not require dense matrix operations; the major computational cost lies in solving the (often sparse) linear systems in (11) and (12). This contrasts with Gramian-based

model reduction approaches such as balanced truncation [25,24], optimal Hankel norm approximation [12] and singular perturbation approximation [21] where large-scale Lyapunov equations need to be solved. Moreover, these computational advantages have been enhanced for standard first order state-space realizations by strategies for optimal selection of tangential interpolation data, see [16].

2.1. Inexact interpolatory model reduction

The basic framework for interpolatory model reduction presumes that the key Eqs. (11) and (12) may be solved exactly or nearly so, at least to an accuracy associated with machine precision. Direct solution methods, employing sparse factorization strategies, for example, are capable of handling systems of significantly large order. However since the need for ever greater modeling detail and fidelity can drive system order to the order of millions, the use of direct solvers for the linear systems (11) and (12) often becomes infeasible and iterative methods must be employed that terminate with possibly coarse approximate solutions to the linear systems. We consider and evaluate issues related to these approaches here.

Suppose $\{\widehat{\mathbf{v}}_1, \ \ldots, \ \widehat{\mathbf{v}}_r\}$ and $\{\widehat{\mathbf{w}}_1, \ \ldots, \ \widehat{\mathbf{w}}_r\}$ are linearly independent sets in $\mathbb{C}^n$ and define

$$\widehat{\mathbf{V}}_r = [\, \widehat{\mathbf{v}}_1, \ \ldots, \ \widehat{\mathbf{v}}_r \,] \qquad \widehat{\mathbf{W}}_r^T = \begin{bmatrix} \widehat{\mathbf{w}}_1^T \\ \vdots \\ \widehat{\mathbf{w}}_r^T \end{bmatrix}. \tag{15}$$

$\widehat{\mathbf{w}}_i$ and $\widehat{\mathbf{v}}_j$ will be viewed as approximate solutions to the linear systems (11) and (12) and accordingly we will refer to them as "inexact" solutions to (11) and (12). Nonetheless, unless otherwise stated, these vectors can be any arbitrarily chosen linearly independent vectors in $\mathbb{C}^n$.

Define residuals, $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_j$, corresponding to $\widehat{\mathbf{w}}_i$ and $\widehat{\mathbf{v}}_j$, as

$$\boldsymbol{\xi}_i = \mathcal{K}(\mu_i)^T \, \widehat{\mathbf{w}}_i - \mathcal{C}(\mu_i)^T \, \mathbf{c}_i \quad \text{and} \quad \boldsymbol{\eta}_j = \mathcal{K}(\sigma_j)\widehat{\mathbf{v}}_j - \mathcal{B}(\sigma_j)\mathbf{b}_j. \tag{16}$$

The deviations from the corresponding exact solutions are then

$$\delta\mathbf{w}_i = \widehat{\mathbf{w}}_i - \mathbf{w}_i = \mathcal{K}(\mu_i)^{-T} \, \boldsymbol{\xi}_i \quad \text{and} \quad \delta\mathbf{v}_j = \widehat{\mathbf{v}}_j - \mathbf{v}_j = \mathcal{K}(\sigma_j)^{-1} \, \boldsymbol{\eta}_j. \tag{17}$$

The resulting (inexact) basis matrices destined for use in a reduced order model are

$$\widehat{\mathbf{W}}_r = \mathbf{W}_r + [\delta\mathbf{w}_1, \ \ldots, \ \delta\mathbf{w}_r] \tag{18}$$

$$\widehat{\mathbf{V}}_r = \mathbf{V}_r + [\delta\mathbf{v}_1, \ \ldots, \ \delta\mathbf{v}_r]. \tag{19}$$

Define reduced order maps associated with these inexact bases:

$$\widehat{\mathcal{K}}_r(s) = \widehat{\mathbf{W}}_r^T \mathcal{K}(s) \widehat{\mathbf{V}}_r, \quad \widehat{\mathcal{B}}_r(s) = \widehat{\mathbf{W}}_r^T \mathcal{B}(s), \quad \text{and} \quad \widehat{\mathcal{C}}_r(s) = \mathcal{C}(s)\widehat{\mathbf{V}}_r, \tag{20}$$

together with the associated inexact reduced order transfer function

$$\widehat{\mathcal{H}}_r(s) = \widehat{\mathcal{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathcal{B}}_r(s).$$

Notice that we are free to make any choice for bases for the subspaces, $\widehat{\mathcal{V}}_r$ and $\widehat{\mathcal{W}}_r$, in defining $\widehat{\mathcal{H}}_r(s)$; no change in the definition of (20) is necessary. As a practical matter, it is generally prudent to choose well conditioned bases in computation.

3. Forward error

3.1. Interpolation error

Inexactness in the solution of the key linear systems (11) and (12) produces a computed reduced order transfer function, $\widehat{\mathcal{H}}_r(s)$ that no longer interpolates $\mathcal{H}(s)$; typically, the reduced order system response will no longer match any component of the full-order system response at any of the complex frequencies $\{\mu_i\}_{i=1}^r$ and $\{\sigma_i\}_{i=1}^r$ that have been specified. How much response error has been introduced at these points?

The particular realization taken for a transfer function can create innate sensitivities to perturbations associated with that representation. Define perturbed transfer functions,

$$\mathcal{H}_{\delta\mathcal{B}}(s) = \mathcal{C}(s)\mathcal{K}(s)^{-1}(\mathcal{B}(s) + \delta\mathcal{B}) \quad \text{and} \quad \mathcal{H}_{\delta\mathcal{C}}(s) = (\mathcal{C}(s) + \delta\mathcal{C})\mathcal{K}(s)^{-1}\mathcal{B}(s).$$

In discussing perturbations in system response caused by $\delta\mathcal{B}$ and $\delta\mathcal{C}$ at $s = \sigma$, it is natural to introduce the following quantities:

$$\text{cond}_{\mathcal{B}}(\mathcal{H}(\sigma)) = \frac{\|\mathcal{C}(\sigma)\mathcal{K}(\sigma)^{-1}\|\,\|\mathcal{B}(\sigma)\|}{\|\mathcal{H}(\sigma)\|}$$
$$\text{cond}_{\mathcal{C}}(\mathcal{H}(\sigma)) = \frac{\|\mathcal{C}(\sigma)\|\,\|\mathcal{K}(\sigma)^{-1}\mathcal{B}(\sigma)\|}{\|\mathcal{H}(\sigma)\|}$$

to be *condition numbers of the transfer function response*, by way of analogy to the condition number of algebraic linear systems. (Unless otherwise noted, norms will always refer to the Euclidean 2-norm for vectors or the naturally induced spectral norm for matrices). It is straightforward to show that these quantities measure the relative sensitivity of the system with respect to perturbations in $\mathcal{B}$ and $\mathcal{C}$, respectively:

$$\frac{\|\mathcal{H}_{\delta\mathcal{B}}(\sigma) - \mathcal{H}(\sigma)\|}{\|\mathcal{H}(\sigma)\|} \leqslant \text{cond}_{\mathcal{B}}(\mathcal{H}(\sigma))\frac{\|\delta\mathcal{B}\|}{\|\mathcal{B}(\sigma)\|} \quad \text{and}$$
$$\frac{\|\mathcal{H}_{\delta\mathcal{C}}(\sigma) - \mathcal{H}(\sigma)\|}{\|\mathcal{H}(\sigma)\|} \leqslant \text{cond}_{\mathcal{C}}(\mathcal{H}(\sigma))\frac{\|\delta\mathcal{C}\|}{\|\mathcal{C}(\sigma)\|}.$$

For values of s such that $\mathcal{K}_r(s)$ and $\widehat{\mathcal{K}}_r(s)$ are nonsingular, define the matrix-valued functions,

$$\begin{aligned}
&\mathcal{P}_r(s) = \mathcal{K}(s)\mathbf{V}_r\mathcal{K}_r(s)^{-1}\mathbf{W}_r^T, \quad \mathcal{Q}_r(s) = \mathbf{V}_r\mathcal{K}_r(s)^{-1}\mathbf{W}_r^T\mathcal{K}(s),\\
&\widehat{\mathcal{P}}_r(s) = \mathcal{K}(s)\widehat{\mathbf{V}}_r\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathbf{W}}_r^T, \quad \text{and} \quad \widehat{\mathcal{Q}}_r(s) = \widehat{\mathbf{V}}_r\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathbf{W}}_r^T\mathcal{K}(s)
\end{aligned} \tag{21}$$

where defined, $\mathcal{P}_r(s)$, $\mathcal{Q}_r(s)$, $\widehat{\mathcal{P}}_r(s)$, and $\widehat{\mathcal{Q}}_r(s)$ are differentiable (indeed, analytic) with respect to s, having derivatives that satisfy:

$$\begin{aligned}
&\widehat{\mathcal{P}}_r'(s) = \left(\mathbf{I} - \widehat{\mathcal{P}}_r\right)\mathcal{K}'(s)\mathcal{K}(s)^{-1}\widehat{\mathcal{P}}_r\\
&\text{and}\\
&\widehat{\mathcal{Q}}_r'(s) = \widehat{\mathcal{Q}}_r\mathcal{K}(s)^{-1}\mathcal{K}'(s)\left(\mathbf{I} - \widehat{\mathcal{Q}}_r\right)
\end{aligned} \tag{22}$$

with equivalent expressions for $\mathcal{P}_r'(s)$ and $\mathcal{Q}_r'(s)$. We will make a series of observations about properties of $\widehat{\mathcal{P}}_r(s)$ and $\widehat{\mathcal{Q}}_r(s)$ which will have immediately apparent parallels to properties for $\mathcal{P}_r(s)$ and $\mathcal{Q}_r(s)$.

Observe first that $\widehat{\mathcal{P}}_r^2 = \widehat{\mathcal{P}}_r$ and $\widehat{\mathcal{Q}}_r^2 = \widehat{\mathcal{Q}}_r$ so both $\widehat{\mathcal{P}}_r(s)$ and $\widehat{\mathcal{Q}}_r(s)$ are skew projectors. These projectors are of interest because the pointwise error in the transfer function can be expressed as

$$\begin{aligned}
\mathcal{H}(s) - \widehat{\mathcal{H}}_r(s) &= \mathcal{C}(s)\left(\mathcal{K}(s)^{-1} - \widehat{\mathbf{V}}_r\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathbf{W}}_r^T\right)\mathcal{B}(s)\\
&= \mathcal{C}(s)\mathcal{K}(s)^{-1}\left(\mathbf{I} - \widehat{\mathcal{P}}_r(s)\right)\mathcal{B}(s).
\end{aligned}$$

Similarly,

$$\mathcal{H}(s) - \widehat{\mathcal{H}}_r(s) = \mathcal{C}(s)\,(\mathbf{I} - \widehat{\Omega}_r(s))\,\mathcal{K}(s)^{-1}\mathcal{B}(s)$$

and

$$\mathcal{H}(s) - \widehat{\mathcal{H}}_r(s) = \mathcal{C}(s)\,(\mathbf{I} - \widehat{\Omega}_r(s))\,\mathcal{K}(s)^{-1}\,(\mathbf{I} - \widehat{\mathcal{P}}_r(s))\,\mathcal{B}(s).$$

The derivative of this last expression can be computed with the aid of (22) and observing $\mathcal{K}(s)^{-1}\widehat{\mathcal{P}}_r(s) = \widehat{\Omega}_r(s)\mathcal{K}(s)^{-1}$:

$$\begin{aligned}\mathcal{H}'(s) - \widehat{\mathcal{H}}_r'(s) = &\frac{d}{ds}\left[\mathcal{C}(s)\mathcal{K}(s)^{-1}\right](\mathbf{I} - \widehat{\mathcal{P}}_r(s))\,\mathcal{B}(s)\\ &+ \mathcal{C}(s)\,(\mathbf{I} - \widehat{\Omega}_r(s))\,\frac{d}{ds}\left[\mathcal{K}(s)^{-1}\mathcal{B}(s)\right]\\ &- \mathcal{C}(s)\,(\mathbf{I} - \widehat{\Omega}_r(s))\,\frac{d}{ds}\left[\mathcal{K}(s)^{-1}\right](\mathbf{I} - \widehat{\mathcal{P}}_r(s))\,\mathcal{B}(s).\end{aligned} \tag{23}$$

We introduce the following (s-dependent) subspaces:

$$\begin{aligned}&\mathfrak{P}_r(s) = \operatorname{Ran}\mathcal{P}_r(s) = \operatorname{Ran}\mathcal{K}(s)\mathbf{V}_r, \quad \mathfrak{Q}_r(s) = \operatorname{Ker}\left(\mathbf{W}_r^T\mathcal{K}(s)\right)^{\perp},\\ &\widehat{\mathfrak{P}}_r(s) = \operatorname{Ran}\widehat{\mathcal{P}}_r(s) = \operatorname{Ran}\mathcal{K}(s)\widehat{\mathbf{V}}_r, \quad \widehat{\mathfrak{Q}}_r(s) = \operatorname{Ker}\left(\widehat{\mathbf{W}}_r^T\mathcal{K}(s)\right)^{\perp},\\ &\mathfrak{B}_m(s) = \operatorname{Ran}\mathcal{K}(s)^{-1}\mathcal{B}(s), \quad \mathfrak{C}_p(s) = \operatorname{Ker}\left(\mathcal{C}(s)\mathcal{K}(s)^{-1}\right)^{\perp}.\end{aligned}$$

$\widehat{\mathcal{P}}_r(s)$ maps vectors in $\mathbb{C}^n$ onto $\widehat{\mathfrak{P}}_r(s)$ along $\widehat{\mathcal{W}}_r^{\perp}$ and $\widehat{\Omega}_r$ maps vectors in $\mathbb{C}^n$ onto $\widehat{\mathcal{V}}_r$ along $\widehat{\mathfrak{Q}}_r(s)^{\perp}$.

Given two subspaces of $\mathbb{C}^n$, say $\mathcal{M}$ and $\mathcal{N}$, we express the proximity of one to the other in terms of the angle between the subspaces, $\Theta(\mathcal{M},\mathcal{N}) \in [0, \frac{\pi}{2}]$ defined as

$$\sup_{\mathbf{x}\in\mathcal{M}}\inf_{\mathbf{y}\in\mathcal{N}}\frac{\|\mathbf{y}-\mathbf{x}\|}{\|\mathbf{x}\|} = \sin\Theta(\mathcal{M},\mathcal{N}).$$

$\Theta(\mathcal{M},\mathcal{N})$ is the largest canonical angle between $\mathcal{M}$ and a "closest" subspace $\widehat{\mathcal{N}}$ of $\mathcal{N}$ having dimension equal to $\dim\mathcal{M}$. Notice that if $\dim\mathcal{N} < \dim\mathcal{M}$ then $\Theta(\mathcal{M},\mathcal{N}) = \frac{\pi}{2}$ and $\Theta(\mathcal{M},\mathcal{N}) = 0$ if and only if $\mathcal{M}\subset\mathcal{N}$. $\Theta(\mathcal{M},\mathcal{N})$ is asymmetrically defined with respect to $\mathcal{M}$ and $\mathcal{N}$, however if $\dim\mathcal{N} = \dim\mathcal{M}$ then $\Theta(\mathcal{M},\mathcal{N}) = \Theta(\mathcal{N},\mathcal{M})$. If $\mathbf{\Pi}_{\mathcal{M}}$ and $\mathbf{\Pi}_{\mathcal{N}}$ denote orthogonal projectors onto $\mathcal{M}$ and $\mathcal{N}$, respectively, then $\sin\Theta(\mathcal{M},\mathcal{N}) = \|(\mathbf{I}-\mathbf{\Pi}_{\mathcal{M}})\mathbf{\Pi}_{\mathcal{N}}\|$.

The spectral norm of a skew projector can be expressed in terms of the angle between its range and cokernel [27]. In particular,

$$\|\widehat{\mathcal{P}}_r(s)\| = \|\mathbf{I} - \widehat{\mathcal{P}}_r(s)\| = \frac{1}{\cos\Theta(\widehat{\mathfrak{P}}_r(s),\ \widehat{\mathcal{W}}_r)}, \tag{24}$$

$$\|\widehat{\Omega}_r(s)\| = \|\mathbf{I} - \widehat{\Omega}_r(s)\| = \frac{1}{\cos\Theta(\widehat{\mathfrak{Q}}_r(s),\ \widehat{\mathcal{V}}_r)}. \tag{25}$$

Theorem 3.1. *Given the full-order model $\mathcal{H}(s) = \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)$, interpolation points $\{\sigma_j\}\subset\mathbb{C}$, $\{\mu_i\}\subset\mathbb{C}$ and corresponding tangential directions, $\{\mathbf{b}_j\}\subset\mathbb{C}^m$ and $\{\mathbf{c}_i\}\subset\mathbb{C}^p$, let the inexact interpolatory*

reduced model $\widehat{\mathcal{H}}_r(s) = \widehat{\mathcal{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathcal{B}}_r(s)$ *be constructed as defined in (15)-(20). The (tangential) interpolation error at* μ_i *and* σ_j *is*

$$\frac{\|\widehat{\mathcal{H}}_r(\sigma_j)\mathbf{b}_j - \mathcal{H}(\sigma_j)\mathbf{b}_j\|}{\|\mathcal{H}(\sigma_j)\mathbf{b}_j\|} \leqslant cond_{\mathcal{B}}(\mathcal{H}(\sigma_j)\mathbf{b}_j)\,\frac{\sin\Theta(\mathcal{C}_p(\sigma_j),\ \widehat{\mathcal{W}}_r)}{\cos\Theta(\widehat{\mathfrak{P}}_r(\sigma_j),\ \widehat{\mathcal{W}}_r)}\,\frac{\|\boldsymbol{\eta}_j\|}{\|\mathcal{B}(\sigma_j)\mathbf{b}_j\|}, \tag{26}$$

$$\frac{\|\mathbf{c}_i^T\widehat{\mathcal{H}}_r(\mu_i) - \mathbf{c}_i^T\mathcal{H}(\mu_i)\|}{\|\mathbf{c}_i^T\mathcal{H}(\mu_i)\|} \leqslant cond_{\mathcal{C}}(\mathbf{c}_i^T\mathcal{H}(\mu_i))\frac{\sin\Theta\,(\mathcal{B}_m(\mu_i),\ \widehat{\mathcal{V}}_r)}{\cos\Theta\,(\widehat{\Omega}_r(\mu_i),\ \widehat{\mathcal{V}}_r)}\,\frac{\|\boldsymbol{\xi}_i\|}{\|\mathbf{c}_i^T\mathcal{C}(\mu_i)\|}. \tag{27}$$

If $\mu_i = \sigma_i$ *then,*

$$|\mathbf{c}_i^T\widehat{\mathcal{H}}_r(\mu_i)\mathbf{b}_i - \mathbf{c}_i^T\mathcal{H}(\mu_i)\mathbf{b}_i| \leqslant \frac{\|\mathcal{K}(\mu_i)^{-1}\|\,\|\boldsymbol{\eta}_i\|\,\|\boldsymbol{\xi}_i\|}{\max(\cos\Theta(\widehat{\mathfrak{P}}_r(\mu_i),\ \widehat{\mathcal{W}}_r)\,,\cos\Theta\,(\widehat{\Omega}_r(\mu_i),\ \widehat{\mathcal{V}}_r))}. \tag{28}$$

and

$$\begin{aligned}|\mathbf{c}_i^T\mathcal{H}'(\mu_i)\mathbf{b}_i - \mathbf{c}_i^T\widehat{\mathcal{H}}_r{}'(\mu_i)\mathbf{b}_i| \leqslant M\Bigg(&\frac{\|\boldsymbol{\eta}_i\|}{\cos\Theta(\widehat{\mathfrak{P}}_r(\mu_i),\ \widehat{\mathcal{W}}_r)} + \frac{\|\boldsymbol{\xi}_i\|}{\cos\Theta(\widehat{\Omega}_r(\mu_i),\ \widehat{\mathcal{V}}_r)}\\ &+ \frac{\|\boldsymbol{\eta}_i\|}{\cos\Theta(\widehat{\mathfrak{P}}_r(\mu_i),\ \widehat{\mathcal{W}}_r)}\,\frac{\|\boldsymbol{\xi}_i\|}{\cos\Theta(\widehat{\Omega}_r(\mu_i),\ \widehat{\mathcal{V}}_r)}\Bigg)\end{aligned} \tag{29}$$

with $M = max(\left\|\frac{d}{ds}\left[\mathbf{c}_i^T\mathcal{C}\,\mathcal{K}^{-1}\right]\Big|_{\mu_i}\right\|, \left\|\frac{d}{ds}\left[\mathcal{K}^{-1}\mathcal{B}\mathbf{b}_i\right]\Big|_{\mu_i}\right\|, \left\|\frac{d}{ds}\left[\mathcal{K}^{-1}\right]\Big|_{\mu_i}\right\|)$.

Proof. From (17), $\widehat{\mathbf{v}}_j = \mathcal{K}(\sigma_j)^{-1}(\mathcal{B}(\sigma_j)\mathbf{b}_j + \boldsymbol{\eta}_j)$, which implies then that $\mathcal{K}(\sigma_j)\widehat{\mathbf{v}}_j = \mathcal{B}(\sigma_j)\mathbf{b}_j + \boldsymbol{\eta}_j \in \widehat{\mathfrak{P}}_r(\sigma_j)$ and $\left(\mathbf{I} - \widehat{\mathcal{P}}_r(\sigma_j)\right)\left(\mathcal{B}(\sigma_j)\mathbf{b}_j + \boldsymbol{\eta}_j\right) = 0$, which may be rearranged to obtain

$$\left(\mathbf{I} - \widehat{\mathcal{P}}_r(\sigma_j)\right)\mathcal{B}(\sigma_j)\mathbf{b}_j = -\left(\mathbf{I} - \widehat{\mathcal{P}}_r(\sigma_j)\right)\boldsymbol{\eta}_j. \tag{30}$$

Let $\widehat{\boldsymbol{\Pi}}$ be the orthogonal projector taking $\mathbb{C}^n$ onto $\widehat{\mathcal{W}}_r = \mathrm{Ker}\left(\widehat{\mathcal{P}}_r(s)\right)^{\perp}$. One may directly verify that $\mathbf{I} - \widehat{\mathcal{P}}_r(s) = (\mathbf{I} - \widehat{\boldsymbol{\Pi}})\left(\mathbf{I} - \widehat{\mathcal{P}}_r(s)\right)$, and

$$\begin{aligned}\widehat{\mathcal{H}}_r(\sigma_j)\mathbf{b}_j - \mathcal{H}(\sigma_j)\mathbf{b}_j &= -\mathcal{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}\left(\mathbf{I} - \widehat{\mathcal{P}}_r(\sigma_j)\right)\mathcal{B}(\sigma_j)\mathbf{b}_j\\ &= \mathcal{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}\left(\mathbf{I} - \widehat{\mathcal{P}}_r(\sigma_j)\right)\boldsymbol{\eta}_j\\ &= \mathcal{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}(\mathbf{I} - \widehat{\boldsymbol{\Pi}})\left(\mathbf{I} - \widehat{\mathcal{P}}_r(\sigma_j)\right)\boldsymbol{\eta}_j.\end{aligned} \tag{31}$$

Now suppose $\boldsymbol{\Gamma}$ is an orthogonal projector onto $\mathcal{C}_p(\sigma_j)$. We have then that $\mathrm{Ran}(\mathbf{I} - \boldsymbol{\Gamma}) = \mathrm{Ker}(\mathcal{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}$, so that $\mathcal{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}) = \mathcal{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}\boldsymbol{\Gamma}$ and

$$\widehat{\mathcal{H}}_r(\sigma_j)\mathbf{b}_j - \mathcal{H}(\sigma_j)\mathbf{b}_j = \mathcal{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}\boldsymbol{\Gamma}\,(\mathbf{I} - \widehat{\boldsymbol{\Pi}})\left(\mathbf{I} - \widehat{\mathcal{P}}_r(\sigma_j)\right)\boldsymbol{\eta}_j.$$

Taking norms, we obtain an estimate yielding (26):

$$\begin{aligned}\|\widehat{\mathcal{H}}_r(\sigma_j)\mathbf{b}_j - \mathcal{H}(\sigma_j)\mathbf{b}_j\| &\leqslant \|\,(\mathbf{I} - \widehat{\boldsymbol{\Pi}})\,\boldsymbol{\Gamma}\left(\mathcal{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}\right)^T\| \cdot \|\mathbf{I} - \widehat{\mathcal{P}}_r(\sigma_j)\| \cdot \|\boldsymbol{\eta}_j\|\\ &\leqslant \|\mathcal{C}(\sigma_j)\mathcal{K}(\sigma_j)^{-1}\| \cdot \frac{\sin\Theta(\mathcal{C}_p(\sigma_j),\ \widehat{\mathcal{W}}_r)}{\cos\Theta(\widehat{\mathfrak{P}}_r(\sigma_j),\ \widehat{\mathcal{W}}_r)} \cdot \|\boldsymbol{\eta}_j\|\end{aligned}$$

(27) is shown similarly, noting first that

$$\mathbf{c}_i^T \mathcal{C}(\mu_i) \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) = -\boldsymbol{\xi}_i^T \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right). \tag{32}$$

Defining the orthogonal projector, $\widehat{\Xi}$, that takes $\mathbb{C}^n$ onto $\widehat{\mathcal{V}}_r = \text{Ran}\left(\widehat{\Omega}_r(s)\right)$, one observes next $\mathbf{I} - \widehat{\Omega}_r(s) = \left(\mathbf{I} - \widehat{\Omega}_r(s)\right)\left(\mathbf{I} - \widehat{\Xi}\right)$ so that

$$\begin{aligned}
\left\|\mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i) - \mathbf{c}_i^T \mathcal{H}(\mu_i)\right\| &= \|\mathbf{c}_i^T \mathcal{C}(\mu_i) \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) \mathcal{K}(\mu_i)^{-1} \mathcal{B}(\mu_i)\| \\
&\leqslant \|\boldsymbol{\xi}_i^T \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) \left(\mathbf{I} - \widehat{\Xi}\right) \mathcal{K}(\mu_i)^{-1} \mathcal{B}(\mu_i)\| \\
&\leqslant \|\boldsymbol{\xi}_i\| \cdot \|\mathbf{I} - \widehat{\Omega}_r(\mu_i)\| \cdot \|\left(\mathbf{I} - \widehat{\Xi}\right) \mathcal{K}(\mu_i)^{-1} \mathcal{B}(\mu_i)\| \\
&\leqslant \|\mathcal{K}(\mu_i)^{-1} \mathcal{B}(\mu_i)\| \cdot \frac{\sin\Theta(\mathcal{B}_m(\mu_i),\ \widehat{\mathcal{V}}_r)}{\cos\Theta(\widehat{\Omega}_r(\mu_i),\ \widehat{\mathcal{V}}_r)} \cdot \|\boldsymbol{\xi}_i\|.
\end{aligned}$$

When $\mu_i = \sigma_i$, we have

$$\begin{aligned}
\mathbf{c}_i^T \mathcal{H}(\mu_i)\mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i)\mathbf{b}_i &= \mathbf{c}_i^T \mathcal{C}(\mu_i) \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) \mathcal{K}(\mu_i)^{-1} \left(\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)\right) \mathcal{B}(\mu_i)\mathbf{b}_i \\
&= \boldsymbol{\xi}_i^T \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) \mathcal{K}(\mu_i)^{-1} \left(\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)\right) \boldsymbol{\eta}_i \\
&= \begin{cases} \boldsymbol{\xi}_i^T \mathcal{K}(\mu_i)^{-1} \left(\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)\right) \boldsymbol{\eta}_i, & \text{or} \\ \boldsymbol{\xi}_i^T \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) \mathcal{K}(\mu_i)^{-1} \boldsymbol{\eta}_i, \end{cases}
\end{aligned}$$

leading then to two estimates:

$$|\mathbf{c}_i^T \mathcal{H}(\mu_i)\mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i)\mathbf{b}_i| \leqslant \|\boldsymbol{\xi}_i\| \cdot \|\boldsymbol{\eta}_i\| \cdot \|\mathcal{K}(\mu_i)^{-1}\| \cdot \|\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)\|$$

and

$$|\mathbf{c}_i^T \mathcal{H}(\mu_i)\mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r(\mu_i)\mathbf{b}_i| \leqslant \|\boldsymbol{\xi}_i\| \cdot \|\boldsymbol{\eta}_i\| \cdot \|\mathcal{K}(\mu_i)^{-1}\| \cdot \|\mathbf{I} - \widehat{\Omega}_r(\mu_i)\|.$$

These can be combined to yield (28).

The last inequality comes from using (23) with $s = \mu_i$:

$$\begin{aligned}
\mathbf{c}_i^T \mathcal{H}'(\mu_i)\mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r'(\mu_i)\mathbf{b}_i = &\frac{d}{ds}\left[\mathbf{c}_i^T \mathcal{C}\, \mathcal{K}^{-1}\right]\Big|_{\mu_i} \left(\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)\right) \mathcal{B}(\mu_i)\mathbf{b}_i \\
&+ \mathbf{c}_i^T \mathcal{C}(\mu_i) \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) \frac{d}{ds}\left[\mathcal{K}^{-1}\mathcal{B}\mathbf{b}_i\right]\Big|_{\mu_i} \\
&- \mathbf{c}_i^T \mathcal{C}(\mu_i) \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) \frac{d}{ds}\left[\mathcal{K}^{-1}\right]\Big|_{\mu_i} \left(\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)\right) \mathcal{B}(\mu_i)\mathbf{b}_i.
\end{aligned}$$

Then from (30), (32), and the Cauchy–Schwarz inequality

$$\begin{aligned}
\left|\mathbf{c}_i^T \mathcal{H}'(\mu_i)\mathbf{b}_i - \mathbf{c}_i^T \widehat{\mathcal{H}}_r'(\mu_i)\mathbf{b}_i\right| \leqslant &\left|\frac{d}{ds}\left[\mathbf{c}_i^T \mathcal{C}\, \mathcal{K}^{-1}\right]\Big|_{\mu_i} \left(\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)\right) \boldsymbol{\eta}_i\right| \\
&+ \left|\boldsymbol{\xi}_i^T \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) \frac{d}{ds}\left[\mathcal{K}^{-1}\mathcal{B}\mathbf{b}_i\right]\Big|_{\mu_i}\right| \\
&+ \left|\boldsymbol{\xi}_i^T \left(\mathbf{I} - \widehat{\Omega}_r(\mu_i)\right) \frac{d}{ds}\left[\mathcal{K}^{-1}\right]\Big|_{\mu_i} \left(\mathbf{I} - \widehat{\mathcal{P}}_r(\mu_i)\right) \boldsymbol{\eta}_i\right| \\
\leqslant &\left\|\frac{d}{ds}\left[\mathbf{c}_i^T \mathcal{C}\, \mathcal{K}^{-1}\right]\Big|_{\mu_i}\right\| \cdot \frac{\|\boldsymbol{\eta}_i\|}{\cos\Theta(\widehat{\mathcal{P}}_r(\mu_i),\ \widehat{\mathcal{W}}_r)} \\
&+ \frac{\|\boldsymbol{\xi}_i\|}{\cos\Theta(\widehat{\Omega}_r(\mu_i),\ \widehat{\mathcal{V}}_r)} \cdot \left\|\frac{d}{ds}\left[\mathcal{K}^{-1}\mathcal{B}\mathbf{b}_i\right]\Big|_{\mu_i}\right\| \\
&+ \left\|\frac{d}{ds}\left[\mathcal{K}^{-1}\right]\Big|_{\mu_i}\right\| \cdot \frac{\|\boldsymbol{\eta}_i\|}{\cos\Theta(\widehat{\mathcal{P}}_r(\mu_i),\ \widehat{\mathcal{W}}_r)} \frac{\|\boldsymbol{\xi}_i\|}{\cos\Theta(\widehat{\Omega}_r(\mu_i),\ \widehat{\mathcal{V}}_r)}
\end{aligned}$$

which yields the conclusion. □

Consider the effect of solving (11) and (12) approximately with successively increasing levels of accuracy that force the residual norms to zero, $\|\boldsymbol{\eta}_j\| \to 0$ and $\|\boldsymbol{\xi}_i\| \to 0$. The multiplicative behavior of the error bound (28) with respect to $\|\boldsymbol{\eta}_j\|$ and $\|\boldsymbol{\xi}_i\|$ contrasts with the additive behavior seen in (26) and (27) and suggests some potential benefit in using the same interpolation points for both left and right interpolation, i.e., choosing $\mu_i = \sigma_i$ for $i = 1, \ldots, r$. Note that this choice also forces convergent (bitangential) derivative interpolation as shown in (29). Indeed, choosing $\mu_i = \sigma_i$ for $i = 1, \ldots, r$ is a *necessary* condition for forming $\mathcal{H}_2$-optimal interpolatory reduced order models for first-order descriptor realizations, as we discuss in Section 5 (see also [16]). Beyond this, there can be notable computational advantages in choosing $\mu_i = \sigma_i$, since the linear systems to be solved in (11) and (12) then have the same coefficient matrix; allowing one potentially to reuse factorizations and preconditioners.

Certain applications require the retention of structural properties such as symmetry in passing from $\mathcal{K}$ to $\widehat{\mathcal{K}}_r$ and one is compelled to choose $\widehat{\mathbf{W}}_r = \widehat{\mathbf{V}}_r$ ("one-sided" model reduction), so the vectors $\{\widehat{\mathbf{w}}_1, \ldots, \widehat{\mathbf{w}}_r\}$ might not be approximate solutions to (11) in the usual sense. Nonetheless, the behavior of the interpolation error is still governed by (26) and (27). We explore this in the following numerical example.

We illustrate the character of the results given in Theorem 3.1, bounding the response error at the nominal interpolation points caused by inexact solves in (11) and (12). To this end, we consider a delay differential equation of the form introduced in (4) taking $n = 2000$, $m = p = 1$ and $\tau_{inp} = \tau_{out} = 0$. The coefficient matrices for the full-order model in (4) were taken from [6]. We construct multiple reduced models all of order $r = 3$, solving (11) and (12) with different levels of accuracy. We chose three logarithmically spaced values, $\sigma_1 = 0.001$, $\sigma_2 = 0.0316$, $\sigma_3 = 1.0$, and fixed them as interpolation points. We then obtained approximate solutions of varying accuracy to (11) and (12) in a manner described in more detail below, assembled the inexact interpolation basis matrices, $\widehat{\mathbf{V}}_r$ and $\widehat{\mathbf{W}}_r$, and obtained reduced models of order $r = 3$ having the same internal delay structure as the original system:

$$\begin{aligned}\widehat{\mathcal{H}}_r(s) &= \widehat{\mathcal{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathcal{B}}_r(s)\\ &= \mathbf{C}\widehat{\mathbf{V}}_r\left(s\,\widehat{\mathbf{W}}_r^T\mathbf{E}\widehat{\mathbf{V}}_r - \widehat{\mathbf{W}}_r^T\mathbf{A}_0\widehat{\mathbf{V}}_r - e^{-s\,\tau_{sys}}\,\widehat{\mathbf{W}}_r^T\mathbf{A}_1\widehat{\mathbf{V}}_r\right)^{-1}\widehat{\mathbf{W}}_r^T\mathbf{B}.\end{aligned}$$

We considered both the usual "two-sided" model reduction process that involves approximate solution of both (11) and (12) and the "one-sided" process that involves approximate solutions only to (12) to generate $\widehat{\mathbf{V}}_r$ and then assigning $\widehat{\mathbf{W}}_r = \widehat{\mathbf{V}}_r$. Linear systems were solved with GMRES terminating with a final relative residual below a uniform tolerance denoted by ε.

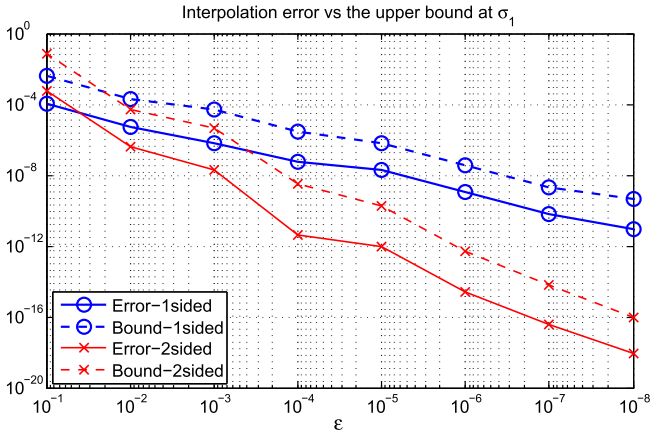


Fig. 1. Behavior of interpolation error and upper bounds vs. ε.

We generated reduced order models in this way, varying the relative residual tolerance ε from 10^{-1} down to 10^{-8}. Fig. 1 below shows the resulting interpolation errors $|\mathcal{H}(\sigma_1) - \widehat{\mathcal{H}}_r(\sigma_1)|$ and bounds from Eqs. (26) and (28) for one-sided and two-sided cases, respectively, as ε varies. Observe that the bounds in Theorem 3.1 predict the convergence behavior of the true error quite well; the rates (slopes) are matched almost exactly. Note also that the interpolation error decays much faster for two-sided reduction than for one-sided reduction. Indeed, the ratio of the two errors is close to ε, i.e., for a given tolerance ε, the interpolation error for two-sided reduction is approximately ε times smaller than the interpolation error for one-sided reduction.

Analogous results regarding behavior of the bounds and interpolation error are observed at σ_2 and σ_3 and so are omitted for brevity.

3.2. Global error bounds

Thus far we have focussed on the extent to which interpolation properties are lost in the computed reduced models when inexact solves are introduced into the process, considering in effect *local* error bounds. Clearly, it is important to understand the effect of inexact solves on the overall *global* quality of the reduced order model. There are two commonly used measures for closeness of two conforming dynamical systems (i.e., those with the same input and output dimensions):

$$\text{the } \mathcal{H}_2\text{-norm:} \quad \|\mathcal{H} - \mathcal{G}\|_{\mathcal{H}_2} = \frac{1}{2\pi}\int_{-\infty}^{\infty} \|\mathcal{H}(\iota\omega) - \mathcal{G}(\iota\omega)\|_F^2 \, d\omega,$$
$$\text{the } \mathcal{H}_\infty\text{-norm:} \quad \|\mathcal{H} - \mathcal{G}\|_{\mathcal{H}_\infty} = \max_{\omega\in\mathbb{R}} \|\mathcal{H}(\iota\omega) - \mathcal{G}(\iota\omega)\|_2 \, .$$

Since reduced models are completely determined by the subspaces, $\mathcal{V}_r$ and $\mathcal{W}_r$, as shown in (8), we first evaluate (in Theorem 3.2) how much inexact interpolatory subspaces, $\widehat{\mathcal{V}}_r$ and $\widehat{\mathcal{W}}_r$, can deviate from the corresponding true subspaces, $\mathcal{V}_r$ and $\mathcal{W}_r$, as a result of inexact solves. The effect of this deviation on the resulting model reduction (forward) error will be shown in Theorem 3.3. In this way, we are able to connect model reduction error to observable quantities that are associated with inexact solves, such as the relative stopping criterion ε.

Theorem 3.2. *Let the columns of $\mathbf{V}_r$ and $\widehat{\mathbf{V}}_r$ be exact and approximate solutions to (12) and the columns of $\mathbf{W}_r$ and $\widehat{\mathbf{W}}_r$ be exact and approximate solutions to (11). Suppose approximate solutions are computed to a relative residual tolerance of $\varepsilon > 0$, so that $\|\boldsymbol{\eta}_i\| \leqslant \varepsilon \|\mathcal{B}(\sigma_i)\mathbf{b}_i\|$ and $\|\boldsymbol{\xi}_i\| \leqslant \varepsilon \, \|\mathcal{C}(\mu_i)^T \mathbf{c}_i\|$, where the residuals $\boldsymbol{\eta}_i$ and $\boldsymbol{\xi}_i$ are defined in (16).*

Denoting the associated subspaces as $\mathcal{V}_r$, $\widehat{\mathcal{V}}_r$, $\mathcal{W}_r$ and $\widehat{\mathcal{W}}_r$ then

$$\sin\Theta(\widehat{\mathcal{V}}_r,\ \mathcal{V}_r) \leqslant \frac{\varepsilon\sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{V}}_r\mathbf{D}_v)}, \tag{33}$$
$$\sin\Theta(\widehat{\mathcal{W}}_r,\ \mathcal{W}_r) \leqslant \frac{\varepsilon\sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{W}}_r\mathbf{D}_w)} \tag{34}$$

where $\mathbf{D}_v$ and $\mathbf{D}_w$ are diagonal scaling matrices defined as

$$\mathbf{D}_v = \mathit{diag}\left((\|\mathcal{K}(\sigma_1)^{-1}\|\,\|\mathcal{B}(\sigma_1)\mathbf{b}_1\|)^{-1},\ \ldots,\ (\|\mathcal{K}(\sigma_r)^{-1}\|\,\|\mathcal{B}(\sigma_r)\mathbf{b}_r\|)^{-1}\right) \text{ and}$$
$$\mathbf{D}_w = \mathit{diag}\left((\|\mathcal{K}(\mu_1)^{-1}\|\,\|\mathcal{C}(\mu_1)^T\mathbf{c}_1\|)^{-1},\ \ldots,\ (\|\mathcal{K}(\mu_r)^{-1}\|\,\|\mathcal{C}(\mu_r)^T\mathbf{c}_r\|)^{-1}\right)$$

and $\varsigma_{\min}(\mathbf{M})$ denotes the smallest singular value of the matrix $\mathbf{M}$.

Proof. We prove (33). The proof of (34) is similar.

Write $\widehat{\mathbf{V}}_r = \mathbf{V}_r + \mathbf{E}$ with $\mathbf{E} = [\mathcal{K}(\sigma_1)^{-1}\boldsymbol{\eta}_1, \ldots, \mathcal{K}(\sigma_r)^{-1}\boldsymbol{\eta}_r]$. Then

$$\begin{aligned}
\sin\Theta(\widehat{\mathcal{V}}_r,\ \mathcal{V}_r) &= \max_{\hat{\mathbf{v}}\in\widehat{\mathcal{V}}_r}\min_{\mathbf{v}\in\mathcal{V}_r}\frac{\|\mathbf{v}-\hat{\mathbf{v}}\|}{\|\hat{\mathbf{v}}\|}\\
&= \max_{x_i}\min_{z_i}\frac{\|\sum_{i=1}^r z_i\mathcal{K}(\sigma_i)^{-1}\mathcal{B}(\sigma_i)\mathbf{b}_i - \sum_{i=1}^r x_i\widehat{\mathbf{v}}_i\|}{\|\sum_{i=1}^r x_i\widehat{\mathbf{v}}_i\|}\\
&= \max_{x_i}\min_{z_i}\frac{\|\sum_{i=1}^r (z_i-x_i)\mathcal{K}(\sigma_i)^{-1}\mathcal{B}(\sigma_i)\mathbf{b}_i - x_i\mathcal{K}(\sigma_i)^{-1}\boldsymbol{\eta}_i\|}{\|\sum_{i=1}^r x_i\widehat{\mathbf{v}}_i\|}\\
&\leqslant \max_{x_i}\frac{\|\sum_{i=1}^r x_i\mathcal{K}(\sigma_i)^{-1}\boldsymbol{\eta}_i\|}{\|\sum_{i=1}^r x_i\widehat{\mathbf{v}}_i\|} = \max_{\mathbf{x}}\frac{\|\mathbf{E}\mathbf{x}\|}{\|\widehat{\mathbf{V}}_r\mathbf{x}\|} = \max_{\mathbf{x}}\frac{\|\mathbf{E}\mathbf{D}\mathbf{x}\|}{\|\widehat{\mathbf{V}}_r\mathbf{D}\mathbf{x}\|},
\end{aligned}$$

where $\mathbf{D} = \mathrm{diag}(d_1,\ \ldots,\ d_r)$ is a diagonal matrix with positive diagonal entries, $d_i > 0$, that are fixed but for the moment unspecified.

Note that

$$\begin{aligned}
\|\mathbf{E}\mathbf{D}\mathbf{x}\| &\leqslant \|\mathbf{E}\mathbf{D}\|\,\|\mathbf{x}\| \leqslant \sqrt{r}\,\|\mathbf{x}\|\ \max_i\left(d_i\|\mathcal{K}(\sigma_i)^{-1}\boldsymbol{\eta}_i\|\right)\\
&\leqslant \sqrt{r}\,\|\mathbf{x}\|\ \max_i\left(d_i\|\mathcal{K}(\sigma_i)^{-1}\|\,\|\boldsymbol{\eta}_i\|\right).
\end{aligned}$$

Thus we have,

$$\sin\Theta(\widehat{\mathcal{V}}_r,\ \mathcal{V}_r) \leqslant \sqrt{r}\,\frac{\max_i\left(d_i\|\mathcal{K}(\sigma_i)^{-1}\|\,\|\boldsymbol{\eta}_i\|\right)}{\min_{\mathbf{x}}\left(\|\widehat{\mathbf{V}}_r\mathbf{D}\mathbf{x}\|/\|\mathbf{x}\|\right)} = \sqrt{r}\,\frac{\max_i\left(d_i\|\mathcal{K}(\sigma_i)^{-1}\|\,\|\boldsymbol{\eta}_i\|\right)}{\varsigma_{\min}(\widehat{\mathbf{V}}_r\mathbf{D})}. \tag{35}$$

This bound is valid for any choice of diagonal scalings, $\mathbf{D}$, so we can minimize the right hand side of (35) with respect to $d_1,\ \ldots,\ d_r$. The *Column Equilibration Theorem* of van der Sluis [28] asserts that the optimal choice of $d_1,\ \ldots,\ d_r$ is such that $d_i\|\mathcal{K}(\sigma_i)^{-1}\|\,\|\boldsymbol{\eta}_i\| = C$, independent of $i = 1,\ \ldots,\ r$. If inexact solves terminate with residuals satisfying $\|\boldsymbol{\eta}_i\| \approx \varepsilon\,\|\mathcal{B}(\sigma_i)\mathbf{b}_i\|$ then we may take $C = \varepsilon$ and $d_i = \left(\|\mathcal{K}(\sigma_i)^{-1}\|\,\|\mathcal{B}(\sigma_i)\mathbf{b}_i\|\right)^{-1}$ to achieve the best bound possible with the information given. This leads to (33). □

As a practical matter, the column scalings used in (33) and (34) will not be computationally feasible in realistic settings. If instead we scale the columns of $\widehat{\mathbf{V}}_r$ and $\widehat{\mathbf{W}}_r$ to have unit norm (cheap !) – taking $\widetilde{\mathbf{D}}_v = \mathrm{diag}\,(1/\|\widehat{\mathbf{v}}_1\|,\ \ldots,\ 1/\|\widehat{\mathbf{v}}_r\|)$ and $\widetilde{\mathbf{D}}_w = \mathrm{diag}\,(1/\|\widehat{\mathbf{w}}_1\|,\ \ldots,\ 1/\|\widehat{\mathbf{w}}_r\|)$, the bound for (33) degrades to

$$\sin\Theta(\widehat{\mathcal{V}}_r,\ \mathcal{V}_r) \leqslant \max_i \kappa_2\,(\mathcal{K}(\sigma_i), \widehat{\mathbf{v}}_i)\,\frac{\varepsilon\,\sqrt{r}}{\varsigma_{\min}(\widehat{\mathbf{V}}_r\widetilde{\mathbf{D}}_v)},$$

where $\kappa_2\,(\mathcal{K}(\sigma_i), \widehat{\mathbf{v}}_i) = \dfrac{\|\mathcal{K}(\sigma_i)^{-1}\|\,\|\mathcal{B}(\sigma_i)\mathbf{b}_i\|}{\|\widehat{\mathbf{v}}_i\|} > 1$ is the *condition number of the linear system* (12). A similar expression holds for $\sin\Theta(\widehat{\mathcal{W}}_r,\ \mathcal{W}_r)$. In many cases, these condition numbers have only modest magnitude and the bounds (33) and (34) remain descriptive.

Theorem 3.3. *Let the columns of $\mathbf{V}_r$ and $\widehat{\mathbf{V}}_r$ be exact and approximate solutions to (12) and the columns of $\mathbf{W}_r$ and $\widehat{\mathbf{W}}_r$ be exact and approximate solutions to (11). Let the associated subspaces be denoted as $\mathcal{V}_r$, $\widehat{\mathcal{V}}_r$, $\mathcal{W}_r$ and $\widehat{\mathcal{W}}_r$ and the associated reduced order systems be denoted as $\mathcal{H}_r(s)$ (exact) and $\widehat{\mathcal{H}}_r(s)$ (inexact). Then*

$$\frac{\|\mathcal{H}_r - \widehat{\mathcal{H}}_r\|_{\mathcal{H}_\infty}}{\frac{1}{2}\left(\|\mathcal{H}_r\|_{\mathcal{H}_\infty} + \|\widehat{\mathcal{H}}_r\|_{\mathcal{H}_\infty}\right)} \leqslant M\ \max\left(\sin\Theta(\widehat{\mathcal{V}}_r,\ \mathcal{V}_r),\ \sin\Theta(\widehat{\mathcal{W}}_r,\ \mathcal{W}_r)\right),$$

where

$$M = 2\,\max\left(\frac{\max_{\omega\in\mathbb{R}}\, cond_{\mathcal{C}}(\mathcal{H}_r(\iota\omega))}{\min_{\omega\in\mathbb{R}}\cos\Theta(\widehat{\Omega}_r(\iota\omega),\ \widehat{\mathcal{V}}_r)},\ \frac{\max_{\omega\in\mathbb{R}}\, cond_{\mathcal{B}}(\widehat{\mathcal{H}}_r(\iota\omega))}{\min_{\omega\in\mathbb{R}}\cos\Theta(\mathfrak{P}_r(\iota\omega),\ \mathcal{W}_r)}\right)$$

and

$$cond_{\mathcal{B}}(\widehat{\mathcal{H}}_r(s)) = \frac{\|\widehat{\mathcal{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathbf{W}}_r^T\|\ \|\mathcal{B}(s)\|}{\|\widehat{\mathcal{H}}_r(s)\|}$$
$$cond_{\mathcal{C}}(\mathcal{H}_r(s)) = \frac{\|\mathcal{C}(s)\|\ \|\mathbf{V}_r\mathcal{K}_r(s)^{-1}\mathcal{B}_r(s)\|}{\|\mathcal{H}_r(s)\|}.$$

Proof. Note that for all $s\in\mathbb{C}$ for which $\mathcal{H}_r$ and $\widehat{\mathcal{H}}_r$ are both analytic,

$$\begin{aligned}\|\mathcal{H}_r(s)-\widehat{\mathcal{H}}_r(s)\| &= \|\mathcal{C}(s)\left(\mathbf{V}_r\mathcal{K}_r(s)^{-1}\mathbf{W}_r^T-\widehat{\mathbf{V}}_r\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathbf{W}}_r^T\right)\mathcal{B}(s)\|\\ &= \|\mathcal{C}(s)\left(\Omega_r(s)-\widehat{\Omega}_r(s)\right)\mathcal{K}(s)^{-1}\mathcal{B}(s)\|\\ &= \|\mathcal{C}(s)\left(\left(\mathbf{I}-\widehat{\Omega}_r(s)\right)\Omega_r(s)-\widehat{\Omega}_r(s)\left(\mathbf{I}-\Omega_r(s)\right)\right)\mathcal{K}(s)^{-1}\mathcal{B}(s)\|.\end{aligned}$$

So,

$$\begin{aligned}\|\mathcal{H}_r(s)-\widehat{\mathcal{H}}_r(s)\| &\leqslant \|\mathcal{C}(s)\left(\mathbf{I}-\widehat{\Omega}_r(s)\right)\Omega_r(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)\|\\ &\quad+\|\mathcal{C}(s)\widehat{\Omega}_r(s)\left(\mathbf{I}-\Omega_r(s)\right)\mathcal{K}(s)^{-1}\mathcal{B}(s)\|\\ &\leqslant \|\mathcal{C}(s)\left(\mathbf{I}-\widehat{\Omega}_r(s)\right)\Omega_r(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)\|\\ &\quad+\|\mathcal{C}(s)\mathcal{K}(s)^{-1}\widehat{\mathcal{P}}_r(s)\left(\mathbf{I}-\mathcal{P}_r(s)\right)\mathcal{B}(s)\|\\ &\leqslant \|\mathcal{C}(s)\left(\mathbf{I}-\widehat{\Omega}_r(s)\right)\ \left(\mathbf{I}-\widehat{\Xi}\right)\Xi\,\Omega_r(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)\|\\ &\quad+\|\mathcal{C}(s)\mathcal{K}(s)^{-1}\widehat{\mathcal{P}}_r(s)\,\widehat{\mathbf{\Pi}}\left(\mathbf{I}-\mathbf{\Pi}\right)\ \left(\mathbf{I}-\mathcal{P}_r(s)\right)\mathcal{B}(s)\|\\ &\leqslant \|\mathcal{C}(s)\|\,\|\mathbf{I}-\widehat{\Omega}_r(s)\|\ \|\left(\mathbf{I}-\widehat{\Xi}\right)\Xi\|\ \|\Omega_r(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)\|\\ &\quad+\|\mathcal{C}(s)\mathcal{K}(s)^{-1}\widehat{\mathcal{P}}_r(s)\|\,\|\widehat{\mathbf{\Pi}}\left(\mathbf{I}-\mathbf{\Pi}\right)\|\,\|\mathbf{I}-\mathcal{P}_r(s)\|\,\|\mathcal{B}(s)\|\\ &\leqslant \|\mathcal{C}(s)\|\frac{\sin\Theta(\widehat{\mathcal{V}}_r,\ \mathcal{V}_r)}{\cos\Theta(\widehat{\Omega}_r(s),\ \widehat{\mathcal{V}}_r)}\|\Omega_r(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)\|\\ &\quad+\|\mathcal{C}(s)\mathcal{K}(s)^{-1}\widehat{\mathcal{P}}_r(s)\|\,\frac{\sin\Theta(\widehat{\mathcal{W}}_r,\ \mathcal{W}_r)}{\cos\Theta(\mathfrak{P}_r(s),\ \mathcal{W}_r)}\,\|\mathcal{B}(s)\|\\ &\leqslant \mathrm{cond}_{\mathcal{C}}(\mathcal{H}_r(s))\frac{\sin\Theta(\widehat{\mathcal{V}}_r,\ \mathcal{V}_r)}{\cos\Theta(\widehat{\Omega}_r(s),\ \widehat{\mathcal{V}}_r)}\|\mathcal{H}_r(s)\|\\ &\quad+\|\widehat{\mathcal{H}}_r(s)\|\,\frac{\sin\Theta(\widehat{\mathcal{W}}_r,\ \mathcal{W}_r)}{\cos\Theta(\mathfrak{P}_r(s),\ \mathcal{W}_r)}\,\mathrm{cond}_{\mathcal{B}}(\widehat{\mathcal{H}}_r(s)).\end{aligned}$$

Maximizing over $s=\iota\omega$ with $\omega\in\mathbb{R}$ gives

$$\begin{aligned}\|\mathcal{H}_r-\widehat{\mathcal{H}}_r\|_{\mathcal{H}_\infty} &\leqslant \max_{\omega\in\mathbb{R}}\mathrm{cond}_{\mathcal{C}}(\mathcal{H}_r(\iota\omega))\frac{\sin\Theta(\widehat{\mathcal{V}}_r,\ \mathcal{V}_r)}{\min_{\omega\in\mathbb{R}}\cos\Theta(\widehat{\Omega}_r(s),\ \widehat{\mathcal{V}}_r)}\|\mathcal{H}_r\|_{\mathcal{H}_\infty}\\ &\quad+\max_{\omega\in\mathbb{R}}\mathrm{cond}_{\mathcal{B}}(\widehat{\mathcal{H}}_r(\iota\omega))\frac{\sin\Theta(\widehat{\mathcal{W}}_r,\ \mathcal{W}_r)}{\min_{\omega\in\mathbb{R}}\cos\Theta(\mathfrak{P}_r(s),\ \mathcal{W}_r)}\,\|\widehat{\mathcal{H}}_r\|_{\mathcal{H}_\infty}\end{aligned}$$

which leads immediately to the conclusion. □

Table 2
The relative error $\dfrac{\|\mathcal{H}_r - \widehat{\mathcal{H}}_r\|_{\mathcal{H}_\infty}}{\|\mathcal{H}_r\|_{\mathcal{H}_\infty}}$ as ε varies.

ε	$\mathcal{H}_2$-optimal $\{\sigma_i\}$	*Ad hoc* $\{\sigma_i\}$
10^{-1}	7.22×10^{-1}	5.05×10^{-1}
10^{-2}	2.00×10^{-1}	1.64×10^{-1}
10^{-3}	4.27×10^{-2}	4.11×10^{-1}
10^{-4}	1.07×10^{-2}	2.38×10^{-1}
10^{-5}	2.76×10^{-4}	5.62×10^{-1}
10^{-6}	2.56×10^{-5}	2.13×10^{-2}
10^{-7}	2.91×10^{-6}	3.52×10^{-3}
10^{-8}	1.51×10^{-7}	6.18×10^{-5}
10^{-9}	2.07×10^{-8}	1.76×10^{-5}
10^{-10}	2.17×10^{-9}	5.15×10^{-6}

3.3. Illustrative examples

The process to be modeled arises in cooling within a rolling mill and is modeled as boundary control of a two dimensional heat equation. A finite element discretization results in a descriptor system of the form

$$\mathbf{E}\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad y(t) = \mathbf{C}\mathbf{x}(t),$$

where $\mathbf{A}, \mathbf{E} \in \mathbb{R}^{5177\times 5177}$, $\mathbf{B} \in \mathbb{R}^{5177\times 7}$, $\mathbf{C} \in \mathbb{R}^{6\times 5177}$. For simplicity, we focus on a SISO full-order subsystem that relates the sixth input to the second output. For details regarding the modeling, discretization, optimal control design, and model reduction, see [8,9].

We show the results of interpolatory model reduction using an *ad hoc* choice of interpolation points: 6 logarithmically spaced points between $10^{0.5}$ and 10; and an $\mathcal{H}_2$-optimal choice of interpolation points obtained by the method of [16]. For each case, we reduce the system order to $r = 6$ using first *exact* interpolatory model reduction (i.e., the linear systems are solved directly) and then with inexact model reduction with varying choices of termination criteria. The resulting reduced-order models are denoted by $\mathcal{H}_r(s)$ and $\widehat{\mathcal{H}}_r(s)$, respectively.

To see the effect of the choice of interpolation points on the underlying model reduction problem, we vary the relative residual termination tolerance, ε between 10^{-1} and 10^{-10} and show how quickly $\widehat{\mathcal{H}}_r(s)$ converges to $\mathcal{H}_r(s)$ for both the *ad hoc* selection and the $\mathcal{H}_2$-optimal selection of interpolation points. Table 2 shows the relative $\mathcal{H}_\infty$ error between $\widehat{\mathcal{H}}_r(s)$ and $\mathcal{H}_r(s)$ as ε decreases. For the $\mathcal{H}_2$-optimal choice of interpolation points, $\widehat{\mathcal{H}}_r(s)$ converges to $\mathcal{H}_r(s)$ as ε decreases, for the *ad hoc* choice of points, there is almost no improvement in accuracy until $\varepsilon = 10^{-6}$.

The behavior exhibited in Table 2 becomes clearer once we inspect the subspace angles between the exact interpolatory subspaces $\mathcal{V}_r$, $\mathcal{W}_r$ and the inexact ones $\widehat{\mathcal{V}}_r$ and $\widehat{\mathcal{W}}_r$. Table 3 shows the sine of the angle between the exact and inexact interpolatory subspaces as ε varies. While the gap decreases significantly as ε decreases for an $\mathcal{H}_2$-optimal selection of interpolation points, there is a much smaller improvement in the gap with respect to ε for an *ad hoc* choice of points. This behavior will be re-visited in more detail in Section 4.2 revealing that the $\mathcal{H}_2$-optimal (or good) interpolation points are expected to produce reduced order models that are more robust with respect to perturbations due to inexact solves.

4. Backward error

Instead of seeking bounds on how much an inexactly computed reduced model differs from an exactly computed counterpart, one may view an inexactly computed reduced order model as an exactly computed reduced order model of a perturbed full-order system. That is, we wish to find a full-order system

$$\widetilde{\mathcal{H}}(s) = \widetilde{\mathcal{C}}(s)\widetilde{\mathcal{K}}(s)^{-1}\widetilde{\mathcal{B}}(s) \tag{36}$$

Table 3
$r = 6$; $\sin\Theta(\mathcal{V}_r, \widehat{\mathcal{V}}_r)$ and $\sin\Theta(\mathcal{W}_r, \widehat{\mathcal{V}}_r)$ as ε varies.

	$\sin\Theta(\mathcal{V}_r, \widehat{\mathcal{V}}_r)$		$\sin\Theta(\mathcal{W}_r, \widehat{\mathcal{W}}_r)$	
ε	$\mathcal{H}_2$-optimal $\{\sigma_i\}$	*ad hoc* $\{\sigma_i\}$	$\mathcal{H}_2$-optimal $\{\sigma_i\}$	*ad hoc* $\{\sigma_i\}$
10^{-1}	9.85×10^{-1}	9.99×10^{-1}	9.99×10^{-1}	9.99×10^{-1}
10^{-2}	1.99×10^{-1}	9.99×10^{-1}	9.97×10^{-1}	9.93×10^{-1}
10^{-3}	2.36×10^{-2}	9.99×10^{-1}	4.87×10^{-1}	9.83×10^{-1}
10^{-4}	4.39×10^{-3}	9.60×10^{-1}	6.38×10^{-2}	9.99×10^{-1}
10^{-5}	2.72×10^{-4}	5.80×10^{-1}	7.09×10^{-3}	7.20×10^{-1}
10^{-6}	2.90×10^{-5}	4.57×10^{-2}	9.88×10^{-4}	1.19×10^{-1}
10^{-7}	3.46×10^{-6}	6.90×10^{-3}	6.87×10^{-5}	2.00×10^{-2}
10^{-8}	3.85×10^{-7}	7.92×10^{-4}	6.71×10^{-6}	2.26×10^{-3}
10^{-9}	3.63×10^{-8}	1.01×10^{-4}	9.16×10^{-7}	2.60×10^{-4}
10^{-10}	2.71×10^{-9}	1.28×10^{-5}	6.35×10^{-8}	3.10×10^{-5}

so that the inexactly computed reduced model for $\mathcal{H}(s) = \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)$ would be an *exactly* computed interpolatory reduced model for $\widetilde{\mathcal{H}}(s)$. Given left and right tangential interpolation data as in (9) and (10) that has contributed toward producing the inexactly computed interpolatory reduced model $\widehat{\mathcal{H}}_r(s)$, find $\widetilde{\mathcal{H}}(s)$ as in (36) so that

$$\mathbf{c}_i^T\widetilde{\mathcal{H}}(\mu_i) = \mathbf{c}_i^T\widehat{\mathcal{H}}_r(\mu_i) \quad \text{for } i = 1, \ \ldots, \ r, \text{ and}$$
$$\widetilde{\mathcal{H}}(\sigma_j)\mathbf{b}_j = \widehat{\mathcal{H}}_r(\sigma_j)\mathbf{b}_j \quad \text{for } j = 1, \ \ldots, \ r$$

and so that $\widehat{\mathcal{H}}_r$ could have been computed from the perturbed system $\widetilde{\mathcal{H}}$ from the given tangential interpolation data via an exact computation. Specifically, given computed (inexact) projecting bases

$$\widehat{\mathbf{V}}_r = [\, \widehat{\mathbf{v}}_1, \ \ldots, \ \widehat{\mathbf{v}}_r \,] \qquad \widehat{\mathbf{W}}_r^T = \begin{bmatrix} \widehat{\mathbf{w}}_1^T \\ \vdots \\ \widehat{\mathbf{w}}_r^T \end{bmatrix},$$

as in (15), and a resulting (inexact) reduced order coprime realization

$$\widehat{\mathcal{H}}_r(s) = \widehat{\mathcal{C}}_r(s)\widehat{\mathcal{K}}_r(s)^{-1}\widehat{\mathcal{B}}_r(s),$$

find a full-order system $\widetilde{\mathcal{H}}(s) = \widetilde{\mathcal{C}}(s)\widetilde{\mathcal{K}}(s)^{-1}\widetilde{\mathcal{B}}(s)$ so that left and right interpolation conditions hold:

$$\mathbf{c}_i^T\widetilde{\mathcal{C}}(\mu_i) = \widehat{\mathbf{w}}_i^T\widetilde{\mathcal{K}}(\mu_i) \quad \text{for } i = 1, \ \ldots, \ r, \tag{37}$$
$$\widetilde{\mathcal{K}}(\sigma_j)\widehat{\mathbf{v}}_j = \widetilde{\mathcal{B}}(\sigma_j)\mathbf{b}_j \quad \text{for } j = 1, \ \ldots, \ r, \tag{38}$$

and so that

$$\widehat{\mathcal{K}}_r(s) = \widehat{\mathbf{W}}_r^T\widetilde{\mathcal{K}}(s)\widehat{\mathbf{V}}_r, \quad \widehat{\mathcal{B}}_r(s) = \widehat{\mathbf{W}}_r^T\widetilde{\mathcal{B}}(s), \quad \text{and} \quad \widehat{\mathcal{C}}_r(s) = \widetilde{\mathcal{C}}(s)\widehat{\mathbf{V}}_r. \tag{39}$$

There (typically) will be an infinite number of possible systems, $\widetilde{\mathcal{H}}$, that are consistent with the computed reduced system $\widehat{\mathcal{H}}_r$ in this sense — we are interested in those that are *close* to the original system $\mathcal{H}$ with respect to a convenient system norm such as $\mathcal{H}_\infty$ or $\mathcal{H}_2$. In order to proceed, it is convenient to restrict the class of backwardly compatible systems, $\widetilde{\mathcal{H}}$. We consider those that have realizations that are *constant* perturbations from the corresponding original system factors:

$$\widetilde{\mathcal{K}}(s) = \mathcal{K}(s) + \mathbf{F}, \quad \widetilde{\mathcal{B}}(s) = \mathcal{B}(s) + \mathbf{E}, \quad \text{and} \quad \widetilde{\mathcal{C}}(s) = \mathcal{C}(s) + \mathbf{G}, \tag{40}$$

where **E**, **F**, and **G** are *constant* matrices. The conditions (37), (38), and (39) impose constraints on **E**, **F**, and **G**. Indeed, (37) and (38) imply that

$$\widehat{\mathbf{w}}_i^T\mathbf{F} + \boldsymbol{\xi}_i^T = \mathbf{c}_i^T\mathbf{E} \quad \text{for } i = 1, \ \ldots, \ r, \text{ and}$$
$$\mathbf{F}\widehat{\mathbf{v}}_j + \boldsymbol{\eta}_j = \mathbf{G}\mathbf{b}_j \quad \text{for } j = 1, \ \ldots, \ r.$$

(39) implies that

$$\widehat{\mathbf{W}}_r^T\mathbf{F}\widehat{\mathbf{V}}_r = \mathbf{0}, \quad \widehat{\mathbf{W}}_r^T\mathbf{G} = \mathbf{0}, \quad \text{and} \quad \mathbf{E}\widehat{\mathbf{V}}_r = \mathbf{0}.$$

Taken together, we find that backward perturbations of the form (40) can exist only if

$$\boldsymbol{\xi}_i^T\widehat{\mathbf{V}}_r = \mathbf{0} \quad \text{for } i = 1, \ \ldots, \ r, \text{ and } \widehat{\mathbf{W}}_r^T\boldsymbol{\eta}_j = \mathbf{0} \quad \text{for } j = 1, \ \ldots, \ r. \tag{41}$$

Thus, we find constraints on the inexact interpolation residuals $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_j$ in order for a backwardly compatible system of the form (40) to exist. More complicated perturbation classes than (40) may be considered that would allow us to remove the conditions (41), of course, but instead we choose to focus on a computational framework that guarantees (41). The Biconjugate Gradient Algorithm (BiCG) will be an example of an iterative solution strategy that fits this framework [1,4]; others can be constructed without difficulty, although many standard strategies such as GMRES, do not fit this framework.

4.1. The Petrov–Galerkin framework for inexact solves

We have observed above that (41) is necessary for there to be a well-defined backward error of the form (40) to exist. The simplest framework within which one may generate reduced order models that are guaranteed to satisfy this condition involves a Petrov–Galerkin formalism for producing approximate solutions to (11) and (12). For simplicity, we restrict our discussion to the case that $\mu_i = \sigma_i$ (identical left and right interpolation points).

Let $\mathcal{P}_N$ and $\mathcal{Q}_N$ be N-dimensional subspaces of $\mathbb{C}^n$ satisfying a nondegeneracy condition: $(\mathcal{K}(\sigma_i)\mathcal{P}_N)^{\perp} \cap \mathcal{Q}_N = \{0\}$ for all shifts, σ_i to be considered. The *Petrov–Galerkin framework* for generating approximate solutions to the interpolation conditions (11) and (12) proceeds as follows:

$$\begin{aligned} &\text{Find } \widetilde{\mathbf{v}}_j \in \mathcal{P}_N \text{ so that } \mathcal{K}(\sigma_j)\widetilde{\mathbf{v}}_j - \mathcal{B}(\sigma_j)\mathbf{b}_j \perp \mathcal{Q}_N \quad \text{and} \\ &\text{find } \widetilde{\mathbf{w}}_j \in \mathcal{Q}_N \text{ so that } \mathcal{K}(\sigma_j)^T\widetilde{\mathbf{w}}_j - \mathcal{C}(\sigma_j)^T\mathbf{c}_i \perp \mathcal{P}_N \end{aligned} \tag{42}$$

Computed quantities generated within a Petrov–Galerkin framework will be denoted with a "tilde" to distinguish them from earlier "hat" quantities where no structure was assumed in the inexact solves. The following theorem asserts that if a reduced order model is computed within a Petrov–Galerkin framework (42), then one can obtain a structured backward error that throws the effect of inexact solves back onto a perturbation on the original dynamical system.

Theorem 4.1. *Given a full-order model $\mathcal{H}(s) = \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathcal{B}(s)$, interpolation points $\{\sigma_j\}_{j=1}^r$, and tangent directions $\{\mathbf{b}_i\}_{i=1}^r$ and $\{\mathbf{c}_i\}_{i=1}^r$, let the inexact solutions $\widetilde{\mathbf{v}}_j$ for $\mathcal{K}(\sigma_j)^{-1}\mathcal{B}(\sigma_j)\mathbf{b}_j$ and $\widetilde{\mathbf{w}}_j$ for $\mathcal{K}(\sigma_j)^{-T}\mathcal{C}(\sigma_j)^T\mathbf{c}_j$ be obtained in a Petrov–Galerkin framework as in (42). Let $\widetilde{\mathbf{V}}_r$ and $\widetilde{\mathbf{W}}_r$ denote the corresponding inexact interpolatory bases; i.e.,*

$$\widetilde{\mathbf{V}}_r = [\ \widetilde{\mathbf{v}}_1, \ \ \ldots, \ \ \widetilde{\mathbf{v}}_r\ ] \quad \text{and} \quad \widetilde{\mathbf{W}}_r = [\ \widetilde{\mathbf{w}}_1, \ \ \ldots, \ \ \widetilde{\mathbf{w}}_r\ ]. \tag{43}$$

Define residuals

$$\boldsymbol{\eta}_j = \mathcal{K}(\sigma_j)\widetilde{\mathbf{v}}_j - \mathcal{B}(\sigma_j)\mathbf{b}_j \quad \text{and} \quad \boldsymbol{\xi}_j = \mathcal{K}(\sigma_j)^T\widetilde{\mathbf{w}}_j - \mathcal{C}(\sigma_j)^T\mathbf{c}_j,$$

residual matrices

$$\mathbf{R_b} = [\eta_1,\ \eta_2,\ \ldots,\ \eta_r], \qquad \mathbf{R_c} = [\xi_1,\ \xi_2,\ \ldots,\ \xi_r], \tag{44}$$

and the rank $2r$ *matrix*

$$\mathbf{F}_{2r} = \mathbf{R_b}(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\widetilde{\mathbf{W}}_r^T + \widetilde{\mathbf{V}}_r(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\mathbf{R}_\mathbf{c}^T. \tag{45}$$

Let $\widetilde{\mathcal{H}}_r(s) = \widetilde{\mathcal{C}}_r(s)\widetilde{\mathcal{K}}_r(s)^{-1}\widetilde{\mathcal{B}}_r(s)$ *denote the computed inexact reduced model via the Petrov–Galerkin process where*

$$\widetilde{\mathcal{K}}_r(s) = \widetilde{\mathbf{W}}_r^T\mathcal{K}(s)\widetilde{\mathbf{V}}_r,\ \ \widetilde{\mathcal{B}}_r(s) = \widetilde{\mathbf{W}}_r^T\mathcal{B}(s), \quad \textit{and} \quad \widetilde{\mathcal{C}}_r(s) = \mathcal{C}(s)\widetilde{\mathbf{V}}_r. \tag{46}$$

Then, $\widetilde{\mathcal{H}}_r(s)$ *exactly tangentially interpolates the perturbed full-order model*

$$\widetilde{\mathcal{H}}(s) = \mathcal{C}(s)(\mathcal{K}(s) + \mathbf{F}_{2r})^{-1}\mathcal{B}(s), \tag{47}$$

at each σ_i*:*

$$\widetilde{\mathcal{H}}(\sigma_i)\mathbf{b}_i = \widetilde{\mathcal{H}}_r(\sigma_i)\mathbf{b}_i, \quad \mathbf{c}_i^T\widetilde{\mathcal{H}}(\sigma_i) = \mathbf{c}_i^T\widetilde{\mathcal{H}}_r(\sigma_i),$$
$$\textit{and} \quad \mathbf{c}_i^T\widetilde{\mathcal{H}}'(\sigma_i)\mathbf{b}_i = \mathbf{c}_i^T\widetilde{\mathcal{H}}_r'(\sigma_i)\mathbf{b}_i \quad \textit{for each } i = 1, \ldots, r.$$

Proof. The computed model, $\widetilde{\mathcal{H}}_r(s)$, will (exactly) tangentially interpolate a perturbed model $\widetilde{\mathcal{H}}(s) = \mathcal{C}(s)(\mathcal{K}(s) + \mathbf{F})^{-1}\mathcal{B}(s)$ provided the following interpolation conditions hold:

$$(\mathcal{K}(\sigma_i) + \mathbf{F})\,\tilde{\mathbf{v}}_i = \mathcal{B}(\sigma_i)\mathbf{b}_i \text{ and } \tilde{\mathbf{w}}_i^T\,(\mathcal{K}(\sigma_i) + \mathbf{F}) = \mathbf{c}_i^T\,\mathcal{C}(\sigma_i) \quad \text{for } i = 1, \ldots, r.$$

Equivalently, these can be interpreted as conditions on the perturbation $\mathbf{F}$. Rewriting this using notation defined above, $\mathbf{F}$ must satisfy

$$\mathbf{F}\widetilde{\mathbf{V}}_r = \mathbf{R_b} \quad \text{and} \quad \widetilde{\mathbf{W}}_r^T\mathbf{F} = \mathbf{R}_\mathbf{c}^T. \tag{48}$$

The Petrov–Galerkin framework guarantees $\widetilde{\mathbf{W}}_r^T\mathbf{R_b} = \mathbf{0}$ and $\mathbf{R}_\mathbf{c}^T\widetilde{\mathbf{V}}_r = \mathbf{0}$. Substitution of $\mathbf{F}_{2r}$ from (45) into (48) verifies that $\mathbf{F}_{2r}$ is a perturbation to $\mathcal{K}(s)$ for which the computed (inexact) vectors become (exact) interpolation vectors.

Note that since $\widetilde{\mathbf{W}}_r^T\mathbf{F}_{2r}\widetilde{\mathbf{V}}_r = \mathbf{0}$,

$$\widetilde{\mathcal{K}}_r(s) = \widetilde{\mathbf{W}}_r^T\mathcal{K}(s)\widetilde{\mathbf{V}}_r = \widetilde{\mathbf{W}}_r^T(\mathcal{K}(s) + \mathbf{F}_{2r})\widetilde{\mathbf{V}}_r.$$

Consequently, the reduced model $\widetilde{\mathcal{H}}_r(s)$ obtained by inexact solves in (46) is what one would have obtained by exact interpolatory model reduction of $\widetilde{\mathcal{H}}(s)$. □

Theorem 4.2. *Assume the hypotheses of Theorem 4.1 and that* $\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r$ *is nonsingular. Define an oblique projector,* $\widetilde{\mathbf{\Phi}}_r = \widetilde{\mathbf{V}}_r(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\widetilde{\mathbf{W}}_r^T$. *The backward perturbation* $\mathbf{F}_{2r}$ *given in Theorem 4.1 satisfies*

$$\|\mathbf{F}_{2r}\|_F \leqslant \sqrt{r}\,\|\widetilde{\mathbf{\Phi}}_r\| \cdot \left(\max_i \frac{\|\eta_i\|}{\|\tilde{\mathbf{v}}_i\|}\varsigma_{\min}(\widetilde{\mathbf{V}}_r\mathbf{D})^{-1} + \max_i \frac{\|\xi_i\|}{\|\tilde{\mathbf{w}}_i\|}\varsigma_{\min}(\widetilde{\mathbf{W}}_r\mathbf{D})^{-1}\right)$$

where $\varsigma_{\min}$ *denotes the smallest singular value and* $\|\mathbf{M}\|_F = \sqrt{\mathrm{trace}(\mathbf{M}^T\mathbf{M})}$ *denotes the Frobenius norm of a matrix,* $\mathbf{M}$.

Proof. Note that

$$\|\mathbf{F}_{2r}\|_F \leqslant \|\mathbf{R_b}(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\widetilde{\mathbf{W}}_r^T\|_F + \|\widetilde{\mathbf{V}}_r(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\mathbf{R}_\mathbf{c}^T\|_F.$$

Let $\widetilde{\mathbf{V}}_r$ have an orthogonal factorization as $\widetilde{\mathbf{V}}_r = \mathbf{Q}_v\mathbf{L}_v$ with $\mathbf{Q}_v^*\mathbf{Q}_v = \mathbf{I}$. Then

$$\begin{aligned}
\|\mathbf{R_b}(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\widetilde{\mathbf{W}}_r^T\|_F &= \|\mathbf{R_b}\mathbf{L}_v^{-1}\mathbf{L}_v(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\widetilde{\mathbf{W}}_r^T\|_F \\
&\leqslant \|\mathbf{R_b}\mathbf{L}_v^{-1}\|_F \cdot \|\mathbf{L}_v(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\widetilde{\mathbf{W}}_r^T\| \\
&\leqslant \|\mathbf{R_b}\mathbf{L}_v^{-1}\|_F \cdot \|\widetilde{\mathbf{\Phi}}_r\| \\
&\leqslant \|\mathbf{R_b}\widetilde{\mathbf{D}}_v(\mathbf{L}_v\widetilde{\mathbf{D}}_v)^{-1}\|_F \cdot \|\widetilde{\mathbf{\Phi}}_r\| \\
&\leqslant \|\mathbf{R_b}\widetilde{\mathbf{D}}_v\|_F \cdot \|(\mathbf{L}_v\widetilde{\mathbf{D}}_v)^{-1}\| \cdot \|\widetilde{\mathbf{\Phi}}_r\|,
\end{aligned}$$

where we have introduced a diagonal scaling matrix

$$\widetilde{\mathbf{D}}_v = \mathrm{diag}(1/\|\widetilde{\mathbf{v}}_1\|,\ 1/\|\widetilde{\mathbf{v}}_2\|,\ \ldots,\ 1/\|\widetilde{\mathbf{v}}_r\|).$$

Easily one sees $\|\mathbf{R_b}\widetilde{\mathbf{D}}_v\|_F \leqslant \sqrt{r}\max_i \dfrac{\|\boldsymbol{\eta}_i\|}{\|\widetilde{\mathbf{v}}_i\|}$. For the remaining term, note that

$$\|(\mathbf{L}_v\widetilde{\mathbf{D}}_v)^{-1}\| = \left(\min_{\mathbf{x}} \frac{\|\widetilde{\mathbf{V}}_r\widetilde{\mathbf{D}}_v\mathbf{x}\|}{\|\mathbf{x}\|}\right)^{-1} = \varsigma_{\min}(\widetilde{\mathbf{V}}_r\widetilde{\mathbf{D}}_v)^{-1}$$

A similar bound for $\|\widetilde{\mathbf{V}}_r(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\mathbf{R}_\mathbf{c}^T\|_F$ is produced by an analogous process, which leads then to the final estimate for $\|\mathbf{F}_{2r}\|_F$. □

Note that the perturbation $\mathbf{F}_{2r}$ is completely determined by accessible, computed quantities. Hence, one can use $\mathbf{F}_{2r}$ to determine how accurately one must solve the underlying linear systems in order to assure system fidelity of a given order.

Theorem 4.3. *If* $\|\mathbf{F}_{2r}\| < 1/\|\mathcal{K}(s)^{-1}\|_{\mathcal{H}_\infty}$ *then*

$$\|\mathcal{H}(s) - \widetilde{\mathcal{H}}(s)\|_{\mathcal{H}_2} \leqslant \frac{\|\mathcal{C}(s)\mathcal{K}(s)^{-1}\|_{\mathcal{H}_2}\,\|\mathcal{K}(s)^{-1}\mathcal{B}(s)\|_{\mathcal{H}_\infty}}{1 - \|\mathcal{K}(s)^{-1}\|_{\mathcal{H}_\infty}\,\|\mathbf{F}_{2r}\|}\,\|\mathbf{F}_{2r}\|.$$

Proof. The system-wise backward error associated with inexact solves may be written as

$$\begin{aligned}
\mathcal{H}(s) - \widetilde{\mathcal{H}}(s) &= \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathcal{B}(s) - \mathcal{C}(s)\left(\mathcal{K}(s) + \mathbf{F}_{2r}\right)^{-1}\mathcal{B}(s) \\
&= \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathbf{F}_{2r}\left(\mathcal{K}(s) + \mathbf{F}_{2r}\right)^{-1}\mathcal{B}(s) \\
&= \mathcal{C}(s)\mathcal{K}(s)^{-1}\mathbf{F}_{2r}\left(\mathbf{I} + \mathcal{K}(s)^{-1}\mathbf{F}_{2r}\right)^{-1}\mathcal{K}(s)^{-1}\mathcal{B}(s).
\end{aligned}$$

Define $\mathcal{M}(s) = \mathbf{F}_{2r}\left(\mathbf{I} + \mathcal{K}(s)^{-1}\mathbf{F}_{2r}\right)^{-1}$ and observe that

$$\begin{aligned}
\|\mathcal{H}(s) - \widetilde{\mathcal{H}}(s)\|_{\mathcal{H}_2}^2 &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \|\mathcal{C}(\imath\omega)\mathcal{K}(\imath\omega)^{-1}\,\mathcal{M}(\imath\omega)\,\mathcal{K}(\imath\omega)^{-1}\mathcal{B}(\imath\omega)\|_F^2\,d\omega \\
&\leqslant \frac{1}{2\pi}\int_{-\infty}^{\infty} \|\mathcal{C}(\imath\omega)\mathcal{K}(\imath\omega)^{-1}\|_F^2 \cdot \|\,\mathcal{M}(\imath\omega)\,\|^2 \cdot \|\mathcal{K}(\imath\omega)^{-1}\mathcal{B}(\imath\omega)\|^2\,d\omega \\
&\leqslant \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \|\mathcal{C}(\imath\omega)\mathcal{K}(\imath\omega)^{-1}\|_F^2\,d\omega\right) \cdot \max_{\omega\in\mathbb{R}} \|\mathcal{M}(\imath\omega)\,\|^2 \cdot \max_{\omega\in\mathbb{R}} \|\mathcal{K}(\imath\omega)^{-1}\mathcal{B}(\imath\omega)\|^2 \\
&\leqslant \|\mathcal{C}(s)\mathcal{K}(s)^{-1}\|_{\mathcal{H}_2}^2 \cdot \|\mathcal{K}(s)^{-1}\mathcal{B}(s)\|_{\mathcal{H}_\infty}^2 \cdot \|\mathcal{M}(s)\|_{\mathcal{H}_\infty}^2.
\end{aligned}$$

To estimate $\|\mathcal{M}(s)\|_{\mathcal{H}_\infty}$, a rearrangement of the definition of $\mathcal{M}(s)$ provides

$$\mathcal{M}(s) = \left(\mathbf{I} - \mathcal{M}(s)\mathcal{K}(s)^{-1}\right)\mathbf{F}_{2r}.$$

So we have immediately,

$$\begin{aligned}\|\mathcal{M}(s)\|_{\mathcal{H}_\infty} &= \max_{\omega\in\mathbb{R}} \|\mathcal{M}(\iota\omega)\| \leqslant \max_{\omega\in\mathbb{R}} \|\mathbf{I} - \mathcal{M}(\iota\omega)\mathcal{K}(\iota\omega)^{-1}\| \cdot \|\mathbf{F}_{2r}\| \\ &\leqslant \left(1 + \max_{\omega\in\mathbb{R}} \|\mathcal{M}(\iota\omega)\mathcal{K}(\iota\omega)^{-1}\|\right) \|\mathbf{F}_{2r}\| \\ &\leqslant \left(1 + \|\mathcal{M}(s)\|_{\mathcal{H}_\infty} \|\mathcal{K}(s)^{-1}\|_{\mathcal{H}_\infty}\right) \|\mathbf{F}_{2r}\|.\end{aligned}$$

Since $\|\mathcal{K}(s)^{-1}\|_{\mathcal{H}_\infty} \|\mathbf{F}_{2r}\| < 1$, this last expression can be rearranged to obtain

$$\|\mathcal{M}(s)\|_{\mathcal{H}_\infty} \leqslant \frac{\|\mathbf{F}_{2r}\|}{1 - \|\mathcal{K}(s)^{-1}\|_{\mathcal{H}_\infty} \|\mathbf{F}_{2r}\|}$$

which implies the conclusion. □

By combining Theorem 3.3 with Theorem 3.2 or combining Theorem 4.2 with Theorem 4.3, we approach our goal of connecting quantities that we have control over, such as the termination threshold, ε, to relevant system theoretic errors, $\|\mathcal{H}_r - \widehat{\mathcal{H}}_r\|$ and $\|\mathcal{H} - \widetilde{\mathcal{H}}\|$, which are quantities we would like to control.

One may use these expressions as a basis to devise and investigate different, effective stopping criteria in large-scale numerical settings. For example, while ε appears explicitly in Theorem 3.2 in a way that suggests its use as a relative residual norm threshold; Theorem 4.2 suggests a scaling of the residual norm by the norm of the solution vector as another possible stopping criterion. These and related ideas are the focus of on-going work.

4.2. Quantities of interest in derived bounds

By combining Theorem 4.2 with Theorem 4.3, one observes that perturbation effects of the inexact solves on the system theoretical (model reduction related) measures critically depend on the four quantities: The norm of the oblique projector $\widetilde{\mathbf{\Phi}}_r = \widetilde{\mathbf{V}}_r(\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\widetilde{\mathbf{W}}_r^T$ of the underlying model reduction problem, reciprocals of the minimum singular values of the scaled primitive bases $\widetilde{\mathbf{V}}_r\mathbf{D}$ and $\widetilde{\mathbf{W}}_r\mathbf{D}$; and the stopping criterion ε for the inexact solves, (which affects $\max_i \dfrac{\|\boldsymbol{\eta}_i\|}{\|\widetilde{\mathbf{v}}_i\|}$ and $\max_i \dfrac{\|\boldsymbol{\xi}_i\|}{\|\widetilde{\mathbf{w}}_i\|}$).

The ε term is associated directly with inexact solves and is under the control of the user. The remaining quantities $\varsigma_{\min}(\widetilde{\mathbf{V}}_r\mathbf{D})^{-1}$, $\varsigma_{\min}(\widetilde{\mathbf{W}}_r\mathbf{D})^{-1}$ and $\|\widetilde{\mathbf{\Phi}}_r\|$, depend largely on the selection of interpolation points $\{\sigma_i\}$ and tangent directions, but the influence of interpolation data on the magnitude of these quantities is difficult to anticipate.

In this section, we will investigate experimentally the effects of the interpolation point selection on the three quantities of interest, $\varsigma_{\min}(\widetilde{\mathbf{V}}_r\mathbf{D})^{-1}$, $\varsigma_{\min}(\widetilde{\mathbf{W}}_r\mathbf{D})^{-1}$ and $\|\widetilde{\mathbf{\Phi}}_r\|$, appearing in the derived bounds. These quantities are continuous with respect to the primitive basis vectors, $\{\widetilde{\mathbf{v}}_1, \ldots, \widetilde{\mathbf{v}}_r\}$ and $\{\widetilde{\mathbf{w}}_1, \ldots, \widetilde{\mathbf{w}}_r\}$ in neighborhoods where $\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r$ is nonsingular (i.e., where the projector $\widetilde{\mathbf{\Phi}}_r$ is well defined). Thus it will be sufficient to examine how the magnitudes of the quantities of interest depend on interpolation data presuming that the necessary linear solves are done *exactly*; for modest convergence thresholds, the effect of inexact solves on these magnitudes is secondary to the effect of interpolation point location.

For our numerical study, we use the International Space Station 12A Module as the full-order model. The model has order $n = 1412$. We examine a single-input single-output subsystem, $\mathcal{H}(s)$, reducing the order from 1412 to order r with r varying from 2 to 70 in increments of two. For each reduced order, we

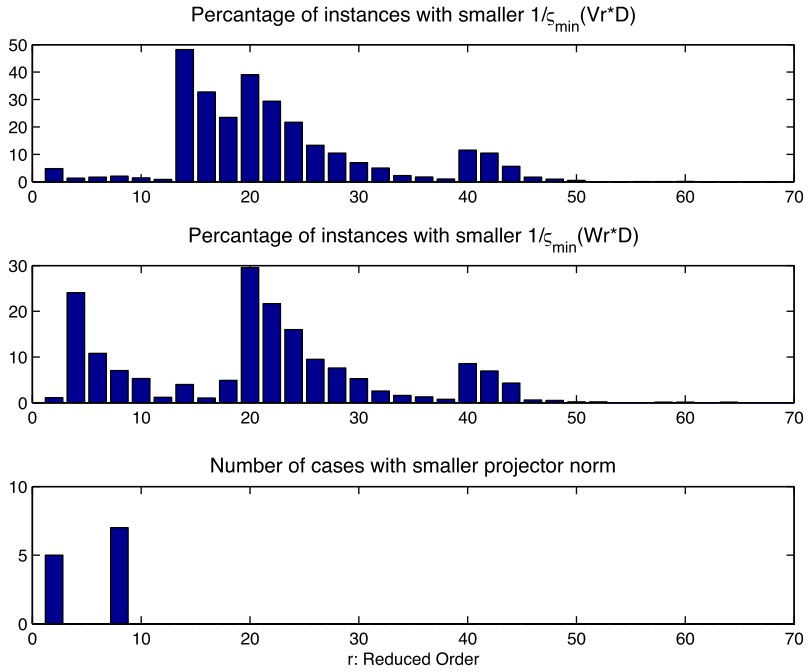


Fig. 2. Comparison of $\varsigma_{\min}(\mathbf{V}_r\mathbf{D})^{-1}$, $\varsigma_{\min}(\mathbf{W}_r\mathbf{D})^{-1}$ and $\|\mathbf{\Phi}_r\|$ for random shift selections relative to values for $\mathcal{H}_2$-optimal shifts.

chose 2000 random shift selections and computed $\varsigma_{\min}(\mathbf{V}_r\mathbf{D})^{-1}$, $\varsigma_{\min}(\mathbf{W}_r\mathbf{D})^{-1}$ and $\|\mathbf{\Phi}_r\|$. For each r, $r/2$ shifts were sampled from a uniform distribution on a rectangular region in the positive half-plane: $\left\{ z \in \mathbb{C} \;\middle|\; \begin{array}{l} \min_\lambda |\mathrm{Re}(\lambda)| \leqslant \mathrm{Re}(z) \leqslant \max_\lambda |\mathrm{Re}(\lambda)| \\ |\mathrm{Im}(z)| \leqslant \max_\lambda |\mathrm{Im}(\lambda)| \end{array} \right\}$, where the max and min are taken over all the poles of the system. The remaining $r/2$ shifts were taken to be the complex conjugates of this random sample, so as to produce a shift configuration that was closed under conjugation. Additionally for each r, we applied model reduction using the $\mathcal{H}_2$-optimal interpolation points generated by the method of [16]. Then, for each r, out of the 2000 randomly generated shift selections, we counted the number of cases where the random shift selection yielded smaller values of $\varsigma_{\min}(\mathbf{V}_r\mathbf{D})^{-1}$, $\varsigma_{\min}(\mathbf{W}_r\mathbf{D})^{-1}$ and $\|\mathbf{\Phi}_r\|$. The results are shown in Fig. 2. Fig. 2(a) and (b) show that for most of the cases, the $\mathcal{H}_2$-optimal interpolation points yield smaller values for $\varsigma_{\min}(\mathbf{V}_r\mathbf{D})^{-1}$, $\varsigma_{\min}(\mathbf{W}_r\mathbf{D})^{-1}$. Indeed, for $r \geqslant 48$, the $\mathcal{H}_2$-optimal points produced smaller values in more than 99% of the cases. Also, for the last three cases: $r = 66, r = 68,$ and $r = 70,$ the $\mathcal{H}_2$-optimal interpolation points always yielded smaller quantities. The results are even more dramatic for the projector norm, which is important in scaling the perturbation effects caused by inexact solves, see Theorem 4.2: Out of 70,000 cases (2000 selections for each r value), the $\mathcal{H}_2$-optimal interpolation point selection produced smaller condition numbers in all except 7 instances: 5 instances for $r = 2$, and 2 instances for $r = 8$. These numerical results illustrate that $\mathcal{H}_2$-optimal interpolation points can be expected to yield smaller values for $\varsigma_{\min}(\mathbf{V}_r\mathbf{D})^{-1}$, $\varsigma_{\min}(\mathbf{W}_r\mathbf{D})^{-1}$ and $\|\mathbf{\Phi}_r\|$, and hence should produce reduced order models that are more robust with respect to perturbations.

Fig. 2 also shows that for $r = 14$, 48% of the randomly selected shifts yielded smaller values of $\varsigma_{\min}(\mathbf{V}_r\mathbf{D})^{-1}$. However, when we inspected the 2000 randomly selected shift sets for $r = 14$ in more detail, we observed some interesting additional features. We computed the three quantities $\varsigma_{\min}(\mathbf{V}_r\mathbf{D})^{-1}$, $\varsigma_{\min}(\mathbf{W}_r\mathbf{D})^{-1}$ and $\|\mathbf{\Phi}_r\|$ for each of the 2000 randomly selected shift sets, and compared them with the corresponding value derived from an $\mathcal{H}_2$-optimal shift selection. The results are shown in Fig. 3. The top plot shows $\varsigma_{\min}(\mathbf{V}_r\mathbf{D})/\varsigma_{\min}(\mathbf{V}_r^{\mathrm{opt}}\mathbf{D})$ where $\mathbf{V}_r^{\mathrm{opt}}$ stands for the primitive interpolatory

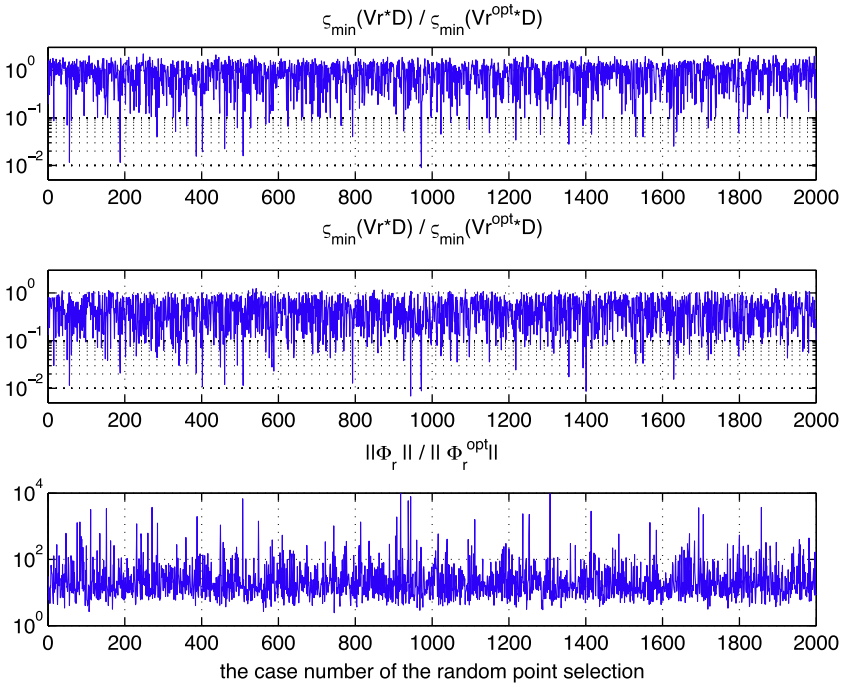


Fig. 3. Detailed comparison for $r = 14$.

basis for the $\mathcal{H}_2$-optimal points. The bigger this ratio, the better the random shift selection. Even though for 48% of the cases, the random selection was better, the highest this ratio becomes is 2.20, i.e., the random shifts were never much better than a factor of 2 better than what $\mathcal{H}_2$-optimal shifts provided. For the remaining 52% of the cases, the randomly selected shifts were worse, and often worse by a factor of 100 or more. The situation for $\mathbf{W}_r$ is shown in the middle plot. Once more, the situation is dramatically in the favor of the $\mathcal{H}_2$-optimal interpolation points when the projector norm is inspected; the bottom plot in Fig. 3 which depicts the ratio $\|\mathbf{\Phi}_r\|/\|\mathbf{\Phi}_r^{\text{opt}}\|$ where $\mathbf{\Phi}_r^{\text{opt}}$ denotes the projector for the $\mathcal{H}_2$-optimal points. As illustrated in Fig. 2, there are no random shift cases yielding a smaller projector norm. Furthermore, in many cases the projector norm for the random shift selection is almost 4 order of magnitudes higher than that of the $\mathcal{H}_2$-optimal points. Indeed, on average the projector norm for the random points is 8.19×10^1 times higher. These numbers change more in the favor of the $\mathcal{H}_2$-optimal points as r increases. For example, for $r = 50$, while the ratio $\varsigma_{\min}(\mathbf{V}_r\mathbf{D})/\varsigma_{\min}(\mathbf{V}_r^{\text{opt}}\mathbf{D})$ becomes only as high as 1.48, it becomes as low as 2.89×10^{-4} for some random selections; and the ratio, $\|\mathbf{\Phi}_r\|/\|\mathbf{\Phi}_r^{\text{opt}}\|$, can reach as high as 2.91×10^5. For $r = 70$, $\|\mathbf{\Phi}_r\|$ for random selection is 1.73×10^2 times higher than $\|\mathbf{\Phi}_r^{\text{opt}}\|$ on average. The three quantities we have been investigating appear to be extremely well conditioned for $\mathcal{H}_2$-optimal interpolation points. Even for $r = 70$, both $\varsigma_{\min}(\mathbf{V}_r^{\text{opt}}\mathbf{D})^{-1}$, $\varsigma_{\min}(\mathbf{W}_r^{\text{opt}}\mathbf{D})^{-1}$ remain smaller than 10 and $\|\mathbf{\Phi}_r^{\text{opt}}\|$ is smaller than 7.

5. Inexact solves in optimal interpolatory approximation

The quality of the reduced-order model in interpolatory model reduction clearly depends on the selection of interpolation points and tangent directions. Until recently, this selection process was mostly *ad hoc*, and this factor had been the principal disadvantage of interpolatory model reduction. For systems in standard first-order state-space form, Gugercin et al. [16] have produced that an

$\mathcal{H}_2$-optimal interpolation point / tangent direction selection strategy and proposed an Iterative Rational Krylov Algorithm (IRKA) to generate interpolatory reduced-order models that are (locally) optimal with respect to the $\mathcal{H}_2$ norm. (An $\mathcal{H}_2$-optimal interpolation point selection strategy is still unknown for the general coprime factorization framework.) In this section, we investigate the behavior of inexact solves within the $\mathcal{H}_2$-optimal interpolatory approximation setting, specifically examining the behavior when inexact solves are employed in IRKA. In the rest of this section, we briefly review the optimal $\mathcal{H}_2$ approximation problem and the method of [16]. We then show how inexact solves can be employed effectively in this setting and discuss observed effects on optimality of the final reduced model. Our discussion focuses on systems in first-order descriptor form:

$$\mathcal{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}, \tag{49}$$

where $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$ and $\mathbf{C} \in \mathbb{R}^{p\times n}$.

5.1. Optimal $\mathcal{H}_2$ approximation problem

Given the full-order system as in (49), the goal of the optimal $\mathcal{H}_2$ model reduction problem is to find a reduced-order model $\mathcal{H}_r(s)$ that minimizes the $\mathcal{H}_2$ error; i.e.,

$$\|\mathcal{H} - \mathcal{H}_r\|_{\mathcal{H}_2} = \min_{\substack{G_r \text{ stable} \\ \dim(G_r)=r}} \|\mathcal{H} - G_r\|_{\mathcal{H}_2}\,. \tag{50}$$

Many researchers have worked on this problem. These efforts can be grouped into two categories: Lyapunov-based optimal $\mathcal{H}_2$ methods such as [31,26,17,18,30,32]; and interpolation-based optimal $\mathcal{H}_2$ methods such as [23,16,15,29,10,14,20,5,7]. Here, we will focus on the interpolation-based approach. However we note that Gugercin et al. [16] has shown that these two frameworks are theoretically equivalent; hence motivating the use of interpolatory approaches to optimal $\mathcal{H}_2$ approximation since they are numerically superior to the Lyapunov-based approaches.

Since the optimization problem (50) is nonconvex, obtaining a global minimizer is a hard task and can be intractable. The usual approach is to find reduced order models that satisfy first-order necessary optimality conditions. Meier and Luenberger [23] introduced interpolation-based $\mathcal{H}_2$-optimality conditions for SISO systems. Analogous $\mathcal{H}_2$-optimality conditions for MIMO systems have recently been developed by [16,10,29] which in turn have led to analogous algorithms for the MIMO case; see [16,10] for more details.

Theorem 5.1. *Given $\mathcal{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$, let $\mathcal{H}_r(s) = \sum_{i=1}^{r} \frac{1}{s-\widehat{\lambda}_i}\mathbf{c}_i\mathbf{b}_i^T$ be the best r^{th} order approximation of $\mathcal{H}$ with respect to the $\mathcal{H}_2$ norm. Then*

$$\begin{aligned} &(a)\ \mathcal{H}(-\widehat{\lambda}_k)\mathbf{b}_k = \mathcal{H}_r(-\widehat{\lambda}_k)\mathbf{b}_k, \quad (b)\ \mathbf{c}_k^T\mathcal{H}(-\widehat{\lambda}_k) = \mathbf{c}_k^T\mathcal{H}_r(-\widehat{\lambda}_k), \text{ and} \\ &(c)\ \mathbf{c}_k^T\mathcal{H}'(-\widehat{\lambda}_k)\mathbf{b}_k = \mathbf{c}_k^T\mathcal{H}_r'(-\widehat{\lambda}_k)\mathbf{b}_k \quad \text{for } k = 1,\ 2,\ \ldots,\ r. \end{aligned} \tag{51}$$

5.1.1. An algorithm for interpolatory optimal $\mathcal{H}_2$ model reduction

Theorem 5.1 reveals that any $\mathcal{H}_2$ optimal reduced-order model $\mathcal{H}_r(s)$ is a bitangential Hermite interpolant to $\mathcal{H}(s)$ at mirror images of the reduced-order poles. However, since the interpolation points and the tangent directions (and consequently, $\mathbf{V}_r$ and $\mathbf{W}_r$), depend on the final reduced-model to be computed, they are not known *a priori*. The *Iterative Rational Krylov Algorithm* (IRKA) of [16] resolves this problem by iteratively correcting the interpolation points and the directions as outlined in Algorithm 1: The reduced-order poles are reflected across the imaginary axis to become the next set of interpolation points; the tangent directions are corrected using residue directions from the current reduced model. Upon convergence, the resulting interpolatory reduced-order model satisfies the necessary conditions of Theorem 5.1. For further details on IRKA, see [16].

Algorithm 1. IRKA for MIMO $\mathcal{H}_2$ optimal tangential interpolation

1. Make an initial r-fold shift selection: $\{\sigma_1, \ldots, \sigma_r\}$ and initial tangent directions $\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_r$ and $\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_r$.
2. $\mathbf{V}_r = \left[(\sigma_1 \mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\hat{\mathbf{b}}_1 \ \cdots \ (\sigma_r \mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\hat{\mathbf{b}}_r \right]$,
   $\mathbf{W}_r = \left[(\sigma_1 \mathbf{E} - \mathbf{A}^T)^{-1}\mathbf{C}^T\hat{\mathbf{c}}_1 \ \cdots \ (\sigma_r \mathbf{E} - \mathbf{A}^T)^{-1}\mathbf{C}^T\hat{\mathbf{c}}_1 \right]$.
3. While (not converged):
   (a) $\mathbf{A}_r = \mathbf{W}_r^T\mathbf{A}\mathbf{V}_r, \mathbf{E}_r = \mathbf{W}_r^T\mathbf{E}\mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T\mathbf{B}$, and $\mathbf{C}_r = \mathbf{C}\mathbf{V}_r$.
   (b) Compute $\mathbf{Y}^*\mathbf{A}_r\mathbf{X} = diag(\tilde{\lambda}_i)$ and $\mathbf{Y}^*\mathbf{E}_r\mathbf{X} = \mathbf{I}_r$ where $\mathbf{Y}^*$ and $\mathbf{X}$ are the left and right eigenvector matrices for $\lambda\mathbf{E}_r - \mathbf{A}_r$.
   (c) $\sigma_i \longleftarrow -\lambda_i(\mathbf{A}_r, \mathbf{E}_r)$ for $i = 1, \ldots, r$, $\hat{\mathbf{b}}_i^* \longleftarrow \mathbf{e}_i^T\mathbf{Y}^*\mathbf{B}_r$ and $\hat{\mathbf{c}}_i \longleftarrow \mathbf{C}_r\mathbf{X}\mathbf{e}_i$.
   (d) $\mathbf{V}_r = \left[(\sigma_1 \mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\hat{\mathbf{b}}_1 \ \cdots \ (\sigma_r \mathbf{E} - \mathbf{A})^{-1}\mathbf{B}\hat{\mathbf{b}}_r \right]$.
   (e) $\mathbf{W}_r = \left[(\sigma_1 \mathbf{E} - \mathbf{A}^T)^{-1}\mathbf{C}^T\hat{\mathbf{c}}_1 \ \cdots \ (\sigma_r \mathbf{E} - \mathbf{A}^T)^{-1}\mathbf{C}^T\hat{\mathbf{c}}_1 \right]$.
4. $\mathbf{A}_r = \mathbf{W}_r^T\mathbf{A}\mathbf{V}_r, \mathbf{E}_r = \mathbf{W}_r^T\mathbf{E}\mathbf{V}_r, \mathbf{B}_r = \mathbf{W}_r^T\mathbf{B}, \mathbf{C}_r = \mathbf{C}\mathbf{V}_r$.

*5.2. Inexact Iterative Rational Krylov Algorithm (*InxIRKA*)*

For large system order, one may see from Algorithm 1, that the main cost of IRKA will generally be solving $2r$ large linear systems at each step. If the IRKA iteration converges in k steps, a total of $2rk$ linear systems will need to be solved. In settings where system dimension reaches into the millions, iterative linear system solvers become necessary and inexact linear system solves must be incorporated into IRKA. We refer to the modified algorithm as the *Inexact Iterative Rational Krylov Algorithm* (InxIRKA) and describe it in Algorithm 2 below. We employ the Petrov–Galerkin framework for the inexact solves. In Algorithm 2, the function $\mathbf{F}_{\mathsf{PG}}$ in

$$[\widetilde{\mathbf{v}}_i, \widetilde{\mathbf{w}}_i] = \mathbf{F}_{\mathsf{PG}}\left(\mathbf{A}, \mathbf{E}, \mathbf{B}, \sigma_i, \mathbf{b}_i, \mathbf{c}_i, \mathbf{v}^{(0)}, \mathbf{w}^{(0)}, \varepsilon\right)$$

denotes an inexact solve using a Petrov–Galerkin framework to approximately solve the linear systems $(\sigma_i\mathbf{E} - \mathbf{A})\mathbf{v}_i = \mathbf{B}\mathbf{b}_i$ and $(\sigma_i\mathbf{E} - \mathbf{A})^T\mathbf{w}_i = \mathbf{C}^T\mathbf{c}_i$ with initial guesses $\mathbf{v}^{(0)}$ and $\mathbf{w}^{(0)}$, respectively, and a relative residual termination tolerance ε, i.e., at the end,

$$\frac{\|(\sigma_i\mathbf{E} - \mathbf{A})\widetilde{\mathbf{v}}_i - \mathbf{B}\mathbf{b}_i\|}{\|\mathbf{B}\mathbf{b}_i\|} \leqslant \varepsilon \quad \text{and} \quad \frac{\|(\sigma_i\mathbf{E} - \mathbf{A})^T\widetilde{\mathbf{w}}_i - \mathbf{C}^T\mathbf{c}_i\|}{\|\mathbf{C}^T\mathbf{c}_i\|} \leqslant \varepsilon$$

.

Algorithm 2. InxIRKA for MIMO $\mathcal{H}_2$ optimal tangential interpolation

1. Make an initial r-fold shift selection: $\{\sigma_1, \ldots, \sigma_r\}$ and initial tangent directions $\hat{\mathbf{b}}_1, \ldots, \hat{\mathbf{b}}_r$ and $\hat{\mathbf{c}}_1, \ldots, \hat{\mathbf{c}}_r$.
2. $[\widetilde{\mathbf{v}}_i, \widetilde{\mathbf{w}}_i] = \mathbf{F}_{\mathsf{PG}}(\mathbf{A}, \mathbf{E}, \mathbf{B}, \sigma_i, \mathbf{b}_i, \mathbf{c}_i, \mathbf{0}, \mathbf{0}, \varepsilon)$ for $i = 1, \ldots, r$.
3. $\widetilde{\mathbf{V}}_r = [\ \widetilde{\mathbf{v}}_1,\ \widetilde{\mathbf{v}}_2,\ \ldots,\ \widetilde{\mathbf{v}}_r\ ]$ and $\widetilde{\mathbf{W}}_r = [\ \widetilde{\mathbf{w}}_1,\ \widetilde{\mathbf{w}}_2,\ \ldots,\ \widetilde{\mathbf{w}}_r\ ]$.
4. While (not converged):
   (a) $\widetilde{\mathbf{A}}_r = \widetilde{\mathbf{W}}_r^T\mathbf{A}\widetilde{\mathbf{V}}_r, \widetilde{\mathbf{E}}_r = \widetilde{\mathbf{W}}_r^T\mathbf{E}\widetilde{\mathbf{V}}_r, \widetilde{\mathbf{B}}_r = \widetilde{\mathbf{W}}_r^T\mathbf{B}$, and $\widetilde{\mathbf{C}}_r = \mathbf{C}\widetilde{\mathbf{V}}_r$.
   (b) Compute $\mathbf{Y}^*\widetilde{\mathbf{A}}_r\mathbf{X} = diag(\tilde{\lambda}_i)$ and $\mathbf{Y}^*\widetilde{\mathbf{E}}_r\mathbf{X} = \mathbf{I}_r$ where $\mathbf{Y}^*$ and $\mathbf{X}$ are the left and right eigenvector matrices of $\lambda\mathbf{E}_r - \mathbf{A}_r$.
   (c) $\sigma_i \longleftarrow -\lambda_i(\widetilde{\mathbf{A}}_r, \widetilde{\mathbf{E}}_r)$ for $i = 1, \ldots, r$, $\hat{\mathbf{b}}_i^* \longleftarrow \mathbf{e}_i^T\mathbf{Y}^*\widetilde{\mathbf{B}}_r$ and $\hat{\mathbf{c}}_i \longleftarrow \widetilde{\mathbf{C}}_r\mathbf{X}\mathbf{e}_i$.
   (d) $[\widetilde{\mathbf{v}}_i, \widetilde{\mathbf{w}}_i] = \mathbf{F}_{\mathsf{PG}}(\mathbf{A}, \mathbf{E}, \mathbf{B}, \sigma_i, \mathbf{b}_i, \mathbf{c}_i, \widetilde{\mathbf{v}}_i, \widetilde{\mathbf{w}}_i, \varepsilon)$ for $i = 1, \ldots, r$.
   (e) $\widetilde{\mathbf{V}}_r = [\ \widetilde{\mathbf{v}}_1,\ \widetilde{\mathbf{v}}_2,\ \ldots,\ \widetilde{\mathbf{v}}_r\ ]$ and $\widetilde{\mathbf{W}}_r = [\ \widetilde{\mathbf{w}}_1,\ \widetilde{\mathbf{w}}_2,\ \ldots,\ \widetilde{\mathbf{w}}_r\ ]$.
5. $\widetilde{\mathbf{A}}_r = \widetilde{\mathbf{W}}_r^T\mathbf{A}\widetilde{\mathbf{V}}_r, \widetilde{\mathbf{E}}_r = \widetilde{\mathbf{W}}_r^T\mathcal{E}\widetilde{\mathbf{V}}_r, \widetilde{\mathbf{B}}_r = \widetilde{\mathbf{W}}_r^T\mathbf{B}$, and $\widetilde{\mathbf{C}}_r = \mathbf{C}\widetilde{\mathbf{V}}_r$.

As discussed and illustrated in [16,3], in most cases IRKA converges rapidly; that is, the interpolation points and directions at the k^{th} step of IRKA stagnate rapidly with respect to k. Let $\sigma_i^{(k)}$ and $\mathbf{b}_i^{(k)}$ denote the i^{th} interpolation point and right-tangential direction, respectively, at the k^{th} step. Then we expect that as k increases, the solution $\mathbf{v}_i^{(k)}$ of the linear system $(\sigma_i^{(k)}\mathbf{E} - \mathbf{A})\mathbf{v}^{(k)} = \mathbf{B}\mathbf{b}_i^{(k}$ from the k^{th} step approaches to the solution $\mathbf{v}_i^{(k+1)}$ of the linear system $(\sigma_i^{(k+1)}\mathbf{E} - \mathbf{A})\mathbf{v}^{(k+1)} = \mathbf{B}\mathbf{b}_i^{(k+1)}$ at the $(k+1)^{\text{st}}$ step. This is precisely the reason that in Step 4.(d) of Algorithm 2, we use $\mathbf{v}_i^{(k)}$ as an initial guess in solving $(\sigma_i^{(k+1)}\mathbf{E} - \mathbf{A})\mathbf{v}^{(k+1)} = \mathbf{B}\mathbf{b}_i^{(k+1)}$ at the $(k+1)^{\text{st}}$. We expect that this initialization strategy will speed-up the convergence of the iterative solves.

The development of effective stopping criteria based rationally on system theoretic error measures as we have introduced them here is the focus of on-going work. Similar approaches toward the design of effective preconditioning techniques and reuse of preconditioners tailored for interpolatory model reduction and especially for optimal $\mathcal{H}_2$ approximation are also under investigation.

5.3. Effect of inexact solves in the IrkaIRKA setting

The first question to answer in IrkaIRKA is whether a statement can be made about the optimality as in the exact IRKA case. Employing the Petrov–Galerkin framework makes this possible:

Corollary 5.1. *Let $\widetilde{\mathcal{H}}_r(s)$ be obtained by Algorithm 2. Then $\widetilde{\mathcal{H}}_r(s)$ satisfies the necessary conditions for optimal $\mathcal{H}_2$ approximation of a nearby full-order model $\widetilde{\mathcal{H}}(s) = \mathbf{C}(s\mathbf{E} - (\mathbf{A} + \mathbf{F}_{2r}))^{-1}\mathbf{B}$ where $\mathbf{F}_{2r}$ is the rank-2r perturbation matrix defined in (45).*

Corollary 5.1 shows that with the help of the underlying Petrov–Galerkin framework, we state that the final reduced model of IrkaIRKA is an optimal $\mathcal{H}_2$ approximation to a nearby full-order model.

As we discussed in Section 4.2, for a good selection of interpolation points, interpolatory model reduction is expected to be robust with respect to perturbations due to inexact solves. Hence, if one feeds the optimal interpolation points from IRKA into an inexact interpolation framework, we expect that the resulting reduced model will be close to the optimal reduced model of IRKA. However, the optimal interpolation points are not known initially and IrkaIRKA will be initiated with a nonoptimal initial shift selection. If the initial interpolation points and directions are poorly selected, at the early stages of the iteration, perturbations due to inexact solves might be magnified by this poor selection. One can avoid this scenario by using a small termination threshold ε in the early steps of IrkaIRKA, and then gradually increase ε as the iteration starts to converge. However, we note that in our numerical experiments using random initialization strategies, IrkaIRKA performed robustly and yielded high fidelity reduced models that are also close to the true optimal reduced model. This is illustrated in Section 5.4 below. Effective initialization strategies are discussed in [16] as well.

5.4. Numerical results for IrkaIRKA

Here we illustrate the usage of inexact solves in the optimal $\mathcal{H}_2$ approximation setting by comparing IRKA with IrkaIRKA. We use the example of Section 3.3, but with a finer discretization leading to a state-space dimension of $n = 20{,}209$. We focus on a MIMO version using 2-inputs and 2-outputs.

We reduce the order to $r = 6$ using both IRKA and IrkaIRKA. In IrkaIRKA, the dual linear systems are solved in a Petrov–Galerkin framework using BiCG [4] where we use three different values for the relative residual termination threshold of ε: 10^{-5}, 10^{-3}, and 10^{-1}. In all cases, the behavior of IrkaIRKA is virtually indistinguishable from that of IRKA. Starting with the same initial conditions, both IRKA and IrkaIRKA converge within 10 iteration steps in all 5 cases. The evolution of the $\mathcal{H}_2$ errors $\|\mathcal{H} - \mathcal{H}_r\|_{\mathcal{H}_2}$ and $\|\mathcal{H} - \widetilde{\mathcal{H}}_r\|_{\mathcal{H}_2}$ during the course of IRKA and IrkaIRKA, respectively, are depicted in the top plot of Fig. 4. The figure shows that IrkaIRKA behavior is almost an exact replica of that of IRKA. The deviation from the exact IRKA is noticeable in the graph only for $\varepsilon = 10^{-1}$. To illustrate how much $\mathcal{H}_r$ deviates from $\widetilde{\mathcal{H}}_r$ as IRKA and IrkaIRKA evolve, we show the progress of $\|\mathcal{H}_r - \widetilde{\mathcal{H}}_r\|_{\mathcal{H}_2}$ in the bottom plot of Fig. 4. For this example, we initialized both IRKA and IrkaIRKA with an initial reduced-order

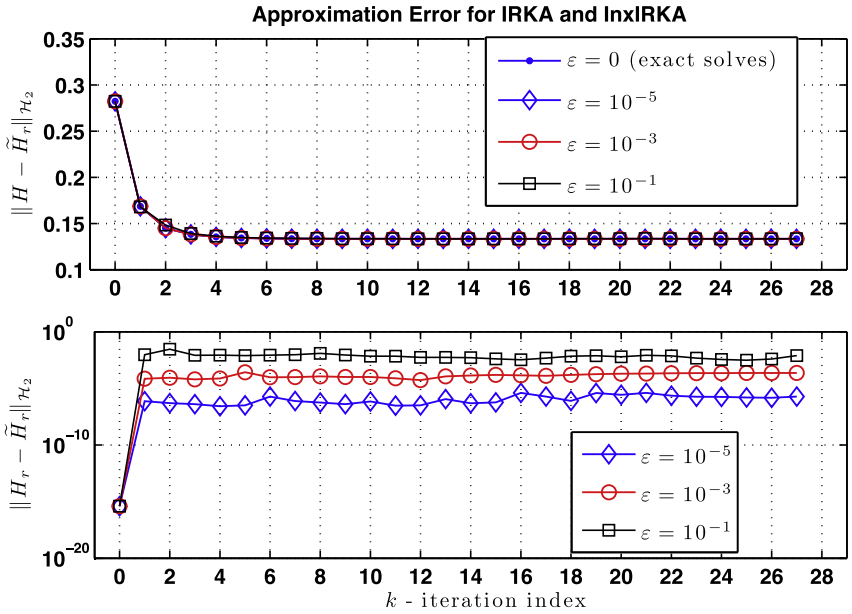


Fig. 4. Evolution of the $\mathcal{H}_2$ error during IRKA and InxIRKA.

Table 4
Evolution of the model reduction errors as ε varies.

ε	$\mathcal{H}_2$ error	$\mathcal{H}_\infty$ error
0	$3.708415753 \times 10^{-4}$	$1.084442854 \times 10^{-2}$
10^{-5}	$3.708415754 \times 10^{-4}$	$1.084425703 \times 10^{-2}$
10^{-4}	$3.708415778 \times 10^{-4}$	$1.084282001 \times 10^{-2}$
10^{-3}	$3.708418102 \times 10^{-4}$	$1.082437228 \times 10^{-2}$
10^{-2}	$3.708621743 \times 10^{-4}$	$1.064836300 \times 10^{-2}$
10^{-1}	$3.716780975 \times 10^{-4}$	$1.055441476 \times 10^{-2}$

Table 5
Evolution of the perturbation error as ε varies.

ε	$\|\mathcal{H}_r - \widetilde{\mathcal{H}}_r\|_{\mathcal{H}_2}$	$\|\mathcal{H}_r - \widetilde{\mathcal{H}}_r\|_{\mathcal{H}_\infty}$
10^{-5}	5.1921×10^{-9}	2.7776×10^{-7}
10^{-4}	5.7156×10^{-8}	2.4611×10^{-6}
10^{-3}	6.3982×10^{-7}	2.1043×10^{-5}
10^{-2}	5.9277×10^{-6}	2.0910×10^{-4}
10^{-1}	2.2056×10^{-5}	2.9228×10^{-3}

model (as opposed to specifying initial interpolation points and tangent directions). Thus, $\mathcal{H}_r = \widetilde{\mathcal{H}}_r$ initially and no linear solvers are involved in the first ($k = 0$) step. One could expect that perturbation errors due to inexact solves might accumulate over the course of the InxIRKA iteration, but this does not appear to be the case as this figure illustrates. The magnitude of $\|\mathcal{H}_r - \widetilde{\mathcal{H}}_r\|_{\mathcal{H}_2}$ remains relatively constant throughout the iteration at a magnitude proportional to the termination criterion.

The resulting $\mathcal{H}_2$ and $\mathcal{H}_\infty$ model reduction errors, $\|\mathcal{H} - \mathcal{H}_r\|_{\mathcal{H}_2}$ and $\|\mathcal{H} - \mathcal{H}_r\|_{\mathcal{H}_\infty}$ (with $\mathcal{H}_r$ obtained from IRKA), vs. $\|\mathcal{H} - \widetilde{\mathcal{H}}_r\|_{\mathcal{H}_2}$ and $\|\mathcal{H} - \widetilde{\mathcal{H}}_r\|_{\mathcal{H}_\infty}$ (with $\widetilde{\mathcal{H}}_r$ obtained from InxIRKA) are given as ε varies in Table 4 below. The row corresponding to $\varepsilon = 0$ represents the errors due to exact IRKA. These numbers demonstrate that employing inexact solves in InxIRKA does not degrade the model reduction performance. We also measure the difference between $\mathcal{H}_r$ and $\widetilde{\mathcal{H}}_r$ in both $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms as ε varies. These results are tabulated in Table 5: Note that while $\|\mathcal{H} - \mathcal{H}_r\|_{\mathcal{H}_2}$ and $\|\mathcal{H} - \mathcal{H}_r\|_{\mathcal{H}_\infty}$ are respectively $\mathcal{O}(10^{-4})$ and $\mathcal{O}(10^{-2})$, the contributions attributable to $\mathcal{H}_r - \widetilde{\mathcal{H}}_r$ are much smaller

Table 6
Optimal interpolations points as ε varies.

σ_i(IRKA)	σ_i(InxIRKA), $\varepsilon = 10^{-3}$	σ_i(InxIRKA), $\varepsilon = 10^{-1}$
1.0802×10^{-5}	1.0800×10^{-5}	1.2396×10^{-5}
9.7164×10^{-4}	9.7080×10^{-4}	9.5860×10^{-4}
6.6310×10^{-3}	6.6246×10^{-3}	6.5923×10^{-3}
5.7925×10^{-2}	5.7938×10^{-2}	5.7929×10^{-2}
9.0460×10^{-1}	9.0419×10^{-1}	8.9877×10^{-1}
1.4127×10^{0}	1.4126×10^{0}	1.4104×10^{0}

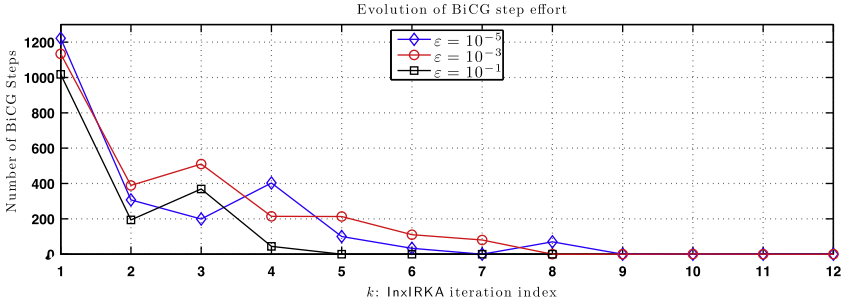


Fig. 5. Evolution of BiCG effort during InxIRKA for shift closest to the imaginary axis.

in magnitude and do not alter the resulting (optimal) model reduction performance in any significant way. If one were to convert the perturbation errors in Table 5 to relative error (as opposed to the displayed absolute error), both $\|\mathcal{H}_r - \widetilde{\mathcal{H}}_r\|_{\mathcal{H}_2}$ and $\|\mathcal{H}_r - \widetilde{\mathcal{H}}_r\|_{\mathcal{H}_\infty}$ starts at $\mathcal{O}(10^{-6})$ for $\varepsilon = 10^{-5}$, and increases linearly by one order as ε increases by the same amount.

We finally list, in Table 6, the final exact and inexact optimal interpolation points due to IRKA, and InxIRKA for $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-1}$: Not surprisingly, the resulting interpolation points are very close to each other (though not the same). This can be viewed as another illustration of the fact that $\widetilde{\mathcal{H}}_r$ is an $\mathcal{H}_2$ optimal approximation to a nearby full-order system.

As discussed above, in the implementation of InxIRKA, we used the solution vectors from the previous step as the initial guess for the linear system in the next step taking advantage of the convergence in the interpolation points and tangent directions. To illustrate the effectiveness of this simple approach, throughout InxIRKA we monitor the number of BiCG steps required to solve each linear system. We illustrate the behavior only for one of the interpolation points. We choose the interpolation points closest to the imaginary axis since these produce the hardest linear systems to solve and invariably contribute most to the cost of inexact solves. Fig. 5 depicts the number of BiCG steps required as InxIRKA proceeds for these interpolation points using three different stopping criteria $\varepsilon = 10^{-5}$, $\varepsilon = 10^{-3}$ and $\varepsilon = 10^{-1}$. The figure clearly illustrates that re-using the solutions from the previous steps works very effectively in reducing the overall cost of the BiCG. The number of BiCG steps goes from 1200 down to 200 in 3 to 4 steps.

6. Structure-preserving interpolation for descriptor systems

The backward error analysis of Section 4 has been presented for the transfer functions in the generalized coprime factorization form as in (2). In this section, we show that stronger conclusions on the structure of the reduced system can be drawn in the case the system has a realization as a descriptor system, that is,

$$\mathcal{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}, \tag{52}$$

where $\mathbf{E}, \mathbf{A} \in \mathbb{R}^{n\times n}$, $\mathbf{B} \in \mathbb{R}^{n\times m}$, and $\mathbf{C} \in \mathbb{R}^{p\times n}$ are constant matrices. In this case, for the interpolation points $\{\sigma_j\}_{j=1}^r$, and the tangent directions $\{\mathbf{b}_j\}_{j=1}^r$ and $\{\mathbf{c}_j\}_{j=1}^r$, the associated primitive interpolatory

bases $\mathbf{V}_r$ and $\mathbf{W}_r$ can be obtained from (13) and (14) using $\mathcal{K}(s) = s\mathbf{E} - \mathbf{A}$, $\mathcal{B}(s) = \mathbf{B}$ (constant matrix) and $\mathcal{C}(s) = \mathbf{C}$ (constant matrix). Then, the resulting reduced-order model is given by

$$\mathcal{H}_r(s) = \mathbf{C}_r(s\mathbf{E}_r - \mathbf{A}_r)^{-1}\mathbf{B}_r, \tag{53}$$

where

$$\mathbf{E}_r = \mathbf{W}_r^T\mathbf{E}\mathbf{V}_r, \quad \mathbf{A}_r = \mathbf{W}_r^T\mathbf{A}\mathbf{V}_r, \quad \mathbf{B}_r = \mathbf{W}_r^T\mathbf{B}, \quad \text{and} \quad \mathbf{C}_r = \mathbf{C}\mathbf{V}_r. \tag{54}$$

Let the set $\mathcal{S} = \{\sigma_i, \mathbf{b}_i, \mathbf{c}_i\}$ denote given tangential interpolation data. Define the matrices $\mathbb{L}[\mathcal{H}, \mathcal{S}] \in \mathbb{C}^{r\times r}$ and $\mathbb{M}[\mathcal{H}, \mathcal{S}] \in \mathbb{C}^{r\times r}$ corresponding to the dynamical system $\mathcal{H}(s)$ and interpolation data $\mathcal{S}$:

$$(\mathbb{L}[\mathcal{H}, \mathcal{S}])_{i,j} := \begin{cases} \dfrac{\mathbf{c}_i^T\left(\mathcal{H}(\sigma_i) - \mathcal{H}(\sigma_j)\right)\mathbf{b}_j}{\sigma_i - \sigma_j} & \text{if } \ i \neq j \\ \\ \mathbf{c}_i^T\mathcal{H}'(\sigma_i)\mathbf{b}_i & \text{if } \ i = j \end{cases} \tag{55}$$

$$(\mathbb{M}[\mathcal{H}, \mathcal{S}])_{i,j} := \begin{cases} \dfrac{\mathbf{c}_i^T\left(\sigma_i\mathcal{H}(\sigma_i) - \sigma_j\mathcal{H}(\sigma_j)\right)\mathbf{b}_j}{\sigma_i - \sigma_j} & \text{if } \ i \neq j \\ \\ \mathbf{c}_i^T\left[s\mathcal{H}(s)\right]'\big|_{s=\sigma_i}\mathbf{b}_i & \text{if } \ i = j \end{cases} \tag{56}$$

$\mathbb{L}[\mathcal{H}, \mathcal{S}]$ is the *Loewner matrix* associated with the interpolation data $\mathcal{S}$ and the dynamical system $\mathcal{H}(s)$, $\mathbb{M}[\mathcal{H}, \mathcal{S}]$ is the *shifted Loewner matrix* associated with the interpolation data $\mathcal{S}$ and the system $s\mathcal{H}(s)$, see [3,22]. The next theorem presents a canonical structure for the exact interpolatory reduced-order model (53) and (54).

Theorem 6.1 [22]. *Given a full-order model $\mathcal{H}(s) = \mathbf{C}(s\mathbf{E} - \mathbf{A})^{-1}\mathbf{B}$ and tangential interpolation data $\mathcal{S} = \{\sigma_i, \mathbf{b}_i, \mathbf{c}_i\}$, then the reduced-order quantities in (54) satisfy*

$$\begin{aligned} \mathbf{E}_r &= -\mathbb{L}[\mathcal{H}, \mathcal{S}], \\ \mathbf{A}_r &= -\mathbb{M}[\mathcal{H}, \mathcal{S}], \end{aligned} \qquad \mathbf{B}_r = \begin{bmatrix} \mathbf{c}_1^T\mathcal{H}(\sigma_1) \\ \vdots \\ \mathbf{c}_r^T\mathcal{H}(\sigma_r) \end{bmatrix}, \tag{57}$$

$$\mathbf{C}_r = \left[\ \mathcal{H}(\sigma_1)\mathbf{b}_1, \ \ldots, \ \mathcal{H}(\sigma_r)\mathbf{b}_r\ \right].$$

6.1. The Petrov–Galerkin framework and structure preservation

Theorem 6.1 presents a canonical form for the exact bitangential Hermite interpolant in the case of standard state-space model. Next we show that if a Petrov–Galerkin framework is employed in the solution of the linear systems, the inexact reduced-model will have exactly the same form as the exact one. The result is a direct consequence of Theorems 4.1 and 6.1.

Corollary 6.1. *Given the standard full-order model $\mathcal{H}(s) = \mathbf{C}(s\mathbf{E}-\mathbf{A})^{-1}\mathbf{B}$ together with the interpolation data $\mathcal{S} = \{\sigma_i, \mathbf{b}_i, \mathbf{c}_i\}$, let the inexact solutions $\widetilde{\mathbf{v}}_j$ for $(\sigma_j\mathbf{E}-\mathbf{A})^{-1}\mathbf{B}\mathbf{b}_j$ and $\widetilde{\mathbf{w}}_j$ for $(\sigma_j\mathbf{E}-\mathbf{A})^{-T}\mathbf{C}^T\mathbf{c}_j$ be obtained in a Petrov–Galerkin framework as in (42). Let $\widetilde{\mathbf{V}}_r$ and $\widetilde{\mathbf{W}}_r$ denote the corresponding inexact Krylov bases as in (43). Define the residuals*

$$\widetilde{\boldsymbol{\eta}}_j = (\sigma_j\mathbf{E} - \mathbf{A})\widetilde{\mathbf{v}}_j - \mathbf{B}\mathbf{b}_j \quad \text{and} \quad \widetilde{\boldsymbol{\xi}}_j = (\sigma_j\mathbf{E} - \mathbf{A})^T\widetilde{\mathbf{w}}_j - \mathbf{C}^T\mathbf{c}_j.$$

Let the residual matrices $\mathbf{R_b}$ *and* $\mathbf{R_c}$*, and the rank* $2r$ *matrix* $\mathbf{F}_{2r}$ *be as defined in (44) and (45), respectively. Then, the inexact interpolatory reduced-order model*

$$\widetilde{\mathcal{H}}_r(s) = \widetilde{\mathbf{C}}_r(s\widetilde{\mathbf{E}}_r - \widetilde{\mathbf{A}}_r)^{-1}\widetilde{\mathbf{B}}_r \tag{58}$$

is an exact Hermite bitangential interpolant for the perturbed full-order model

$$\widetilde{\mathcal{H}}(s) = \mathbf{C}(s\mathbf{E} - (\mathbf{A} + \mathbf{F}_{2r}))^{-1}\mathbf{B}. \tag{59}$$

Moreover, the reduced-order quantities satisfy

$$\begin{matrix}\widetilde{\mathbf{E}}_r = -\mathbb{L}[\widetilde{\mathcal{H}}, \mathcal{S}], \\ \widetilde{\mathbf{A}}_r = -\mathbb{M}[\widetilde{\mathcal{H}}, \mathcal{S}],\end{matrix} \quad \widetilde{\mathbf{B}}_r = \begin{bmatrix} \mathbf{c}_1^T\widetilde{\mathcal{H}}(\sigma_1) \\ \vdots \\ \mathbf{c}_r^T\widetilde{\mathcal{H}}(\sigma_r) \end{bmatrix}, \quad \widetilde{\mathbf{C}}_r = [\ \widetilde{\mathcal{H}}(\sigma_1)\mathbf{b}_1,\ \ \ldots,\ \widetilde{\mathcal{H}}(\sigma_r)\mathbf{b}_r\ ]. \tag{60}$$

where $\mathbb{L}[\widetilde{\mathcal{H}}, \mathcal{S}]$ *and* $\mathbb{M}[\widetilde{\mathcal{H}}, \mathcal{S}]$ *are the Loewner matrices associated with the dynamical systems* $\widetilde{\mathcal{H}}(s)$ *and* $s\widetilde{\mathcal{H}}(s)$ *respectively, and the interpolation data* $\mathcal{S}$ *as defined in (55) and (56).*

Corollary 6.1 reveals that the inexact reduced-order model quantities have exactly the same structure as their exact counterparts. The interpolation data $\mathcal{S}$ is the same in both cases; the only difference is that $\mathcal{H}(s)$ is replaced by $\widetilde{\mathcal{H}}(s)$ in the construction that yields the Loewner-matrix structure. The preservation of this structure is independent of the accuracy to which the linear systems are solved. In the case where $\mathbf{E} = \mathbf{I}$, the structure of the exact and inexact reduced-models becomes even simpler:

Corollary 6.2. *Assume the hypotheses of Theorem 6.1 with* $\mathbf{E} = \mathbf{I}$*. Then the exact interpolant* $\mathcal{H}_r(s) = \mathbf{C}_r(s\mathbf{I}_r - \mathbf{A}_r)^{-1}\mathbf{B}_r$ *satisfies*

$$\mathbf{A}_r = \mathbf{\Sigma} - \mathbf{Q}\mathbf{B}, \quad \mathbf{B}_r = \mathbf{Q}, \quad \text{and} \quad \mathbf{C}_r = [\ \mathcal{H}(\sigma_1)\mathbf{b}_1,\ \ \ldots,\ \mathcal{H}(\sigma_r)\mathbf{b}_r\ ], \tag{61}$$

where

$$\mathbf{Q} = (\mathbf{W}_r^T\mathbf{V}_r)^{-1}\mathbf{W}_r^T\mathbf{B}, \quad \mathbf{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \quad \text{and} \quad \mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_r]. \tag{62}$$

Assume the hypotheses of Corollary 6.1 with $\mathbf{E} = \mathbf{I}$*. Then, the inexact interpolant* $\widetilde{\mathcal{H}}_r(s) = \widetilde{\mathbf{C}}_r(s\mathbf{I}_r - \widetilde{\mathbf{A}}_r)^{-1}\widetilde{\mathbf{B}}_r$ *satisfies*

$$\widetilde{\mathbf{A}}_r = \mathbf{\Sigma} - \widetilde{\mathbf{Q}}\mathbf{B}, \quad \widetilde{\mathbf{B}}_r = \widetilde{\mathbf{Q}}, \quad \text{and} \quad \widetilde{\mathbf{C}}_r = [\ \widetilde{\mathcal{H}}(\sigma_1)\mathbf{b}_1,\ \ \ldots,\ \widetilde{\mathcal{H}}(\sigma_r)\mathbf{b}_r\ ], \tag{63}$$

where

$$\widetilde{\mathbf{Q}} = (\widetilde{\mathbf{W}}_r^T\widetilde{\mathbf{V}}_r)^{-1}\widetilde{\mathbf{W}}_r^T\mathbf{B}, \tag{64}$$

$\widetilde{\mathcal{H}}(s)$ *is the perturbed full-order model as in (59) with* $\mathbf{E} = \mathbf{I}$*, and* $\mathbf{\Sigma}$ *and* $\mathbf{B}$ *are as defined in (62).*

Corollary 6.2 illustrates that in the case of $\mathbf{E} = \mathbf{I}$, both of the reduced system matrices, $\mathbf{A}_r$ and $\widetilde{\mathbf{A}}_r$, are perturbations of rank $\min(r, m, p)$ to the diagonal matrix of interpolation points, $\mathbf{\Sigma}$.

References

[1] Kapil Ahuja, Recycling Bi-Lanczos algorithms: BiCG, CGS, BiCGSTAB, Master's thesis, Virginia Tech, Blacksburg, Virginia, August 2009.
[2] A.C. Antoulas, Approximation of Large-Scale Dynamical Systems (Advances in Design and Control), Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2005.

[3] A.C. Antoulas, C.A. Beattie, S. Gugercin, Interpolatory model reduction of large-scale dynamical systems, in: J. Mohammadpour, K. Grigoriadis (Eds.), Efficient Modeling and Control of Large-Scale Systems, Springer-Verlag, 2010.
[4] R. Barrett, M. Berry, T.F. Chan, J. Demmel, J.M. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, H. Van der Vorst, Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods,, Society for Industrial Mathematics, 1994, pp. .
[5] C.A. Beattie, S. Gugercin, Krylov-based minimization for optimal $\mathcal{H}_2$ model reduction, in: 46th IEEE Conference on Decision and Control, December 2007, pp. 4385–4390.
[6] C.A. Beattie, S. Gugercin, Interpolatory projection methods for structure-preserving model reduction, Systems Control Lett. 58 (3) (2009) 225–232.
[7] C.A. Beattie, S. Gugercin, A trust region method for optimal $\mathcal{H}_2$ model reduction, in: 48th IEEE Conference on Decision and Control, December 2009.
[8] P. Benner, Solving large-scale control problems, IEEE Control Systems Mag. 24 (1) (2004) 44–59.
[9] P. Benner, J. Saak, Efficient numerical solution of the LQR-problem for the heat equation, Proc. Appl. Math. Mech. 4 (1) (2004) 648–649.
[10] A. Bunse-Gerstner, D. Kubalinska, G. Vossen, D. Wilczek, $\mathcal{H}_2$-optimal model reduction for large scale discrete dynamical MIMO systems, J. Comput. Appl. Math. (2009), doi:10.1016/j.cam.2008.12.029.
[11] K. Gallivan, A. Vandendorpe, P. Van Dooren, Model reduction via truncation: an interpolation point of view, Linear Algebra Appl. 375 (2003) 115–134.
[12] K. Glover, All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds, Int. J. Control 39 (6) (1984) 1115–1193.
[13] E. Grimme, Krylov projection methods for model reduction, Ph.D. thesis, Coordinated-Science Laboratory, University of Illinois at Urbana-Champaign, 1997.
[14] S. Gugercin, An iterative rational Krylov algorithm (IRKA) for optimal $\mathcal{H}_2$ model reduction, in: Householder Symposium XVI, Seven Springs Mountain Resort, PA, USA, May 2005.
[15] S. Gugercin, A.C. Antoulas, C.A. Beattie, A rational Krylov iteration for optimal $\mathcal{H}_2$ model reduction, in: Proceedings of MTNS, vol. 2006, 2006.
[16] S. Gugercin, A.C. Antoulas, C.A. Beattie, $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems, SIAM J. Matrix Anal. Appl. 30 (2) (2008) 609–638.
[17] Y. Halevi, Frequency weighted model reduction via optimal projection, IEEE Trans. Automat. Control 37 (10) (1992) 1537–1542.
[18] D. Hyland, D. Bernstein, The optimal projection equations for model reduction and the relationships among the methods of Wilson, Skelton, and Moore, IEEE Trans. Automat. Control 30 (12) (1985) 1201–1211.
[19] J.G. Korvink, E.B. Rudnyi, Oberwolfach benchmark collection, in: Dimension Reduction of Large-scale Systems: Proceedings of a Workshop Held in Oberwolfach, Germany, October 19–25, 2003, Springer-Verlag, 2005, pp. 311.
[20] D. Kubalinska, A. Bunse-Gerstner, G. Vossen, D. Wilczek, $\mathcal{H}_2$-optimal interpolation based model reduction for large-scale systems, in: Proceedings of the 16th International Conference on System Science, Poland, 2007.
[21] Y. Liu, B.D.O. Anderson, Singular perturbation approximation of balanced systems, Int. J. Control 50 (4) (1989) 1379–1405.
[22] A.J. Mayo, A.C. Antoulas, A framework for the solution of the generalized realization problem, Linear Algebra Appl. 425 (2–3) (2007) 634–662.
[23] L. Meier III, D. Luenberger, Approximation of linear constant systems, IEEE Trans. Automat. Control 12 (5) (1967) 585–588.
[24] B. Moore, Principal component analysis in linear systems: controllability, observability, and model reduction, IEEE Trans. Automat. Control 26 (1) (1981) 17–32.
[25] C. Mullis, R. Roberts, Synthesis of minimum roundoff noise fixed point digital filters, IEEE Trans. Circuits Systems 23 (9) (1976) 551–562.
[26] J.T. Spanos, M.H. Milman, D.L. Mingori, A new algorithm for L_2 optimal model reduction, Automatica J. IFAC 28 (5) (1992) 897–909.
[27] D.B. Szyld, The many proofs of an identity on the norm of oblique projections, Numer. Algorithms 42 (3) (2006) 309–323.
[28] A. van der Sluis, Condition numbers and equilibration of matrices, Numer. Math. 14 (1) (1969) 14–23.
[29] P. van Dooren, K.A. Gallivan, P.A. Absil, $\mathcal{H}_2$-optimal model reduction of MIMO systems, Appl. Math. Lett. 21 (12) (2008) 1267–1273.
[30] D.A. Wilson, Optimum solution of model-reduction problem, Proc. IEE 117 (6) (1970) 1161–1165.
[31] W.Y. Yan, J. Lam, An approximate approach to $\mathcal{H}_2$ optimal model reduction, IEEE Trans. Automat. Control 44 (7) (1999) 1341–1358.
[32] D. Zigic, L.T. Watson, C. Beattie, Contragradient transformations applied to the optimal projection equations, Linear Algebra Appl. 188 (1993) 665–676.