Vietnam Journal of Mathematics (2020) 48:963–985
https://doi.org/10.1007/s10013-020-00424-0

ORIGINAL ARTICLE

Revisiting IRKA: Connections with Pole Placement and Backward Stability

Christopher Beattie[1] · **Zlatko Drmač**[2] · **Serkan Gugercin**[1]

Received: 13 November 2019 / Accepted: 30 March 2020 / Published online: 5 August 2020

Abstract
The iterative rational Krylov algorithm (IRKA) is a popular approach for producing locally optimal reduced-order $\mathcal{H}_2$-approximations to linear time-invariant (LTI) dynamical systems. Overall, IRKA has seen significant practical success in computing high fidelity (locally) optimal reduced models and has been successfully applied in a variety of large-scale settings. Moreover, IRKA has provided a foundation for recent extensions to the systematic model reduction of bilinear and nonlinear dynamical systems. Convergence of the basic IRKA iteration is generally observed to be rapid—but not always; and despite the simplicity of the iteration, its convergence behavior is remarkably complex and not well understood aside from a few special cases. The overall effectiveness and computational robustness of the basic IRKA iteration is surprising since its algorithmic goals are very similar to a pole assignment problem, which can be notoriously ill-conditioned. We investigate this connection here and discuss a variety of nice properties of the IRKA iteration that are revealed when the iteration is framed with respect to a primitive basis. We find that the connection with pole assignment suggests refinements to the basic algorithm that can improve convergence behavior, leading also to new choices for termination criteria that assure backward stability.

Keywords Interpolation · Model reduction · $\mathcal{H}_2$-optimality · Pole placement · Backward stability

Mathematics Subject Classification (2010) 15A12 · 41A05 · 49K15 · 93A15 · 93C05 · 93B55

1 Introduction

The iterative rational Krylov algorithm (IRKA) was introduced in [29] as an approach for producing locally optimal reduced-order $\mathcal{H}_2$-approximations to linear time-invariant (LTI)

Dedicated to Volker Mehrmann on the occasion of his 65th birthday.

✉ Serkan Gugercin
gugercin@vt.edu

Extended author information available on the last page of the article.

dynamical systems given, say, as

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{b}u(t), \qquad y(t) = \mathbf{c}^T\mathbf{x}(t), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$, and $\mathbf{b}, \mathbf{c} \in \mathbb{R}^n$. We will assume that the dynamical system is stable, i.e., all the eigenvalues of $\mathbf{A}$ have negative real parts. The cases of interest will be when n is very large, and we seek a substantially lower order dynamical system, say,

$$\dot{\mathbf{x}}_r(t) = \mathbf{A}_r\mathbf{x}_r(t) + \mathbf{b}_r\, u(t), \qquad y_r(t) = \mathbf{c}_r^T\mathbf{x}_r(t), \tag{2}$$

with $\mathbf{A}_r \in \mathbb{R}^{r\times r}$, and $\mathbf{b}_r, \mathbf{c}_r \in \mathbb{R}^r$. One seeks a realization (2) so that the reduced system order $r \ll n$ and the reduced system output $y_r \approx y$ uniformly well over all inputs $u \in \mathcal{L}_2$ with $\int_0^\infty |u(t)|^2\, dt \le 1$.

Projection-based model reduction is a common framework to obtain reduced models: Given the full model (1), construct two model reduction bases $\mathbf{V}, \mathbf{W} \in \mathbb{C}^{n\times r}$ with $\mathbf{W}^T\mathbf{V}$ invertible. Then the reduced model quantities in (2) are given by

$$\mathbf{A}_r = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{A}\mathbf{V}, \quad \mathbf{b}_r = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{b}, \quad \text{and} \quad \mathbf{c}_r = \mathbf{c}\mathbf{V}. \tag{3}$$

The following question arises: How to choose $\mathbf{V}$ and $\mathbf{W}$ so that the reduced model is a high-fidelity approximation to the original one? There are many different ways to construct $\mathbf{V}$ and $\mathbf{W}$, and we refer the reader to [3, 4, 14] for detailed descriptions of such methods for linear dynamical systems. Here we focus on constructing optimal interpolatory reduced models.

To discuss interpolation and optimality, we first need to define the concept of transfer function. Let $\mathcal{Y}(s)$, $\mathcal{Y}_r(s)$, and $\mathcal{U}(s)$ denote the Laplace transforms of $y(t)$, $y_r(t)$, and $u(t)$, respectively. Taking the Laplace transforms of (1) and (2) yields

$$\begin{aligned} \mathcal{Y}(s) &= H(s)\mathcal{U}(s) \quad \text{where } H(s) = \mathbf{c}^T(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \quad \text{and} \\ \mathcal{Y}_r(s) &= H_r(s)\mathcal{U}(s) \quad \text{where } H_r(s) = \mathbf{c}_r^T(s\mathbf{I}_r - \mathbf{A}_r)^{-1}\mathbf{b}_r. \end{aligned}$$

The rational functions $H(s)$ and $H_r(s)$ are the transfer functions associated with the full model (1) and the reduced model (2). While $H(s)$ is a degree-n rational function, $H_r(s)$ is of degree-r.

Interpolatory model reduction aims to construct an $H_r(s)$ that interpolates $H(s)$ at selected points in the complex plane. Indeed, we will focus on Hermite interpolation, as this will be tied to optimality later. Suppose we are given r mutually distinct interpolation points (also called *shifts*), $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_r\}$, in the complex plane. We will assume that the shifts have positive real parts and that are closed (as a set) under conjugation, i.e., there exists an index permutation $(i_1, i_2, \ldots, i_r)$ such that $\overline{\boldsymbol{\sigma}} = \{\overline{\sigma_1}, \overline{\sigma_2}, \ldots, \overline{\sigma_r}\} = \{\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_r}\}$.

Given $\boldsymbol{\sigma}$, construct the model reduction bases $\mathbf{V} \in \mathbb{C}^{n\times r}$ and $\mathbf{W} \in \mathbb{C}^{n\times r}$ such that

$$\mathsf{Range}(\mathbf{V}) = \mathsf{span}\left\{(\sigma_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \ldots, (\sigma_r\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\right\} \quad \text{and} \tag{4}$$

$$\mathsf{Range}(\mathbf{W}) = \mathsf{span}\left\{(\sigma_1\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}, \ldots, (\sigma_r\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}\right\}. \tag{5}$$

Then, the reduced model (2) constructed as in (3) satisfies

$$H_r(\sigma_i) = H(\sigma_i) \quad \text{and} \quad H_r'(\sigma_i) = H'(\sigma_i) \quad \text{for } i = 1, 2, \ldots, r. \tag{6}$$

In other words, $H_r(s)$ is a rational Hermite interpolant to $H(s)$ at the specified interpolation points. However, this construction requires knowing the interpolation points. How should one choose them to guarantee a high-fidelity reduced model?

We will measure fidelity using the $\mathcal{H}_2$ norm: The $\mathcal{H}_2$ norm of a dynamical system with transfer function $H(s)$ is defined as

$$\|H\|_{\mathcal{H}_2} = \sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} |H(\mathrm{i}\omega)|^2 d\omega}.$$

For the full model (1) and the reduced model (2), the output error satisfies

$$\|y - y_r\|_{L_\infty} \leq \|H - H_r\|_{\mathcal{H}_2} \|u\|_{L_2},$$

where $\|y - y_r\|_{L_\infty} = \sup_{t\geq 0} |y(t) - y_r(t)|$ and $\|u\|_{L_2} = \sqrt{\int_0^\infty |u(t)|^2 dt}$. So, a reduced model that minimizes the $\mathcal{H}_2$ distance $\|H - H_r\|_{\mathcal{H}_2}$ is guaranteed to yield uniformly good approximations over finite energy inputs. Therefore, it is desirable to find a reduced model with transfer function $H_r(s)$ that minimizes the $\mathcal{H}_2$ distance, i.e., to find $H_r(s)$ such that

$$\|H - H_r\|_{\mathcal{H}_2} = \min_{\substack{G_r \text{ stable} \\ \text{order } G_r \leq r}} \|H - G_r\|_{\mathcal{H}_2}$$

at least locally in a neighborhood of H_r. This is a heavily studied topic; see, e.g., [13, 33, 41, 44, 45] for Sylvester-equation formulation and [9, 16, 29, 34, 38, 42, 46] for interpolation formulation. Indeed, these two formulations are equivalent as shown in [29] and we focus on the interpolatory formulation.

How does the $\mathcal{H}_2$ optimality relate to Hermite interpolation? Let $\mu_1, \ldots, \mu_r$ be the eigenvalues of $\mathbf{A}_r$, assumed simple. If $H_r(s)$ is an $\mathcal{H}_2$-optimal approximation to $H(s)$, then it is a Hermite interpolant to $H(s)$ at the points $\sigma_i = -\mu_i$, i.e.,

$$H_r(-\mu_i) = H(-\mu_i) \quad \text{and} \quad H_r'(-\mu_i) = H'(-\mu_i), \qquad \text{for } i = 1, 2, \ldots, r. \tag{7}$$

These conditions are known as Meier–Luenberger conditions for optimality [38]. However, one cannot simply use $\sigma_i = -\mu_i$ in constructing $\mathbf{V}$ and $\mathbf{W}$ in (4)–(5) since μ_is are not known a priori. This requires an iteratively corrected algorithm. The iterative rational Krylov algorithm IRKA [29] as outlined in Algorithm 1 precisely achieves this task. It reflects the intermediate interpolation points until the required optimality criterion, i.e., $\sigma_i = -\mu_i$ is met. Upon convergence, the reduced model is a locally optimal $\mathcal{H}_2$-approximation to (2). IRKA has been successful in producing locally optimal reduced models at a modest cost and many variants have been proposed; see, e.g., [7–9, 15, 27, 28, 32, 39, 40, 43]. Moreover, it has been successfully extended to model reduction of bilinear [11, 26] and quadratic-bilinear systems [12], two important classes of structured nonlinear systems.

Algorithm 1 $(\mathbf{A}_r, \mathbf{b}_r, \mathbf{c}_r) = \mathsf{IRKA}(\mathbf{A}, \mathbf{b}, \mathbf{c}, r)$.

1: Initialize shifts $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_r) \subset \mathbb{C}_+ \equiv \{z \in \mathbb{C} : \mathrm{Re}(z) > 0\}$ that are closed (as a set) under under conjugation;
2: **repeat**
3: Compute a basis of $\mathsf{span}((\sigma_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}, \ldots, (\sigma_r\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}) \to \mathbf{V}$;
4: Compute a basis of $\mathsf{span}((\sigma_1\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}, \ldots, (\sigma_r\mathbf{I} - \mathbf{A}^T)^{-1}\mathbf{c}) \to \mathbf{W}$;
5: $\mathbf{A}_r = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{A}\mathbf{V}$;
6: Compute the eigenvalues (reduced poles) $\boldsymbol{\lambda}(\mathbf{A}_r) = (\lambda_1(\mathbf{A}_r), \ldots, \lambda_r(\mathbf{A}_r))$;
7: Compute the (matching) distance ζ between the sets $\boldsymbol{\lambda}(\mathbf{A}_r)$ and $-\boldsymbol{\sigma}$;
8: $\sigma_i \longleftarrow -\lambda_i(\mathbf{A}_r)$, $i = 1, \ldots, r$;
9: **until** ζ *sufficiently small*
10: $\mathbf{b}_r = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{b}$; $\mathbf{c}_r = \mathbf{V}^T\mathbf{c}$;
11: The reduced order model is $(\mathbf{A}_r, \mathbf{b}_r, \mathbf{c}_r)$.

Our goal in this paper is not to compare model reduction techniques, nor is it to illustrate the effectiveness of IRKA and its variants. For example, reduced models produced by IRKA, as specified in Algorithm 1, are not *a priori* guaranteed to be asymptotically stable although there is overwhelming numerical evidence that one should expect this [4, 5]. In practice, one might consider only mirroring the stable eigenvalues in Step 8 of Algorithm 1. We will ignore these issues here and refer instead to sources cited above for supporting analyses. Our main goal here is to revisit IRKA in its original form and reveal new connections to the pole placement problem (Section 3) by a thorough analysis of the quantities involved in a special basis (Section 2). This will lead to a backward stability formulation relating then to new stopping criteria (Section 4).

In order to keep the discussion concise, we focus here on single-input/single-output dynamical systems, i.e., $u(t), y(t), y_r(t) \in \mathbb{R}$. For detailed discussion of $\mathcal{H}_2$-optimal model reduction in the complementary multi-input/multi-output case, see [4, 5, 14].

2 Structure in the Primitive Bases

In Steps 3 and 4 of IRKA as laid out in Algorithm 1 above, the matrices $\mathbf{V}$ and $\mathbf{W}$ are each chosen as bases for a pair of rational Krylov subspaces. The reduced model is independent of the particular bases chosen and one usually constructs them to be orthonormal. We consider a different choice in this section, and show that if $\mathbf{V}$ and $\mathbf{W}$ are chosen instead as *primitive* bases, i.e., if

$$\mathbf{V} = \left[\ (\sigma_1\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\ \ \ldots\ \ (\sigma_r\mathbf{I} - \mathbf{A})^{-1}\mathbf{b}\ \right] \quad \text{and} \quad \mathbf{W} = \left[\ (\sigma_1\mathbf{I} - \mathbf{A})^{-T}\mathbf{c}\ \ \ldots\ \ (\sigma_r\mathbf{I} - \mathbf{A})^{-T}\mathbf{c}\ \right], \quad (8)$$

then the state-space realization of the reduced model exhibits an important structure which forms the foundation of our analysis that follows in Sections 3 and 4. Therefore, in the rest of the paper, we use primitive bases for $\mathbf{V}$ and $\mathbf{W}$ as given in (8). We emphasize that this does not change the reduced model $H_r(s)$; it is simply a change of basis that reveals nontrivial structure that can be exploited both in the theoretical analysis of the algorithm and for its efficient software implementation.

It is easy to check that ([29]), for $\mathbf{V}$ and $\mathbf{W}$ as primitive bases (8), the matrices $\mathbf{W}^T\mathbf{A}\mathbf{V}$ and $\mathbf{W}^T\mathbf{V}$ are symmetric; but not necessarily Hermitian. Moreover, one may directly verify

that ([5]), $\mathbf{W}^T\mathbf{V}$ is the Loewner matrix whose (i, j)th entry, for $i, j = 1, \ldots, r$, is given by

$$(\mathbf{W}^T\mathbf{V})_{ij} = [\sigma_i, \sigma_j]H := \frac{H(\sigma_i) - H(\sigma_j)}{\sigma_i - \sigma_j}, \tag{9}$$

with the convention that $[\sigma_i, \sigma_i]H = H'(\sigma_i)$.

Lemma 1 *Let $\omega_r(z) = (z - \sigma_1)(z - \sigma_2)\ldots(z - \sigma_r)$ be the nodal polynomial associated with the shifts $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \ldots, \sigma_r\}$. For any monic polynomial $p_r \in \mathcal{P}_r$, define the vector*

$$\mathbf{q} = (q_1, \ldots, q_r)^T, \quad q_i = \frac{p_r(\sigma_i)}{\omega_r'(\sigma_i)}, \quad i = 1, \ldots, r,$$

and the matrix $\mathbf{A}_r = \boldsymbol{\Sigma}_r - \mathbf{q}\mathbf{e}^T$ with $\boldsymbol{\Sigma}_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Then $\det(z\mathbf{I} - \mathbf{A}_r) = p_r(z)$ and

$$\mathbf{A}\mathbf{V} - \mathbf{V}\mathbf{A}_r = -p_r(\mathbf{A})[\omega_r(\mathbf{A})]^{-1}\mathbf{b}\mathbf{e}^T, \tag{10}$$

$$\mathbf{W}^T\mathbf{A} - \mathbf{A}_r^T\mathbf{W}^T = -\mathbf{e}\mathbf{c}^T p_r(\mathbf{A})[\omega_r(\mathbf{A})]^{-1}. \tag{11}$$

Proof Pick any index $1 \le k \le r$ and consider $f_k(z) = p_r(z) - z \cdot \prod_{i \ne k}(z - \sigma_i)$. Evidently, $f_k \in \mathcal{P}_{r-1}$ and so the Lagrange interpolant on $\sigma_1, \sigma_2, \ldots, \sigma_r$ is exact:

$$f_k(z) = p_r(z) - z \cdot \prod_{i \ne k}(z - \sigma_i) = \sum_{i=1}^{r} f_k(\sigma_i)\frac{\omega_r(z)}{(z - \sigma_i)\omega_r'(\sigma_i)}.$$

Divide by $\omega_r(z)$ and rearrange to obtain

$$\frac{z}{\sigma_k - z} - \sum_{i=1}^{r}\left(-\frac{f_k(\sigma_i)}{\omega_r'(\sigma_i)}\right)\frac{1}{\sigma_i - z} = -\frac{p_r(z)}{\omega_r(z)}. \tag{12}$$

Let Γ be a Jordan curve that separates $\mathbb{C}$ into two open, simply-connected sets, $\mathcal{C}_1, \mathcal{C}_2$ with $\mathcal{C}_1$ containing all the eigenvalues of $\mathbf{A}$ and $\mathcal{C}_2$ containing both the point at ∞ and the shifts $\{\sigma_1, \ldots, \sigma_r\}$. For any function $f(z)$ that is analytic in a compact set containing $\mathcal{C}_1$, $f(\mathbf{A})$ can be defined as $f(\mathbf{A}) = \frac{1}{2\pi\mathrm{i}}\int_\Gamma f(z)(z\mathbf{I} - \mathbf{A})^{-1}dz$. Applying this to (12) gives

$$\mathbf{A}(\sigma_k\mathbf{I} - \mathbf{A})^{-1} - \sum_{i=1}^{r}\left(-\frac{f_k(\sigma_i)}{\omega_r'(\sigma_i)}\right)(\sigma_i\mathbf{I} - \mathbf{A})^{-1} = -p_r(\mathbf{A})[\omega_r(\mathbf{A})]^{-1}.$$

Postmultiplication by $\mathbf{b}$ provides the kth column of (10), while premultiplication by $\mathbf{c}^T$ (and since $(\sigma_k\mathbf{I} - \mathbf{A})^{-1}$ commutes with $\mathbf{A}$) provides the kth row of (11).

To compute the characteristic polynomial of $\mathbf{A}_r$, we use the alternative factorizations[1],

$$\begin{pmatrix} z\mathbf{I} - \boldsymbol{\Sigma}_r & \mathbf{q} \\ \mathbf{e}^T & -1 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & -\mathbf{q} \\ \mathbf{0}^T & 1 \end{pmatrix}\begin{pmatrix} z\mathbf{I} - \mathbf{A}_r & \mathbf{0} \\ \mathbf{0}^T & -1 \end{pmatrix}\begin{pmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{e}^T & 1 \end{pmatrix}$$

$$\begin{pmatrix} z\mathbf{I} - \boldsymbol{\Sigma}_r & \mathbf{q} \\ \mathbf{e}^T & -1 \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{e}^T(z\mathbf{I} - \boldsymbol{\Sigma}_r)^{-1} & 1 \end{pmatrix}\begin{pmatrix} z\mathbf{I} - \boldsymbol{\Sigma}_r & \mathbf{0} \\ \mathbf{0}^T & -a(z) \end{pmatrix}\begin{pmatrix} \mathbf{I} & (z\mathbf{I} - \boldsymbol{\Sigma}_r)^{-1}\mathbf{q} \\ \mathbf{0} & 1 \end{pmatrix},$$

[1]For the reader's convenience, here we actually reproduce the proof of the Sherman–Morrison determinant formula.

where $a(z) = 1 + \mathbf{e}^T(z\mathbf{I} - \mathbf{\Sigma}_r)^{-1}\mathbf{q}$. Then we have that

$$\begin{aligned}\det(z\mathbf{I} - \mathbf{A}_r) &= \det(z\mathbf{I} - \mathbf{\Sigma}_r) \cdot a(z) = \omega_r(z) \cdot \left(1 + \mathbf{e}^T(z\mathbf{I} - \mathbf{\Sigma}_r)^{-1}\mathbf{q}\right) \\ &= \omega_r(z) + \sum_{i=1}^{r} p_r(\sigma_i)\left(\frac{\omega_r(z)}{\omega_r'(\sigma_i)(z - \sigma_i)}\right) = p_r(z), \end{aligned} \tag{13}$$

where the last equality follows by observing that the penultimate expression describes a monic polynomial of degree r that interpolates p_r at $\sigma_1, \sigma_2, \ldots, \sigma_r$. □

Lemma 1, and more specifically (10) and (11), reveal the rational Krylov structure arising from the choice of $\mathbf{V}$ and $\mathbf{W}$ in (8). At first, connection of the involved quantities such as the vector $\mathbf{q}$ to IRKA quantities might not be clear. In the following result, we make these connections precise.

Lemma 2 *Let the reduced model $H_r(s) = \mathbf{c}_r^T(s\mathbf{I} - \mathbf{A}_r)^{-1}\mathbf{b}_r$ be obtained by projection as in (3) using the primitive bases $\mathbf{W}$ and $\mathbf{V}$ in (8). Let*

$$p_r(z) = \det(z\mathbf{W}^T\mathbf{V} - \mathbf{W}^T\mathbf{A}\mathbf{V})/\det(\mathbf{W}^T\mathbf{V}).$$

Then (10) and (11) hold with $\mathbf{A}_r = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{A}\mathbf{V}$. In particular, $\mathbf{b}_r = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{b} = \mathbf{q}$ and $\mathbf{A}_r = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{A}\mathbf{V} = \mathbf{\Sigma}_r - \mathbf{e}\mathbf{q}^T$. Moreover, if μ_ℓ is an eigenvalue of $\mathbf{A}_r$, then

$$\mathbf{x}_\ell = (\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1}\,\mathbf{q} \tag{14}$$

is an associated (right) eigenvector of $\mathbf{A}_r$ for $\ell = 1, 2, \ldots, r$. Similarly, the vector

$$\mathbf{x}_\ell^T\mathbf{W}^T\mathbf{V} = v_\ell\mathbf{e}^T\,(\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1} \tag{15}$$

is an associated left eigenvector of $\mathbf{A}_r$, for $\ell = 1, 2, \ldots, r$, with $v_\ell = \hat{\phi}_\ell\frac{p_r'(\mu_\ell)}{\omega_r(\mu_\ell)}$, and $\hat{\phi}_\ell$ is the residue of $H_r(s)$ at $s = \mu_\ell$, i.e., $\hat{\phi}_\ell = \lim_{s\to\mu_\ell}(s - \mu_\ell)H_r(s)$.

Proof Choose a monic polynomial $\hat{p}_r \in \mathcal{P}_r$ so that $\mathbf{W}^T\hat{p}_r(\mathbf{A})[\omega_r(\mathbf{A})]^{-1}\mathbf{b} = 0$. Then (10) and (11) hold with an associated $\mathbf{A}_r = \mathbf{\Sigma}_r - \mathbf{q}\mathbf{e}^T$ as given in Lemma 1. But then applying $\mathbf{W}^T$ to (10) yields $\mathbf{A}_r = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{A}\mathbf{V}$. This in turn implies $\hat{p}_r(z) = p_r(z)$.

Suppose that μ_ℓ is an eigenvalue of $\mathbf{A}_r$. Directly substitute $\mathbf{x}_\ell = (\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1}\mathbf{q}$ and use (13) with $z = \mu_\ell$ to obtain

$$\begin{aligned}\mathbf{A}_r\mathbf{x}_\ell &= \left(\mathbf{\Sigma}_r - \mathbf{q}\mathbf{e}^T\right)(\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1}\,\mathbf{q} = \left(\mathbf{\Sigma}_r - \mu_\ell\mathbf{I} - \mathbf{q}\mathbf{e}^T + \mu_\ell\mathbf{I}\right)(\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1}\,\mathbf{q} \\ &= \mathbf{q}\left(1 - \mathbf{e}^T\,(\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1}\,\mathbf{q}\right) + \mu_\ell\,(\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1}\,\mathbf{q} = \mu_\ell\mathbf{x}_\ell.\end{aligned}$$

Thus, $\mathbf{x}_\ell$ is a right eigenvector for $\mathbf{A}_r$ associated with μ_ℓ. Note that $\mathbf{x}_\ell$ also solves the generalized eigenvalue problems:

$$\text{(a) } \mathbf{W}^T\mathbf{A}\mathbf{V}\mathbf{x}_\ell = \mu_\ell\mathbf{W}^T\mathbf{V}\mathbf{x}_\ell \quad \text{and} \quad \text{(b) } \mathbf{x}_\ell^T\mathbf{W}^T\mathbf{A}\mathbf{V} = \mu_\ell\,\mathbf{x}_\ell^T\mathbf{W}^T\mathbf{V}. \tag{16}$$

(16a) is immediate from the definition of $\mathbf{A}_r$. (16b) is obtained by transposition of (16a) and using the facts that $\mathbf{W}^T\mathbf{A}\mathbf{V}$ and $\mathbf{W}^T\mathbf{V}$ are symmetric. Notice that (16b) shows that $\mathbf{x}_\ell^T\mathbf{W}^T\mathbf{V}$ is a left eigenvector for $\mathbf{A}_r$ associated with μ_ℓ. On the other hand, direct substitution also shows that

$$\left[\mathbf{e}^T\,(\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1}\right]\mathbf{A}_r = \left[\mathbf{e}^T\,(\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1}\right]\left(\mathbf{\Sigma}_r - \mathbf{q}\mathbf{e}^T\right) = \mu_\ell\left[\mathbf{e}^T\,(\mathbf{\Sigma}_r - \mu_\ell\mathbf{I})^{-1}\right],$$

so $\mathbf{e}^T\,(\mathbf{\Sigma}_r - \mu_\ell \mathbf{I})^{-1}$ is also a left eigenvector of $\mathbf{A}_r$ associated with μ_ℓ. We must have then

$$\mathbf{x}_\ell^T \mathbf{W}^T \mathbf{V} = \nu_\ell \mathbf{e}^T\,(\mathbf{\Sigma}_r - \mu_\ell \mathbf{I})^{-1}$$

for some scalar ν_ℓ which we now determine. Using (9) and (14), the jth component of each side of the equation can be expressed as

$$\left(\mathbf{x}_\ell^T \mathbf{W}^T \mathbf{V}\right)_j = \sum_{i=1}^{r} \frac{p_r(\sigma_i)}{\omega_r'(\sigma_i)} \frac{[\sigma_i, \sigma_j]H_r}{\sigma_i - \mu_\ell} = \frac{\nu_\ell}{\sigma_j - \mu_\ell}. \tag{17}$$

Define the function $f(z) = p_r(z)[z, \sigma_j]H_r$. It is easily checked that $f(z)$ is a polynomial of degree $r-1$ and so Lagrange interpolation on $\{\sigma_1, \sigma_2, \ldots, \sigma_r\}$ is exact:

$$f(z) = \sum_{i=1}^{r} \left(p_r(\sigma_i)[\sigma_i, \sigma_j]H_r\right) \frac{\omega_r(z)}{(z-\sigma_i)\omega_r'(\sigma_i)}.$$

Now evaluate this expression at $z = \mu_\ell$:

$$\sum_{i=1}^{r} \frac{(p_r(\sigma_i)[\sigma_i, \sigma_j]H_r)\omega_r(\mu_\ell)}{(\mu_\ell - \sigma_i)\omega_r'(\sigma_i)} = f(\mu_\ell) = \lim_{z\to\mu_\ell} p_r(z) \frac{H_r(z) - H_r(\sigma_j)}{z - \sigma_j} = \frac{p_r'(\mu_\ell)\hat{\phi}_\ell}{\mu_\ell - \sigma_j},$$

where we observe that $\lim_{z\to\mu_\ell} p_r(z)H_r(z) = \hat{\phi}_\ell \prod_{i\neq\ell}(\mu_\ell - \mu_i) = \hat{\phi}_\ell p_r'(\mu_\ell)$. Comparing this expression to (17) we find $\nu_\ell = \hat{\phi}_\ell \frac{p_r'(\mu_\ell)}{\omega_r(\mu_\ell)}$. □

Lemma 2 illustrates that if the primitive bases $\mathbf{V}$ and $\mathbf{W}$ in (8) are employed in IRKA, then the reduced matrix $\mathbf{A}_r$ at every iteration step is a rank-1 perturbation of the diagonal matrix of shifts. This matrix $\mathbf{A}_r = \mathbf{\Sigma}_r - \mathbf{q}\mathbf{e}^T$ is known as the generalized companion matrix. This special structure allows explicit computation of the left and right eigenvectors of $\mathbf{A}_r$ as well. The next corollary gives further details about the spectral decomposition of $\mathbf{A}_r$.

Corollary 1 *Consider the setup in Lemma 2. Define the $r \times r$ Cauchy matrix $\mathbf{C} = \mathbf{C}(\boldsymbol{\sigma}, \boldsymbol{\mu})$ as*

$$\mathbf{C}_{ij} = \frac{1}{\sigma_i - \mu_j} \quad \text{for } i, j = 1, 2, \ldots, r, \tag{18}$$

and the $r \times r$ diagonal matrix $\mathbf{D_q} = \operatorname{diag}(q_1, q_2, \ldots, q_r)$. Then $\mathbf{A}_r = \mathbf{\Sigma} - \mathbf{q}\mathbf{e}^T$ has the spectral decomposition

$$\mathbf{A}_r = \mathbf{X}\mathbf{M}\mathbf{X}^{-1} \quad \text{where } \mathbf{M} = \operatorname{diag}(\mu_1, \mu_2, \ldots, \mu_r) \quad \text{and} \quad \mathbf{X} = \mathbf{D_q}\mathbf{C}. \tag{19}$$

Moreover, $\mathbf{A}_r^T = \mathbf{D_q}^{-1}\mathbf{A}_r\mathbf{D_q}$ and its spectral decomposition is

$$\mathbf{A}_r^T = \mathbf{\Sigma} - \mathbf{e}\mathbf{q}^T = \mathbf{D_q}^{-1}\mathbf{A}_r\mathbf{D_q} = \mathbf{C}\mathbf{M}\mathbf{C}^{-1}. \tag{20}$$

Proof The spectral decomposition of $\mathbf{A}_r$ in (19) directly follows from (14) by observing that $\mathbf{x}_\ell = (\mathbf{\Sigma}_r - \mu_\ell \mathbf{I})^{-1}\mathbf{q} = \begin{bmatrix} \frac{q_1}{\sigma_1-\mu_\ell} & \frac{q_2}{\sigma_2-\mu_\ell} & \cdots & \frac{q_r}{\sigma_r-\mu_\ell} \end{bmatrix}^T$. Therefore, the eigenvector matrix $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_r \end{bmatrix}$ can be written as $\mathbf{X} = \mathbf{D_q}\mathbf{C}$, proving (19). The spectral decomposition of $\mathbf{A}_r^T$ in (20) can be proved similarly using (15), i.e., the fact that $(\mathbf{\Sigma}_r - \mu_\ell \mathbf{I})^{-T}\mathbf{e}$ is an eigenvector of $\mathbf{A}_r^T$. Finally, $\mathbf{D_q}^{-1}\mathbf{A}_r\mathbf{D_q} = \mathbf{D}_q^{-1}(\mathbf{\Sigma}_r - \mathbf{q}\mathbf{e}^T)\mathbf{D_q} = \mathbf{\Sigma}_r - \mathbf{D}_q^{-1}\mathbf{q}\mathbf{e}^T\mathbf{D_q}$ since both $\mathbf{D}_q$ and $\mathbf{\Sigma}_r$ are diagonal. Moreover, it follows from the definition of $\mathbf{D_q} = \operatorname{diag}(q_1, q_2, \ldots, q_r)$ that $\mathbf{D_q}^{-1}\mathbf{q} = \mathbf{e}$ and $\mathbf{e}^T\mathbf{D_q} = \mathbf{q}$, thus completing the proof. □

3 A Pole Placement Connection

The main goal of this paper is to reveal the structure of the iterations in Algorithm 1, and in particular to study the limiting behaviour of the sequence of the shifts $\boldsymbol{\sigma}^{(k)}$, $k = 1, 2, \ldots$. In this section, we explore an intriguing idea to recast the computation of the shifts in Algorithm 1 in a pole placement framework, and then to examine its potential for improving the convergence.

As Lemma 2 illustrates, if the primitive bases (8) are employed in IRKA, then at every step of IRKA, we have $\mathbf{A}_r = \boldsymbol{\Sigma}_r - \mathbf{q}\mathbf{e}^T$. Then, in the kth step, we start with the shifts $\sigma_1^{(k)}, \ldots, \sigma_r^{(k)}$ and use them to build the matrix

$$\mathbf{A}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)}) = \operatorname{diag}\left(\sigma_1^{(k)}, \ldots, \sigma_r^{(k)}\right) - \mathbf{q}^{(k+1)}\mathbf{e}^T, \tag{21}$$

where the vector $\mathbf{q}^{(k+1)}$ (the reduced input in step k) ensures that the Hermite interpolation conditions are fulfilled, see (6). If $\sigma_i^{(k)}$ is real, then $q_i^{(k+1)}$ is real as well; if for some $i \neq j \sigma_i^{(k)} = \overline{\sigma_j^{(k)}}$, then $q_i^{(k+1)} = \overline{q_j^{(k+1)}}$. As a consequence, $\mathbf{A}_r^{(k+1)}$ is similar to a real matrix and its eigenvalues will remain closed under complex conjugation. Further, if some $q_i^{(k+1)} = 0$, then the corresponding $\sigma_i^{(k)}$ is an eigenvalue of $\mathbf{A}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)})$; thus, if we assume that the shifts are in the open right half-plane and that $\mathbf{A}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)})$ is stable, then $q_i^{(k+1)} \neq 0$ for all i. Then, the new set of shifts is defined as

$$\boldsymbol{\sigma}^{(k+1)} = -\boldsymbol{\mu}^{(k+1)}, \quad \text{where } \boldsymbol{\mu}^{(k+1)} = \operatorname{eig}(\mathbf{A}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)})), \tag{22}$$

where eig(·)is a numerical algorithm that computes the eigenvalues and returns them in some order.[2] In the limit as $k \to \infty$ the shifts should satisfy (6) and (7).

3.1 Measuring Numerical Convergence

Numerical convergence in an implementation of Algorithm 1 is declared if $\boldsymbol{\sigma}^{(k+1)} \approx \boldsymbol{\sigma}^{(k)}$, where the distance between two consecutive sets of shifts is measured using the optimal matching[3]

$$d(\boldsymbol{\sigma}^{(k+1)}, \boldsymbol{\sigma}^{(k)}) = \min_{\pi \in \mathbb{S}_r} \max_{i=1:r} |\boldsymbol{\sigma}_{\pi(i)}^{(k+1)} - \boldsymbol{\sigma}_i^{(k)}|, \quad \text{where } \mathbb{S}_r \text{ denotes the symmetric group.}$$

In an implementation, it is convenient to use the easier to compute Hausdorff distance

$$h(\boldsymbol{\sigma}^{(k+1)}, \boldsymbol{\sigma}^{(k)}) = \max\left\{ \max_j \min_i \left|\boldsymbol{\sigma}^{(k+1)_j} - \boldsymbol{\sigma}_i^{(k)}\right|, \ \max_i \min_j \left|\boldsymbol{\sigma}_i^{(k)} - \boldsymbol{\sigma}_j^{(k+1)}\right| \right\}$$

for which $h(\boldsymbol{\sigma}^{(k+1)}, \boldsymbol{\sigma}^{(k)}) \leq d(\boldsymbol{\sigma}^{(k+1)}, \boldsymbol{\sigma}^{(k)})$, so the stopping criterion (Line 9 in Algorithm 1) must be first satisfied in the Hausdorff metric.

Numerical evidence shows that many scenarios are possible during the iterations in Algorithm 1—from swift to slow convergence. Characterizing the limit behavior in general is an open problem; in the case of symmetric systems local convergence is established in [25]. Moreover, we have also encountered misconvergence in form of the existence of at least two accumulation points that seem to indicate existence of periodic points of the mapping (22). This is illustrated in the following example.

[2] Since the matrix is a rank one perturbation of the diagonal matrix, all eigenvalues can be computed in $O(r^2)$ operations by specially tailored algorithms.

[3] The shifts (eigenvalues) are naturally considered as equivalence classes in $\mathbb{C}^r/\mathbb{S}_r$.

Example 1 We take the matrix $\mathbf{A} \in \mathbb{R}^{120\times 120}$ from the CD player benchmark example [18, 19] from the NICONET benchmark collection [17] and set $\mathbf{b} = \mathbf{c} = \mathbf{e}$. With a particular set of $r = 29$ initial shifts, we obtained separate behaviours for the odd and the even iterates, as shown in Fig. 1.

Hence, it is of both theoretical and practical interest to explore possibilities for improving the convergence. Supplying good initial shifts is certainly beneficial, and in [20, 21] we show that the less expensive Vector Fitting algorithm can be used for preprocessing/preconditioning to generate good shifts that are then forwarded to IRKA to advance them to a local optimum.

An alternative course of action is to deploy an additional control in the iterations which will keep steering the shifts toward the desired positions. In fact, an example of such an intervention has been already used in the numerical implementation of Algorithm 1. Namely, it can happen that in some steps the matrix (21) is not stable and some of its eigenvalues (22) are in the right half-plane. To correct this situation, the unstable ones (real or complex-conjugate pair(s)) are flipped across the imaginary axis, so that the new shifts $\boldsymbol{\sigma}^{(k+1)}$ stay in the right-half plane.

This is an explicit (brute force) post-festum reassignment of the shifts to correct for stability. In [20], we showed that such a step (in the framework of Vector Fitting) can be recast as pole placement. Now that we have resorted (implicitly) to the pole placement mechanism, we can think of using it as a proactive strategy for improving convergence. In

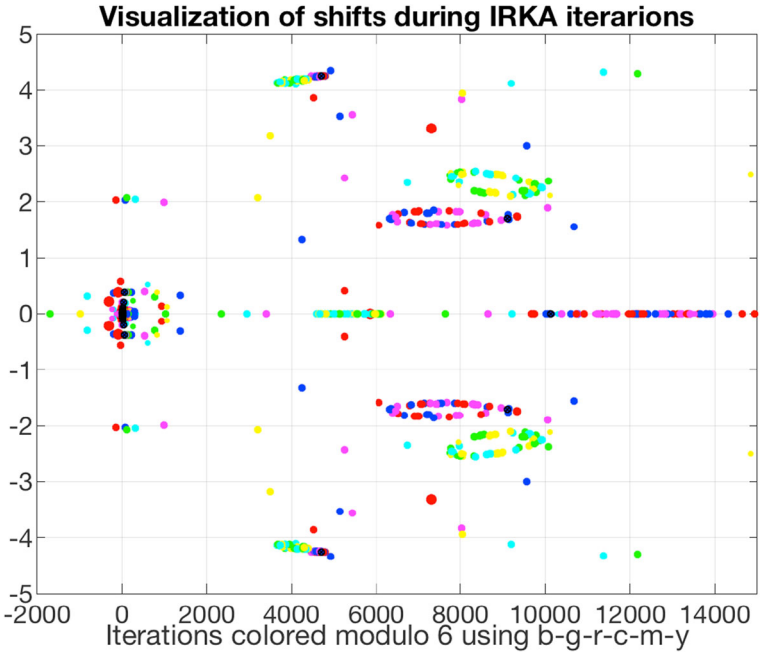

Fig. 1 (Example 1) The history of the shifts obtained using Algorithm 1. The iterations are colored using six colors periodically as follows: ● ● ● ● ● ● Note how the odd and the even iterates build two separated pairs of "smoke rings" (abscissa range [6000, 10000]); more smaller rings can be identified in the abscissa range [3000, 5000]. The shifts do not converge to a fixed point, but the difference between the two sub-sequences (the even and the odd indices) converges to a nonzero value

the rest of this section, we explore this idea and discover interesting connections with some variations of IRKA.

3.2 Reduced Input-to-State Vector as a Pole Placement Feedback Vector

Motivated by the above discussion, we reinterpret (a posteriori) the vector $\mathbf{q}^{(k+1)}$ in (21) as the feedback vector that reallocates the eigenvalues[4] of $\mathrm{diag}(\boldsymbol{\sigma}^{(k)})$ into $\boldsymbol{\mu}^{(k+1)}$; in other words we view (21) as a pole-placement problem. Then we can use the uniqueness argument and write $\mathbf{q}^{(k+1)}$ explicitly as (see [36, 37])

$$-q_i^{(k+1)} = \frac{\prod_{j=1}^{r}(\sigma_i^{(k)} - \mu_j^{(k+1)})}{\prod_{\substack{j=1\\ j\neq i}}^{r}(\sigma_i^{(k)} - \sigma_j^{(k)})} = \left(\sigma_i^{(k)} - \mu_i^{(k+1)}\right)\prod_{\substack{j=1\\ j\neq i}}^{r}\frac{\sigma_i^{(k)} - \mu_j^{(k+1)}}{\sigma_i^{(k)} - \sigma_j^{(k)}}, \quad i = 1, \ldots, r. \tag{23}$$

On the other hand, for the fulfillment of the necessary conditions for optimality, besides the Hermite interpolation built in (23), the additional fixed point condition should hold:

$$\mathrm{eig}\left(\mathbf{A}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)})\right) \approx -\boldsymbol{\sigma}^{(k)}. \tag{24}$$

The latter is what we hope to reach with the equality in the limit as $k \to \infty$, and in practice up to a reasonable tolerance, see Section 4.1. If we consider the condition (24) as an eigenvalue assignment problem, and think of the vector $\mathbf{q}^{(k+1)}$ in (21) simply as the feedback vector, then (24) can be satisfied in one step, provided we drop the interpolation condition and use an appropriate feedback $\mathbf{f}^{(k+1)}$ vector instead of $\mathbf{q}^{(k+1)}$. The feedback $\mathbf{f}^{(k+1)}$ can be constructed explicitly (see [36, 37]) as

$$f_i^{(k+1)} = -2\sigma_i^{(k)}\prod_{\substack{j=1\\ j\neq i}}^{r}\frac{\sigma_i^{(k)} + \sigma_j^{(k)}}{\sigma_i^{(k)} - \sigma_j^{(k)}}, \quad i = 1, \ldots, r. \tag{25}$$

Of course, the above formula is a special case of (23), where we reflect the poles. However, if at a particular step some of the eigenvalues are unstable, we should not reflect the corresponding shifts. This means that in an implementation, we may apply only partial pole placement. Hence, altogether, it would make sense to interweave interpolation and eigenvalue assignment by combining $\mathbf{f}^{(k+1)}$ and $\mathbf{q}^{(k+1)}$ using an appropriately chosen parameter $\alpha_k \in [0, 1]$, and thus obtain a modified iteration step, as outlined in Algorithm 2. To incorporate this new shift-updating scheme into IRKA, Algorithm 2 should replace Step 8 in Algorithm 1.

Algorithm 2 IRKA + pole placement for shift updates; kth step.

1: Compute the reduced input vector $\mathbf{q}^{(k+1)}$.
2: Compute the feedback vector $\mathbf{f}^{(k+1)}$ using (25). (Keep track of stability.)
3: $\breve{\mathbf{q}}^{(k+1)} = \alpha_k \mathbf{q}^{(k+1)} + (1 - \alpha_k)\mathbf{f}^{(k+1)}$, with an appropriate $\alpha_k \in [0, 1]$.
4: $\breve{\mathbf{A}}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)}) = \mathrm{diag}(\sigma_i^{(k)})_{i=1}^{r} - \breve{\mathbf{q}}^{(k+1)}\mathbf{e}^T$.
5: The new shifts are $\boldsymbol{\sigma}^{(k+1)} = -\mathrm{eig}(\breve{\mathbf{A}}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)}))$.

[4] We tacitly assume that throughout the iterations all shifts are simple.

Proposition 1 *For the real LTI system (1), and $\boldsymbol{\sigma}^{(k)}$ closed under complex conjugation, the matrix $\breve{\mathbf{A}}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)})$ is similar to a real matrix and, thus, $\boldsymbol{\sigma}^{(k+1)}$ remains closed under complex conjugation.*

Proof From (25), we conclude that $f_i^{(k+1)}$ is real if $\sigma_i^{(k)}$ is real. Further, if for some $i \neq j$ $\sigma_i^{(k)} = \overline{\sigma_j^{(k)}}$, then $f_i^{(k+1)} = \overline{f_j^{(k+1)}}$. We have already concluded that $\mathbf{q}^{(k+1)}$ has an analogous structure. Since α_k is real, $\breve{\mathbf{A}}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)})$ is similar to a real matrix. □

It remains an open problem how to chose the coefficients α_k adaptively and turn Algorithm 2 into a robust black-box scheme. We now show an interesting connection that might provide some guidelines.

3.3 Connection to the Krajewski–Viaro Scheme

Improving the convergence of fixed point iterations is an important topic. In general, the fixed point problem $f(x) = x$ can be equivalently solved as the problem

$$f_\alpha(x) = x, \quad \text{where } f_\alpha(x) = \alpha f(x) + (1-\alpha)x,\ \alpha \neq 0,$$

where the parameter α is used, e.g., to modify eigenvalues of the corresponding Jacobian. This is a well-known technique (Mann iteration), with many variations. In the context of $\mathcal{H}_2$ model reduction, this scheme has been successfully applied by Ferrante, Krajewski, Lepschy and Viarro [24], and Krajewski and Viaro [35]. Concretely, Krajewski and Viaro [35, Algorithm 4] propose a modified step for the IRKA procedure, outlined in Algorithm 3:

Algorithm 3 Krajewski–Viaro scheme for shift updates; kth step.

1: Let $\breve{\boldsymbol{\wp}}_r^{(k)}$ be the modified (monic) polynomial, whose reflected zeros are the shifts $\boldsymbol{\sigma}^{(k)}$, i.e. $\breve{\boldsymbol{\wp}}_r^{(k)}(-\sigma_i^{(k)}) = 0, i = 1, \ldots, r$.
2: Compute $\mathbf{A}_r^{(k+1)} = (\mathbf{W}_k^T\mathbf{V}_k)^{-1}\mathbf{W}_k^T\mathbf{A}\mathbf{V}_k$ and the coefficients of its characteristic polynomial $\boldsymbol{\wp}_r^{(k+1)}$. Write this mapping as $\boldsymbol{\wp}_r^{(k+1)} = \Phi(\breve{\boldsymbol{\wp}}_r^{(k)})$.
3: Define the new polynomial

$$\breve{\boldsymbol{\wp}}_r^{(k+1)} = \alpha\boldsymbol{\wp}_r^{(k+1)} + (1-\alpha)\breve{\boldsymbol{\wp}}_r^{(k)} \equiv \Phi_\alpha(\breve{\boldsymbol{\wp}}_r^{(k)}), \tag{26}$$

where the linear combination of the polynomials is formed using their coefficients.
4: The new shifts are then the reflected roots of $\breve{\boldsymbol{\wp}}_r^{(k+1)}$.

Krajewski and Viaro [35] do not elaborate on the details of computing the coefficients of the characteristic polynomials of the reduced matrix $(\mathbf{W}_k^T\mathbf{V}_k)^{-1}\mathbf{W}_k^T\mathbf{A}\mathbf{V}_k$. From a numerical point of view, this is not feasible, not even for moderate dimensions. Computing coefficients of the characteristic polynomial by, e.g., the classical Faddeev–Leverrier trace formulas is both too expensive ($O(r^4)$) and too ill-conditioned. A modern approach would reduce the matrix to a Schur or Hessenberg form and then exploit the triangular or, respectively, Hessenberg structure. However, after completing all those (tedious) tasks, the zeros of $\breve{\boldsymbol{\wp}}_r^{(k+1)}$ in Line 4 are best computed by transforming the problem into an eigenvalue computation for an appropriate companion matrix. Ultimately, this approach is only conceptually interesting as a technique for improving convergence, and in this form it is applicable only for small values of r.

We now show that when represented in a proper basis, this computation involving characteristic polynomials becomes rather elegant and simple, and further provides interesting insights. In particular, in this proper basis, Algorithm 3 is equivalent to Algorithm 2.

Theorem 1 *In the Lagrange basis of $\omega_r^{(k)} + \mathcal{P}_{r-1}$, the Krajewski–Viaro iteration is equivalent to the "IRKA + pole placement" iteration of Algorithm 2.*

Proof Note that by Lemma 1 we can write $\mathbf{A}_r^{(k+1)} = \operatorname{diag}(\sigma_i^{(k)}) - \mathbf{q}^{(k+1)}\mathbf{e}^T$, where $q_i^{(k+1)} = \wp_r^{(k+1)}(\sigma_i)/(\omega_r^{(k+1)})'(\sigma_i)$, $i = 1, \ldots, r$, and $\wp_r^{(k+1)}(z)$ is the characteristic polynomial of $\mathbf{A}_r^{(k+1)}$. Further, using Lemma 1, we can write $\wp_r^{(k+1)}(z)$ as

$$\begin{aligned}\wp_r^{(k+1)}(z) &= \omega_r^{(k)}(z) + \sum_{i=1}^{r} \wp_r^{(k+1)}(\sigma_i^{(k)}) \frac{\omega_r^{(k)}(z)}{(\omega_r^{(k)})'(\sigma_i^{(k)})(z-\sigma_i^{(k)})} \\ &= \omega_r^{(k)}(z) + \sum_{i=1}^{r} q_i^{(k+1)} \frac{\omega_r^{(k)}(z)}{z-\sigma_i^{(k)}}, \quad \omega_r^{(k)}(z) = \prod_{i=1}^{r}(z-\sigma_i^{(k)}).\end{aligned}$$

If we consider the monic polynomials of degree r as the linear manifold $\omega_r^{(k)} + \mathcal{P}_{r-1}$, and fix in $\mathcal{P}_{r-1}$ the Lagrange basis with the nodes $\sigma_i^{(k)}$, $i = 1, \ldots, r$, then

$$\begin{aligned}\breve{\wp}_r^{(k)}(z) &= \omega_r^{(k)}(z) + \sum_{i=1}^{r} \breve{\wp}_k(\sigma_i^{(k)}) \frac{\omega_r^{(k)}(z)}{(\omega_r^{(k)})'(\sigma_i^{(k)})(z-\sigma_i^{(k)})} \\ &= \omega_r^{(k)}(z) + \sum_{i=1}^{r} \frac{\prod_{j=1}^{r}(\sigma_i^{(k)} + \sigma_j)^{(k)}}{\prod_{\substack{j=1 \\ j\neq i}}^{r}(\sigma_i^{(k)} - \sigma_j^{(k)})} \frac{\omega_r^{(k)}(z)}{z-\sigma_i^{(k)}} = \omega_r^{(k)}(z) + \sum_{i=1}^{r} f_i^{(k+1)} \frac{\omega_r^{(k)}(z)}{z-\sigma_i^{(k)}},\end{aligned}$$

where we used that $\breve{\wp}_r^{(k)}(\sigma_i^{(k)}) = \prod_{j=1}^{r}(\sigma_i^{(k)} + \sigma_j^{(k)})$, $(\omega_r^{(k)})'(\sigma_i^{(k)}) = \prod_{\substack{j=1 \\ j\neq i}}^{r}(\sigma_i^{(k)} - \sigma_j^{(k)})$ and that the feedback vector (25) can be written as

$$\frac{\prod_{j=1}^{r}(\sigma_i^{(k)} + \sigma_j^{(k)})}{\prod_{\substack{j=1 \\ j\neq i}}^{r}(\sigma_i^{(k)} - \sigma_j^{(k)})} = 2\sigma_i^{(k)} \prod_{\substack{j=1 \\ j\neq i}}^{r} \frac{(\sigma_i^{(k)} + \sigma_j^{(k)})}{(\sigma_i^{(k)} - \sigma_j^{(k)})} = f_i^{(k+1)}.$$

Hence, $\breve{\wp}_r^{(k)}$ is the characteristic polynomial of $\operatorname{diag}(\boldsymbol{\sigma}^{(k)}) - \mathbf{f}^{(k+1)}\mathbf{e}^T$, and we have further

$$\begin{aligned}\breve{\wp}_r^{(k+1)}(z) &= \omega_r^{(k)}(z) + \sum_{i=1}^{r} (\alpha_k q_i^{(k+1)} + (1-\alpha_k) f_i^{(k+1)}) \frac{\omega_r^{(k)}(z)}{z-\sigma_i^{(k)}} \\ &= \omega_r^{(k)}(z) + \sum_{i=1}^{r} \frac{\alpha_k \wp_r^{(k+1)}(\sigma_i^{(k)}) + (1-\alpha_k)\breve{\wp}_r^{(k)}(\sigma_i^{(k)})}{(\omega_r^{(k)})'(\sigma_i^{(k)})} \frac{\omega_r^{(k)}(z)}{z-\sigma_i^{(k)}} \\ &= \omega_r^{(k)}(z) + \sum_{i=1}^{r} \frac{\breve{p}_r^{(k+1)}(\sigma_i^{(k)})}{(\omega_r^{(k)})'(\sigma_i^{(k)})} \frac{\omega_r^{(k)}(z)}{z-\sigma_i^{(k)}},\end{aligned}$$

which implies that $\breve{\wp}_r^{(k+1)}$ is the characteristic polynomial of the matrix $\breve{\mathbf{A}}_r^{(k+1)}(\boldsymbol{\sigma}^{(k)})$, represented by the vector $\breve{\mathbf{q}}^{(k+1)}$ from Lines 3 and 4 of Algorithm 2. This follows from the proof of Lemma 1. □

Krajewski and Viaro [35] use a fixed value of the parameter α, and show that different (fixed) values may lead to quite different convergence behavior. This modification can

turn a non-convergent process into a convergent one, but it can also slow down an already convergent one.[5] Following the discussion from [24], α is best chosen to move the smallest eigenvalue of the Jacobian of Φ_α (evaluated at the fixed point of Φ) into the interval $(-1, 1)$. This does not seem to be a simple task as it requires estimates of the eigenvalues of the (estimated) Jacobian. Another option is to try different values of α in an iterative reduced order model design.

This equivalence of the schemes in Algorithm 2 and in Algorithm 3 reveals a problem that is not easily seen in the framework of Algorithm 3. Now we may clearly see that part of the "energy" in Algorithm 3 is put into reflecting the shifts, and this, at least in some situations, may be wasted effort. Although the optimal reduced order model is guaranteed to be stable, the iterates, generally, are not. This means that some shifts $\sigma_i^{(k)}$ may be in the left-half plane, and the $\mathbf{f}^{(k+1)}$ component of the modified $\check{\mathbf{q}}^{(k+1)}$ will tend to reflect them to the right-half plane. This in turn forces the new reduced matrix $\check{\mathbf{A}}_r^{(k+1)}$ to have some eigenvalues in the right-half plane, thus creating a vicious cycle. Hence, in the first step (Line 1), one should correct $\boldsymbol{\sigma}^{(k)}$, if necessary.

Remark 1 The facts that pole placement may be extremely ill-conditioned [31], where (depending on the distribution of the target values) even as small as $r = 15$ could lead to ill-conditioning, and that IRKA is actually doing pole placement in disguise, opens many nontrivial issues. For instance, what is a reasonable threshold for the stopping criterion? Will we be able to actually test it (and detect convergence) in finite precision? What are relevant condition numbers in the overall process? Do IRKA iterations drive the shifts to well-conditioned configurations for which the feedback vector (the reduced input) is reasonably small (in norm, as illustrated in the right panels of Figs. 2 and 3 in Example 2 below) and successful in achieving numerical convergence? If yes, what is the underlying principle/driving mechanism? In the next section, we touch upon some of these issues.

4 Perturbation Effects and Backward Stability in IRKA

We turn our attention now to numerical considerations in the implementation of IRKA. We focus on two issues: (i) What are the perturbative effects of finite-precision arithmetic in terms of system-theoretic quantities? (ii) What are the effects of "numerical convergence" on the reduced model?

4.1 Limitations of Finite Precision Arithmetic

Suppose that we are given magic shifts $\boldsymbol{\sigma}$ so that the eigenvalues of $\mathbf{A}_r = \boldsymbol{\Sigma}_r - \mathbf{q}\mathbf{e}^T$ are exactly $\boldsymbol{\lambda} = -\boldsymbol{\sigma}$, or $\boldsymbol{\lambda} \approx -\boldsymbol{\sigma}$ up to a small tolerance $\boldsymbol{\epsilon}$. However, in floating point computations, the vector $\mathbf{q} = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T\mathbf{b}$ is computed up to an error $\boldsymbol{\delta}\mathbf{q}$, and therefore instead of $\mathbf{A}_r$, we have $\widetilde{\mathbf{A}}_r = \boldsymbol{\Sigma}_r - (\mathbf{q} + \boldsymbol{\delta}\mathbf{q})\mathbf{e}^T$. In practice, the source of $\boldsymbol{\delta}\mathbf{q}$ is twofold: First, in large-scale settings, the primitive Krylov bases $\mathbf{V}$ and $\mathbf{W}$ are usually computed by an iterative method which uses restricted information from suitably chosen subspaces and thus generates a truncation error; see, e.g., [1, 2, 10]. In addition, computation is polluted by

[5] We should point out here that the dimensions n and r are rather small in all reported numerical experiments in [35].

omnipresent rounding errors of finite precision arithmetic. How the size of $\boldsymbol{\delta}\mathbf{q}$ influences the other components of the IRKA is relevant information that we investigate in this subsection.

Assume for the moment that $\boldsymbol{\delta}\mathbf{q}$ is the only perturbation in one step of IRKA. We want to understand how the eigenvalues of $\mathbf{A}_r$ and $\widetilde{\mathbf{A}}_r$ differ as a function of $\boldsymbol{\delta}\mathbf{q}$. In particular, we want to discover the relevant condition numbers that play a role in this perturbation analysis.

Theorem 2 *Let* $\mathbf{A}_r = \boldsymbol{\Sigma}_r - \mathbf{q}\mathbf{e}^T$ *and* $\widetilde{\mathbf{A}}_r = \boldsymbol{\Sigma}_r - \widetilde{\mathbf{q}}\mathbf{e}^T$ *be diagonalizable, where* $\widetilde{\mathbf{q}} = \mathbf{q} + \boldsymbol{\delta}\mathbf{q}$. *Let* $\mathbf{A}_r$ *and* $\widetilde{\mathbf{A}}_r$ *have the spectral decompositions* $\mathbf{A}_r = \mathbf{X}\mathbf{M}\mathbf{X}^{-1}$ *and* $\widetilde{\mathbf{A}}_r = \widetilde{\mathbf{X}}\widetilde{\mathbf{M}}\widetilde{\mathbf{X}}^{-1}$, *where* $\mathbf{M} = \mathrm{diag}(\mu_i)_{i=1}^r$, $\widetilde{\mathbf{M}} = \mathrm{diag}(\tilde{\mu}_i)_{i=1}^r$ *and the eigenvector matrices* $\mathbf{X} = \mathbf{D}_\mathbf{q}\mathbf{C}$ *and* $\widetilde{\mathbf{X}} = \widetilde{\mathbf{D}}_\mathbf{q}\widetilde{\mathbf{C}}$ *as described in Corollary 1. Then there exists a permutation* π *such that*

$$\sqrt{\sum_{i=1}^{r}\left|\frac{\mu_i - \tilde{\mu}_{\pi(i)}}{\mu_i}\right|^2} \leq \|\mathbf{C}\|_2\|(\mathbf{C}\mathbf{M})^{-1}\|_2\kappa_2(\widetilde{\mathbf{C}})\|\boldsymbol{\delta}\mathbf{q}\mathbf{e}^T\|_2. \tag{26}$$

Proof Note that we can equivalently compare the spectra of $\mathbf{A}_r^T$ and $\widetilde{\mathbf{A}}_r^T$. From (15), we know that the spectral decomposition of $\mathbf{A}_r^T$ is given by

$$\mathbf{A}_r^T = \boldsymbol{\Sigma}_r - \mathbf{e}\mathbf{q}^T = \mathbf{D}_\mathbf{q}^{-1}\mathbf{A}_r\mathbf{D}_\mathbf{q} = \mathbf{C}\mathbf{M}\mathbf{C}^{-1}. \tag{27}$$

Similarly for $\widetilde{\mathbf{A}}_r^T$, we obtain

$$\widetilde{\mathbf{A}}_r^T = \boldsymbol{\Sigma}_r - \mathbf{e}\widetilde{\mathbf{q}}^T = \widetilde{\mathbf{D}}_\mathbf{q}^{-1}\widetilde{\mathbf{A}}_r\widetilde{\mathbf{D}}_\mathbf{q} = \widetilde{\mathbf{C}}\widetilde{\mathbf{M}}\widetilde{\mathbf{C}}^{-1}. \tag{28}$$

Next we employ the perturbation results of Elsner and Friedland [23], and Eisenstat and Ipsen [22], while taking into account the special structure of both matrices. Write $\widetilde{\mathbf{A}}_r^T = \mathbf{A}_r^T + \boldsymbol{\delta}\mathbf{A}_r^T$, where $\boldsymbol{\delta}\mathbf{A}_r^T = -\mathbf{e}\boldsymbol{\delta}\mathbf{q}^T$. Using the spectral decompositions (20) and (28), the matrix $\mathbf{A}_r^{-T}\widetilde{\mathbf{A}}_r^T - \mathbf{I} = \mathbf{A}_r^{-T}\boldsymbol{\delta}\mathbf{A}_r^T$ can be transformed into

$$\mathbf{M}^{-1}(\mathbf{C}^{-1}\widetilde{\mathbf{C}})\widetilde{\mathbf{M}} - (\mathbf{C}^{-1}\widetilde{\mathbf{C}}) = \mathbf{C}^{-1}\mathbf{A}_r^{-T}\boldsymbol{\delta}\mathbf{A}_r^T\widetilde{\mathbf{C}}. \tag{29}$$

Set $\mathbf{Y} = \mathbf{C}^{-1}\widetilde{\mathbf{C}}$ and take the absolute value of y_{ij}, an arbitrary entry of $\mathbf{Y}$ at the (i, j)th position, to obtain

$$|y_{ij}|\left|\frac{\tilde{\mu}_j}{\mu_i} - 1\right| = |(\mathbf{C}^{-1}\mathbf{A}_r^{-T}\boldsymbol{\delta}\mathbf{A}_r^T\widetilde{\mathbf{C}})_{ij}| \quad \text{for } 1 \leq i, j \leq r.$$

Hence

$$\sum_{i=1}^{r}\sum_{j=1}^{r}|y_{ij}|^2\left|\frac{\tilde{\mu}_j}{\mu_i} - 1\right|^2 = \|\mathbf{C}^{-1}\mathbf{A}_r^{-T}\boldsymbol{\delta}\mathbf{A}_r^T\widetilde{\mathbf{C}}\|_F^2,$$

where the Hadamard product matrix $\mathbf{Y} \circ \overline{\mathbf{Y}} = (|y_{ij}|^2)_{i,j=1}^r$ is entry-wise bounded by

$$\sigma_{\min}(\mathbf{Y})^2\mathbf{S}_{ij} \leq (\mathbf{Y} \circ \overline{\mathbf{Y}})_{ij} \leq \sigma_{\max}(\mathbf{Y})^2\mathbf{S}_{ij},$$

where $\mathbf{S}$ is a doubly-stochastic matrix; see [23]. Hence

$$\sum_{i=1}^{r}\sum_{j=1}^{r}\mathbf{S}_{ij}\left|\frac{\tilde{\mu}_j}{\mu_i} - 1\right|^2 \leq \|\mathbf{Y}^{-1}\|_2^2\|\mathbf{C}^{-1}\mathbf{A}_r^{-T}\boldsymbol{\delta}\mathbf{A}_r^T\widetilde{\mathbf{C}}\|_F^2 \leq \kappa_2(\mathbf{C})^2\kappa_2(\widetilde{\mathbf{C}})^2\|\mathbf{A}_r^{-T}\boldsymbol{\delta}\mathbf{A}_r^T\|_F^2. \tag{30}$$

The expression on the left-hand side of (30) can be considered as a function defined on the convex polyhedral set of doubly-stochastic matrices, whose extreme points are the

permutation matrices. Thus, for some permutation π, we obtain

$$\sum_{i=1}^{r} \left| \frac{\tilde{\mu}_{\pi(i)} - \mu_i}{\mu_i} \right|^2 \leq \|\mathbf{Y}^{-1}\|_2^2 \|\mathbf{C}^{-1}\mathbf{A}_r^{-T}\boldsymbol{\delta}\mathbf{A}_r^T\widetilde{\mathbf{C}}\|_F^2.$$

Then, using (20), the spectral decomposition of $\mathbf{A}_r^{-T}$, and the definition of $\boldsymbol{\delta}\mathbf{A}_r$ complete the proof. □

Remark 2 One can write (29) as $\mathbf{Y}\widetilde{\mathbf{M}} - \mathbf{M}\mathbf{Y} = \mathbf{C}^{-1}\boldsymbol{\delta}\mathbf{A}_r^T\widetilde{\mathbf{C}}$ and conclude that there exists a permutation p such that (see [23])

$$\sqrt{\sum_{i=1}^{n} |\tilde{\mu}_{p(i)} - \mu_i|^2} \leq \kappa_2(\mathbf{C})\kappa_2(\tilde{\mathbf{C}})\|\boldsymbol{\delta}\mathbf{q}\mathbf{e}^T\|_2.$$

Remark 3 The right-hand side in relation (26) can also be bounded by

$$\sqrt{r}\kappa_2(C)\kappa_2(\widetilde{\mathbf{C}})\frac{\|\boldsymbol{\delta}\mathbf{q}\|_2}{\min_i |\mu_i|}.$$

From the numerical point of view, Theorem 2 cannot be good news—Cauchy matrices can be ill-conditioned. A few random trials will quickly produce a 10×10 Cauchy matrix with condition number greater than 10^{10}. The most notorious example is the Hilbert matrix, which at the dimension 100 has a condition number larger than 10^{150}. No function of the matrix that is influenced by that condition number of the eigenvectors can be satisfactorily computed in 32 bit machine arithmetic. The 64 bit double precision allows only slightly larger dimensions before the condition number takes over the machine double precision.

On the other hand, we note that our goal is not to place the shifts at any predefined locations in the complex plane. Instead, we are willing to let them go wherever they want, under the condition that they remain closed under conjugation and stationary at those positions. It should also be noted that the distribution of the shifts obviously plays a role in this considerations. The following example will illustrate this; especially the impact of the optimal $\mathcal{H}_2$ interpolation points.

Example 2 As in Example 1, we first take the CD player model [18, 19] of order $n = 120$ and apply IRKA as in Algorithm 1 for $r = 2$, $r = 16$, and $r = 26$. In each case, IRKA is initialized by randomly assigned shifts. The condition numbers of $\mathbf{C}$ for each case, recorded throughout the iterations, are shown on the left panel of Fig. 2. IRKA drastically reduces the condition number of $\mathbf{C}$ throughout the iteration, more than 15 order of magnitudes for $r = 16$ and $r = 26$ cases. Therefore, IRKA keeps assigning shifts in such a way that $\mathbf{C}$ becomes better and better conditioned; thus in affect limiting the perturbation effects predicted by Theorem 2. Moreover, we can observe that the reduced input vectors $\mathbf{q}^{(k)}$, which act also as feedback vectors that steer the shifts, diminish in norm over the iterations; see the right panel of Fig. 2.

We have observed the same effect in all the examples we have tried. For brevity, we include only one more such result using the International Space Station 1R Module [6, 30] of order $n = 270$. As for the CD Player model, we reduce this model with IRKA using random initial shifts and this time chose reduced orders of $r = 10$, $r = 20$, and $r = 30$. The results depicted in Fig. 3 reveal the same behavior: The condition number $\kappa_2(\mathbf{C})$ is reduced significantly during IRKA as shifts converge to the optimal shifts; the same holds for the reduced input norms $\|\mathbf{q}^{(k)}\|_2$. This observation raises intriguing theoretical questions about

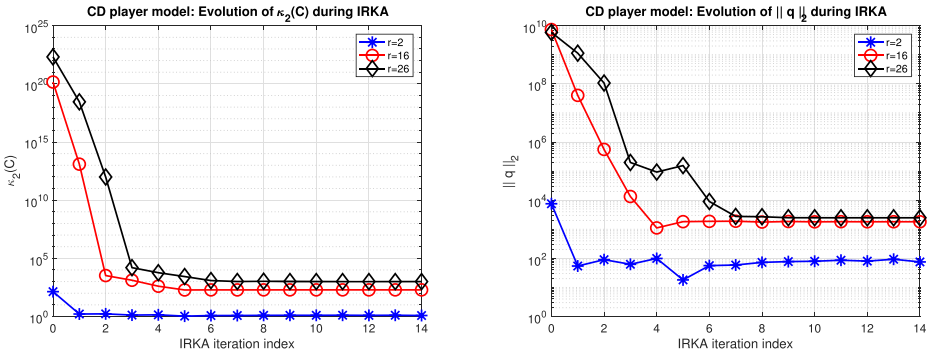

Fig. 2 $\kappa_2(\mathbf{C})$ and $\|\mathbf{q}^{(k)}\|_2$ during IRKA for the CD Player example

the distribution of the $\mathcal{H}_2$-optimal shifts, as their impact mimics that of Chebyshev points (as opposed to linearly spaced ones) in polynomial interpolation. These issues will not be studied here and are left to future papers.

4.2 Stopping Criterion and Backward Stability

Analytically, $\mathcal{H}_2$ optimality is satisfied when $\boldsymbol{\sigma} = -\boldsymbol{\mu}$. However, in practice Algorithm 1 will be terminated once a numerical convergence threshold is met. In this section, we will investigate the impact of numerical convergence on the resulting reduced model. The pole-placement connection we established in Section 3 will play a fundamental role in answering this question.

Suppose that in Algorithm 1 a numerical stopping (convergence) criterion has been satisfied, i.e., the eigenvalues $\mu_1, \ldots, \mu_k$ of $\mathbf{A}_r$ *are close to the reflected shifts*. Both the shifts $\boldsymbol{\sigma}$ and the computed eigenvalues $\boldsymbol{\mu}$ are unordered r-tuples of complex numbers, and we measure their distance using optimal matching, see Section 3.1. Hence, we define the indexing of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_r)$ so that

$$\|\boldsymbol{\mu} - (-\boldsymbol{\sigma})\|_\infty = \min_{\pi \in \mathbb{S}_r} \max_{k=1:r} |\mu_{\pi(k)} - (-\sigma_k)|.$$

Recall that the shifts $\boldsymbol{\sigma} = \{\sigma_1, \ldots, \sigma_r\}$ are closed under conjugation, with strictly positive real parts, and all assumed to be simple. The eigenvalues $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_r\}$ are assumed

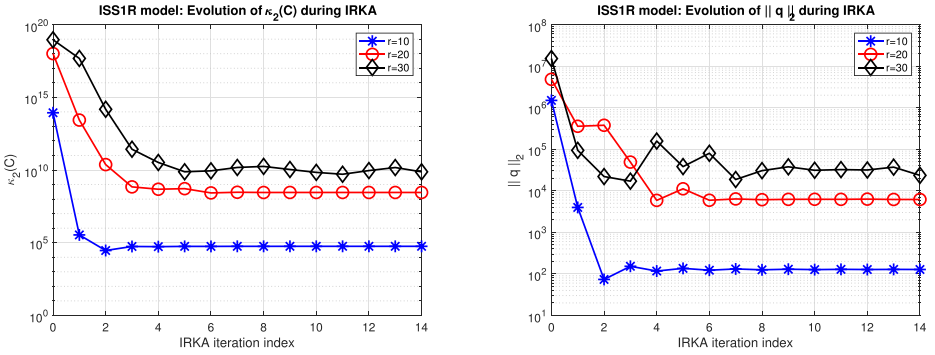

Fig. 3 $\kappa_2(\mathbf{C})$ and $\|\mathbf{q}^{(k)}\|_2$ during IRKA for the IRR 1R example

also simple, and they are obviously closed under conjugation. With this setup, we write

$$\mu_k = -\sigma_k + \varepsilon_k, \quad k = 1, \ldots, r, \tag{31}$$

where we note that the ε_k's are closed under conjugation as well. Our goal is to relate the ε_k's and the quality of the computed reduced order model identified by the triplet $(\mathbf{A}_r, \mathbf{b}_r, \mathbf{c}_r)$ in the sense of backward error. In particular, we need a yardstick to determine when an ε_k is small.

There is a caveat here: for given shifts $\boldsymbol{\sigma}$, the vector $\boldsymbol{\mu}$ consists of the computed eigenvalues, thus possibly highly ill-conditioned and computed with large errors. Our analysis here considers the computed eigenvalues as exact but for a slightly changed input data,[6] and we focus on the stopping criterion and how to justify it through a backward stability statement. This means that we want to interpret the computed reduced order model as an exact reduced order model for an LTI close to the original one (1).

The representation of the reduced order model in the primitive basis presented in Section 2 yields an elegant structure and provides theoretical insights. On the other hand, having numerical computation in mind, the same structure gives reasons to exercise caution. This caution is particularly justified because, as we showed in Section 3, the ultimate step of the iteration is an eigenvalue assignment problem: find the shifts $\boldsymbol{\sigma}$ such that the eigenvalues of $\boldsymbol{\Sigma}_r - \mathbf{q}(\boldsymbol{\sigma})\mathbf{e}^T$ are the reflected shifts.

Using (23), (31), and Lemma 1, we know that the vector $\mathbf{q} = \begin{bmatrix} q_1 & q_2 & \ldots & q_r \end{bmatrix}^T$ satisfies

$$q_i = (\sigma_i - \mu_i) \prod_{\substack{k=1 \\ k \neq i}}^{r} \frac{\sigma_i - \mu_k}{\sigma_i - \sigma_k} = (2\sigma_i - \varepsilon_i) \prod_{\substack{k=1 \\ k \neq i}}^{r} \frac{\sigma_i + \sigma_k - \varepsilon_k}{\sigma_i - \sigma_k}, \qquad i = 1, \ldots, r. \tag{32}$$

We now do the following *Gedankenexperiment*. Consider the true reflections of the shifts, $\mu_i^\bullet = -\sigma_i$. Since the pair $(\boldsymbol{\Sigma}_r, \mathbf{e})$ is controllable[7], there exists $\mathbf{q}^\bullet$ such that the eigenvalues of $\boldsymbol{\Sigma}_r - \mathbf{q}^\bullet \mathbf{e}^T$ are precisely $\mu_1^\bullet, \ldots, \mu_r^\bullet$. In fact, by the formula (25), the feedback $\mathbf{q}^\bullet$ is explicitly given as

$$q_i^\bullet = 2\sigma_i \prod_{\substack{k=1 \\ k \neq i}}^{r} \frac{\sigma_i + \sigma_k}{\sigma_i - \sigma_k}, \qquad i = 1, \ldots, r. \tag{33}$$

Comparing (32) and (33), we see that our computed vector $\mathbf{q}$ satisfies

$$q_i = q_i^\bullet \prod_{k=1}^{r} \left(1 - \frac{\varepsilon_k}{\sigma_i + \sigma_k}\right) \equiv q_i^\bullet (1 - \eta_i), \qquad i = 1, \ldots, r. \tag{34}$$

Since all the shifts are in the right-half-plane, $|\sigma_i + \sigma_k|$ is bounded from below by $2 \min_j \mathrm{Re}(\sigma_j) > 0$. We find it desirable to have our actually computed $\mathbf{q}$ close to $\mathbf{q}^\bullet$, because $\mathbf{q}^\bullet$ does exactly what we would like the computed $\mathbf{q}$ to achieve in the limit. This indicates one possible stopping criterion for the iterations—the maximal allowed distance between the new and the old shifts should guarantee small $|\varepsilon_k/(\sigma_i + \sigma_k)|$ for all i, k.

If we define $\mathbf{A}_r^\bullet \equiv \boldsymbol{\Sigma}_r - \mathbf{q}^\bullet \mathbf{e}^T$, then its eigenvalues are the reflections of the shifts. Comparing this outcome with the actually computed $\mathbf{A}_r = \boldsymbol{\Sigma}_r - \mathbf{q}\mathbf{e}^T$, we obtain the following result.

[6] This is the classical backward error interpretation.

[7] Let $\mathbf{A} \in \mathbb{C}^{n \times n}$ and $\mathbf{b} \in \mathbb{C}^n$. Then, the pair $(\mathbf{A}, \mathbf{b})$ is called controllable if $\mathrm{rank} \begin{bmatrix} \mathbf{b} & \mathbf{A}\mathbf{b} & \ldots & \mathbf{A}^{n-1}\mathbf{b} \end{bmatrix} = n$.

Proposition 2 *Let Algorithm 1 be stopped with computed* $\mathbf{A}_r = \mathbf{\Sigma}_r - \mathbf{q}\mathbf{e}^T$, *and let the eigenvalues of* $\mathbf{A}_r$ *be* $\mu_k = -\sigma_k + \varepsilon_k$, $k = 1, \ldots, r$. *Let*

$$\boldsymbol{\varepsilon}^\bullet \equiv \max_{1\le i\le r} \left| \prod_{k=1}^{r} \left(1 - \frac{\varepsilon_k}{\sigma_i + \sigma_k}\right) - 1 \right| < 1, \qquad \boldsymbol{\varepsilon} \equiv \max_{1\le i\le r} \left| \prod_{k=1}^{r} \left(1 + \frac{\varepsilon_k}{\sigma_i - \mu_k}\right) - 1 \right| < 1.$$

Then there exists $\mathbf{A}_r^\bullet = \mathbf{A}_r + \boldsymbol{\delta}\mathbf{A}_r$ *with eigenvalues* $-\sigma_1, \ldots, -\sigma_r$, *and* $\mathbf{q}^\bullet = \mathbf{q} - \boldsymbol{\delta}\mathbf{q}$ *such that* $\mathbf{A}_r^\bullet = \mathbf{\Sigma}_r - \mathbf{q}^\bullet\mathbf{e}^T$; $\|\boldsymbol{\delta}\mathbf{q}\|_2 \le \boldsymbol{\varepsilon}\|\mathbf{q}\|_2$, $\|\boldsymbol{\delta}\mathbf{q}\|_2 \le \boldsymbol{\varepsilon}^\bullet\|q^\bullet\|_2$; *and*

$$\|\boldsymbol{\delta}\mathbf{A}_r\|_2 \le 2\boldsymbol{\varepsilon}^\bullet\|\mathbf{A}_r^\bullet\|_2, \qquad \|\boldsymbol{\delta}\mathbf{A}_r\|_2 \le \boldsymbol{\varepsilon}\left(\|\mathbf{A}_r\|_2 + \left(1 + \max_k \left|\frac{\varepsilon_k}{\mu_k}\right|\right)\|\mathbf{A}_r\|_2\right).$$

Proof Define $\mathbf{q}^\bullet$ using (33) and write $\mathbf{q} = \mathbf{q}^\bullet + \boldsymbol{\delta}\mathbf{q}$. Write the actually computed reduced matrix $\mathbf{A}_r = \mathbf{\Sigma}_r - \mathbf{q}\mathbf{e}^T$ as

$$\mathbf{A}_r = \mathbf{\Sigma}_r - \mathbf{q}^\bullet\mathbf{e}^T - \boldsymbol{\delta}\mathbf{q}\mathbf{e}^T \quad \text{or} \quad \mathbf{A}_r^\bullet = \mathbf{\Sigma}_r - \mathbf{q}^\bullet\mathbf{e}^T, \quad \text{where } \mathbf{A}_r^\bullet = \mathbf{A}_r + \boldsymbol{\delta}\mathbf{q}\mathbf{e}^T. \tag{35}$$

Note that $\|\mathbf{\Sigma}_r\|_2 = \mathrm{spr}(\mathbf{A}_r^\bullet) \le \|\mathbf{A}_r^\bullet\|_2$. Further, using $\mathbf{q}^\bullet\mathbf{e}^T = \mathbf{\Sigma}_r - \mathbf{A}_r^\bullet$ and taking the norm we get

$$\sqrt{r}\|\mathbf{q}^\bullet\|_2 \le \mathrm{spr}(\mathbf{A}_r^\bullet) + \|\mathbf{A}_r^\bullet\|_2 \le 2\|\mathbf{A}_r^\bullet\|_2,$$

and thus the norm of $\boldsymbol{\delta}\mathbf{A}_r = \boldsymbol{\delta}\mathbf{q}\mathbf{e}^T$ can be estimated as

$$\|\boldsymbol{\delta}\mathbf{A}_r\|_2 = \sqrt{r}\|\boldsymbol{\delta}\mathbf{q}\|_2 \le \sqrt{r}\boldsymbol{\varepsilon}^\bullet\|\mathbf{q}^\bullet\|_2 \le 2\boldsymbol{\varepsilon}^\bullet\|\mathbf{A}_r^\bullet\|_2,$$

completing the proof. □

Remark 4 We conclude that in the vicinity of our computed data (reduced quantities) $\mathbf{A}_r$ and $\mathbf{q}$, there exist $\mathbf{A}_r^\bullet$ and $\mathbf{q}^\bullet$ that satisfy the stopping criterion exactly. Both $\|\mathbf{A}_r - \mathbf{A}_r^\bullet\|_2$ and $\|\mathbf{q} - \mathbf{q}^\bullet\|_2$ are estimated by the size of $\|\boldsymbol{\delta}\mathbf{q}\|_2$. But there is a subtlety here: we cannot use $\|\boldsymbol{\delta}\mathbf{q}\|_2$ as the stopping criterion. In other words, if we compute $\mathbf{q}$ and conclude that $\|\mathbf{q} - \mathbf{q}^\bullet\|_2$ is small, it does not mean that the μ_k's are close to the reflections of the σ_k's. There is difference between continuity and forward stability.

Our next goal is to interpret $\boldsymbol{\delta}\mathbf{A}_r$ and $\boldsymbol{\delta}\mathbf{q}$ as the results of backward perturbations in the initial data $\mathbf{A}$, $\mathbf{b}$.

Theorem 3 *Under the assumptions of Proposition 2, there exist backward perturbations* $\boldsymbol{\delta}\mathbf{A}$ *and* $\boldsymbol{\delta}\mathbf{b}$ *such that the reduced order system*

$$\begin{array}{l|l} \mathbf{A}_r^\bullet = \mathbf{\Sigma}_r - \mathbf{q}^\bullet\mathbf{e}^T = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T(\mathbf{A} + \boldsymbol{\delta}\mathbf{A})\mathbf{V} & \mathbf{b}_r^\bullet \equiv \mathbf{q}^\bullet = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T(\mathbf{b} - \delta\mathbf{b}) \\ \hline \mathbf{c}_r^\bullet = \mathbf{c}_r & \end{array}$$

corresponds to exact model reduction of the perturbed full-order model described by the triplet of matrices $(\mathbf{A} + \boldsymbol{\delta}\mathbf{A}, \mathbf{b} - \boldsymbol{\delta}\mathbf{b}, \mathbf{c})$ *and has its poles at the reflected shifts. Let* $G_r^\bullet(s) = \mathbf{c}_r^T(s\mathbf{I}_r - \mathbf{A}_r^\bullet)^{-1}\mathbf{b}_r^\bullet$ *and* $G^\bullet(s) = \mathbf{c}^T(s\mathbf{I}_n - (\mathbf{A} + \boldsymbol{\delta}\mathbf{A}))^{-1}(\mathbf{b} - \boldsymbol{\delta}\mathbf{b})$ *denote the transfer functions of this reduced order system, and the backward perturbed original system, respectively. Then,* $G_r^\bullet(\sigma_i) = G^\bullet(\sigma_i)$, $i = 1, \ldots, r$. *The backward perturbations satisfy*

$$\|\boldsymbol{\delta}\mathbf{b}\|_2 \le \frac{\kappa_2(\mathbf{V})}{\cos\angle(\mathcal{V}, \mathcal{W})}\boldsymbol{\varepsilon}\|\mathbf{b}\|_2$$

and

$$\|\boldsymbol{\delta}\mathbf{A}\|_2 \le \frac{\kappa_2(\mathbf{V})^2}{\cos\angle(\mathcal{V}, \mathcal{W})}\frac{2\boldsymbol{\varepsilon}^\bullet}{1 - 2\boldsymbol{\varepsilon}^\bullet}\|\mathbf{A}\|_2, \quad \textit{provided that } \boldsymbol{\varepsilon}^\bullet < 1/2,$$

where $\mathcal{V} = \mathit{Range}(\mathbf{V})$ *and* $\mathcal{W} = \mathit{Range}(\mathbf{W})$.

Proof First, recall that $\mathbf{q} = \mathbf{U}^T\mathbf{b}$, where $\mathbf{U}^T = (\mathbf{W}^T\mathbf{V})^{-1}\mathbf{W}^T$. Since $\mathbf{U}^T$ has full row-rank, we can determine $\boldsymbol{\delta}\mathbf{b}$ such that $\boldsymbol{\delta}\mathbf{q} = \mathbf{U}^T\boldsymbol{\delta}\mathbf{b}$. (The unique $\boldsymbol{\delta}\mathbf{b}$ of minimal Euclidean norm is $\boldsymbol{\delta}\mathbf{b} = (\mathbf{U}^T)^\dagger\boldsymbol{\delta}\mathbf{q} \in \mathcal{W}$.) Using (35) and $\mathbf{A}_r = \mathbf{U}^T\mathbf{A}\mathbf{V}$ we can write then

$$\mathbf{U}^T\mathbf{A}\mathbf{V} + \mathbf{U}^T\boldsymbol{\delta}\mathbf{b}\mathbf{e}^T = \boldsymbol{\Sigma}_r - \mathbf{U}^T(\mathbf{b} - \boldsymbol{\delta}\mathbf{b})\mathbf{e}^T,$$

where $\|\boldsymbol{\delta}\mathbf{b}\|_2 \le \|\mathbf{U}^\dagger\|_2\|\delta\mathbf{q}\|_2 \le \kappa_2(\mathbf{U})\|\mathbf{b}\|_2\boldsymbol{\varepsilon}$. Since we can express $\mathbf{e}$ as $\mathbf{e} = \mathbf{V}^T\mathbf{f}$ with smallest possible $\mathbf{f} = (\mathbf{V}^T)^\dagger\mathbf{e} \in \mathcal{V}$, we obtain

$$\mathbf{A}_r^\bullet = \mathbf{U}^T(\mathbf{A} + \boldsymbol{\delta}\mathbf{b}\mathbf{f}^T)\mathbf{V} = \boldsymbol{\Sigma}_r - \mathbf{U}^T(\mathbf{b} - \boldsymbol{\delta}\mathbf{b})\mathbf{e}^T.$$

Set $\boldsymbol{\delta}\mathbf{A} = \boldsymbol{\delta}\mathbf{b}\mathbf{f}^T$ and note that $\|\boldsymbol{\delta}\mathbf{A}\|_2 = \|\boldsymbol{\delta}\mathbf{b}\|_2\|\mathbf{f}\|_2$. From Proposition 2, under the mild assumption that $\boldsymbol{\varepsilon}^\bullet < 1/2$, we conclude that

$$\|\boldsymbol{\delta}\mathbf{q}\|_2 \le \frac{2\boldsymbol{\varepsilon}^\bullet}{\sqrt{r}(1-2\boldsymbol{\varepsilon}^\bullet)}\|\mathbf{A}_r\|_2, \quad \text{and thus} \quad \|\boldsymbol{\delta}\mathbf{b}\|_2 \le \frac{2\|\mathbf{U}^\dagger\|_2\boldsymbol{\varepsilon}^\bullet}{\sqrt{r}(1-2\boldsymbol{\varepsilon}^\bullet)}\|\mathbf{A}_r\|_2.$$

Since $\|\mathbf{U}^\dagger\|_2 \le \|\mathbf{V}\|_2$ and $\|\mathbf{f}\|_2 \le \sqrt{r}\|\mathbf{V}^\dagger\|_2$, we have

$$\|\delta\mathbf{A}\|_2 \le \kappa_2(\mathbf{V})\frac{2\boldsymbol{\varepsilon}^\bullet}{1-2\boldsymbol{\varepsilon}^\bullet}\|\mathbf{A}_r\|_2, \qquad \text{where } \|\mathbf{A}_r\| \le \frac{\kappa_2(\mathbf{V})}{\cos\angle(\mathcal{V},\mathcal{W})}\|\mathbf{A}\|_2.$$

Further, it holds that

$$\mathbf{V}\boldsymbol{\Sigma}_r - (\mathbf{A} + \boldsymbol{\delta}\mathbf{b}\mathbf{f}^T)\mathbf{V} = \mathbf{V}\boldsymbol{\Sigma}_r - \mathbf{A}\mathbf{V} - \boldsymbol{\delta}\mathbf{b}\mathbf{f}^T\mathbf{V} = \mathbf{b}\mathbf{e}^T - \boldsymbol{\delta}\mathbf{b}\mathbf{e}^T = (\mathbf{b} - \boldsymbol{\delta}\mathbf{b})\mathbf{e}^T,$$

and this implicitly enforces the interpolation conditions. □

Remark 5 To claim Hermite interpolation, the only freedom left is to change $\mathbf{c}$ into $\mathbf{c} + \boldsymbol{\delta}\mathbf{c}$ to guarantee that $(\sigma_i\mathbf{I} - (\mathbf{A}^T + \mathbf{f}\boldsymbol{\delta}\mathbf{b}^T))^{-1}(\mathbf{c} + \boldsymbol{\delta}\mathbf{c}) \in \mathcal{W}$ for $i = 1, \ldots, r$. In other words, with some $r \times r$ matrix $\boldsymbol{\Omega}$, we should have

$$\mathbf{W}\boldsymbol{\Omega}\boldsymbol{\Sigma} - (\mathbf{A}^T + \mathbf{f}\boldsymbol{\delta}\mathbf{b}^T)\mathbf{W}\boldsymbol{\Omega} = (\mathbf{c} + \boldsymbol{\delta}\mathbf{c})\mathbf{e}^T.$$

If $\boldsymbol{\Omega}$ commutes with $\boldsymbol{\Sigma}$, then $\boldsymbol{\delta}\mathbf{c}\mathbf{e}^T = \mathbf{c}\mathbf{e}^T(\boldsymbol{\Omega} - \mathbf{I}) - \mathbf{f}\boldsymbol{\delta}\mathbf{b}^T\mathbf{W}\boldsymbol{\Omega}$. We can take $\boldsymbol{\Omega} = \mathbf{I}$ and instead of the equality (which is not possible to to obtain), we can choose $\boldsymbol{\delta}\mathbf{c} = -(1/r)\mathbf{f}\boldsymbol{\delta}\mathbf{b}^T\mathbf{W}\mathbf{e}$, which is the least squares approximation. Even though this least-squares construction might provide a near-Hermite interpolation, a more elaborate construction is needed to obtain exact Hermite interpolation for a backward perturbed system. The framework that Beattie

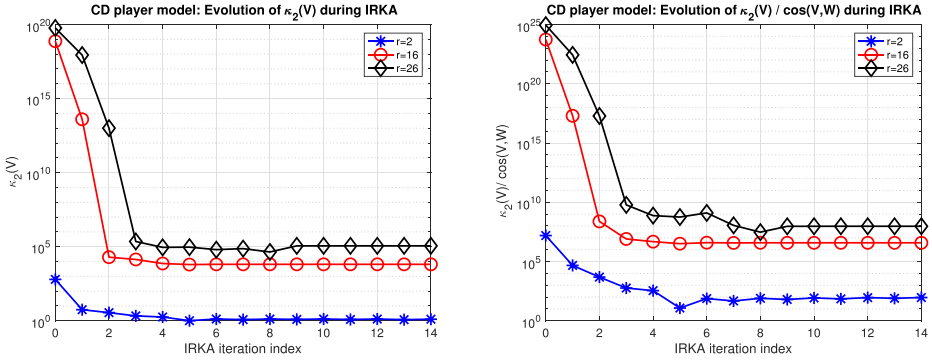

Fig. 4 $\kappa_2(\mathbf{V})$ and $\kappa_2(\mathbf{V})/\cos(\mathcal{V},\mathcal{W})$ during IRKA for the CD Player example

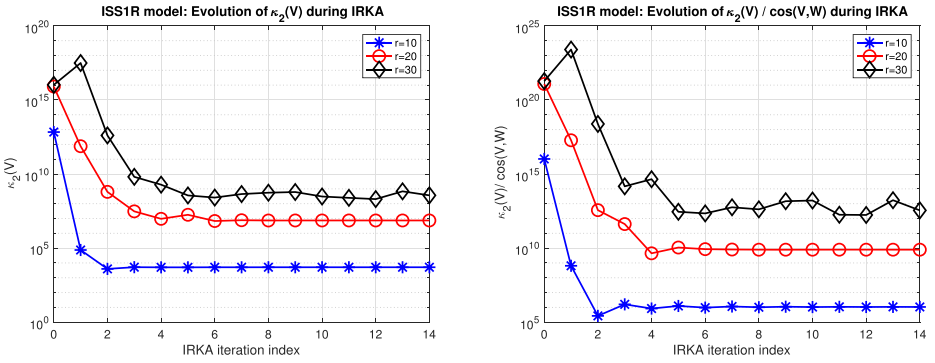

Fig. 5 $\kappa_2(\mathbf{V})$ and $\kappa_2(\mathbf{V})/\cos(\mathcal{V},\mathcal{W})$ during IRKA for ISS1R Example

et al. [10] provided for Hermite interpolation of a backward perturbed system in the special case of inexact solves might prove helpful in this direction.

Our analysis in Section 4.1, specifically Theorem 2, illustrated that the condition number of the Cauchy matrix $\mathbf{C}$ plays a crucial role in the perturbation analysis. And the numerical examples showed that despite Cauchy matrices are known to be extremely ill-conditioned, the IRKA iterations drastically reduced these conditions numbers as the shifts converge to the optimal ones, i.e., as IRKA converges. Our analysis in this section now reveals another important quantity measure: $\frac{\kappa_2(\mathbf{V})}{\cos\angle(\mathcal{V},\mathcal{W})}$. Next, we will repeat the same numerical examples of Section 4.1 and inspect how $\kappa_2(\mathbf{V})$ and $\frac{\kappa_2(\mathbf{V})}{\cos\angle(\mathcal{V},\mathcal{W})}$ vary during IRKA.

Example 3 We use the same models and experiments from Example 2. During the reduction of the CD player model to $r = 2$, $r = 16$, and $r = 26$ via IRKA, we record the evolution of $\kappa_2(\mathbf{V})$ and $\frac{\kappa_2(\mathbf{V})}{\cos\angle(\mathcal{V},\mathcal{W})}$. The results depicted in Fig. 4 show a similar story: Both quantities are drastically reduced during the iteration thus leading to significantly smaller backward errors $\|\boldsymbol{\delta}\mathbf{q}\|$ and $\|\boldsymbol{\delta}\mathbf{A}\|$ in Theorem 3.

We repeat the same experiments for the ISS 1R model and the results are shown in Fig. 5. The conclusion is the same: $\kappa_2(\mathbf{V})$ and $\frac{\kappa_2(\mathbf{V})}{\cos\angle(\mathcal{V},\mathcal{W})}$ are reduced ten orders of magnitudes during IRKA.

5 Conclusions

By employing primitive rational Krylov bases, we have provided here an analysis for the structure of reduced order quantities appearing in IRKA that reveals a deep connection to the classic pole-placement problem. We exploited this connection to motivate algorithmic modifications to IRKA and developed a complementary backward stability analysis. Several numerical examples demonstrate IRKA's remarkable tendency to realign shifts (interpolation points) in a way that drastically reduces the condition numbers of the quantities involved, thus minimizing perturbative effects and accounting in some measure for IRKA's observed robustness.

Acknowledgements The work of Beattie was supported in parts by NSF through Grant DMS-1819110. The work of Drmač was supported in parts by the Croatian Science Foundation Grant IP-2019-04-6268 and

the DARPA Contracts HR0011-16-C-0116 and HR0011-18-9-0033. The work of Gugercin was supported in parts by NSF through Grant DMS-1720257 and DMS-1819110.

References

1. Ahuja, K., Benner, P., de Sturler, E., Feng, L.: Recycling BiCGSTAB with an application to parametric model order reduction. SIAM J. Sci. Comput. **37**, S429–S446 (2015)
2. Ahuja, K., de Sturler, E., Gugercin, S., Chang, E.: Recycling BiCG with an application to model reduction. SIAM J. Sci. Comput. **34**, A1925–A1949 (2012)
3. Antoulas, A.C.: Approximation of Large-Scale Dynamical Systems. Advances in Design and Control. SIAM, Philadelphia (2005)
4. Antoulas, A.C., Beattie, C.A., Gugercin, S.: Interpolatory Methods for Model Reduction. Computational Science and Engineering, vol. 21. SIAM, Philadelphia (2020)
5. Antoulas, A.C., Beattie, C.A., Gugercin, S.: Interpolatory model reduction of large-scale dynamical systems. In: Mohammadpour, J., Grigoriadis, K.M. (eds.) Efficient Modeling and Control of Large-Scale Systems, pp. 2–58. Springer, Boston, MA (2010)
6. Antoulas, A.C., Sorensen, D.C., Gugercin, S.: A survey of model reduction methods for large-scale systems. Contemp. Math. **280**, 193–219 (2001)
7. Beattie, C., Gugercin, S.: Krylov-based minimization for optimal $\mathcal{H}_2$ model reduction. In: Proceedings of 46th IEEE Conference on Decision and Control, pp. 4385–4390. IEEE, Los Alamitos (2007)
8. Beattie, C., Gugercin, S.: A trust region method for optimal $\mathcal{H}_2$ model reduction. In: Proceedings of the 48th IEEE Conference on Decision and Control, pp. 5370–5375. IEEE, Los Alamitos (2009)
9. Beattie, C., Gugercin, S.: Realization-independent $\mathcal{H}_2$-approximation. In: Proceedings of 51st IEEE Conference on Decision and Control, pp. 4953–4958 (2012)
10. Beattie, C., Gugercin, S., Wyatt, S.: Inexact solves in interpolatory model reduction. Linear Algebra Appl. **436**, 2916–2943 (2012)
11. Benner, P., Breiten, T.: Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems. SIAM J. Matrix Anal. Appl. **33**, 859–885 (2012)
12. Benner, P., Goyal, P., Gugercin, S.: $\mathcal{H}_2$-quasi-optimal model order reduction for quadratic-bilinear control systems. SIAM J. Matrix Anal. Appl. **39**, 983–1032 (2018)
13. Benner, P., Köhler, M., Saak, J.: Sparse-dense Sylvester equations in $\mathcal{H}_2$-model order reduction. Tech. Rep. MPIMD/11-11, Max Planck Institute Magdeburg Preprints (2011)
14. Benner, P., Ohlberger, M., Cohen, A., Willcox, K. (eds.): Model Reduction and Approximation. Theory and Algorithms. SIAM, Philadelphia (2017)
15. Breiten, T., Beattie, C., Gugercin, S.: Near-optimal frequency-weighted interpolatory model reduction. Syst. Control Lett. **78**, 8–18 (2015)
16. Bunse-Gerstner, A., Kubalińska, D., Vossen, G., Wilczek, D.: $\mathcal{H}_2$-norm optimal model reduction for large scale discrete dynamical MIMO systems. J. Comput. Appl. Math. **233**, 1202–1216 (2010)
17. Chahlaoui, Y., Van Dooren, P.: A collection of benchmark examples for model reduction of linear time invariant dynamical systems. Tech. rep., SLICOT Working Note, 2002–2 (2002)
18. Chahlaoui, Y., Van Dooren, P.: Benchmark examples for model reduction of linear time-invariant dynamical systems. In: Benner, P., Sorensen, D.C., Mehrmann, V. (eds.) Dimension Reduction of Large-Scale Systems, pp. 379–392. Springer, Berlin (2005)
19. Draijer, W., Steinbuch, M., Bosgra, O.: Adaptive control of the radial servo system of a compact disc player. Automatica **28**, 455–462 (1992)
20. Drmač, Z., Gugercin, S., Beattie, C.: Quadrature-based vector fitting for discretized $\mathcal{H}_2$ approximation. SIAM J. Sci. Comput. **37**, A625–A652 (2015)
21. Drmač, Z., Gugercin, S., Beattie, C.: Vector fitting for matrix-valued rational approximation. SIAM J. Sci. Comput. **37**, A2346–A2379 (2015)
22. Eisenstat, S., Ipsen, I.: Three absolute perturbation bounds for matrix eigenvalues imply relative bounds. SIAM J. Matrix Anal. Appl. **20**, 149–158 (1998)
23. Elsner, L., Friedland, S.: Singular values, doubly stochastic matrices, and applications. Linear Algebra Appl. **220**, 161–169 (1995)
24. Ferrante, A., Krajewski, W., Lepschy, A., Viaro, U.: Convergent algorithm for l_2 model reduction. Automatica **35**, 75–79 (1999)
25. Flagg, G., Beattie, C., Gugercin, S.: Convergence of the iterative rational Krylov algorithm. Syst. Control Lett. **61**, 688–691 (2012)

26. Flagg, G., Gugercin, S.: Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems. SIAM J. Matrix Anal. Appl. **36**, 549–579 (2015)
27. Goyal, P., Redmann, M.: Time-limited $\mathcal{H}_2$-optimal model order reduction. Appl. Math. Comput. **355**, 184–197 (2019)
28. Gugercin, S.: An iterative SVD-krylov based method for model reduction of large-scale dynamical systems. Linear Algebra Appl. **428**, 1964–1986 (2008)
29. Gugercin, S., Antoulas, A.C., Beattie, C.: $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems. SIAM J. Matrix Anal. Appl. **30**, 609–638 (2008)
30. Gugercin, S., Antoulas, A.C., Bedrossian, M.: Approximation of the international space station 1R and 12A models. In: Proceedings of the 40th IEEE Conference on Decision and Control, pp. 1515–1516. IEEE, Los Alamitos (2001)
31. He, C., Laub, A., Mehrmann, V.: Placing Plenty of Poles is Pretty Preposterous. Tech. Rep., Preprint SPC 95-17 . Forschergruppe Scientific Parallel Computing, Fakultt für Mathematik, TU Chemnitz-Zwickau (1995)
32. Hokanson, J.M., Magruder, C.C.: $\mathcal{H}_2$-optimal model reduction using projected nonlinear least squares. arXiv:1811.11962 (2018)
33. Hyland, D., Bernstein, D.: The optimal projection equations for model reduction and the relationships among the methods of Wilson, Skelton, and Moore. IEEE Trans. Autom. Control **30**, 1201–1211 (1985)
34. Krajewski, W., Lepschy, A., Redivo-Zaglia, M., Viaro, U.: A program for solving the l_2 reduced-order model problem with fixed denominator degree. Numer. Algor. **9**, 355–377 (1995)
35. Krajewski, W., Viaro, U.: Iterative-interpolation algorithms for l_2 model reduction. Control Cybern. **38**, 543–554 (2009)
36. Mehrmann, V., Xu, H.: An analysis of the pole placement problem I. The single-input case. Electron. Trans. Numer. Anal. **4**, 89–105 (1996)
37. Mehrmann, V., Xu, H.: Choosing poles so that the single-input pole placement problem is well conditioned. SIAM J. Matrix Anal. Appl. **19**, 664–681 (1998)
38. Meier, L.I., Luenberger, D.: Approximation of linear constant systems. IEEE Trans. Autom. Control **12**, 585–588 (1967)
39. Panzer, H.K.F., Jaensch, S., Wolf, T., Lohmann, B.: A greedy rational Krylov method for $\mathcal{H}_2$-pseudooptimal model order reduction with preservation of stability. In: 2013 American Control Conference, pp. 5512–5517 (2013)
40. Poussot-Vassal, C.: An iterative SVD-tangential interpolation method for medium-scale MIMO systems approximation with application on flexible aircraft. In: 2011 50th IEEE Conference on Decision and Control and European Control Conference, pp. 7117–7122 (2011)
41. Spanos, J.T., Milman, M.H., Mingori, D.L.: A new algorithm for l_2 optimal model reduction. Automatica **28**, 897–909 (1992)
42. Van Dooren, P., Gallivan, K.A., Absil, P.A.: $\mathcal{H}_2$-optimal model reduction of MIMO systems. Appl. Math. Lett. **21**, 1267–1273 (2008)
43. Vuillemin, P., Poussot-Vassal, C., Alazard, D.: $\mathcal{H}_2$ optimal and frequency limited approximation methods for large-scale LTI dynamical systems. IFAC Proceedings Volumes **46**, 719–724 (2013)
44. Žigić, D., Watson, L.T., Beattie, C.: Contragredient transformations applied to the optimal projection equations. Linear Algebra Appl. **188–189**, 665–676 (1993)
45. Wilson, D.A.: Optimum solution of model-reduction problem. Proc. Inst. Electr. Eng. **117**, 1161–1165 (1970)
46. Xu, Y., Zeng, T.: Optimal $\mathcal{H}_2$ model reduction for large scale MIMO systems via tangential interpolation. Int. J. Numer. Anal. Model. **8**, 174–188 (2011)

Affiliations

Christopher Beattie[1] · Zlatko Drmač[2] · Serkan Gugercin[1]

Christopher Beattie
beattie@vt.edu

Zlatko Drmač
drmac@math.hr

[1] Department of Mathematics and Division of Computational Modeling and Data Analytics, Virginia Tech, Blacksburg, VA, 24061, USA

[2] Faculty of Science, Department of Mathematics, University of Zagreb, Zagreb, 10000, Croatia