INEXACT LINEAR SOLVES IN MODEL REDUCTION OF BILINEAR DYNAMICAL SYSTEMS

Rajendra Choudhary and Kapil Ahuja

Abstract

Bilinear dynamical systems are commonly used in science and engineering because they form a bridge between linear and non-linear systems. However, simulating them is still a challenge because of their large size. Hence, a lot of research is currently being done for reducing such bilinear dynamical systems (termed as bilinear model order reduction or bilinear MOR). Bilinear iterative rational Krylov algorithm (BIRKA) is a very popular, standard and mathematically sound algorithm for bilinear MOR, which is based upon interpolatory projection technique. An efficient variant of BIRKA, Truncated BIRKA (or TBIRKA) has also been recently proposed.

Like for any MOR algorithm, these two algorithms also require solving multiple linear systems as part of the model reduction process. For reducing very large dynamical systems, which is now-a-days becoming a norm, scaling of such linear systems with respect to input dynamical system size is a bottleneck. For efficiency, these linear systems are often solved by an iterative solver, which introduces approximation errors. Hence, stability analysis of MOR algorithms with respect to inexact linear solves is important. In our past work, we have shown that under mild conditions, BIRKA is stable (in the sense as discussed above). Here, we look at stability of TBIRKA in the same context.

Besides deriving the conditions for a stable TBIRKA, our other novel contribution is the more intuitive methodology for achieving this. This approach exploits the fact that in TBIRKA a bilinear dynamical system can be represented by a finite set of functions, which was not possible

Key Words: Model order reduction, Bilinear dynamical systems, Interpolatory projection, Stability analysis, Backward stability, Perturbation analysis, Volterra series interpolation.

2010 Mathematics Subject Classification: Primary 34D10, 34D20, 41A05, 65F10.
in BIRKA (because infinite such functions were needed there). The stability analysis techniques that we propose here can be extended to many other methods for doing MOR of bilinear dynamical systems, e.g., using balanced truncation or the ADI methods.

1 Introduction

A dynamical system, usually represented by a set of differential equations, can be linear or non-linear [1, 2, 3]. Linear dynamical systems have been studied more than the non-linear ones because of the obvious ease in working with them. Bilinear dynamical systems form a good bridge between the linear and the non-linear cases, and are usually approximated by a varying degree of bilinearity [4, 5, 6]. In this manuscript, we focus on bilinear dynamical systems.

Dynamical systems coming from different real world applications are very large in size. Thus, simulation and computation with such systems is prohibitively expensive. Model Order Reduction (MOR) techniques provide a smaller system that besides being cheaper to work with, also replicates the input-output behaviour of the original system to a great extent [7, 8, 9, 10, 11, 12, 13, 14, 15, 16]. Since, bilinear dynamical systems have been recently studied, the techniques for reducing them are also recent.

Out of the many methods available for performing bilinear MOR [14, 15, 17, 18, 19, 20, 21, 22], we focus on a commonly used interpolatory projection method. BIRKA (Bilinear Iterative Rational Krylov Algorithm) [15] is a very popular algorithm based upon this technique for reducing first-order bilinear dynamical systems*. BIRKA’s biggest drawback is that it does not scale well in time (with respect to increase in the size of the input dynamical system). A cheaper variant of BIRKA, called TBIRKA (Truncated Bilinear Iterative Rational Krylov Algorithm) [21, 22] has also been proposed.

Like in any other MOR algorithm, in BIRKA and TBIRKA also, people often use direct methods like LU-factorization, etc., to solve the arising linear systems, which are too expensive (i.e., computational complexity of $O(n^3)$, where $n$ is the original system size). A common solution to this scaling problem is to use iterative methods like the Krylov subspace methods, etc., which have a reduced computational complexity (i.e., $O(n \times \text{nnz})$, where $\text{nnz}$ is the number of nonzeros in the system matrix) [23, 24]. Although iterative methods are cheap, they are inexact too. Hence, studying stability of the underlying MOR algorithm (here BIRKA and TBIRKA) with respect to such approximate linear solves becomes important [25, 26].

*First-order implies that the highest derivative of the state variable in the dynamical system is one. Second-order and higher-orders are similarly defined.
One of the first works that performed such a stability analysis focused on popular MOR algorithms for \textit{first-order} linear dynamical systems [27]. Here, the authors briefly mention that their analysis would be easily carried from the \textit{first-order} to the \textit{second-order} case. A detailed stability analysis focusing on reducing \textit{second-order} linear dynamical systems has been done in [28]. A different kind of stability analysis for MOR of \textit{second-order} linear dynamical systems has been done in [29]. In this, the authors first show that the SOAR algorithm (second order Arnoldi) is unstable with respect to the machine precision errors (and not inexact linear solves). Then, they propose a Two-level orthogonal Arnoldi (TOAR) algorithm that cures this instability of SOAR. An extended stability analysis for BIRKA (as above, a popular MOR algorithm for \textit{first-order} bilinear dynamical systems) has been recently done in [30]. For rest of this manuscript, whenever stability analysis is referred, we mean it with respect to inexact linear solves.

We follow the stability analysis framework of BIRKA from [30] and propose equivalent theorems for TBIRKA. The approach here is slightly different, which forms our most novel contribution. Norm of the dynamical system plays an important role in stability analysis (the kind of norm is discussed later). In BIRKA stability analysis, a single expression for bilinear dynamical system norm is used (involving a Volterra series). In TBIRKA stability analysis, a similar single expression (involving a truncated Volterra series) leads to complications. Alternatively, in TBIRKA, because of truncation, the bilinear dynamical system can be represented by a finite set of functions. This was not possible in BIRKA where infinite such functions were needed. Thus, in TBIRKA stability analysis, we use norm of all such functions leading to easier derivations. Our stability analysis, as done for BIRKA earlier and for TBIRKA here, can be easily extended to other MOR algorithms for bilinear dynamical systems, e.g., projection based [14], implicit Volterra series [17], balanced truncation [18], gramian based [20], etc.

The rest of the paper is divided into three more parts. In Section 2, we first give a brief overview of MOR for bilinear dynamical systems using a projection method. Next, we review the stability analysis of BIRKA from [30]. A stability analysis typically involves satisfying two conditions. Hence, finally in this section, we study the first condition for a stable TBIRKA. We analyze the second condition of stability in TBIRKA’s context in Section 3. In Section 4, we give conclusions and discuss the future work. For the rest of this paper, we use the terms and notations as listed below.

a. The $H_2$–norm is a functional norm defined as [27, 21, 22]

$$
\|H_k\|_{H_2}^2 = \left(\frac{1}{2\pi}\right)^k \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \|H_k(\imath\omega_1, \ldots, \imath\omega_k)\|^2_F \, d\omega_1 \ldots d\omega_k, \quad (1)
$$
where \( i \) denotes \( \sqrt{-1} \). Here, we assume that all \( H_2 \)-norms computed further exist. In other words, the improper integrals defined by the \( H_2 \)-norm give finite value. This is a reasonable assumption because this happens often in practice (see [27], where stability analysis of IRKA is done as well as [30], where stability analysis of BIRKA is done).

b. The \( H_\infty \)-norm is also a functional norm, defined as [27, 21, 22]

\[
\|H_k\|_{H_\infty} = \max_{\omega_1, \ldots, \omega_k \in \mathbb{R}} \|H_k(i\omega_1, \ldots, i\omega_k)\|_2.
\]

c. The Kronecker product between two matrices \( P \) (of size \( m \times n \)), and \( Q \) (of size \( s \times t \)) is defined as

\[
P \otimes Q = \begin{bmatrix} p_{11}Q & \cdots & p_{1n}Q \\ \vdots & \ddots & \vdots \\ p_{m1}Q & \cdots & p_{mn}Q \end{bmatrix},
\]

where \( p_{ij} \) is an element of matrix \( P \) and order of \( P \otimes Q \) is \( ms \times nt \).

d. vec operator on a matrix \( P \) is defined as

\[
\text{vec}(P) = (p_{11}, \ldots, p_{m1}, p_{12}, \ldots, p_{m2}, \ldots, \ldots, p_{1n}, \ldots, p_{mn})^T.
\]

e. Also, \( I_n \) denotes an identity matrix of size \( n \times n \) and \( \mathbb{R} \) denotes the set of real numbers.

2 Background

A first-order bilinear dynamical system is usually represented as [14, 15]

\[
\zeta : \begin{cases} \dot{x}(t) = Ax(t) + Nx(t)u(t) + bu(t), \\ y(t) = cx(t), \end{cases}
\]

where \( A, N \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^{n \times 1} \) and \( c \in \mathbb{R}^{1 \times n} \). Also, \( u(t) : \mathbb{R} \to \mathbb{R}, y(t) : \mathbb{R} \to \mathbb{R} \) and \( x(t) : \mathbb{R} \to \mathbb{R}^n \), represent input, output and state of the bilinear dynamical system, respectively. This is a Single Input Single Output (SISO) system, which we have chosen for ease of our analysis. We plan to look at Multiple Input Multiple Output (MIMO) systems as part of our future work.

A bilinear dynamical system can also be represented by a series of transfer functions, i.e.,

\[
\zeta = \lim_{k \to \infty} \zeta^k,
\]
Rajendra Choudhary and Kapil Ahuja

where \( \zeta^k = \{ H_1(s_1), H_2(s_1, s_2), \ldots, H_k(s_1, s_2, \ldots, s_k) \} \). Here, \( H_k(s_1, s_2, \ldots, s_k) \) is called the \( k^{th} \) order transfer function of the bilinear dynamical system and is defined as

\[
H_k(s_1, s_2, \ldots, s_k) = c(s_kI - A)^{-1} N(s_{k-1}I - A)^{-1} \cdots N(s_1I - A)^{-1} b. \tag{4}
\]

After reduction, the bilinear dynamical system (2) can be represented as

\[
\zeta_r: \begin{cases}
\dot{x}_r(t) = A_rx_r(t) + N_rx_r(t)u(t) + b_ru(t), \\
y_r(t) = c_r x_r(t),
\end{cases}
\]  

where \( A_r, N_r \in \mathbb{R}^{r \times r}, b_r \in \mathbb{R}^{r \times 1} \) and \( c_r \in \mathbb{R}^{1 \times r} \) with \( r \ll n \). The main goal of model reduction is to approximate \( \zeta \) by \( \zeta_r \) in an appropriate norm, such that for all admissible inputs, \( y_r(t) \) is nearly same to \( y(t) \).

As mentioned earlier, we use interpolatory projection for performing model reduction. The two common and standard algorithms here, BIRKA and TBIRKA, use a Petrov-Galerkin projection. Let \( V_r \) and \( W_r \) be the two \( r \)-dimensional subspaces, such that \( V_r = \text{Range}(V_r) \) and \( W_r = \text{Range}(W_r) \), where \( V_r \in \mathbb{R}^{n \times r} \) and \( W_r \in \mathbb{R}^{n \times r} \) are matrices. Also, let \( (W_r^T V_r) \) be invertible\(^1\). Applying the projection \( x(t) = V_r x_r(t) \), and enforcing the Petrov-Galerkin conditions [15, 22] on the original bilinear dynamical system (2), we get the reduced system as

\[
W_r^T (V_r \dot{x}_r(t) - AV_r x_r(t) - NV_r x_r(t)u(t) - bu(t)) = 0,
\]

\[
y(t) = cV_r x_r(t).
\]

Comparing the above two equations with their respective equations in (5), we get a relation between the original system matrices and the reduced system matrices, i.e.,

\[
A_r = (W_r^T V_r)^{-1} W_r^T AV_r, \quad N_r = (W_r^T V_r)^{-1} W_r^T NV_r,
\]

\[
b_r = (W_r^T V_r)^{-1} W_r^T b, \quad c_r = cV_r.
\]

One way of obtaining subspaces \( V_r \) and \( W_r \) is to use Volterra series interpolation. Further, to decide where to interpolate so as to obtain an optimal reduced model, an \( H_2 \)-optimization problem is commonly solved (Theorem 4.7 from [21]).

\textbf{Theorem 1.} [21] Let \( \zeta \) be a bilinear system of order \( n \). Let \( \zeta_r \) be an \( H_2 \)-optimal approximation of order \( r \). Then, \( \zeta_r \) satisfies the following multi-point

---

\(^1\)Obtaining such an invertible matrix is not difficult [15, 22].
Volterra series interpolation conditions:
\[
\sum_{k=1}^{\infty} \sum_{i_1=1}^{r} \cdots \sum_{i_k=1}^{r} \phi_{i_1, i_2, \ldots, i_k} \frac{H_k(-\lambda_{i_1}, -\lambda_{i_2}, \ldots, -\lambda_{i_k})}{H_{r_k}(-\lambda_{i_1}, -\lambda_{i_2}, \ldots, -\lambda_{i_k})} = 0,
\]
and
\[
\sum_{k=1}^{\infty} \sum_{i_1=1}^{r} \cdots \sum_{i_k=1}^{r} \phi_{i_1, i_2, \ldots, i_k} \left( \sum_{j=1}^{k} \frac{\partial}{\partial s_j} H_k(-\lambda_{i_1}, -\lambda_{i_2}, \ldots, -\lambda_{i_k}) \right) = 0.
\]

where \( \phi_{i_1, i_2, \ldots, i_k} \) and \( \lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_k} \) are residues and poles of the transfer function \( H_{r_k} \) associated with \( \zeta_r \), respectively.

BIRKA is designed in such a way that at convergence, the conditions of Theorem 1 are satisfied leading to a locally \( H_2 \)-optimal reduced model. Algorithm 1 lists BIRKA.

\begin{algorithm}[H]
\caption{BIRKA [15, 21, 22]}
1: Given an input bilinear dynamical system \( A, N, b, c \).
2: Select an initial guess for the reduced system as \( \hat{A}, \hat{N}, \hat{b}, \hat{c} \). Also select stopping tolerance \( btol \).
3: While (relative change in eigenvalues of \( \hat{A} \geq btol \))
   a. \( R\Lambda R^{-1} = \hat{A}, \quad \hat{b} = \hat{b}^T R^{-T}, \quad \hat{c} = \hat{c}^T \hat{R}, \quad \hat{N} = R^T \hat{N} R^{-T} \).
   b. \( \text{vec}(V) = \left( -\Lambda \otimes I_n - I_r \otimes A - \hat{N}^T \otimes \hat{N} \right)^{-1} (\hat{b}^T \otimes b) \).
   c. \( \text{vec}(W) = \left( -\Lambda \otimes I_n - I_r \otimes A^T - \hat{N}^T \otimes \hat{N} \right)^{-1} (\hat{c}^T \otimes c^T) \).
   d. \( V_r = \text{orth}(V), \quad W_r = \text{orth}(W) \).
   e. \( \hat{A} = (W_r^T V_r)^{-1} W_r^T A V_r, \quad \hat{N} = (W_r^T V_r)^{-1} W_r^T N V_r, \quad \hat{b} = (W_r^T V_r)^{-1} W_r^T b, \quad \hat{c} = c V_r. \)
4: \( A_r = \hat{A}, \quad N_r = \hat{N}, \quad b_r = \hat{b}, \quad c_r = \hat{c}. \)
\end{algorithm}

TBIRKA is similar to BIRKA in most aspects except that it performs a truncated Volterra series interpolation. Here, instead of \( \zeta \) in (2)-(3), they work
with \( \zeta^M \), which is defined as
\[
\zeta^M = \{ H_1(s_1), H_2(s_1, s_2), H_3(s_1, s_2, s_3), \ldots, H_M(s_1, \ldots, s_M) \},
\]
with \( H_k(s_1, \ldots, s_k) \) for \( k = 1, \ldots, M \) is given by (4). Equivalent of Theorem 1 above is as follows (Theorem 4.8 from [21]):

**Theorem 2.** [21] Let \( \zeta = (A, N, b, c) \) be an order \( n \) bilinear system and \( \zeta^M \) be the polynomial system determined by \( \zeta \). Let \( \zeta^r = (A_r, N_r, b_r, c_r) \) be a bilinear system of order \( r \), and define \( \zeta^M_r \) as the polynomial system determined by \( \zeta^r \). Suppose that \( \zeta^M_r \) is an \( H_2 \)-optimal approximation to \( \zeta^M \). Then \( \zeta^M_r \) satisfies
\[
\sum_{k=1}^{M} \sum_{l_1=1}^{r} \ldots \sum_{l_k=1}^{r} \phi_{l_1, l_2, \ldots, l_k} H_k(-\lambda_{l_1}, -\lambda_{l_2}, \ldots, -\lambda_{l_k}) =
\]
\[
\sum_{k=1}^{M} \sum_{l_1=1}^{r} \ldots \sum_{l_k=1}^{r} \phi_{l_1, l_2, \ldots, l_k} H_{r_k}(-\lambda_{l_1}, -\lambda_{l_2}, \ldots, -\lambda_{l_k}),
\]
and
\[
\sum_{k=1}^{M} \sum_{l_1=1}^{r} \ldots \sum_{l_k=1}^{r} \phi_{l_1, l_2, \ldots, l_k} \left( \sum_{j=1}^{k} \frac{\partial}{\partial s_j} H_k(-\lambda_{l_1}, -\lambda_{l_2}, \ldots, -\lambda_{l_k}) \right) =
\]
\[
\sum_{k=1}^{M} \sum_{l_1=1}^{r} \ldots \sum_{l_k=1}^{r} \phi_{l_1, l_2, \ldots, l_k} \left( \sum_{j=1}^{k} \frac{\partial}{\partial s_j} H_{r_k}(-\lambda_{l_1}, -\lambda_{l_2}, \ldots, -\lambda_{l_k}) \right),
\]
where \( \phi_{l_1, l_2, \ldots, l_k} \) and \( \lambda_{l_1}, \lambda_{l_2}, \ldots, \lambda_{l_k} \) are residues and poles of the transfer function \( H_{r_k} \) associated with \( \zeta^M_r \), respectively.

Algorithm 2 lists TBIRKA.

Both BIRKA and TBIRKA in turn require solving large sparse linear systems of equations. If we compare Algorithm 1 and 2, we realize that the cost of linear solvers at each step of the \textbf{while} loop in the former is \( \mathcal{O}(nr \times nr) \) and in the latter is \( \mathcal{O}(M \times nr \times nr) \). This makes it seem that TBIRKA is more expensive than BIRKA. However, TBIRKA is implemented in such a way that the Kronecker products are avoided leading to cost of linear solves at each step of the \textbf{while} loop to be \( \mathcal{O}(M \times r \times (n \times n)) \). Thus, if \( M < r \), which is usually the case (e.g., \( M = 3 \) or 4 and \( r \gg M \)), then TBIRKA is more efficient than BIRKA. For further details on this see chapter 4 in [21] and Section 5.3 in [22]. These implementation details do not affect our stability analysis, and hence, we use Algorithm 2 in the current form as our base.

As mentioned earlier, using iterative methods for solving such linear systems introduces approximation errors. We have done a detailed stability analysis of BIRKA with respect to the inexact linear solves in [30], and we briefly
Algorithm 2 TBIRKA [21, 22]

1: Given an input bilinear dynamical system $A$, $N$, $b$, $c$.
2: Select an initial guess for the reduced system as $\hat{A}$, $\hat{N}$, $\hat{b}$, $\hat{c}$. Also select the truncation index $M$ and stopping tolerance $tbtol$.
3: While (relative change in eigenvalues of $\hat{A}$ $\geq$ tbtol)
   a. $R\Lambda R^{-1} = \hat{A}$, $\hat{b} = \hat{b}^T R^{-T}$, $\hat{c} = \hat{c} R$, $\hat{N} = R^T N R^{-T}$.
   b. Compute
      $$vec(V_1) = (-\Lambda \otimes I_n - I_r \otimes A)^{-1} \left(\hat{b}^T \otimes b\right),$$
      $$vec(W_1) = (-\Lambda \otimes I_n - I_r \otimes A^T)^{-1} \left(\hat{c}^T \otimes c^T\right).$$
   c. For $j = 2, \ldots, M$, solve
      $$vec(V_j) = (-\Lambda \otimes I_n - I_r \otimes A)^{-1} \left(\hat{N}^T \otimes N\right) vec(V_{j-1}),$$
      $$vec(W_j) = (-\Lambda \otimes I_n - I_r \otimes A^T)^{-1} \left(\hat{N} \otimes N^T\right) vec(W_{j-1}).$$
   d. $V = \sum_{j=1}^{M} V_j$, $W = \sum_{j=1}^{M} W_j$.
   e. $V_r = orth(V)$, $W_r = orth(W)$.
   f. $\hat{A} = (W_r^T V_r)^{-1} W_r^T A V_r$, $\hat{N} = (W_r^T V_r)^{-1} W_r^T N V_r$,
      $\hat{b} = (W_r^T V_r)^{-1} W_r^T b$, $\hat{c} = c V_r$.
4: $A_r = \hat{A}$, $N_r = \hat{N}$, $b_r = \hat{b}$, $c_r = \hat{c}$.

revisit this next. Generally, accuracy is the metric that tells about the correctness in the output of an inexact algorithm. Due to unavailability of the exact output, it is not possible to determine accuracy [25, 30]. A more easier metric is stability. Backward stability is one such notation, which says “A backward stable algorithm gives exactly the right output to nearly the right input” [25]. In our context, theoretically we obtain two reduced systems. One by applying an inexact MOR algorithm (with iterative linear solves) on the original full model, and other by applying the same MOR algorithm but exactly (with direct linear solves) on a perturbed full model (the perturbation is introduced in the original full model as part of stability analysis, and is an unknown quantity). If these two reduced systems are equal (first condition), with the difference between the original full model and the perturbed full model equal
to the order of perturbation (second condition), then the MOR algorithm under consideration is called backward stable. The two theorems summarizing this stability analysis for BIRKA are listed below.

**Theorem 3.** [30] If the inexact linear solves in BIRKA (lines 3b. and 3c. of Algorithm 1) are solved using a Petrov-Galerkin framework, then BIRKA satisfies the first condition of backward stability with respect to these solves.

**Theorem 4.** [30] Let
\[
\hat{Q} = \left( - \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} \otimes \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} - \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \otimes \begin{bmatrix} A & 0 \\ 0 & A \end{bmatrix} - \begin{bmatrix} N & 0 \\ 0 & N \end{bmatrix} \otimes \begin{bmatrix} N & 0 \\ 0 & N \end{bmatrix} \right),
\]
where \( I_n \) is an identity matrix of size \( n \times n \) and \( \otimes \) denotes the standard Kronecker product. Also, let \( \hat{F} = (I_{2n} \otimes \hat{F} + \hat{F} \otimes I_{2n}) \) with \( \hat{F} = \begin{bmatrix} 0 & 0 \\ 0 & F \end{bmatrix} \), where \( F \) is the perturbation introduced in \( A \) matrix of the input dynamical system and \( I_{2n} \) is an identity matrix of size \( 2n \times 2n \). If \( \hat{Q} \) is invertible, \( \| \hat{Q}^{-1} \|_2 < 1 \), and \( \| \hat{F} \|_2 < 1 \), then BIRKA satisfies the second condition of backward stability with respect to the inexact linear solves.

Next, we analyze the backward stability of TBIRKA. Here, the first condition is satisfied in a way similar to that of BIRKA except that some extra orthogonality conditions are imposed on the linear solver (discussed below).

**Theorem 5.** Let the inexact linear solves in TBIRKA (lines 3b. and 3c. of Algorithm 2) are solved using a Petrov-Galerkin framework with
\[
\sum_{j=1}^{M} W_j \perp R_{b_j} \quad \text{and} \quad \sum_{j=1}^{M} V_j \perp R_{c_j},
\]
where \( R_{b_1} \) and \( R_{b_j} \) are the residuals in the first equations of 3b. and 3c. of Algorithm 2, respectively; \( R_{c_1} \) and \( R_{c_j} \) are the residuals in the second equations of 3b. and 3c. of Algorithm 2, respectively; and \( j = 2, \ldots, M \). Then, TBIRKA satisfies the first condition of backward stability with respect to these solves.

**Proof.** Follows the same pattern as the proof for Theorem 2.1 in [30].

From the above theorem, we infer that the underlying iterative solver should firstly be based upon a Petrov-Galerkin framework. Since BiConjugate Gradient (i.e., BiCG) is one such algorithm [23, 24], we propose its use in
TBIRKA. This is exactly same as for BIRKA. Secondly, this particular solver should also satisfy the extra orthogonalities (7).

These orthogonalities have a form similar to the orthogonalities required while reducing second order linear dynamical systems ((23) and (24) in [28]), and can be easily satisfied by using a recycling variant of the underlying iterative solver. In [28], the ideal iterative solver to be used is Conjugate Gradient (i.e., CG) [23, 24] (due to the use of Galerkin projection). Hence, to satisfy the similar orthogonalities there, without any extra cost, the authors use Recycling Conjugate Gradient (i.e., RCG) [33]. Since here BiCG is the ideal iterative solver (as discussed above), we propose the use of Recycling BiConjugate Gradient (i.e., RBiCG) [32, 31], which would ensure that orthogonalities given by (7) are satisfied without any extra cost.

To satisfy the second condition of backward stability of TBIRKA, we need to show that

\[ \| \zeta^M - \tilde{\zeta}^M \|_{H_2} = O(\|F\|_2), \]  

where \( \zeta^M \) is given by (6) or

\[ \zeta^M = \{ H_1(s_1), H_2(s_1, s_2), \ldots, H_M(s_1, \ldots, s_M) \}, \]  

(9a)

with \( H_k(s_1, \ldots, s_k) \) for \( k = 1, \ldots, M \) given by (4) or

\[ H_k(s_1, \ldots, s_k) = c(s_k I - A)^{-1} N(s_{k-1} I - A)^{-1} \ldots N(s_1 I - A)^{-1} b, \]  

(9b)

\[ \tilde{\zeta}^M = \{ \tilde{H}_1(s_1), \tilde{H}_2(s_1, s_2), \ldots, \tilde{H}_M(s_1, \ldots, s_M) \}, \]  

(10a)

with for \( k = 1, \ldots, M \)

\[ \tilde{H}_k(s_1, \ldots, s_k) = c(s_k I - (A + F))^{-1} N(s_{k-1} I - (A + F))^{-1} \ldots N(s_1 I - (A + F))^{-1} b, \]  

(10b)

and assuming perturbation in \( A \) matrix of the input dynamical system (as for BIRKA stability; see Theorem 4 earlier).

One way to satisfy (8) is to use the definition of the \( H_2 \)-norm of \( \zeta^M - \tilde{\zeta}^M \),
i.e., from Lemma 5.1 of [22]

$$\left\| \zeta_M - \tilde{\zeta}_M \right\|_{H_2}^2 = \left( [c \ -c] \odot [c \ -c] \right) \sum_{j=0}^{M} \left( - \begin{bmatrix} A & 0 \\ 0 & A + F \end{bmatrix} \odot \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \right) \cdot \left( - \begin{bmatrix} A & 0 \\ 0 & A + F \end{bmatrix} \odot \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \right)^{-1} \left( \begin{bmatrix} N & 0 \\ 0 & N \end{bmatrix} \odot \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right)^j$$

$$\left( \begin{bmatrix} b \\ b \end{bmatrix} \odot \begin{bmatrix} b \\ b \end{bmatrix} \right).$$

This approach is followed in satisfying the second condition of backward stability of BIRKA, but for TBIRKA it turns out to be more challenging. The reason for this is that the definition of the $H_2$–norm of $\zeta - \tilde{\zeta}$ used in BIRKA is different from (11), i.e., from Corollary 4.1 of [15] or Theorem 4.5 of [21]

$$\left\| \zeta - \tilde{\zeta} \right\|_{H_2}^2 = \left( [c \ -c] \odot [c \ -c] \right) \left( - \begin{bmatrix} A & 0 \\ 0 & A + F \end{bmatrix} \odot \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \right) \cdot \left( - \begin{bmatrix} A & 0 \\ 0 & A + F \end{bmatrix} \odot \begin{bmatrix} I_n & 0 \\ 0 & I_n \end{bmatrix} \right)^{-1} \left( \begin{bmatrix} b \\ b \end{bmatrix} \odot \begin{bmatrix} b \\ b \end{bmatrix} \right).$$

Another way to satisfy (8) in case of TBIRKA, which turns out to be more easier, is to show that

$$\left\| H_k (s_1) - \tilde{H}_k (s_1) \right\|_{H_2} \propto \mathcal{O}(\|F\|_2),$$

$$\left\| H_2 (s_1, s_2) - \tilde{H}_2 (s_1, s_2) \right\|_{H_2} \propto \mathcal{O}(\|F\|_2),$$

$$\vdots$$

$$\left\| H_M (s_1, \ldots, s_M) - \tilde{H}_M (s_1, \ldots, s_M) \right\|_{H_2} \propto \mathcal{O}(\|F\|_2),$$

where $H_k (s_1, \ldots, s_k)$ for $k = 1, \ldots, M$ is given by (4) and (9b), and $\tilde{H}_k (s_1, \ldots, s_k)$ for $k = 1, \ldots, M$ is given by (10b). This way was not possible in BIRKA because there $M \rightarrow \infty$ (see (2)-(4)).

3 Satisfying the Second Condition of Backward Stability

We use induction to prove (12). To prove this condition, we first abstract out the term containing the perturbation $F$ from the normed difference be-

---

Recall, in BIRKA we work with $\zeta$ rather than $\zeta^M$. 

between the transfer function of the original system \((H_M (s_1, \ldots, s_M))\) and the transfer function of the perturbed system \((\tilde{H}_M (s_1, \ldots, s_M))\) (in Lemma 6). Next, we use induction to prove that this abstracted term is \(O(\|F\|_2)\) (in Lemma 7). Finally, we use the result of these two lemmas to prove (12) (in Theorem 8). Note, that in all our subsequent derivations, we assume that all inverses used exist. This is an acceptable assumption because the inverse of matrices arising here are of the form as in [27] and [30] (the papers that discuss stability of IRKA and BIRKA, respectively).

**Lemma 6.** If

\[
H_M (s_1, \ldots, s_M) = c (s_M I_n - A)^{-1} N (s_{M-1} I_n - A)^{-1} \ldots N (s_1 I_n - A)^{-1} b
\]

and

\[
\tilde{H}_M (s_1, \ldots, s_M) = c (s_M I_n - (A + F))^{-1} N (s_{M-1} I_n - (A + F))^{-1} \ldots N (s_1 I_n - (A + F))^{-1} b,
\]

then the \(H_2\)-norm of their difference

\[
\left\| H_M (s_1, \ldots, s_M) - \tilde{H}_M (s_1, \ldots, s_M) \right\|_{H_2}^2 \leq \left\| c \mathcal{K}^{-1} (s_M) \right\|_{H_2}^2
\]

\[
\left\| \mathcal{K}^{-1} (s_{M-1}) \right\|_{H_2}^2 \ldots \left\| \mathcal{K}^{-1} (s_1) \right\|_{H_2}^2 \left\| U (s_1, \ldots, s_M) \right\|_{H_\infty}^2 \left\| \mathcal{K}^{-1} (s_1) b \right\|_{H_\infty}^2
\]

where, \(\mathcal{K} (s_i) = (s_i I_n - A)\) for \(i = 1, \ldots, M\) and

\[
U (s_1, \ldots, s_M) = \mathcal{K} (s_1) \ldots \mathcal{K} (s_{M-1}) \left( N \mathcal{K}^{-1} (s_{M-1}) \ldots N \mathcal{K}^{-1} (s_2) N
\]

\[
- (I_n - F \mathcal{K}^{-1} (s_M))^{-1} N \mathcal{K}^{-1} (s_{M-1}) (I_n - F \mathcal{K}^{-1} (s_{M-1}))^{-1} \ldots (13)
\]

\[
N \mathcal{K}^{-1} (s_2) (I_n - F \mathcal{K}^{-1} (s_2))^{-1} N (I_n - \mathcal{K}^{-1} (s_1) F)^{-1}
\]

**Proof.** Using the definition of \(H_2\)-norm (1), we get

\[
\left\| H_M (s_1, \ldots, s_M) - \tilde{H}_M (s_1, \ldots, s_M) \right\|_{H_2}^2
\]

\[
= \left( \frac{1}{2\pi} \right)^M \lim_{m \to \infty} \int_{-m}^{m} \ldots \int_{-m}^{m} \left\| c \mathcal{K}^{-1} (i \omega_M) N \mathcal{K}^{-1} (i \omega_{M-1}) \ldots N \mathcal{K}^{-1} (i \omega_1) b
\]

\[
- c (\mathcal{K} (i \omega_M) - F)^{-1} N (\mathcal{K} (i \omega_{M-1}) - F)^{-1} \ldots N (\mathcal{K} (i \omega_2) - F)^{-1}
\]

\[
N (\mathcal{K} (i \omega_1) - F)^{-1} b \right\|_F^2 d\omega_1 \ldots d\omega_M
\]
From the mean value theorem of integration \[34\] we know
\[
\int_{-m}^{m} \cdots \int_{-m}^{m} \left| e^{\mathcal{K}^{-1}(i\omega_M)} \left( N\mathcal{K}^{-1}(i\omega_{M-1}) \cdots N\mathcal{K}^{-1}(i\omega_2) N - (I_n - F\mathcal{K}^{-1}(i\omega_M))^{-1} N\mathcal{K}^{-1}(i\omega_{M-1}) (I_n - F\mathcal{K}^{-1}(i\omega_M))^{-1} \right) \right| \mathcal{K}^{-1}(i\omega_1) b \right|^2_F \, d\omega_1 \cdots d\omega_M
\]
\[
= \left( \frac{1}{2\pi} \right)^{M} \lim_{m \to \infty} \int_{-m}^{m} \cdots \int_{-m}^{m} \left| e^{\mathcal{K}^{-1}(i\omega_M)} \left( N\mathcal{K}^{-1}(i\omega_{M-1}) \cdots N\mathcal{K}^{-1}(i\omega_2) N - (I_n - F\mathcal{K}^{-1}(i\omega_M))^{-1} N\mathcal{K}^{-1}(i\omega_{M-1}) (I_n - F\mathcal{K}^{-1}(i\omega_M))^{-1} \right) \right| \mathcal{K}^{-1}(i\omega_1) b \right|^2_F \, d\omega_1 \cdots d\omega_M.
\]

Using \( U(s_1, \ldots, s_M) \) given by \(13\), \( \|X Y Z\|_F \leq \|X\|_F \|Y Z\|_F \), \( \|Y Z\|_F \leq \|Y\|_F \|Z\|_2 \), and comparison integral inequality\(^5 \) \[34\] for any matrices \( X, Y \) and \( Z \), in the above equation, we have
\[
\left\| H_M(s_1, \ldots, s_M) - \tilde{H}_M(s_1, \ldots, s_M) \right\|_{H_2}^2 \\
\leq \left( \frac{1}{2\pi} \right)^{M} \lim_{m \to \infty} \int_{-m}^{m} \cdots \int_{-m}^{m} \left| e^{\mathcal{K}^{-1}(i\omega_M)} \right|^2_F \|\mathcal{K}^{-1}(i\omega_{M-1})\|^2_F \cdots \|\mathcal{K}^{-1}(i\omega_1)\|^2_F \|U(i\omega_1, \ldots, i\omega_M)\|^2_2 \|\mathcal{K}^{-1}(i\omega_1) b\|^2_2 \, d\omega_1 \cdots d\omega_M.
\]

From the mean value theorem of integration \[34\] we know
\[
\int_{-m}^{m} \int_{-m}^{m} f(i\omega_2) g(i\omega_1, i\omega_2) h(i\omega_1) \, d\omega_1 d\omega_2 = \int_{-m}^{m} f(i\omega_2) \left( \int_{-m}^{m} g(i\omega_1, i\omega_2) h(i\omega_1) \, d\omega_1 \right) \, d\omega_2
\]
\(^5\)This inequality says if \( f(x) \) and \( g(x) \) are integrable over \([a, b]\) and \( f(x) \leq g(x) \), then \( \int_{a}^{b} f(x) \, dx \leq \int_{a}^{b} g(x) \, dx \). Note that although we have improper integrals here, this inequality still holds because of the earlier assumption that such integrals give a finite value.
\[
\begin{align*}
&\leq \int_{-m}^{m} f(i\omega_2) \left( \max_{c \in \mathbb{R}} (g(ic, i\omega_2)) \right) \int_{-m}^{m} h(i\omega_1) \, d\omega_1 \\
&\leq \left( \int_{-m}^{m} f(i\omega_2) \max_{c \in \mathbb{R}} (g(ic, i\omega_2)) \, d\omega_2 \right) \int_{-m}^{m} h(i\omega_1) \, d\omega_1 \\
&\leq \max_{c, d \in \mathbb{R}} (g(ic, id)) \int_{-m}^{m} f(i\omega_2) \, d\omega_2 \int_{-m}^{m} h(i\omega_1) \, d\omega_1.
\end{align*}
\]

Using this property in (14) we get

\[
\|H_M(s_1, \ldots, s_M) - \tilde{H}_M(s_1, \ldots, s_M)\|_{H_2}^2 \\
\leq \left( \frac{1}{2\pi} \right)^M \lim_{m \to \infty} \int_{-m}^{m} \cdots \int_{-m}^{m} \left\| c \mathcal{K}^{-1}(i\omega_M) \right\|_F^2 \left\| \mathcal{K}^{-1}(i\omega_{M-1}) \right\|_F^2 \cdots \left\| \mathcal{K}^{-1}(i\omega_1) \right\|_F^2 \\
\max_{\omega_1 \in \mathbb{R}} \left\| \mathcal{K}^{-1}(i\omega_1) b \right\|_2^2 \\
\leq \left\| c \mathcal{K}^{-1}(s_M) \right\|_{H_2}^2 \left\| \mathcal{K}^{-1}(s_{M-1}) \right\|_{H_2}^2 \cdots \left\| \mathcal{K}^{-1}(s_1) \right\|_{H_2}^2 \\
\left\| U(s_1, \ldots, s_M) \right\|_{H_\infty}^2 \left\| \mathcal{K}^{-1}(s_1) b \right\|_{H_\infty}^2.
\]

\[ \square \]

**Lemma 7.** If \( U_M = U(s_1, \ldots, s_M) \) defined in (13), \( \|F\|_2 < 1 \), and \( \| \mathcal{K}^{-1}(s) \|_{H_\infty} < 1 \), where \( \mathcal{K}(s) = (sI_n - A) \). Then we have

\[ \|U_M\|_{H_\infty} \propto O(\|F\|_2). \]

**Proof.** We prove this by mathematical induction.

**Base Case**

(i) **First subsystem**

This is the linear system case, which has been already proved in [27] (see Theorem 4.3 of [27]).

(ii) **Second subsystem**

Substituting \( M = 2 \) in (13), we get

\[ U_2 = \mathcal{K}(s_1) \left( N - (I_n - F \mathcal{K}^{-1}(s_2))^{-1} N (I_n - \mathcal{K}^{-1}(s_1) F)^{-1} \right). \]

---

As mentioned in Footnote §, the improper integral does not affect application of this mean value theorem because all such integrals are assumed to give a finite value.
Here, for the first inequality we have $-h$; hence, the applicable matrix norm is 2. From [35, page 527], we know that

$$\|I - H_n\| = \max_{\omega \in \mathbb{R}} \left( \|K^{-1}(\omega)\| \right)$$

Taking $H_\infty$ norm on both sides, we get

$$\|U_2\|_{H_\infty} = \left\| \mathcal{K}(s_1) \left( - N - (I_n + F\mathcal{K}^{-1}(s_2)) \right) \right\|_{H_\infty}$$

$$= \max_{\omega_1, \omega_2 \in \mathbb{R}} \left( \|\mathcal{K}(i\omega_1)\| \right) \left( - N\mathcal{K}^{-1}(i\omega_1) - F\mathcal{K}^{-1}(i\omega_2) \right) \left( I_n - \mathcal{K}^{-1}(i\omega_1) \right) \left( I_n - \mathcal{K}^{-1}(i\omega_2) \right) \left( I_n - \mathcal{K}^{-1}(i\omega_1) \right) \left( I_n - \mathcal{K}^{-1}(i\omega_2) \right)$$

Using $\|XY\| \leq \|X\| \|Y\|$ and $\|X + Y\| \leq \|X\| + \|Y\|$, for any two

\[\|\|F\mathcal{K}^{-1}(s_2)\|_{H_\infty} < 1 \text{ and } \|\mathcal{K}^{-1}(s_1)F\|_{H_\infty} < 1, \text{ then by the Neumann series, we get}\]

$$U_2 = \mathcal{K}(s_1) \left( N - \left( I_n + F\mathcal{K}^{-1}(s_2) + (F\mathcal{K}^{-1}(s_2))^2 + \cdots \right) \right)$$

$$= \mathcal{K}(s_1) \left( N - N - N\mathcal{K}^{-1}(s_1)F \left( I_n + \mathcal{K}^{-1}(s_1)F \right) + \cdots \right) -$$

$$\left( I_n + F\mathcal{K}^{-1}(s_2) \right) \left( I_n + F\mathcal{K}^{-1}(s_2) + \cdots \right) N$$

$$= \mathcal{K}(s_1) \left( - N\mathcal{K}^{-1}(s_1)F \left( I_n - \mathcal{K}^{-1}(s_1)F \right)^{-1} -$$

$$F\mathcal{K}^{-1}(s_2) \left( I_n - F\mathcal{K}^{-1}(s_2) \right) \left( I_n - \mathcal{K}^{-1}(s_1)F \right)^{-1} \right)$$

$$= \mathcal{K}(s_1) \left( - N\mathcal{K}^{-1}(s_1)F - F\mathcal{K}^{-1}(s_2) \left( I_n - F\mathcal{K}^{-1}(s_2) \right) \left( I_n - \mathcal{K}^{-1}(s_1)F \right)^{-1} \right)$$

Taking $H_\infty$ norm on both sides, we get

$$\|U_2\|_{H_\infty} = \left\| \mathcal{K}(s_1) \left( - N\mathcal{K}^{-1}(s_1)F - F\mathcal{K}^{-1}(s_2) \left( I_n - F\mathcal{K}^{-1}(s_2) \right) \right)^{-1} \right\|_{H_\infty}$$

$$\left( I_n - \mathcal{K}^{-1}(s_1)F \right)^{-1}$$

\[\text{From [35, page 527], we know that } (I - A)^{-1} = \sum_{k=0}^{\infty} A^k \text{ when } \|A\| < 1 \text{ for any matrix norm. Here, for the first inequality we have } \|F\mathcal{K}^{-1}(s_2)\|_{H_\infty} < 1 \text{ or } \max_{\omega_2 \in \mathbb{R}} \|F\mathcal{K}^{-1}(i\omega_2)\| < 1, \text{ and hence, the applicable matrix norm is }\]
matrices $X$ and $Y$, in the above equation, we get

$$
\|U_2\|_{H_\infty} \leq \max_{\omega_1, \omega_2 \in \mathbb{R}} \left( \|\mathcal{K}(i\omega_1)\|_2 \left( \|N\|_2 \|\mathcal{K}^{-1}(i\omega_1)\|_2 \|F\|_2 + \|F\|_2 \|\mathcal{K}^{-1}(i\omega_2)\|_2 \left( \|I_n - F\mathcal{K}^{-1}(i\omega_2)\|^{-1}_2 \right) \right) \right) \leq \|\mathcal{K}(s_1)\|_{H_\infty} \|N\|_2 \|F\|_2 \left( \|\mathcal{K}^{-1}(s_1)\|_{H_\infty} + \max_{\omega_2 \in \mathbb{R}} \left( \|I_n - F\mathcal{K}^{-1}(i\omega_2)\|^{-1}_2 \right) \right) \right)
$$

(15)

Technically by definition of the $H_\infty$-norm and how $\mathcal{K}(s)$ is defined in our hypotheses, $\|\mathcal{K}(s_1)\|_{H_\infty} = \|\mathcal{K}(s_2)\|_{H_\infty} = \|\mathcal{K}(s)\|_{H_\infty}$, however, for sake of exposition, we keep them separate. Similarly for the $H_\infty$-norm of inverses of $\mathcal{K}(s_1)$ and $\mathcal{K}(s_2)$.

To abstract $\|F\|_2$ out from the above inequality, let us look at $\max_{\omega_2 \in \mathbb{R}} \left( \|I_n - F\mathcal{K}^{-1}(i\omega_2)\|^{-1}_2 \right)$ separately. Recall, while applying Neumann series we assumed that $\|F\mathcal{K}^{-1}(s_2)\|_{H_\infty} < 1$ or $\max_{\omega_2 \in \mathbb{R}} \|F\mathcal{K}^{-1}(i\omega_2)\|_2 < 1$. Since the maximum of such a norm is less than one, we have for all $\omega_2 \in \mathbb{R}$, $\|F\mathcal{K}^{-1}(i\omega_2)\|_2 < 1$. Using this along with Lemma 2.3.3 from [36]** in the above expression, we get

$$
\max_{\omega_2 \in \mathbb{R}} \left( \|I_n - F\mathcal{K}^{-1}(i\omega_2)\|^{-1}_2 \right) \leq \max_{\omega_2 \in \mathbb{R}} \frac{1}{1 - \|F\mathcal{K}^{-1}(i\omega_2)\|_2} \leq \frac{1}{1 - \max_{\omega_2 \in \mathbb{R}} \|F\mathcal{K}^{-1}(i\omega_2)\|_2} \leq \frac{1}{1 - \|F\mathcal{K}^{-1}(s_2)\|_{H_\infty}}.
$$

(16)**

**If $F \in \mathbb{R}^{n \times n}$ and $\|F\|_p < 1$, then $I - F$ is nonsingular and $(I - F)^{-1} = \sum_{k=0}^{\infty} F^k$ with $\|I - F^{-1}\|_p \leq \frac{1}{1 - \|F\|_p}$.**
If we assume $\|F\|_2 < 1$ and $\|\mathcal{X}^{-1}(s)\|_{H_{\infty}} < 1$ (as in our hypotheses), then using earlier used matrix norm properties, we get

\[
\|F\mathcal{X}^{-1}(s_2)\|_{H_{\infty}} = \max_{\omega_2 \in \mathbb{R}} \|F\mathcal{X}^{-1}(i\omega_2)\|_2 \leq \|F\|_2 \max_{\omega_2 \in \mathbb{R}} \|\mathcal{X}^{-1}(i\omega_2)\|_2 \\
\leq \|F\|_2 \|\mathcal{X}^{-1}(s_2)\|_{H_{\infty}} \leq 1,
\]

as assumed for applying Neumann series earlier as well as Lemma 2.3.3 from [36] above. Thus, no extra assumptions beyond those in hypotheses are needed. Further, we also get

\[
1 - \|F\|_2 \|\mathcal{X}^{-1}(s_2)\|_{H_{\infty}} \leq 1 - \|F\|_2 \|\mathcal{X}^{-1}(s_2)\|_{H_{\infty}} \quad \text{or} \quad \frac{1}{1 - \|F\|_2 \|\mathcal{X}^{-1}(s_2)\|_{H_{\infty}}} \leq \frac{1}{1 - \|F\|_2 \|\mathcal{X}^{-1}(s_2)\|_{H_{\infty}}},
\]

(17)

Similarly, we can bound the last term of (15) as follows:

\[
\max_{\omega_1 \in \mathbb{R}} \|\left(I_n - \mathcal{X}^{-1}(i\omega_1)F\right)^{-1}\|_2 \leq \frac{1}{1 - \|\mathcal{X}^{-1}(s_1)\|_{H_{\infty}}} \quad \text{and} \quad \frac{1}{1 - \|\mathcal{X}^{-1}(s_1)\|_{H_{\infty}}} \leq \frac{1}{1 - \|\mathcal{X}^{-1}(s_1)\|_{H_{\infty}}} \|F\|_2.
\]

(18)

(19)

Substituting (16)-(17) and (18)-(19) in (15), we get

\[
\|U_2\|_{H_{\infty}} \leq \|\mathcal{X}(s_1)\|_{H_{\infty}} \|N\|_2 \|F\|_2 \left[\|\mathcal{X}^{-1}(s_1)\|_{H_{\infty}} + \frac{\|\mathcal{X}^{-1}(s_2)\|_{H_{\infty}}}{1 - \|F\|_2 \|\mathcal{X}^{-1}(s_2)\|_{H_{\infty}}} \frac{1}{1 - \|\mathcal{X}^{-1}(s_1)\|_{H_{\infty}}} \|F\|_2\right].
\]

From the above inequality it is clear that if $\|F\|_2 \|\mathcal{X}^{-1}(s_2)\|_{H_{\infty}} < 1$ and $\|\mathcal{X}^{-1}(s_1)\|_{H_{\infty}} \|F\|_2 < 1$, which are true from our hypotheses, then

\[
\|U_2\|_{H_{\infty}} = \mathcal{O}(\|F\|_2).
\]

(iii) Third subsystem

Using $M = 3$ in (13), we get

\[
U_3 = \mathcal{X}(s_1)\mathcal{X}(s_2) \left( N\mathcal{X}^{-1}(s_2)N - (I_n - F\mathcal{X}^{-1}(s_3))^{-1} N\mathcal{X}^{-1}(s_2) \right) (I_n - F\mathcal{X}^{-1}(s_2))^{-1} N(I_n - \mathcal{X}^{-1}(s_1)F)^{-1}.
\]
Following the same steps as in the case of the second subsystem we get

\[
\|U_3\|_{H_\infty} \leq \|\mathcal{K}(s_1)\|_{H_\infty} \|\mathcal{K}(s_2)\|_{H_\infty} \|N\|_2^2 \|F\|_2 \|\mathcal{K}^{-1}(s_2)\|_{H_\infty} \\
\left[ \|\mathcal{K}^{-1}(s_1)\|_{H_\infty} + \frac{\|\mathcal{K}^{-1}(s_2)\|_{H_\infty}}{1 - \|F\|_2 \|\mathcal{K}^{-1}(s_2)\|_{H_\infty}} \right] \\
\left( \frac{1}{1 - \|\mathcal{K}^{-1}(s_1)\|_{H_\infty} \|F\|_2} \right)
\]

Again, from the above inequality it is clear that if \(\|F\|_2 \|\mathcal{K}^{-1}(s_2)\|_{H_\infty}\), \(\|F\|_2 \|\mathcal{K}^{-1}(s_3)\|_{H_\infty}\) and \(\|\mathcal{K}^{-1}(s_1)\|_{H_\infty} \|F\|_2\) are all less than 1, which are true from our hypotheses, then

\[
\|U_3\|_{H_\infty} = O(\|F\|_2).
\]

**Induction Hypothesis**

From (13), we know

\[
U_M = \mathcal{K}(s_1) \ldots \mathcal{K}(s_{M-1}) \left( N\mathcal{K}^{-1}(s_{M-1}) \ldots N\mathcal{K}^{-1}(s_2) \right) N \\
- (I_n - F\mathcal{K}^{-1}(s_M))^{-1} N\mathcal{K}^{-1}(s_{M-1}) (I_n - F\mathcal{K}^{-1}(s_{M-1}))^{-1} \ldots (20) \\
N\mathcal{K}^{-1}(s_2) (I_n - F\mathcal{K}^{-1}(s_2))^{-1} N (I_n - \mathcal{K}^{-1}(s_1) F^{-1}).
\]

Let

\[
\|U_M\|_{H_\infty} \leq \|\mathcal{K}(s_1)\|_{H_\infty} \ldots \|\mathcal{K}(s_{M-1})\|_{H_\infty} \|N\|_2^{M-1} \|F\|_2 \|\mathcal{K}^{-1}(s_{M-1})\|_{H_\infty} \\
\|\mathcal{K}^{-1}(s_2)\|_{H_\infty} \left[ \|\mathcal{K}^{-1}(s_1)\|_{H_\infty} + \frac{\|\mathcal{K}^{-1}(s_2)\|_{H_\infty}}{1 - \|F\|_2 \|\mathcal{K}^{-1}(s_2)\|_{H_\infty}} \right] \\
\vdots \\
\left[ \|\mathcal{K}^{-1}(s_M)\|_{H_\infty} \left( \frac{1}{1 - \|\mathcal{K}^{-1}(s_1)\|_{H_\infty} \|F\|_2} \right) \right] \\
(1 - \|F\|_2 \|\mathcal{K}^{-1}(s_M)\|_{H_\infty}) \ldots (1 - \|F\|_2 \|\mathcal{K}^{-1}(s_2)\|_{H_\infty})
\]

\[
\|U_M\|_{H_\infty} = O(\|F\|_2).
\]
Induction Step
Again, from (13), we know

\[ U_{M+1} = \mathcal{K}(s_1) \cdots \mathcal{K}(s_M) \left( N \mathbf{K}^{-1}(s_M) \cdots N \mathbf{K}^{-1}(s_2) N \right. \]
\[ \left. - \left( I_n - F \mathbf{K}^{-1}(s_{M+1}) \right)^{-1} N \mathbf{K}^{-1}(s_M) \left( I_n - F \mathbf{K}^{-1}(s_M) \right)^{-1} \cdots 
\right. \]
\[ \left. N \mathbf{K}^{-1}(s_2) \left( I_n - F \mathbf{K}^{-1}(s_2) \right)^{-1} N \left( I_n - \mathbf{K}^{-1}(s_1) F \right)^{-1} \right) \].

We first write \( U_{M+1} \) in terms of \( U_M \). Thus, in the above equation using Neumann series, we get

\[ U_{M+1} = \mathcal{K}(s_1) \cdots \mathcal{K}(s_M) \left( N \mathbf{K}^{-1}(s_M) \cdots N \mathbf{K}^{-1}(s_2) N \right. \]
\[ \left. - \left( I_n + F \mathbf{K}^{-1}(s_{M+1}) + (F \mathbf{K}^{-1}(s_{M+1}))^2 + \cdots \right) \right. \]
\[ \left. N \mathbf{K}^{-1}(s_M) \left( I_n - F \mathbf{K}^{-1}(s_M) \right)^{-1} \cdots N \mathbf{K}^{-1}(s_2) \left( I_n - F \mathbf{K}^{-1}(s_2) \right)^{-1} \right. \]
\[ \left. N \left( I_n - \mathbf{K}^{-1}(s_1) F \right)^{-1} \right) \]
\[ = \mathcal{K}(s_1) \cdots \mathcal{K}(s_M) \left( N \mathbf{K}^{-1}(s_M) \cdots N \mathbf{K}^{-1}(s_2) N \right. \]
\[ \left. - N \mathbf{K}^{-1}(s_M) \left( I_n - F \mathbf{K}^{-1}(s_M) \right)^{-1} \cdots N \mathbf{K}^{-1}(s_2) \left( I_n - F \mathbf{K}^{-1}(s_2) \right)^{-1} \right. \]
\[ \left. N \left( I_n - \mathbf{K}^{-1}(s_1) F \right)^{-1} \right. \]
\[ \left. - F \mathbf{K}^{-1}(s_{M+1}) \left( I_n - F \mathbf{K}^{-1}(s_{M+1}) \right)^{-1} \right. \]
\[ \left. N \mathbf{K}^{-1}(s_M) \left( I_n - F \mathbf{K}^{-1}(s_M) \right)^{-1} \cdots N \mathbf{K}^{-1}(s_2) \left( I_n - F \mathbf{K}^{-1}(s_2) \right)^{-1} \right. \]
\[ \left. N \left( I_n - \mathbf{K}^{-1}(s_1) F \right)^{-1} \right) \].

In the above equation, taking \( N \mathbf{K}^{-1}(s_M) \) common from the first two terms
of the bigger bracket, we have

\[
= \mathcal{K}(s_1) \cdots \mathcal{K}(s_M) \left( \mathcal{N} \mathcal{K}^{-1}(s_M) \left( \mathcal{N} \mathcal{K}^{-1}(s_{M-1}) \cdots \mathcal{N} \mathcal{K}^{-1}(s_2) \mathcal{N} \right) - (I_n - F \mathcal{K}^{-1}(s_M))^{-1} \mathcal{N} \mathcal{K}^{-1}(s_2) (I_n - F \mathcal{K}^{-1}(s_M))^{-1} \cdots \right.
\]

\[
N \mathcal{K}^{-1}(s_2) (I_n - F \mathcal{K}^{-1}(s_2))^{-1} N (I_n - \mathcal{K}^{-1}(s_1) F)^{-1} \right)
\]

Now we look at expression of \( U_M \). Multiplying \( \mathcal{K}^{-1}(s_{M-1}) \cdots \mathcal{K}^{-1}(s_1) \) on both the sides of (20) from left, we get

\[
\mathcal{K}^{-1}(s_{M-1}) \cdots \mathcal{K}^{-1}(s_1) U_M
\]

\[
= \left( \mathcal{N} \mathcal{K}^{-1}(s_{M-1}) \cdots \mathcal{N} \mathcal{K}^{-1}(s_2) \mathcal{N} \right) - (I_n - F \mathcal{K}^{-1}(s_M))^{-1} \mathcal{N} \mathcal{K}^{-1}(s_2) (I_n - F \mathcal{K}^{-1}(s_M))^{-1} \cdots \right.
\]

\[
N \mathcal{K}^{-1}(s_2) (I_n - F \mathcal{K}^{-1}(s_2))^{-1} N (I_n - \mathcal{K}^{-1}(s_1) F)^{-1} \right).
\]

\[
\text{Substituting (22) in (21), we get}
\]

\[
U_{M+1} = \mathcal{K}(s_1) \cdots \mathcal{K}(s_M) \left( \mathcal{N} \mathcal{K}^{-1}(s_M) \left( \mathcal{K}^{-1}(s_{M-1}) \cdots \mathcal{K}^{-1}(s_1) U_M \right) - F \mathcal{K}^{-1}(s_{M+1}) (I_n - F \mathcal{K}^{-1}(s_{M+1}))^{-1}
\]

\[
N \mathcal{K}^{-1}(s_M) (I_n - F \mathcal{K}^{-1}(s_M))^{-1} \cdots \mathcal{N} \mathcal{K}^{-1}(s_2) (I_n - F \mathcal{K}^{-1}(s_2))^{-1}
\]

\[
N (I_n - \mathcal{K}^{-1}(s_1) F)^{-1} \right).
\]
Taking $H_\infty$-norm on both sides, we get
\[
\|U_{M+1}\|_{H_\infty} = \\
\|\mathcal{K}(s_1) \ldots \mathcal{K}(s_M) \left( N\mathcal{K}^{-1}(s_M) \mathcal{K}^{-1}(s_{M-1}) \ldots \mathcal{K}^{-1}(s_1) U_M \\
- F\mathcal{K}^{-1}(s_{M+1}) (I_n - F\mathcal{K}^{-1}(s_{M+1}))^{-1} \\
N\mathcal{K}^{-1}(s_M) (I_n - F\mathcal{K}^{-1}(s_M))^{-1} \ldots N\mathcal{K}^{-1}(s_2) (I_n - F\mathcal{K}^{-1}(s_2))^{-1} \\
N (I_n - \mathcal{K}^{-1}(s_1) F)^{-1} \right)\|_{H_\infty}.
\]

As earlier, using the norm inequality properties in the above equation, we get
\[
\|U_{M+1}\|_{H_\infty} \leq \\
\max_{\omega_1, \ldots, \omega_{M+1} \in \mathbb{R}} \left[ \|\mathcal{K}(i\omega_1)\|_2 \ldots \|\mathcal{K}(i\omega_M)\|_2 \left( \|N\|_2 \|\mathcal{K}^{-1}(i\omega_M)\|_2 \ldots \\
\|\mathcal{K}^{-1}(i\omega_1)\|_2 \|U(i\omega_1, \ldots, i\omega_M)\|_2 + \|F\|_2 \|\mathcal{K}^{-1}(i\omega_{M+1})\|_2 \\
\|N\|_2 \|\mathcal{K}^{-1}(i\omega_2)\|_2 \|N\|_2 \|\mathcal{K}^{-1}(i\omega_{M+1})\|_2 \|\mathcal{K}^{-1}(i\omega_2)\|_2 \\
\ldots \|N\|_2 \|\mathcal{K}^{-1}(i\omega_2)\|_2 \|N\|_2 \|\mathcal{K}^{-1}(i\omega_{M+1})\|_2 \|\mathcal{K}^{-1}(i\omega_2)\|_2 \\
\ldots \|\mathcal{K}^{-1}(i\omega_2)\|_2 \|\mathcal{K}^{-1}(i\omega_{M+1})\|_2 \right) \ldots \\
\|\mathcal{K}^{-1}(i\omega_{M+1})\|_2 \|\mathcal{K}^{-1}(i\omega_2)\|_2 \right].
\]

Similar to (16) and (17), here also, using Lemma 2.3.3 from [36] we get
\[
\|U_{M+1}\|_{H_\infty} \leq \|\mathcal{K}(s_1)\|_{H_\infty} \ldots \|\mathcal{K}(s_M)\|_{H_\infty} \|N\|_2 \|\mathcal{K}^{-1}(s_M)\|_{H_\infty} \ldots \\
\|\mathcal{K}^{-1}(s_2)\|_{H_\infty} \left[ \|\mathcal{K}^{-1}(s_1)\|_2 \|U_M\|_{H_\infty} + \\
\|N\|_2 \|\mathcal{K}^{-1}(s_1)\|_{H_\infty} \|U_M\|_{H_\infty} + \\
\frac{\|\mathcal{K}^{-1}(s_1)\|_2 \|\mathcal{K}^{-1}(s_{M+1})\|_{H_\infty}}{(1 - \|F\|_2 \|\mathcal{K}^{-1}(s_{M+1})\|_{H_\infty}) \ldots (1 - \|F\|_2 \|\mathcal{K}^{-1}(s_2)\|_{H_\infty})} \right].
\]

From induction hypothesis we know $\|U_M\|_{H_\infty} \propto \Theta (\|F\|_2)$. Using this we get
\[
\|U_{M+1}\|_{H_\infty} \propto \Theta (\|F\|_2).
\]
Theorem 8. If
\[ H_M(s_1, \ldots, s_M) = c(s_MI_n - A)^{-1}N(s_{M-1}I_n - A)^{-1}\ldots N(s_1I_n - A)^{-1}b, \]
\[ \tilde{H}_M(s_1, \ldots, s_M) = c(s_MI_n - (A + F))^{-1}N(s_{M-1}I_n - (A + F))^{-1}\ldots N(s_1I_n - (A + F))^{-1}b, \]
\[ \|F\|_2 < 1, \text{ and } \|K^{-1}(s)\|_{H_\infty} < 1, \text{ where } K(s) = (sI_n - A), \]
then
\[ \left\| H_M(s_1, \ldots, s_M) - \tilde{H}_M(s_1, \ldots, s_M) \right\|_{H_2}^2 = \mathcal{O} \left( \|F\|_2^2 \right) \]
or
\[ \|
\]
\[ \text{TBIRKA satisfies the second condition of backward stability with respect to inexact linear solves.} \]

4 Conclusions & Future Work

In this paper, we apply iterative linear solvers during model order reduction (MOR) of bilinear dynamical systems. Since such solvers are inexact, the stability of the underlying MOR algorithm, with respect to these approximation errors, is important. Here, we extend the earlier stability analysis done for BIRKA in [30], which is a standard algorithm for obtaining a reduced bilinear dynamical system, to TBIRKA, a cheaper variant of BIRKA.

Proving that an algorithm is stable, typically requires satisfying two conditions. In TBIRKA, fulfilling the first condition for stability leads to constraints on the iterative linear solver, which are similar to those obtained during BIRKA’s stability analysis. The second condition for a stable TBIRKA is satisfied using an approach different than the one used in BIRKA, and is more intuitive.

Our first future direction is to extend our analysis from SISO (Single Input Single Output) to MIMO (Multiple Input Multiple Output) systems. The stability analysis as done for BIRKA earlier and TBIRKA here, all give us sufficiency conditions for a stable underlying MOR algorithm. Hence, second, we plan to derive the necessary conditions for the same. In recent years, there have been a lot of efforts in performing data-driven MOR algorithm (specially using Leowner framework [19]). Similar linear systems and challenges associated with the scaling arise here as well. Our third future direction is to apply this stability analysis to such classes of algorithms as well.

References

[1] H. Kaper, H. Engler. Mathematics And Climate. SIAM, Philadelphia, PA, USA, 2013.
[2] J. D. Meiss. *Differential Dynamical Systems, Revised Edition*. SIAM, Philadelphia, PA, USA, 2017.

[3] J. T. Stuart. Taylor-Vortex flow: A dynamical system. *SIAM Review*, 28(3):315–342, 1986.

[4] P. D’Alessandro, A. Isidori, and A. Ruberti. Realization and structure theory of bilinear dynamical systems. *SIAM J. Control Opt.*, 12(3):517–535, 1974.

[5] R. J. Wilson. *Nonlinear System Theory: The Volterra/ Wiener Approach*. Johns Hopkins Series in Information Sciences and Systems, Johns Hopkins University Press, Baltimore, 1981.

[6] F. Caravetta. Global exact quadratization of continuous-time nonlinear control systems. *SIAM J. Control Opt.*, 53(1):235–261, 2015.

[7] A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. Advances in Design and Control, SIAM, Philadelphia, PA, USA, 2005.

[8] P. Benner, D. C. Sorensen, and V. Mehrmann. *Dimension Reduction of Large-Scale Systems*. Lecture Notes in Computational Science and Engineering, Springer, Berlin, Germany, 2005.

[9] E. J. Grimme. Krylov Projection Methods for Model Reduction. *Ph.D. thesis*, University of Illinois at Urbana-Champaign, Urbana, IL, USA, 1997.

[10] Z. Bai. Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems. *Appl. Num. Math.*, 43(1):9–44, 2002.

[11] K. Willcox and J. Peraire. Balanced model reduction via the proper orthogonal decomposition. *AIAA Journal*, 40(11):2325–2330, 2002.

[12] S. Gugercin, A. C. Antoulas, and C. Beattie. ℋ₂ model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appl.*, 30(2):609–638, 2008.

[13] A. Bunse-Gerstner, D. Kubalińska, G. Vossen, and D. Wilczek. ℋ₂-norm optimal model reduction for large scale discrete dynamical MIMO systems. *J. Comput. Appl. Math.*, 233(5):1202–1216, 2010.

[14] Z. Bai and D. Skoogh. A projection method for model reduction of bilinear dynamical systems. *Linear Algebra Appl.*, 415(2-3):406–425, 2006.
[15] P. Benner and T. Breiten. Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems. *SIAM J. Matrix Anal. Appl.*, 33(3):859–885, 2012.

[16] T. Bonin, H. Faßbender, A. Soppa, and M. Zaeh. A fully adaptive rational global Arnoldi method for the model-order reduction of second-order MIMO systems with proportional damping. *Math. Comp. Simulation.*, 122(C):1–19, 2016.

[17] M. I. Ahmad, U. Baur, and P. Benner. Implicit Volterra series interpolation for model reduction of bilinear systems. *J. Comput. Appl. Math.*, 316:15–28, 2017.

[18] P. Benner and T. Damm. Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems. *SIAM J. Control Optim.*, 49(2):686–711, 2011.

[19] A. C. Antoulas, I. V. Gosea, and A. C. Ionita. Model reduction of bilinear systems in the Loewner framework. *SIAM J. Control Optim.*, 38(5):B889–B916, 2016.

[20] K. L. Xu, Y. L. Jiang, and Z. X. Yang. $H_2$ optimal model order reduction by two-sided technique on Grassmann manifold via the cross-Gramian of bilinear systems. *Inter. J. Control*, 90(3):616–626, 2017.

[21] G. M. Flagg. Interpolation Methods for the Model Reduction of Bilinear Systems. *Ph.D. thesis*, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2012.

[22] G. M. Flagg and S. Gugercin. Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems. *SIAM J. Matrix Anal. Appl.*, 36(2):549–579, 2015.

[23] A. Greenbaum. *Iterative Methods for Solving Linear Systems*, Frontiers in Applied Mathematics. SIAM, Philadelphia, PA, USA, 1997.

[24] Y. Saad. *Iterative Methods for Sparse Linear Systems*. SIAM, Philadelphia, PA, USA, 2003.

[25] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. SIAM, Philadelphia, PA, USA, 1997.

[26] J. W. Demmel. *Applied Numerical Linear Algebra*. SIAM, Philadelphia, PA, USA, 1997.
[27] C. Beattie, S. Gugercin, and S. Wyatt. Inexact solves in interpolatory model reduction. *Linear Algebra Appl.*, 436(8):2916–2943, 2012.

[28] N. P. Singh and K. Ahuja. Stability analysis of inexact solves in moment matching based model reduction. *Preprint arXiv:1803.09283*, 2018.

[29] D. Lu, Y. Su, and Z. Bai. Stability analysis of the two-level orthogonal Arnoldi procedure. *SIAM J. Matrix Anal. Appl.*, 37(1):195–214, 2016.

[30] R. Choudhary and K. Ahuja. Stability analysis of bilinear iterative rational Krylov algorithm. *Linear Algebra Appl.*, 538:56–88, 2018.

[31] K. Ahuja. Recycling Krylov subspaces and preconditioners. *Ph.D. thesis*, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2011.

[32] K. Ahuja, E. D. Sturler, S. Gugercin, and E. R. Chang. Recycling BiCG with an application to model reduction. *SIAM J. Sci. Comput.*, 34(4):A1925–A1949, 2012.

[33] M. L. Parks, E. D. Sturler, G. Mackey, D. D. Johnson, and S. Maiti. Recycling Krylov subspaces for sequences of linear systems. *SIAM J. Sci. Comput.*, 28(5):1651-1674, 2006.

[34] A. Jeffrey. *Handbook of Mathematical Formulas and Integrals, Ed. 3*. Academic Press, 2003.

[35] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, PA, USA, 2000.

[36] G. H. Golub and C. F. Van Loan. *Matrix Computations, Vol. 3*. Johns Hopkins University Press, 2012.