Advances in Computational Mathematics (2019) 45:2241–2271
https://doi.org/10.1007/s10444-019-09695-9

A bilinear $\mathcal{H}_2$ model order reduction approach to linear parameter-varying systems

Peter Benner[1] · **Xingang Cao**[2] · **Wil Schilders**[2]

Received: 4 May 2018 / Accepted: 30 March 2019 /
Published online: 18 April 2019

Abstract
This paper focuses on the model reduction problem for a special class of linear parameter-varying systems. This kind of systems can be reformulated as bilinear dynamical systems. Based on the bilinear system theory, we give a definition of the $\mathcal{H}_2$ norm in the generalized frequency domain. Then, a model reduction method is proposed based on the gradient descent on the Grassmann manifold. The merit of the method is that by utilizing the gradient flow analysis, the algorithm is guaranteed to *converge*, and further speedup of the convergence rate can be achieved as well. Two numerical examples are tested to demonstrate the proposed method.

Keywords Model order reduction · Linear parameter-varying systems · Bilinear dynamical systems · Gradient descent · Grassmann manifold

Mathematics Subject Classification (2010) 14M15 · 34C20 · 65K10 · 93C10

1 Introduction

Linear parameter-varying (LPV) systems are usually used to represent linearizations of nonlinear systems along certain state trajectories. Those trajectories are

Communicated by: Anthony Nouy

✉ Xingang Cao
x.cao.1@tue.nl

[1] Computational Methods in Systems and Control Theory, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr. 1, 39106, Magdeburg, Germany

[2] Centre for Analysis, Scientific Computing and Applications, Eindhoven University of Technology, PO Box 513, 5600 MB, Eindhoven, The Netherlands

unknown in advance, can be *time-dependent* and can only be measured online. One simple example is the aero-elastic dynamics of an aircraft depending on the wind speed and the altitude, which are both unknown and cannot be modeled by any functions of time. A natural idea is to treat the parameters as extra input signals and bring the LPV system into the form of bilinear dynamical systems [3]. Once the system is in such a form, bilinear model order reduction techniques can be applied to reduce it.

The bilinearization approach is first introduced in [3]. In [28], a gradient flow method is applied to solve the bilinear model order reduction problem. However, the authors did not discuss the convergence of the method in terms of the projection matrix. The definition of system Gramians and the associated generalized Lyapunov equations are given there as well. Based on the generalized Lyapunov equations, interpolation-based $\mathcal{H}_2$ optimal methods such as bilinear iterative rational Krylov algorithm (BIRKA) and multipoint Volterra series interpolation methods are developed by [4, 15], respectively. To overcome the stability preservation problem, two approaches are developed [8], which are variations of the BIRKA method. In [9], Bruns considered the geometric nature of the projection matrix and developed a fast gradient flow algorithm (FGFA) and a sequential quadratic programming (SQP) method to find the (locally) optimal projection matrix, which is a generalization of the linear time-invariant (LTI) system cases proposed in [26]. For more works on model order reduction for LTI systems in this direction, i.e., model order reduction by Riemannian optimization methods, we refer to [22, 23].

The work presented in this paper focuses on the gradient flow method proposed by [9]. Since a first-order optimization method is applied, we are interested in speeding up the convergence rate of the optimization algorithm by bounding the line search step size. Following the work for LTI systems [27], two types of upper bounds of the line search step size are proposed. The first one is uniform in every iteration but quite conservative. The second varies over each iteration and speeds up the convergence rate significantly.

The paper is organized as follows. The remaining part of this section briefly reviews the bilinearization approach, which turns an LPV system into a bilinear dynamical system. Section 2 shows the basic system theoretic aspects on stability, system Gramians, and the associated generalized Lyapunov equations. The upper bounds on the Gramians and the definition of the $\mathcal{H}_2$ norm are given in this section as well. Then, the model order reduction problem is formulated as an optimization problem in Section 3. The gradient flow analysis is applied to find the upper bound of the line search step size, which can guarantee convergence of the optimization algorithm. Numerical examples are tested in Section 4. Section 5 concludes the paper.

1.1 Bilinearization of LPV systems

Consider the LPV system as follows:

$$\Sigma(\mathbf{p}) : \begin{cases} \dot{x}(t) = \widehat{A}(\mathbf{p}(t))x(t) + \widehat{B}\hat{u}(t), \\ y(t) = Cx(t), \end{cases} \tag{1.1}$$

with $x \in \mathbb{R}^n$, $\hat{u} \in \mathbb{R}^{\hat{m}}$, $y \in \mathbb{R}^q$, and $\mathbf{p}(t) \in \mathbb{R}^{n_p}$. In many applications, the matrix $\widehat{A}(\mathbf{p}(t))$ has affine dependence on $\mathbf{p}(t)$, i.e.,

$$\widehat{A}(\mathbf{p}(t)) = A + \sum_{i=1}^{n_p} A_i \, p_i(t).$$

If $\widehat{A}(\mathbf{p}(t))$ does not have affine parameter dependence, first-order Taylor expansion can be used to approximate $\widehat{A}(\mathbf{p}(t))$, which results in an affine parameter-dependent system. Better approximation of non-affine parametric matrices can be obtained by empirical matrix interpolation (see, e.g., [6, 19]). To bring the system (1.1) into a bilinear dynamical system, we first augment the input signal as follows:

$$u(t) = \begin{pmatrix} \hat{u}(t) & p_1(t) & \dots & p_{n_p}(t) \end{pmatrix}^\top.$$

Using the following notation,

$$N_j = \begin{cases} \mathbf{0}_{n\times n}, & j = 1, 2, \dots, \hat{m}, \\ A_i, & j = \hat{m} + i, \ i = 1, 2, \dots, n_p, \end{cases} \qquad B = \begin{pmatrix} \widehat{B} & \mathbf{0}_{n\times n_p} \end{pmatrix},$$

a bilinear dynamical system is obtained as follows:

$$\Sigma_{bl} : \begin{cases} \dot{x}(t) = Ax(t) + \sum_{j=1}^{m} N_j x(t) u_j(t) + Bu(t), \\ y(t) = Cx(t), \end{cases} \tag{1.2}$$

where $x \in \mathbb{R}^n$, $u \in \mathbb{R}^m$, and $y \in \mathbb{R}^q$. The constant m is defined as $m := \hat{m} + n_p$.

The model order reduction problem considered in this paper is to find an orthonormal matrix $V \in \mathbb{R}^{n\times r}$ with $r \ll n$ such that the reduced-order system

$$\widehat{\Sigma}_{bl} : \begin{cases} \dot{x}_r(t) = V^\top A V x_r(t) + \sum_{j=1}^{m} V^\top N_j V x_r(t) u_j(t) + V^\top B u(t), \\ y_r(t) = CV x_r(t), \end{cases} \tag{1.3}$$

minimizes the difference $\|\Sigma_{bl} - \widehat{\Sigma}_{bl}\|$ between the original system and the reduced-order one. In this paper, the $\mathcal{H}_2$ norm is considered to quantify the distance above, which will be defined in the following.

The method described in (1.3) is a Galerkin projection approach to model order reduction, while a Petrov–Galerkin-type method would be used when replacing $V^\top$ with some $W^\top$ such that $W^\top V = I_r$. Note that in case A is symmetric negative definite as it is often the case in applications, asymptotic stability of (1.3) is automatically preserved by the Galerkin projection [9, 10] in contrast to the bilinear iterative rational Krylov algorithm (BIRKA) [4] and other Petrov–Galerkin-like methods.

2 Bilinear system theory

Some basic system theoretic aspects are discussed in this section. We review the stability definition and stability criterion in the literature. Then, we focus on the system Gramians, which play a significant role in the model reduction procedure.

Based on the convolution kernels (Volterra kernels) and the generalized transfer functions of the bilinear dynamical system, the $\mathcal{H}_2$ norm is defined, which quantifies the difference between the full-order system and its reduced-order duplicate.

Recall the bilinear dynamical system given by (1.2). Assume that the system input satisfies $u(t) = 0, t < 0$, and the zero initial state $x(0) = 0$. The state of the bilinear system given by (1.2) has a Volterra series expansion [17]:

$$x(t) = \sum_{i=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \sum_{j_1, j_2, \ldots, j_i = 1}^{m} e^{A(t-\tau_1)} N_{j_1} e^{A(\tau_1-\tau_2)} N_{j_2} e^{A(\tau_2-\tau_3)} \cdots N_{j_i-1} e^{A(\tau_{i-1}-\tau_i)} b_{j_i} u_{j_1}(\tau_1) \cdots u_{j_i}(\tau_i) d\tau_1 \cdots d\tau_i, \tag{2.1}$$

where b_{j_i} is the j_ith column of B. It is proved in [28] that if the above Volterra series converges, it converges to the solution of the system in (1.2). For a bounded input, the Volterra series converges on any finite time interval. The system stability can then be interpreted in terms of the Volterra series.

Definition 1 (**BIBO stability**) The bilinear dynamical system Σ_{bl} given by (1.2) is *bounded-input bounded-output* (BIBO) *stable*, if for any bounded-input, the output is bounded on $[0, \infty)$.

Since the output is determined by the Volterra series in (2.1), the system is BIBO stable if for any bounded input, the Volterra series converges on $[0, \infty)$. Unlike for LTI systems, the stability of bilinear systems relates not only to the eigenvalues of A but also to the bound of N_j, $j = 1, 2, \ldots, m$. In general, if the spectrum $\Lambda(A) \subset \mathbb{C}^-$, then there exist two constants $\mu > 0$ and $c > 0$ such that

$$\|e^{At}\|_2 \leq ce^{-\mu t/2}, t \geq 0. \tag{2.2}$$

Assume that $\|u\| := \sqrt{\sum_{j=1}^{m} \|u_j\|^2} \leq M$ with $M > 0$. Let $\kappa = \sum_{j=1}^{m} \|N_j\|$. Then, we are ready to give the stability criterion of a bilinear dynamical system.

Theorem 1 ([28]) *The Volterra series* (2.1) *converges on the time interval* $[0, \infty)$ *for any bounded input if the following two conditions hold,*

i) *The matrix A is stable, i.e., the spectrum $\Lambda(A) \subset \mathbb{C}^-$.*
ii) *The matrices N_j are sufficiently bounded, i.e., $\kappa < \frac{\mu}{cM}$.*

Reachability and observability of bilinear dynamical systems can date back to the early work of [11]. The definitions are quite standard. In simple words, a bilinear system is reachable if for any state in the state space, there exists an $\mathcal{L}_2$ input function to steer the system from the zero state to the desired state in a finite time interval. A bilinear system is observable if any initial state can be uniquely determined from the input-output pair $(u(t), y(t))$ in a finite time interval which contains the initial time t_0. For a more comprehensive discussion of reachability and observability, we point

to [11, 13]. As an alternative, one can also reformulate the bilinear dynamical system as an affine nonlinear control system (ANCS) as follows:

$$\Sigma_{\text{ANCS}} : \begin{cases} \dot{x}(t) = Ax(t) + \sum_{j=1}^{m}(N_j x(t) + B_j)u_j(t), \\ y(t) = Cx(t), \end{cases}$$

and define the reachability and observability in the sense of *distribution algebra* [20].

Recalling the input-to-state Volterra series expansion in (2.1) and changing the variables [7], we can write the input-output mapping as

$$y(t) = \sum_{i=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} \sum_{j_1, j_2, \ldots, j_i = 1}^{m} Ce^{At_i} N_{j_1} e^{At_{i-1}} N_{j_2} e^{At_{i-2}} \cdots N_{j_{i-1}} e^{At_1} \cdot b_{j_i} u_{j_1}(t - t_i) \cdots u_{j_i}(t - t_i - \cdots - t_1) dt_1 \cdots dt_i. \tag{2.3}$$

Correspondingly, the convolution kernels are

$$h_i^{j_1, \ldots, j_i}(t_1, \ldots, t_i) = Ce^{At_i} N_{j_1} e^{At_{i-1}} N_{j_2} e^{At_{i-2}} \cdots N_{j_{i-1}} e^{At_1} b_{j_i}. \tag{2.4}$$

One application of the convolution kernel is to quantify the system energy [5]. The reachability Gramian represents the *input-to-state* energy, which can be written as

$$R = \sum_{i=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} P_i P_i^{\top} dt_1 \cdots dt_i := \sum_{i=1}^{\infty} R_i \tag{2.5}$$

with

$$\begin{aligned} P_1 &= e^{At_1} B, \\ P_i &= e^{At_i} \begin{bmatrix} N_1 & \ldots & N_m \end{bmatrix} (I_m \otimes P_{i-1}), \ i = 2, 3, \ldots, \end{aligned}$$

where $\otimes$ stands for the Kronecker product and I_m is the m-dimensional identity matrix. A close observation of (2.5) shows [28] that R_i, $i = 1, 2, \ldots$, satisfy the following equation:

$$AR_i + R_i A^{\top} + Z_{i-1} = 0, \ i = 1, 2, \ldots, \tag{2.6}$$

with

$$\begin{aligned} Z_0 &= BB^{\top}, \\ Z_i &= N(I_m \otimes R_i)N^{\top}, \ N = \begin{bmatrix} N_1 & \ldots & N_m \end{bmatrix}. \end{aligned}$$

The matrices R_i thus satisfy

$$R_i = \int_0^{\infty} e^{At_i} Z_{i-1} e^{A^{\top} t_i} dt_i. \tag{2.7}$$

Theorem 2 a) *The reachability Gramian R given by (2.5) exists, if the following two conditions hold*

i) *the matrix A is stable, i.e., the spectrum $\Lambda(A) \subset \mathbb{C}^-$.*
ii) *the inequality $\eta < \sqrt{\mu}/c$ holds, where μ quantifies the decay rate of e^{At} as in (2.2), $\eta = \sqrt{\| \sum_{j=1}^{m} N_j N_j^{\top} \|}$.*

b) *If the reachability Gramian R given by (2.5) exists, then it is bounded by*

$$\|R\|_\bullet \le \frac{c^2\|BB^\top\|_\bullet}{\mu - \eta^2 c^2},$$

where $\|\cdot\|_\bullet$ is either the 2,2-induced matrix norm or the Frobenius norm.

Proof Recall the expression for R_i in (2.7). The integrand has the upper bound as follows [27]:

$$\|e^{At_i} Z_{i-1} e^{A^\top t_i}\|_\bullet \le c^2 \|Z_{i-1}\|_\bullet e^{-\mu t_i}. \tag{2.8}$$

Then, we can show that

$$\|R_i\|_\bullet \le \int_0^\infty \|e^{At_i} Z_{i-1} e^{A^\top t_i}\|_\bullet dt_i \le c^2\|Z_{i-1}\|_\bullet \int_0^\infty e^{-\mu t_i} dt_i = \frac{c^2\|Z_{i-1}\|_\bullet}{\mu}.$$

Since R is an infinite sum of R_i, we have the following:

$$\|R\|_\bullet = \|\sum_{i=1}^\infty R_i\|_\bullet \le \sum_{i=1}^\infty \|R_i\|_\bullet \le \frac{c^2 \sum_{i=1}^\infty \|Z_{i-1}\|_\bullet}{\mu}.$$

Noting that $Z_{i-1},\ i = 1, 2, \ldots,$ are given as

$$Z_0 = BB^\top,\ Z_i = N(I_m \otimes R_i)N^\top,$$

we can derive that

$$\begin{aligned}
\|Z_0\|_\bullet &= \|BB^\top\|_\bullet, \\
\|Z_i\|_\bullet &\le \|\sum_{j=1}^m N_j N_j^\top\| \|R_i\|_\bullet \le \eta^2 \frac{c^2\|Z_{i-1}\|_\bullet}{\mu} = \frac{\eta^2 c^2}{\mu}\|Z_{i-1}\|_\bullet.
\end{aligned}$$

Hence, the upper bound for $\|Z_{i-1}\|_\bullet,\ i = 1, 2, \ldots,$ defines a geometric series. It is immediate that if $\frac{\eta^2 c^2}{\mu} < 1$, then,

$$\sum_{i=1}^\infty \|Z_{i-1}\|_\bullet \le \|BB^\top\|_\bullet \frac{1}{1 - \frac{\eta^2 c^2}{\mu}} = \frac{\mu\|BB^\top\|_\bullet}{\mu - \eta^2 c^2}.$$

Then, the upper bound of R is

$$\|R\|_\bullet \le \frac{c^2 \sum_{i=1}^\infty \|Z_{i-1}\|_\bullet}{\mu} \le \frac{c^2\|BB^\top\|_\bullet}{\mu - \eta^2 c^2}.$$

Existence of the above upper bound yields that $\frac{\eta^2 c^2}{\mu} < 1$, i.e., $\eta < \sqrt{\mu}/c$. □

Remark 1 In many applications, the matrix A is unitarily diagonalizable. Thus, $c = 1$ holds. Then, the upper bound of the Gramian R is

$$\|R\|_\bullet \le \frac{\|BB^\top\|_\bullet}{\mu - \eta^2}.$$

For systems where A is not unitarily diagonalizable, the constant c must be computed to determine the upper bound.

From (2.6), each R_i solves a Laypunov equation. Summing up all the $R_i, i = 1, 2, \ldots$, we can show that $R = \sum_{i=1}^{\infty} R_i$ solves a generalized Lyapunov equation.

Proposition 1 ([28]) *Suppose that the matrix A is stable and N_j, $j = 1, 2, \ldots, m$ are sufficiently bounded so that the reachability Gramian R exists. Then,*

a) *R satisfies the generalized Lyapunov equation*

$$AR + RA^\top + \sum_{j=1}^{m} N_j R N_j^\top + BB^\top = 0. \tag{2.9}$$

b) *The bilinear dynamical system given by (1.2) is reachable if and only if R is positive definite.*

Similar to LTI systems, observability is dual to reachability. Let Q denote the observability Gramian. Then, it can be expressed as

$$Q = \sum_{i=1}^{\infty} \int_0^{\infty} \cdots \int_0^{\infty} O_i^\top O_i dt_1 \cdots dt_i := \sum_{i=1}^{\infty} Q_i \tag{2.10}$$

with

$$\begin{aligned} O_1 &= Ce^{At_1}, \\ O_i &= (I_m \otimes O_{i-1})\tilde{N}e^{At_i}, \ i = 2, 3, \ldots, \end{aligned}$$

where $\tilde{N} = \mathrm{col}(N_1, \ldots, N_m)$ is the column concatenation of matrices N_j. The upper bound and the existence of the observability Gramian are stated by the following corollary.

Corollary 1 a) *The observability Gramian given by (2.10) exists if*

i) *the matrix A is stable.*
ii) *the inequality $\eta < \sqrt{\mu}/c$ holds.*

b) *If the Gramian Q exists, it is bounded by*

$$\|Q\|_\bullet \leq \frac{c^2 \|C^\top C\|_\bullet}{\mu - \eta^2 c^2}.$$

Again, the Gramian matrix Q can be computed by solving a generalized Lyapunov equation.

Proposition 2 ([28]) *Suppose A is stable and N_j, $j = 1, 2, \ldots, m$ are sufficiently bounded so that the observability Gramian Q exists. Then,*

i) *Q satisfies the generalized Lyapunov equation*

$$A^\top Q + QA + \sum_{j=1}^{m} N_j^\top Q N_j + C^\top C = 0. \tag{2.11}$$

ii) *The bilinear dynamical system given by (1.2) is observable if and only if Q is positive definite.*

For stable LTI systems, transfer functions are analytic functions in $\mathbb{C}^+$. Hence, the transfer functions live in the Hardy space

$$\mathcal{H}_p^+ := \left\{ f : \mathbb{C}^+ \to \mathbb{C}^q \,\middle|\, \|f\|_{\mathcal{H}_p^+} < \infty,\ f \text{ is analytic.} \right\}.$$

Then, the $\mathcal{H}_2$ norm is defined as $\sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} f(\imath\omega) f^*(\imath\omega) d\omega}$, where $\imath$ is the imaginary unit and * stands for the conjugate transpose. By applying Parceval's identity, one can easily show that the $\mathcal{H}_2$ norm equals the impulse response energy, which is defined in the time domain. For bilinear dynamical systems, the $\mathcal{H}_2$ norm originally given by [28] was defined in the time domain, so it is better to be interpreted as the convolution kernel energy (or Volterra kernel energy). To be more precise, we define the $\mathcal{H}_2$ norm in the (multi-dimensional) frequency domain. Consider the generalized transfer functions obtained by the Laplace transform of the convolution kernels

$$\begin{aligned} &H_i^{j_1,\ldots,j_i}(s_1, s_2, \ldots, s_i) := \mathscr{L}[h_i^{j_1,\ldots,j_i}(t_1, t_2, \ldots, t_i)] \\ &= C(s_i I - A)^{-1} N_{j_1} (s_{i-1} I - A)^{-1} N_{j_2} \cdots (s_2 I - A)^{-1} N_{j_{i-1}} (s_1 I - A)^{-1} b_{j_i}. \end{aligned} \tag{2.12}$$

A stable matrix A guarantees that the matrix functions $(s_i I - A)^{-1}$, $i = 1, 2, \ldots$ are analytic in $\mathbb{C}^+$. Then, the $\mathcal{H}_2$ norm can be defined as follows.

Definition 2 Suppose that the bilinear dynamical system given by (1.2) is stable and the Gramians R and Q exist. Then, the $\mathcal{H}_2$ norm is defined as

$$\begin{aligned} \|\Sigma_{bl}\|_{\mathcal{H}_2}^2 := \sum_{i=1}^{\infty} \frac{1}{(2\pi)^i} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} H_i^j(\imath\omega_1, \ldots, \imath\omega_i) \left(H_i^j(\imath\omega_1, \ldots, \imath\omega_i) \right)^* \\ \times d\omega_1 \ldots d\omega_i, \end{aligned}$$

where j denotes the abbreviation of the multi-index $j_1, \ldots, j_i$.

Applying Parceval's identity, the $\mathcal{H}_2$ norm defined above equals the convolution kernel energy.

Theorem 3 ([28]) *For the bilinear system Σ_{bl} given by (1.2), if the system is stable and the reachability Gramian R (and the observability Gramian Q) exists, the $\mathcal{H}_2$ norm of the system can be computed by*

$$\|\Sigma_{bl}\|_{\mathcal{H}_2} = \sqrt{trace(C R C^\top)} = \sqrt{trace(B^\top Q B)}.$$

3 Model order reduction

As stated in Section 1.1, the model order reduction problem is to find a reduced-order bilinear dynamical system $\widehat{\Sigma}_{bl}$ with a much lower complexity $r \ll n$, such that the distance $\|\Sigma_{bl} - \widehat{\Sigma}_{bl}\|_{\mathcal{H}_2}$ is minimized. Define the error system $\Sigma_{bl} - \widehat{\Sigma}_{bl}$ as

$$\Sigma_{bl,e}: \begin{cases} \begin{pmatrix} \dot{x}(t) \\ \dot{x}_r(t) \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & A_r \end{pmatrix} \begin{pmatrix} x(t) \\ x_r(t) \end{pmatrix} + \sum_{j=1}^{m} \begin{pmatrix} N_j & 0 \\ 0 & N_{rj} \end{pmatrix} \begin{pmatrix} x(t) \\ x_r(t) \end{pmatrix} u_j(t) + Bu(t), \\ e(t) = \begin{pmatrix} C & -C_r \end{pmatrix} \begin{pmatrix} x(t) \\ x_r(t) \end{pmatrix}, \end{cases}$$

with

$$A_r = V^\top A V,\ N_{rj} = V^\top N_j V,\ j = 1, 2, \ldots, m,\ B_r = V^\top B,\ C_r = CV.$$

Note that the reduced-order system $\widehat{\Sigma}_{bl}$ in (1.3) is uniquely defined by the projection matrix V. Since the bilinear dynamical system is input-output invariant under coordinate transformations, the reduced-order system is uniquely determined by the subspace spanned by V rather than the matrix V itself. According to this fact, we reformulate the model order reduction problem as an optimization problem on the Grassmann manifold $\mathcal{G}_{n,r}$. Then, the gradient flow analysis is applied to guarantee the convergence of the optimization algorithm as well as to speed it up. The $\mathcal{H}_2$ norm criterion describes the approximation accuracy in the frequency domain. In the time domain, the approximation is only accurate for impulsive input signals. In order to make the time domain simulation more accurate, we propose to match the steady state if the low-frequency behavior of the system is of more interest or the system reaches the steady state.

3.1 Problem formulation

To solve the $\mathcal{H}_2$ model order reduction problem for a bilinear dynamical system given by (1.2), equivalently, the following problem needs to be solved as follows:

$$\begin{aligned} \min\ & J(V) = \tfrac{1}{2}\mathrm{trace}\left(\begin{pmatrix} C & -CV \end{pmatrix} \begin{pmatrix} R_f & X \\ X^\top & R_r \end{pmatrix} \begin{pmatrix} C^\top \\ -V^\top C^\top \end{pmatrix} \right) \\ s.t.\ & V^\top V = I_r,\ V^\top \xi = \xi^\top V = 0, \text{ and} \\ & \begin{pmatrix} A & 0 \\ 0 & A_r \end{pmatrix} \begin{pmatrix} R_f & X \\ X^\top & R_r \end{pmatrix} + \begin{pmatrix} R_f & X \\ X^\top & R_r \end{pmatrix} \begin{pmatrix} A^\top & 0 \\ 0 & A_r^\top \end{pmatrix} \\ & + \sum_{j=1}^{m} \begin{pmatrix} N_j & 0 \\ 0 & N_{rj} \end{pmatrix} \begin{pmatrix} R_f & X \\ X^\top & R_r \end{pmatrix} \begin{pmatrix} N_j^\top & 0 \\ 0 & N_{rj}^\top \end{pmatrix} + \begin{pmatrix} BB^\top & BB_r^\top \\ B_r B^\top & B_r B_r^\top \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \end{aligned} \tag{3.1}$$

where $R_e = \begin{pmatrix} R_f & X \\ X^\top & R_r \end{pmatrix}$ is the reachability Gramian of the error system $\Sigma_{bl} - \widehat{\Sigma}_{bl}$, and $\xi \in \mathcal{T}_V \mathcal{G}_{n,r}$ denotes an element of the tangent space of the Grassmann manifold $\mathcal{G}_{n,r}$ at $\mathrm{span}\{V\}$. The objective function in (3.1) is derived from Theorem 3, which

computes the half of the squared $\mathcal{H}_2$ norm of the error system $\Sigma_{bl,e}$. Hence, minimizing the objective function results in finding an (at least locally) optimal reduced-order system. The constraint $V^\top V = I_r$ indicates that the matrix V is orthonormal. The second constraint $V^\top \xi = \xi^\top V = 0$ means that the subspace spanned by V is perpendicular to the tangent space at $\operatorname{span}\{V\}$. Equivalently, one can express these two constraints together as $\operatorname{span}\{V\} \in \mathcal{G}_{n,r}$. The third constraint is used to compute the reachability Gramian of the error system. In fact, one also needs to compute the observability Gramian $Q_e = \begin{pmatrix} Q_f & Y \\ Y^\top & Q_r \end{pmatrix}$ of the error system by solving

$$\begin{aligned} &\begin{pmatrix} A^\top & 0 \\ 0 & A_r^\top \end{pmatrix} \begin{pmatrix} Q_f & Y \\ Y^\top & Q_r \end{pmatrix} + \begin{pmatrix} Q_f & Y \\ Y^\top & Q_r \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & A_r \end{pmatrix} \\ &+ \sum_{j=1}^{m} \begin{pmatrix} N_j^\top & 0 \\ 0 & N_{rj}^\top \end{pmatrix} \begin{pmatrix} Q_f & Y \\ Y^\top & Q_r \end{pmatrix} \begin{pmatrix} N_j & 0 \\ 0 & N_{rj} \end{pmatrix} + \begin{pmatrix} C^\top C & -C^\top C_r \\ -C_r^\top C & C_r^\top C_r \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \end{aligned} \tag{3.2}$$

because it is required to compute the gradient.

To solve the optimization problem in (3.1), gradient descent optimization is considered in this paper. Then, the gradient of J with respect to V needs to be computed. An easy way is that we first neglect the geometry of the manifold, i.e., consider $V \in \mathbb{R}^{n\times r}$ and compute the gradient J_V, which is called the Euclidean gradient. Then, the Euclidean gradient is projected onto the tangent space of the Grassmann manifold at $\operatorname{span}\{V\}$, i.e., $\mathcal{T}_V\mathcal{G}_{n,r}$. The obtained gradient ∇J is the actual gradient of J, which is called the Riemannian gradient. For detailed explanations, we refer to [14].

The Euclidean gradient for the objective function J in (3.1) is given by [9] as

$$\begin{aligned} J_V(V) = {} & A^\top V(Y^\top X + Q_r R_r) + AV(X^\top Y + R_r Q_r) \\ & + BB^\top(Y + VQ_r) + C^\top C(-X + VR_r) \\ & + \sum_{j=1}^{m} N_j^\top V(Y^\top N_j X + Q_r V^\top N_j V R_r) \\ & + \sum_{j=1}^{m} N_j V(X^\top N_j^\top Y + R_r V^\top N_j^\top V Q_r). \end{aligned} \tag{3.3}$$

Since for any matrix V satisfying $\operatorname{span}\{V\} \in \mathcal{G}_{n,r}$, $\forall \xi \in \mathcal{T}_V\mathcal{G}_{n,r}$, there holds $V^\top \xi = \xi^\top V = 0$, the Riemannian gradient can be obtained by [14]:

$$\nabla J(V) = (I - VV^\top) J_V(V), \tag{3.4}$$

which means projecting the Euclidean gradient onto the orthogonal complement of V, i.e., the tangent space $\mathcal{T}_V\mathcal{G}_{n,r}$. With the gradient computed, the optimization problem in (3.1) can be solved by the following algorithm.

Algorithm 1 $\mathcal{H}_2$ MOR by gradient descent on $\mathcal{G}_{n,r}$.

Require: initial guess V_0.
Ensure: the optimal projection matrix $\bar{V}$.
1: **for** $k = 0 : K - 1$ **do**
2: Compute the gradient $G_k = \nabla J(V_k)$ by (3.4). Set the search direction $\xi_k = -G_k$.
3: Compute line search step size α_k, such that $J(V(\alpha_k))$ decreases.
4: Set $V_{k+1} = \mathfrak{V}(\alpha_k, V_k, \xi_k)$.
5: **end for**
6: Set $\bar{V} = V_K$

Algorithm 1 requires the computation of the line search step size α_k and an update method $V_{k+1} = \mathfrak{V}(\alpha_k, V_k, \xi_k)$. In general, the line search step size can be obtained by the back-tracking line search [21]. However, such a method can result in a large amount of computational effort because the Gramians R_e and Q_e need to be updated as long as V_k is updated. To compute $V_{k+1} = \mathfrak{V}(\alpha_k, V_k, \xi_k)$, a *retraction* mapping needs to be constructed [1]. By investigating the geometry of the Grassmann manifold $\mathcal{G}_{n,r}$, the geodesic can be used as the *retraction* mapping. Solutions of these two problems lead to the main results of this paper.

Remark 2 If instead of the subspace spanned by V, the matrix V is of interest, one may formulate the optimization problem on the Stiefel manifold. In that case, we only have $V^\top \xi + \xi^\top V = 0$. If the canonical metric is used, the projection of the Euclidean gradient onto the tangent space can be computed by the following [14]:

$$\nabla J(V) = J_V - V J_V^\top V.$$

3.2 Gradient flow analysis

The geodesic on the Grassmann manifold is used to update the projection matrix V_k in each iteration.

Proposition 3 ([14]) *Given a point $V \in \mathbb{R}^{n\times r}$ on the Grassmann manifold $\mathcal{G}_{n,r}$, i.e., $V^\top V = I_r$, and a tangent direction $\xi \in \mathcal{T}_V\mathcal{G}_{n,r}$ with its thin singular value decomposition (SVD) $\xi = USZ^\top$, the geodesic emanating from V in the direction of ξ has the expression*

$$\mathfrak{V}(\alpha, V, \xi) = VZ\cos(S\alpha)Z^\top + U\sin(S\alpha)Z^\top,\ \alpha \in [0, 1]. \tag{3.5}$$

Then, at the kth iteration of Algorithm 1, the matrix V_{k+1} can be updated as follows:

$$V_{k+1} := \mathfrak{V}(\alpha_k, V_k, \xi_k) = V_k Z_k \cos(S_k\alpha_k)Z_k^\top + U_k \sin(S_k\alpha_k)Z_k^\top,$$

where $U_k S_k Z_k^\top = \xi_k$ is the thin SVD of ξ_k. To compute the line search step size, the gradient flow of $\mathfrak{V}(\alpha, V, \xi)$ in terms of the step size α is discussed. $\mathfrak{V}(\alpha, V, \xi)$ will be abbreviated as $\mathfrak{V}(\alpha)$ for notational convenience. Without further specification, $\dot{F}(\alpha)$

stands for the derivative of the function F with respect to α. A simple calculation shows that the geodesic satisfies a first-order ordinary differential equation (ODE).

Theorem 4 *The geodesic equation given by (3.5) is the solution of the first-order constant coefficient ODE*

$$\dot{\mathfrak{V}}(\alpha) := \frac{\partial}{\partial \alpha}\mathfrak{V}(\alpha) = (\xi V^\top - V\xi^\top)\mathfrak{V}(\alpha),\ \alpha \in [0, 1]. \tag{3.6}$$

Proof From (3.5), the first-order derivative of $\mathfrak{V}(\alpha)$ in terms of α is

$$\dot{\mathfrak{V}}(\alpha) = -VZS\sin(S\alpha)Z^\top + US\cos(S\alpha)Z^\top.$$

Notice that $V^\top U = U^\top V = 0$ because $V^\top\xi = \xi^\top V = 0$, we have the following:

$$\begin{aligned} \xi^\top\mathfrak{V}(\alpha) &= ZS\sin(S\alpha)Z^\top \Rightarrow -V\xi^\top\mathfrak{V}(\alpha) = -VZS\sin(S\alpha)Z^\top, \\ V^\top\mathfrak{V}(\alpha) &= Z\cos(S\alpha)Z^\top \Rightarrow \xi V^\top\mathfrak{V}(\alpha) = US\cos(S\alpha)Z^\top, \end{aligned}$$

which imply that

$$\dot{\mathfrak{V}}(\alpha) = -V\xi^\top\mathfrak{V}(\alpha) + \xi V^\top\mathfrak{V}(\alpha) = (\xi V^\top - V\xi^\top)\mathfrak{V}(\alpha). \qquad \square$$

Remark 3 In general, a geodesic is the solution of a second-order ODE [16]. In case of the Grassmann manifold, the corresponding second-order ODE is [14]:

$$\ddot{\mathfrak{V}}(\alpha) + \mathfrak{V}(\alpha)(\dot{\mathfrak{V}}^\top(\alpha)\dot{\mathfrak{V}}(\alpha)) = 0,\ \alpha \in [0, 1].$$

Direct computation shows that (3.5) satisfies the above equation. In our case, since the explicit expression of the geodesic exists and $V^\top\xi = \xi^\top V = 0$ holds, the geodesic equation satisfies a first-order ODE. However, to solve the ODE in (3.6), one still needs the initial value of V and ξ. Hence, the second-order structure is included implicitly in (3.6) as an initial condition requirement.

The Euclidean gradient $J_V(V)$ satisfies a symmetry property, which can be used to simplify the ODE in (3.6) further.

Proposition 4 ([9]) *The Euclidean gradient $J_V(V)$ given by (3.3) has a symmetry property*

$$V^\top J_V(V) = J_V^\top(V)V.$$

At the kth iteration, if the steepest descent direction acts as the tangent direction in the geodesic equation (3.5), the coefficient matrix in (3.6) only relates to V_k and $J_{V_k}(V_k)$, i.e., the Euclidean gradient at V_k. For convenience, we abbreviate $J_{V_k}(V_k)$ as J_{V_k}.

Corollary 2 *Given the starting point of the geodesic as V_k and let $\xi_k = (V_k V_k^\top - I_n) J_{V_k}$ be the tangent direction, then the geodesic equation given by (3.5) solves the following first-order ODE*

$$\dot{\mathfrak{V}}(\alpha) = (V_k J_{V_k}^\top - J_{V_k} V_k^\top)\mathfrak{V}(\alpha) := \Gamma_k \mathfrak{V}(\alpha), \ V(0) = V_k, \ \alpha \in [0, 1]. \tag{3.7}$$

Proof Taking the steepest descent direction as the line search direction, i.e.,

$$\xi_k = (V_k V_k^\top - I_n) J_{V_k},$$

then, we have the following:

$$\begin{aligned} \xi_k V_k^\top &= (V_k V_k^\top - I_n) J_{V_k} V_k^\top = V_k J_{V_k}^\top V_k V_k^\top - J_{V_k} V_k^\top, \\ V_k \xi_k^\top &= V_k J_{V_k}^\top V_k V_k^\top - V_k J_{V_k}^\top. \end{aligned}$$

Hence,

$$\xi_k V_k^\top - V_k \xi_k^\top = V_k J_{V_k}^\top - J_{V_k} V_k^\top := \Gamma_k. \qquad \square$$

Based on the derivative of $\mathfrak{V}(\alpha)$ in Corollary 2, the following results show how the line search step size can be selected to guarantee the convergence of Algorithm 1.

Theorem 5 *Consider the optimization problem in (3.1) with the Euclidean gradient given by (3.3). Let $\mathfrak{V}(\alpha) \in \mathcal{G}_{n,r}, \ \alpha \in [0, 1]$ be a geodesic, which is certainly differentiable with respect to α. Then, the Euclidean gradient $J_V(\alpha) := J_V(\mathfrak{V}(\alpha))$ and its derivative $\dot{J}_V(\alpha) := \partial J_V / \partial \alpha$ have the upper bounds*

$$\|J_V(\alpha)\|_F \leq \zeta_1, \tag{3.8}$$

$$\|\dot{J}_V(\alpha)\|_F \leq \zeta_2 \|\dot{\mathfrak{V}}(\alpha)\|_F, \tag{3.9}$$

where,

$$\zeta_1 := \frac{4\sqrt{r} c^2 \|B\|^2 \|C\|^2 (\mu + c^2 \|A\|)}{(\mu - \eta^2 c^2)^2}, \tag{3.10}$$

$$\zeta_2 := \frac{2\|B\|^2 \|C\|^2}{(\mu - \eta^2 c^2)^3} \Big(3(\mu + \eta^2 c^2 + 2c^2 \|A\|)^2 + c^2 (\mu - \eta^2 c^2)(\mu + 3\eta^2 c^2 + 2c^2 \|A\|) \Big). \tag{3.11}$$

Proof See Appendix A. $\square$

Remark 4 The computation of $\dot{J}_V(\alpha)$ shows that at the kth iteration, $\dot{J}_{V_k}(\alpha) = -\|\Gamma_k\|_F^2 \leq 0$. Hence, the gradient descent algorithm should always converge.

Corollary 3 *Let ζ_1 and ζ_2 be given by Theorem 5. In the optimization process, let V_k denote the projection matrix at the kth iteration and ξ_k be the corresponding steepest*

descent direction. Then, the projection matrix in the $k+1$*st iteration is given by* $\mathfrak{V}(\alpha_k) := \mathfrak{V}(\alpha_k, V_k, \xi_k)$ *in* (3.5) *with* α_k *the step size. If the step size satisfies*

$$0 < \alpha_k < \frac{\sqrt{2}}{\zeta_1 + \sqrt{2}\zeta_2}, \tag{3.12}$$

then the objective function $J(V)$ *decreases, i.e.,* $J(\mathfrak{V}(\alpha_k)) \leq J(V_k),\ k = 0, 1, \ldots$. *The equality holds if and only if* V_k *is the critical point of* $J(V)$.

Proof The proof is similar to the proof of Theorem 4.1 in [27]. According to Taylor's Theorem, there exists a constant $\tau \in [0, \alpha]$ such that

$$J(\mathfrak{V}(\alpha)) \leq J(V_k) + \alpha \dot{J}(V_k) + \frac{\alpha^2}{2}\ddot{J}(\mathfrak{V}(\tau)).$$

Then, applying Corollary 2 and following the proof of Theorem 4.1 in [27], the proof can be completed. □

The step size in (3.12) is uniformly upper bounded by a constant, which is independent of the number of iterations and the projection matrix in each iteration. Sometimes this constant can be really small, which makes the step size too conservative. Although Algorithm 1 is guaranteed to converge under such a step size, the convergence speed is not guaranteed to be fast. Hence, theoretically, one can only conclude that $V_k \to \bar{V}$ if $k \to \infty$.

To obtain a more effective and practical line search step size, the basic idea is to use the higher order Taylor expansion of $J(\mathfrak{V}(\alpha))$ as a function of α, i.e., the step size. Consider the third-order Taylor expansion of the objective function $J(\mathfrak{V}(\alpha))$. By Taylor's Theorem, at the kth iteration, $\exists \tau_k \in \mathbb{R}^+$, for any $0 < \alpha < \tau_k$, we have the following [27]:

$$J(\mathfrak{V}(\alpha)) \leq J(V_k) + \alpha \dot{J}(V_k) + \frac{\alpha^2}{2}\ddot{J}(V_k) + \frac{\alpha^3}{6}\max_{0\leq\alpha\leq\tau_k} |J^{(3)}(\mathfrak{V}(\alpha))|. \tag{3.13}$$

To make sure that the objective function J given by (3.1) decreases, we need to find an α_k such that $J(\mathfrak{V}(\alpha_k)) \leq J(V_k)$. Note that we already know that there exists an upper bound τ_k of α_k. Hence, first we need to find an effective τ_k, which is solved by the following Corollary.

Corollary 4 *Let* $\mathfrak{V}(\alpha),\ \alpha \in [0, 1]$, *be the geodesic equation which satisfies* (3.7), *where* $\mathfrak{V}(\alpha)$ *is orthogonal and* Γ_k *is skew-symmetric, and let* s_k *denote the unique positive real root of the polynomial*

$$2\beta_0 s_k^4 c^2 + 2\beta_1 s_k^3 c^2 + 2\beta_2 s_k^2 c^2 + 2\beta_3 s_k c^2 - \mu + \eta^2 c^2. \tag{3.14}$$

Let the Lie bracket operations be defined as

$$\mathcal{L}_1(A, B) = AB - BA, \quad \mathcal{L}_k(A, B) = \mathcal{L}_{k-1}(A, B)B - B\mathcal{L}_{k-1}(A, B),\ k \geq 2. \tag{3.15}$$

The constants $\beta_i, i = 0, 1, 2, 3$, are given as follows:

$$\beta_0 = \|\mathcal{L}_4(A, \Gamma_k)\| \tag{3.16}$$

$$+\sum_{j=1}^{3}\left(\|N_j\|\|\mathcal{L}_4(N_j, \Gamma_k)\|+4\|\mathcal{L}_1(N_j,\Gamma_k)\|\|\mathcal{L}_3(N_j,\Gamma_k)\|+3\|\mathcal{L}_2(N_j,\Gamma_k)\|^2\right), \tag{3.17}$$

$$\beta_1 = 4\|\mathcal{L}_3(A, \Gamma_k)\|+\sum_{j=1}^{m}\left(4\|N_j\|\|\mathcal{L}_3(N_j, \Gamma_k)\| + 12\|\mathcal{L}_1(N_j, \Gamma_k)\|\|\mathcal{L}_2(N_j, \Gamma_k)\|\right), \tag{3.18}$$

$$\beta_2 = 6\|\mathcal{L}_2(A, \Gamma_k)\| + 6\sum_{j=1}^{m}\left(\|N_j\|\|\mathcal{L}_2(N_j, \Gamma_k)\| + \|\mathcal{L}_1(N_j, \Gamma_k)\|^2\right), \tag{3.19}$$

$$\beta_3 = 3\|\mathcal{L}_1(A, \Gamma_k)\| + 3\sum_{j=1}^{m}\|N_j\|\|\mathcal{L}_1(N_j, \Gamma_k)\|. \tag{3.20}$$

Then, for any given $\tau_k \in (0, s_k)$, $J^{(3)}(\mathfrak{V}(\alpha))$ is upper bounded by

$$\max_{0\le\alpha\le\tau_k} |J^{(3)}(\mathfrak{V}(\alpha))| \le \frac{1}{2}\theta_k. \tag{3.21}$$

Denote $BB^\top$ and $C^\top C$ as $\mathfrak{B}$ and $\mathfrak{C}$, respectively. The constant θ_k is defined by

$$\theta_k = \begin{pmatrix} \|\left(2\Gamma_k^3\mathfrak{C}\ \ \mathcal{L}_3(\mathfrak{C}, \Gamma_k)\right)\|_F \\ 3\|\left(2\mathfrak{C}\Gamma_k^2\ \ \mathcal{L}_2(\mathfrak{C}, \Gamma_k)\right)\|_F \\ 3\|\left(2\mathfrak{C}\Gamma_k\ \ \mathcal{L}_1(\mathfrak{C}, \Gamma_k)\right)\|_F \\ \sqrt{5}\|\mathfrak{C}\|_F \end{pmatrix}^\top \begin{pmatrix} \max_{0\le\alpha\le\tau_k}\|W(\alpha)\|_F \\ \max_{0\le\alpha\le\tau_k}\|\dot{W}(\alpha)\|_F \\ \max_{0\le\alpha\le\tau_k}\|\ddot{W}(\alpha)\|_F \\ \max_{0\le\alpha\le\tau_k}\|W^{(3)}(\alpha)\|_F \end{pmatrix}, \tag{3.22}$$

where $W(\alpha) = \left(X^\top(\alpha)\ \ R_r(\alpha)\right)$ and the upper bounds on $W(\alpha)$ and its first three derivatives are computed from (B.12) in Appendix B. Furthermore, if one chooses a parameter $\rho_1 \in (0, 1)$ and let $\tau_k = \rho_1 s_k$, then there exists a V_k-independent constant ψ such that

$$\theta_k \le \psi\|\Gamma_k\|_F^3.$$

Proof The proof requires technical calculations, so we put it in Appendix B. □

Remark 5 Corollary 4 tells that the upper bound of $\max_{0\le\alpha\le\tau_k}|J^{(3)}(\mathfrak{V}(\alpha))|$ is related to $\|\Gamma_k\|_F^3$, which indicates the speed of convergence of such an upper bound.

Now let's discuss how to derive an upper bound of α_k from the Taylor expansion given in (3.13). To guarantee the decrease of J, we need to find an α such that $J(\mathfrak{V}(\alpha)) - J(V_k) \le 0$. From (3.13), we know that

$$J(\mathfrak{V}(\alpha)) - J(V_k) \le \alpha\dot{J}(V_k) + \frac{\alpha^2}{2}\ddot{J}(V_k) + \frac{\alpha^3}{6}\max_{0\le\alpha\le\tau_k}|J^{(3)}(\mathfrak{V}(\alpha))|.$$

Since,

$$\dot{J}(V_k) = \operatorname{trace}(J_{V_k}^{\top}\dot{\mathfrak{V}}(\alpha))|_{\alpha=0} = \operatorname{trace}(J_{V_k}^{\top}\Gamma_k V_k) = -\frac{1}{2}\operatorname{trace}(\Gamma_k^{\top}\Gamma_k) = -\frac{1}{2}\|\Gamma_k\|_F^2,$$

and for convenience, we denote $\ddot{J}(V_k)$ as $\frac{1}{2}\gamma_k$, then,

$$J(\mathfrak{V}(\alpha)) - J(V_k) \leq \frac{\alpha}{2}(-\|\Gamma_k\|_F^2 + \frac{\alpha}{2}\gamma_k + \frac{\alpha^2}{6}\theta_k) \leq 0.$$

Hence, we need to find a line search step size α_k such that

$$\alpha_k\left(-\|\Gamma_k\|_F^2 + \frac{\alpha_k}{2}\gamma_k + \frac{\alpha_k^2}{6}\theta_k\right) \leq 0. \tag{3.23}$$

Factorizing (3.23), it can be obtained that

$$\frac{\theta_k\alpha_k}{6}\left(\alpha_k - \frac{-3\gamma_k - \sqrt{9\gamma_k^2 + 24\theta_k\|\Gamma_k\|_F^2}}{2\theta_k}\right)\left(\alpha_k - \frac{-3\gamma_k + \sqrt{9\gamma_k^2 + 24\theta_k\|\Gamma_k\|_F^2}}{2\theta_k}\right) \leq 0 \tag{3.24}$$

Let,

$$\phi_k = \frac{-3\gamma_k + \sqrt{9\gamma_k^2 + 24\theta_k\|\Gamma_k\|_F^2}}{2\theta_k} \geq 0, \tag{3.25}$$

which is the unique positive root of (3.24). It is not difficult to derive that when α_k is bounded by the following:

$$0 \leq \alpha_k \leq \phi_k, \tag{3.26}$$

Equation (3.23) is satisfied. Combining this result with the fact that $0 \leq \alpha_k \leq \tau_k$, the following corollary can be stated.

Corollary 5 *Let ϕ_k be given as in (3.25) and $\gamma_k := 2\ddot{J}(V_k)$. Let the constant τ_k be given by Corollary 4, i.e., $0 < \tau_k < s_k$. Then, for any step size α_k which satisfies the following:*

$$\alpha_k \leq \min(\tau_k, \phi_k), \tag{3.27}$$

it can be guaranteed that $J(\mathfrak{V}(\alpha_k)) \leq J(V_k)$, $k = 0, 1, \ldots$, for any initial guess $V_0 \in \mathcal{G}_{n,r}$. Furthermore, for any smaller step size $\alpha_k = \min(\rho_1 s_k, \rho_2\phi_k)$ with $\rho_{1,2} \in (0, 1)$, Algorithm 1 converges.

Proof The first part of the corollary, i.e., (3.27), has already been shown in Corollary 4. We only need to prove the second part here. If there exists some constant ρ_2 such that $\alpha_k \leq \rho_2 \phi_k$, from (3.24), we have

$$\begin{aligned}
&\frac{\theta_k \alpha_k}{6}\left(\alpha_k + \frac{3\gamma_k + \sqrt{9\gamma_k^2 + 24\theta_k \|\Gamma_k\|_F^2}}{2\theta_k}\right)(\alpha_k - \phi_k) \\
\leq\ & \frac{3\gamma_k + \sqrt{9\gamma_k^2 + 24\theta_k \|\Gamma_k\|_F^2}}{12}\alpha_k(\alpha_k - \phi_k) \\
\leq\ & \frac{3\gamma_k + \sqrt{9\gamma_k^2 + 24\theta_k \|\Gamma_k\|_F^2}}{12}\phi_k(\rho_2 - 1)\alpha_k \\
=\ & (\rho_2 - 1)\|\Gamma_k\|_F^2 \alpha_k \leq 0.
\end{aligned}$$

It is immediate that $0 < \rho_2 < 1$ is the condition to guarantee the decrease of the objective function, which means that for $k \to \infty$, $\Gamma_k \to 0$. □

3.3 Matching steady state

The $\mathcal{H}_2$ norm criterion describes how well a system is approximated in the frequency domain. For asymptotically stable LPV systems or nonlinear dynamical systems such as bilinear dynamical systems, usually the time domain simulation is of more interest. Since the $\mathcal{H}_2$ norm is equivalent to the convolution kernel energy in the time domain, the time domain performance can only be guaranteed for impulsive input signals, which is not always the case especially for LPV systems. When the low-frequency behavior of the system is dominant or the system reaches its steady state for a non-impulsive input, the reduced-order system can introduce bias in the time domain simulation. To overcome such a problem and make the reduced-order model more accurate, one can match the steady state by adding an input-dependent feedthrough term to the output equation of (1.3). Consider the bilinear dynamical system in (1.2). Suppose the steady state is $\bar{x}$, then we have

$$\begin{aligned}
\dot{\bar{x}}(t) &= 0 = A\bar{x}(t) + \sum_{j=1}^{m} N_j \bar{x}(t) u_j(t) + Bu(t) \Rightarrow \bar{x}(t) \\
&= -\left(A + \sum_{j=1}^{m} N_j u_j(t)\right)^{-1} Bu(t).
\end{aligned}$$

The steady-state output is $\bar{y}(t) = C\bar{x}(t)$. Similarly, the steady state of the reduced-order system in (1.3) is

$$\bar{x}_r(t) = -\left(A_r + \sum_{j=1}^{m} N_{rj} u_j(t)\right)^{-1} B_r u(t).$$

Correspondingly, the reduced-order steady state output is $\bar{y}_r(t) = C_r \bar{x}_r(t)$. If the steady states of the full-order and the reduced-order systems are matched, we have $\bar{y}(t) = \bar{y}_r(t)$. Then, the next proposition directly follows.

Proposition 5 *Define an input-dependent matrix $D_{ss}(t)$ by*

$$D_{ss}(t) = C_r \left(A_r + \sum_{j=1}^{m} N_{rj} u_j(t) \right)^{-1} B_r - C \left(A + \sum_{j=1}^{m} N_j u_j(t) \right)^{-1} B. \quad (3.28)$$

The steady states of the full-order and the reduced-order systems are matched if the output of the reduced-order system in (1.3) is given by

$$y_r(t) = C_r x_r(t) + D_{ss}(t) u(t).$$

By matching the steady state, the reduced-order system can mimic the low-frequency behavior of the original system. If only the high-frequency behavior of the system is of interest, matching steady state can make the approximation worse, which is not advisable.

Although a large-scale system $(A + \sum_{j=1}^{m} N_j u_j(t))z = -B$ needs to be solved at each time instant, the computational effort is still much less than solving the original bilinear dynamical system. A more effective method would be only computing the steady state of the full-order system at some sampled values of $u(t)$ offline and then applying an interpolation technique online to achieve a speedup, which can be a future research topic.

To close this section, we provide a complete algorithm for reducing a bilinear dynamical system by the proposed method. As for the constant step size method, one can compute the step size by Theorem 5 and Corollary 3 and then simply substitute it into Algorithm 1, the next algorithm only discusses the adaptive step size case.

In the proposed algorithm, $\lambda(\cdot)$ stands for the eigenvalue of a matrix. If $A+A^\top$ is negative definite, $\mu = \lambda_{\min}(-A - A^\top)$ is used. Otherwise, $\mu < 2\lambda_{\min}(-A)$ must hold. For the constant c, only if A is unitarily diagonalizable, $c = 1$ holds. In case that the matrix A is not unitarily diagonalizable, c depends on $\|T\|\|T^{-1}\|$, if $A = T D_A T^{-1}$ for some diagonal matrix D_A. For the case of A having non-trivial Jordan blocks, c cannot be computed numerically. The algorithm is considered to be converged when the norm of the Riemannian gradient G_k is smaller than a user specified tolerance or a maximum number of iterations is reached. Similar to BIRKA, one can also check the eigenvalues of the reduced-order matrix A_r. If the change of the eigenvalues is small enough, the resulting projection matrix V_k is also close to the optimal one.

The computational cost of Algorithm 2 is mainly dominated by solving the second row of the generalized Lyapunov equations in (B.3) to (B.6), i.e., the corresponding generalized Sylvester equations and the reduced-order generalized Lyapunov equations for the matrices X_k, R_{rk}, Y_k, Q_{rk}. Since the reduced-order r is usually much smaller than n, the computational cost mainly comes from solving the generalized Sylvester equations. In this article, we use direct solvers naively, that is, solving the matrix equations by vectorizing them first and then solving the corresponding linear system of equations directly. Hence, the computational cost would be $O(n^3 r^3)$. Other

efficient methods such as the fixed-point method [24], ADI-preconditioned Krylov method [12], and even an extended version of the residual approximation-based iterative Lyapunov solver (RAILS) [2] can be employed to reduce the computational complexity. In both [12, 24], iterative methods such as the fixed-point iteration are introduced to solve the generalized Lyapunov equation. In each iteration, the generalized Lyapunov equation is transformed into the classic Lyapunov equation. Then, in [12], the ADI-preconditioned Krylov subspace method is employed to solve the classic Lyapunov equation in each fixed-point iteration while in [24] the extended Krylov subspace method is used. The RAILS method [2] is a Krylov subspace method as well. When these Krylov subspace methods are employed to solve the classic Lyapunov equation, the original solution is projected onto a low-dimensional subspace. Then, one only needs to solve a small-scale Lyapunov equation. To construct the Krylov subspaces, the computational complexity is linear in the number of non-zeros (nnz) in the A matrix. Namely, if matrix A is sparse, i.e., nnz = $O(n)$, the computational complexity is $O(n)$ (see, e.g., [25]). In case that A is dense, the computational complexity is $O(n^2)$. The total computational cost of combining the fixed point iteration and the Krylov subspace methods is thus either $O(nK)$ or $O(n^2K)$, where K is the number of fixed-point iterations. The numerical examples in [24] show that when A is sparse, their method allows one to solve large-scale problems where n is at the level of 10^5 within a few minutes on a commercially available laptop. Although the aforementioned methods are designated for Lyapunov equations, their concepts can be generalized to Sylvester equations.

Algorithm 2 $\mathcal{H}_2$ MOR by gradient descent on $\mathcal{G}_{n,r}$ with adaptive step size.

Require: State space matrices $A, B, C, N_j, j = 1, 2, \ldots, m$ and initial guess of the projection matrix V_0. Two constants $\rho_{1,2} \in (0, 1)$.

Ensure: $\mathcal{H}_2$ optimal reduced-order system matrices $A_r, B_r, C_r, N_{rj}, j = 1, 2, \ldots, m$ and the optimal projection matrix $\bar{V}$.

1: Compute $\mu \leq 2\lambda_{\min}(-A)$, $\eta = \sqrt{\| \sum_{j=1}^{m} N_j N_j^\top \|}$ and the constant c.
2: Compute the full-order Gramians R_f and Q_f for the evaluation of the objective function $J(V)$.
3: **while** not converged **do**
4: Compute the Gramians X_k, R_{rk}, Y_k, Q_{rk} according to the generalized Lyapunov equations in (3.1) and (3.2), respectively.
5: Compute the Euclidean gradient $J_{V_k}(V_k)$ by (3.3).
6: Compute the Riemannian gradient $G_k = \nabla J(V_k)$ by (3.4). Set the search direction as $\xi_k = -G_k$.
7: Compute $\Gamma_k = V_k J_{V_k}^\top - J_{V_k} V_k^\top$.
8: Compute $\beta_j, j = 0, 1, 2, 3$ by (3.16) – (3.20) and solve (3.14) to derive s_k.
9: Solve the second row of the generalized Lyapunov equations in (B.3) – (B.6).
10: Compute $\gamma_k = 2\ddot{J}(V_k)$ by (B.1c). Compute θ_k by (3.22) and (B.12).
11: Compute ϕ_k by (3.25). Set $\alpha_k = \min(\rho_1 s_k, \rho_2 \phi_k)$.
12: Set $V_{k+1} = \mathfrak{V}(\alpha_k, V_k, \xi_k)$ by (3.5).
13: **end while**
14: Set $\bar{V} = V_K$ and compute the optimal reduced-order system matrices by (1.3).

During the reduction, the reduced-order r needs to be determined. One heuristic method is to investigate the decay rate of $\sqrt{\lambda(Q_f R_f)}$, which can be considered as a generalization of the Hankel singular values of LTI systems. Although these values are not really the Hankel singular values of the system in general, they still provide a hint on how well a system can be approximated at a certain order. Another method would be trial-and-error. By reducing the system to different orders and considering the computational cost, one could determine r. In the numerical examples of this paper, we investigated the decay rate of the aforementioned generalized Hankel singular values. For both examples, a reduced-order system of dimension ten can guarantee the approximation accuracy as well as low computational cost.

Remark 6 An important point we would like to address is that during the model order reduction process, the bilinear dynamical system only serves as an auxiliary system. It is clear that if, e.g., the parameters are constant (like material parameters), the resulting auxiliary input function cannot be an $\mathcal{L}_2$ function unless the parameters were zero. Nevertheless, once we have the form of a bilinear system, we ignore this discrepancy and assume that we have $\mathcal{L}_2$ inputs and that the $\mathcal{H}_2$ norm of the auxiliary bilinear system exists. This allows us to compute an $\mathcal{H}_2$-(sub)optimal reduced-order model. We use the projection subspaces that lead to the reduced bilinear system then for the LPV system to obtain a reduced-order LPV system. There are no claims about stability of the reduced-order LPV system, and also not about optimality with respect to any norm that could measure the approximation quality. The usage of the associated low-dimensional subspaces in order to compute a reduced-order LPV system is only a heuristic, but the numerical experiments in the next section, also in the previous papers introducing this "trick" [8, 9], demonstrate that it often works very well.

4 Numerical examples

Two numerical examples are tested to demonstrate the performance of the proposed method.

4.1 The synthetic example

The first example is the modified version of the synthetic example on *MORwiki* [18]. The system was originally single-input and single-output and given as follows:

$$\begin{aligned} \dot{x}(t) &= (A_0 + \epsilon A_\epsilon)x(t) + Bu(t), \\ y(t) &= Cx(t), \ \epsilon \in (0, 1]. \end{aligned}$$

For the test purpose, the system is modified to

$$\begin{aligned} \dot{x}(t) &= \underbrace{(A_0 + 50 \times A_\epsilon)}_{A} + N_1 u_1(t)x(t) + \underbrace{0.01 \times A_\epsilon}_{N_2} u_2(t)x(t) + \begin{pmatrix} B & 0 \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}, \\ y(t) &= Cx(t), \end{aligned}$$

where $u_1(t)$ is the original input signal and $N_1 = 0$. The second input $u_2(t) \in (0, 100]$ is the time-varying parameter. In this setting, $1000 = \mu > \eta^2 = 100$ and the matrix A is stable. Hence, the Gramians exist. The system dimension is reduced from $n = 100$ to $r = 10$ with the cut-off generalized Hankel singular values smaller than 10^{-9}. In Fig. 1, we show the time domain simulation of the full-order and the reduced-order systems for the parameter variation $u_2(t) = 50+50\sin(0.8\pi t+4\pi/3)$ and the input $u_1(t) = 16 + 20\cos(0.4\pi t)$, which is a sinusoidal signal as well. In this example, the gradient descent optimization stops when the norm of the gradient is less than 10^{-4} and it takes 474 iterations. Let the relative approximation error in the $\mathcal{H}_2$ norm sense be defined as

$$\frac{\|\Sigma_{bl} - \widehat{\Sigma}_{bl}\|_{\mathcal{H}_2}}{\|\Sigma_{bl}\|_{\mathcal{H}_2}}.$$

The relative error for this example is about 0.0066.

Both the input and the parameter variation are low-frequency signals in this test. Hence, matching steady state would be relevant. It can be seen that without matching the steady state, although the dynamics are captured, there is a mismatch in the amplitude of the output. By matching the steady state, the mismatch (in the absolute value) is reduced from 10^{-1} to 10^{-3}. The uniform line search step size in this example is around 1.18×10^{-6}. The minimal value for the adaptive step size is about 0.0197 which is much larger. Hence, although the uniform step size can guarantee convergence, the convergence speed is too slow. By proposing the adaptive step size, the convergence rate has a significant speed up. If we compare the convergence speed in the first 100 iterations, the adaptive step size method is about four *million* times faster than the uniform step size method.

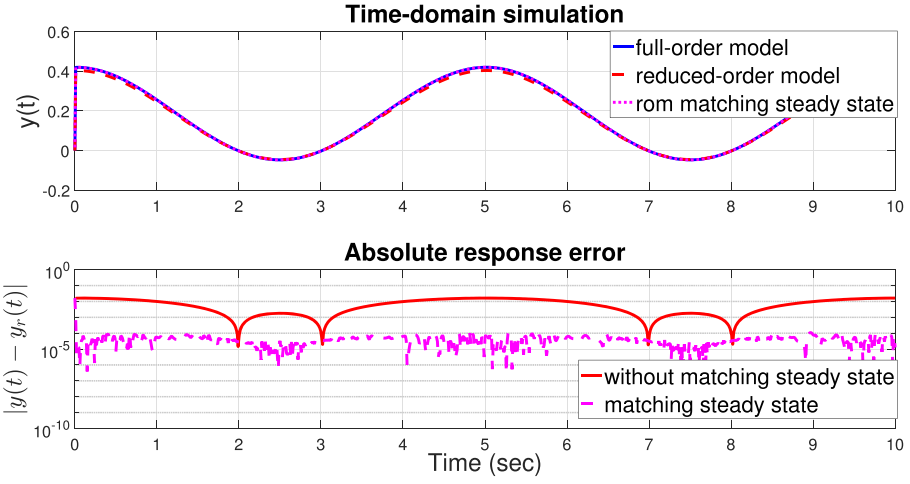

Fig. 1 Time domain simulation and the absolute response error of the synthetic model

4.2 2D heat transfer model

The second example we test is a 2D heat transfer model, which is a modified version of the one in [4]. The system is governed by the partial differential equation

$$\rho C T_t = \kappa (T_{xx} + T_{yy}) + S(x, y), \ \ (x, y) \in [0, 1]^2,$$

with the following boundary conditions

- Dirichlet boundary conditions on left and right, $T = 0$
- Robin boundary conditions on top and bottom, $\mathbf{n} \cdot \nabla T = 0.25(T - 1)S(x, y)$

with $S(x, y)$ the space-dependent source term and $\mathbf{n}$ the normal direction. Spatial discretization leads to a bilinear dynamical system as follows:

$$\begin{aligned} \dot{x}(t) &= Ax(t) + N_1 x(t) u_1(t) + N_2 x(t) u_2(t) + Bu(t) \\ y(t) &= Cx(t). \end{aligned}$$

Suppose that the number of grid points is $n_g = 10$, the system has $10^2 = 100$ states. The output is taken as the sum of the temperature at all the grid points scaled by $1/n_g^2$, i.e., the average temperature. Again, we reduce the system dimension to 10. The cut-off generalized Hankel singular values are all less than 10^{-5}. In 250 iterations, the relative approximation error is reduced from about 0.9717 to about 0.1619. Admittedly, the relative error is rather large, but qualitatively, we obtain very good simulation results.

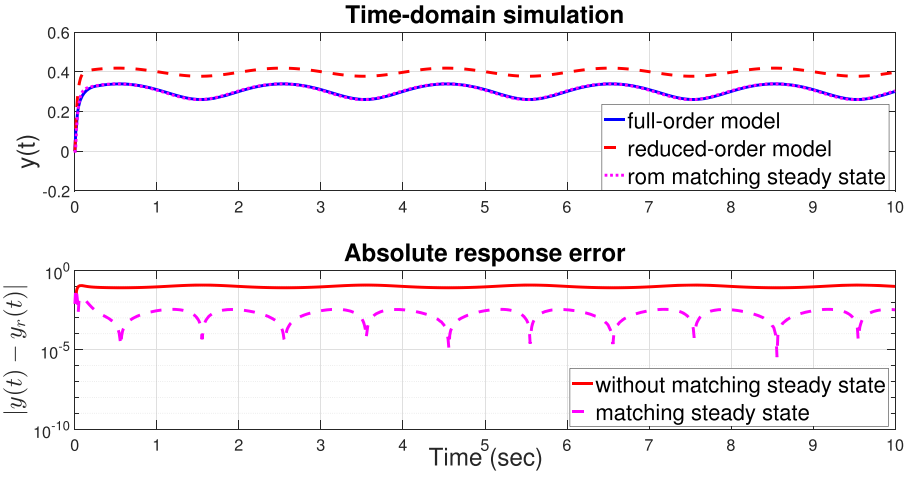

Fig. 2 Time domain simulation and the absolute response error of the heat transfer model

In this example, the uniform line search step size is 0.0015. However, the minimal value of the adaptive step size is about 25.3831, which exceeds 1, but convergence is still guaranteed. We observed that the magnitude of the step size is closely related to the scaling of the output, which deserves further investigation. For testing purposes, the first input channel is set to a step function with amplitude 100. The second input signal is set to a sinusoidal wave $u_2(t) = 10\sin(\pi t) + 15$. Time domain simulation results and the absolute response errors are depicted in Fig. 2. In the numerical test, the system response reaches the steady state and only varies according to the variation of the input $u_2(t)$. Hence, matching the steady state would give a better time domain approximation. It can be seen that without matching the steady state, the mismatch in the output amplitude is really non-negligible, which in this case is at the same level as the output signal (10^{-1}). By matching the steady state, although there is still a mismatch in the amplitude, it is suppressed to the level of 10^{-3}, which results in the relative error at the level of 10^{-2}.

5 Conclusions

This paper discusses a model order reduction method for bilinear dynamical systems. Bilinear dynamical systems can be used to represent a special class of LPV systems. We review the basic system theory for bilinear dynamical systems and propose to define the $\mathcal{H}_2$ norm in the frequency domain, which coincides with the definition of the convolution kernel energy in the literature. As the $\mathcal{H}_2$ norm can be computed from the system Gramians, the model reduction problem amounts to an optimization problem on the Grassmann manifold. We propose to solve the nonlinear, non-convex optimization problem by gradient descent on the Grassmann manifold. To guarantee convergence and speed up the convergence rate, we propose a uniform line search step size and an adaptive step size, which are generalized from the method for LTI systems. Since the $\mathcal{H}_2$ norm is defined in the frequency domain and in the time domain, it is equivalent to the convolution kernel energy, the time domain approximation accuracy is only guaranteed for impulsive input signals. For LPV systems, this is not always the case because the parameters show continuous time-varying behaviors rather than the impulsive behavior. To increase the approximation accuracy, we propose to match the steady state. Although additional computational efforts are needed, the simulation time can still be shortened significantly, especially for multi-query purposes.

Acknowledgments The authors would like to thank the European Cost Action: TD1307-European Model Reduction Network (EU-MORNET) for the funding of a short-term scientific mission, which leads to this work.

Appendix A: Proof of Theorem 5

Recall the Euclidean gradient given by (3.3). Rewrite it into two parts.

$$\begin{aligned} J_{V1} &= A^\top V(Y^\top X + Q_r R_r) + AV(X^\top Y + R_r Q_r) \\ &\quad + BB^\top(Y + VQ_r) + C^\top C(-X + VR_r) \\ J_{V2} &= \sum_{j=1}^{m} N_j^\top V(Y^\top N_j X + Q_r V^\top N_j V R_r) \\ &\quad + \sum_{j=1}^{m} N_j V(X^\top N_j^\top Y + R_r V^\top N_j^\top V Q_r). \end{aligned}$$

Applying Theorem 2 and Corollary 1, we can derive the upper bound of the reduced-order Gramians R_r, Q_r and the solutions of the Sylvester equations X and Y, which are

$$\max(\|X\|_F,\ \|R_r\|_F) \le \frac{\sqrt{r}c^2\|B\|_F^2}{\mu - \eta^2 c^2},\ \max(\|Y\|_F,\ \|Q_r\|_F) \le \frac{\sqrt{r}c^2\|C\|_F^2}{\mu - \eta^2 c^2}; \tag{A.1}$$

$$\max(\|X\|,\ \|R_r\|) \le \frac{c^2\|B\|^2}{\mu - \eta^2 c^2},\ \max(\|Y\|,\ \|Q_r\|) \le \frac{c^2\|C\|^2}{\mu - \eta^2 c^2}. \tag{A.2}$$

Applying the bounds in (A.1) and (A.2), the upper bound of J_{V1} is derived as

$$\|J_{V1}\|_F \le \frac{4\sqrt{r}c^2\|B\|^2\|C\|^2(c^2\|A\| + \mu - \eta^2 c^2)}{(\mu - \eta^2 c^2)^2}.$$

The upper bound of J_{V2} is computed as follows:

$$\begin{aligned} \|J_{V2}\|_F &\le 2\|\sum_{j=1}^{m} N_j^\top V(Y^\top N_j X + Q_r V^\top N_j V R_r)\|_F \\ &\le 2\sum_{j=1}^{m} \|N_j N_j^\top\|(\|Y\|\|X\| + \|Q_r\|\|R_r\|)\|V\|_F \\ &\le \frac{4\eta^2\sqrt{r}c^4\|B\|^2\|C\|^2}{(\mu - \eta^2 c^2)^2}. \end{aligned}$$

As a result, the upper bound of J_V is

$$\|J_V\|_F \le \|J_{V1}\|_F + \|J_{V2}\|_F \le \frac{4\sqrt{r}c^2\|B\|^2\|C\|^2(\mu + c^2\|A\|)}{(\mu - \eta^2 c^2)^2} := \zeta_1.$$

In order to derive $\|\dot{J}_V\|_F$, we need to obtain the upper bound on $\|\dot{X}\|_F$, $\|\dot{Y}\|_F$, $\|\dot{R}_r\|_F$, and $\|\dot{Q}_r\|_F$. Consider (B.4) with Ψ_1 given by (B.8). Applying Theorem 2, it is not difficult to derive that

$$\|\dot{X}\|_F \leq \frac{\|B\|^2(\mu+\eta^2c^2+2c^2\|A\|)}{(\mu-\eta^2c^2)^2}\|\dot{V}\|_F, \tag{A.3}$$

$$\|\dot{R}_r\|_F \leq \frac{2\|B\|^2(\mu+\eta^2c^2+2c^2\|A\|)}{(\mu-\eta^2c^2)^2}\|\dot{V}\|_F. \tag{A.4}$$

In the same way, it can be computed that

$$\|\dot{Y}\|_F \leq \frac{\|C\|^2(\mu+\eta^2c^2+2c^2\|A\|)}{(\mu-\eta^2c^2)^2}\|\dot{V}\|_F, \tag{A.5}$$

$$\|\dot{Q}_r\|_F \leq \frac{2\|C\|^2(\mu+\eta^2c^2+2c^2\|A\|)}{(\mu-\eta^2c^2)^2}\|\dot{V}\|_F. \tag{A.6}$$

Differentiating J_{V1} and computing its Frobenius norm, we obtain

$$\begin{aligned}\|\dot{J}_{V1}\|_F \leq \frac{\|B\|^2\|C\|^2}{(\mu-\eta^2c^2)^3}\Big(&6(\mu-\eta^2c^2+2c^2\|A\|)(\mu+\eta^2c^2+2c^2\|A\|)\\&+2c^2(\mu-\eta^2c^2)^2+4c^2\|A\|(\mu-\eta^2c^2)\Big)\|\dot{V}\|_F.\end{aligned}$$

Now differentiate J_{V2} and compute its Frobenius norm to obtain

$$\|\dot{J}_{V2}\|_F \leq \frac{2c^2\|B\|^2\|C\|^2}{(\mu-\eta^2c^2)^3}\Big(4\eta^2c^2(\mu-\eta^2c^2)+6\eta^2(\mu+\eta^2c^2+2c^2\|A\|)\Big)\|\dot{V}\|_F.$$

Summing them up, we can derive

$$\zeta_2 = \frac{2\|B\|^2\|C\|^2}{(\mu-\eta^2c^2)^3}\Big(3(\mu+\eta^2c^2+2c^2\|A\|)^2+c^2(\mu-\eta^2c^2)(\mu+3\eta^2c^2+2c^2\|A\|)\Big).$$

Note that here, $\dot{V}$ stands for $\dot{\mathfrak{V}}(\alpha)$.

Appendix B: Proof of Corollary 4

The inequality in (3.13) requires up to the third-order derivatives of the objective function $J(\mathfrak{V}(\alpha))$ with respect to α. Denote the matrix $\Phi(\alpha)$ as

$$\Phi(\alpha) = \begin{pmatrix} 0 & X(\alpha) \\ X^\top(\alpha) & R_r(\alpha) \end{pmatrix}.$$

The derivatives of $J(\mathfrak{V}(\alpha))$ are computed as

$$J(V) = \tfrac{1}{2}\mathrm{trace}\left(\mathfrak{C}R_f + \begin{pmatrix} 0 & -\mathfrak{C}V \\ -V^\top\mathfrak{C} & V^\top\mathfrak{C}V \end{pmatrix}\Phi(\alpha)\right), \tag{B.1a}$$

$$\dot{J}(V) = \tfrac{1}{2}\mathrm{trace}\left(\begin{pmatrix} 0 & -\mathfrak{C}\Gamma_k V \\ V^\top\Gamma_k\mathfrak{C} & V^\top\mathcal{L}_1(\mathfrak{C},\Gamma_k)V \end{pmatrix}\Phi(\alpha) + \begin{pmatrix} 0 & -\mathfrak{C}V \\ -V^\top\mathfrak{C} & V^\top\mathfrak{C}V \end{pmatrix}\dot{\Phi}(\alpha)\right), \tag{B.1b}$$

$$\begin{aligned}\ddot{J}(V) = \frac{1}{2}\mathrm{trace}\Bigg(&\begin{pmatrix} 0 & -\mathfrak{C}\Gamma_k^2 V \\ -V^\top\Gamma_k^2\mathfrak{C} & V^\top\mathcal{L}_2(\mathfrak{C},\Gamma_k)V \end{pmatrix}\Phi(\alpha) \\ &+ \begin{pmatrix} 0 & -\mathfrak{C}\Gamma_k V \\ V^\top\Gamma_k\mathfrak{C} & V^\top\mathcal{L}_1(\mathfrak{C},\Gamma_k)V \end{pmatrix}2\dot{\Phi}(\alpha) \\ &+ \begin{pmatrix} 0 & -\mathfrak{C}V \\ -V^\top\mathfrak{C} & V^\top\mathfrak{C}V \end{pmatrix}\ddot{\Phi}(\alpha)\Bigg),\end{aligned} \tag{B.1c}$$

$$\begin{aligned}J^{(3)}(V) = \frac{1}{2}\mathrm{trace}\Bigg(&\begin{pmatrix} 0 & -\mathfrak{C}\Gamma_k^3 V \\ V^\top\Gamma_k\mathfrak{C} & V^\top\mathcal{L}_3(\mathfrak{C},\Gamma_k)V \end{pmatrix}\Phi(\alpha) \\ &+ \begin{pmatrix} 0 & -\mathfrak{C}\Gamma_k^2 V \\ -V^\top\Gamma_k^2\mathfrak{C} & V^\top\mathcal{L}_2(\mathfrak{C},\Gamma_k)V \end{pmatrix}3\dot{\Phi}(\alpha) \\ &+ \begin{pmatrix} 0 & -\mathfrak{C}\Gamma_k V \\ V^\top\Gamma_k\mathfrak{C} & V^\top\mathcal{L}_1(\mathfrak{C},\Gamma_k)V \end{pmatrix}3\ddot{\Phi}(\alpha) \\ &+ \begin{pmatrix} 0 & -\mathfrak{C}V \\ -V^\top\mathfrak{C} & V^\top\mathfrak{C}V \end{pmatrix}\Phi^{(3)}(\alpha)\Bigg),\end{aligned}$$

Note that here, we abbreviate $\mathfrak{V}(\alpha)$ as V for convenience and the matrix Γ_k given by (3.7) is skew-symmetric. Let $W(\alpha) = \begin{pmatrix} X(\alpha)^\top & R_r(\alpha) \end{pmatrix}$ denote the second row of $\Phi(\alpha)$. Since $\|\mathfrak{V}(\alpha)\| = 1$, it is not difficult to derive the following:

$$|J^{(3)}(\alpha)| \le \tfrac{1}{2}\begin{pmatrix} \|W(\alpha)\|_F \\ \|\dot{W}(\alpha)\|_F \\ \|\ddot{W}(\alpha)\|_F \\ \|W^{(3)}(\alpha)\|_F \end{pmatrix}^\top \begin{pmatrix} \|\begin{pmatrix} 2\Gamma_k^3\mathfrak{C} & \mathcal{L}_3(\mathfrak{C},\Gamma_k)\end{pmatrix}\|_F \\ 3\|\begin{pmatrix} 2\mathfrak{C}\Gamma_k^2 & \mathcal{L}_2(\mathfrak{C},\Gamma_k)\end{pmatrix}\|_F \\ 3\|\begin{pmatrix} 2\mathfrak{C}\Gamma_k & \mathcal{L}_1(\mathfrak{C},\Gamma_k)\end{pmatrix}\|_F \\ \sqrt{5}\|\mathfrak{C}\|_F \end{pmatrix}. \tag{B.2}$$

To derive the bound on $|J^{(3)}(\alpha)|$, we need to compute the bound on $\|W(\alpha)\|_F$ and its first three derivatives. Since $W(\alpha)$ depends on the step size, it is reasonable to consider $\max_{0\leq\alpha\leq\tau_k}\|W(\alpha)\|_F$ and its derivatives for some positive scalar τ_k, which varies over the iteration. To do so, we consider the generalized Lyapunov equation associated with $\Phi(\alpha)$ and the derivatives of the generalized Lyapunov equation. Let,

$$A_e = \begin{pmatrix} A & 0 \\ 0 & \mathfrak{V}(\alpha)^\top A\mathfrak{V}(\alpha) \end{pmatrix},\ N_{ej} = \begin{pmatrix} N_j & 0 \\ 0 & \mathfrak{V}(\alpha)^\top N_j\mathfrak{V}(\alpha) \end{pmatrix},$$
$$j = 1, 2, \ldots, m,$$

denote the system matrices of the error system in the kth iteration. The generalized Lyapunov equations associated with $\Phi(\alpha)$, $\dot{\Phi}(\alpha)$, $\ddot{\Phi}(\alpha)$, and $\Phi^{(3)}(\alpha)$ are

$$A_e\Phi + \Phi A_e^\top + \sum_{j=1}^{m} N_{ej}\Phi N_{ej}^\top + \Psi_0 = 0, \tag{B.3}$$

$$A_e\dot{\Phi} + \dot{\Phi}A_e^\top + \sum_{j=1}^{m} N_{ej}\dot{\Phi}N_{ej}^\top + \Psi_1 = 0, \tag{B.4}$$

$$A_e\ddot{\Phi} + \ddot{\Phi}A_e^\top + \sum_{j=1}^{m} N_{ej}\ddot{\Phi}N_{ej}^\top + \Psi_2 = 0, \tag{B.5}$$

$$A_e\Phi^{(3)} + \Phi^{(3)}A_e^\top + \sum_{j=1}^{m} N_{ej}\Phi^{(3)}N_{ej}^\top + \Psi_3 = 0, \tag{B.6}$$

where,

$$\Psi_0 = \begin{pmatrix} 0 & \mathfrak{B}V \\ V^\top\mathfrak{B} & V^\top\mathfrak{B}V \end{pmatrix}, \tag{B.7}$$

$$\begin{aligned}\Psi_1 = &\begin{pmatrix} 0 & \mathfrak{B}\Gamma_k V \\ -V^\top\Gamma_k\mathfrak{B} & V^\top\mathcal{L}_1(\mathfrak{B},\Gamma_k)V \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & V^\top\mathcal{L}_1(A,\Gamma_k)V \end{pmatrix}\Phi \\ &+\Phi\begin{pmatrix} 0 & 0 \\ 0 & V^\top\mathcal{L}_1(A^\top,\Gamma_k)V \end{pmatrix} + \sum_{j=1}^{m}\left(\left(\begin{pmatrix} 0 & 0 \\ 0 & V^\top\mathcal{L}_1(N_j,\Gamma_k)V \end{pmatrix}\Phi N_{ej}^\top\right.\right. \\ &\left.\left.+N_{ej}\Phi\begin{pmatrix} 0 & 0 \\ 0 & V^\top\mathcal{L}_1(N_j^\top,\Gamma_k)V \end{pmatrix}\right)\right),\end{aligned} \tag{B.8}$$

$$
\begin{aligned}
\Psi_2 = &\begin{pmatrix} 0 & \mathfrak{B}\Gamma_k^2 V \\ V^\top \Gamma_k^2 \mathfrak{B} & V^\top \mathcal{L}_2(\mathfrak{B}, \Gamma_k) V \end{pmatrix} + 2 \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(A, \Gamma_k) V \end{pmatrix} \dot{\Phi} \\
&+ 2 \dot{\Phi} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(A^\top, \Gamma_k) V \end{pmatrix} \\
&+ 2 \sum_{j=1}^{m} \left(\begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(N_j, \Gamma_k) V \end{pmatrix} \dot{\Phi} N_{ej}^\top + N_{ej} \dot{\Phi} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(N_j^\top, \Gamma_k) V \end{pmatrix} \right) \\
&+ \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(A, \Gamma_k) V \end{pmatrix} \Phi + \Phi \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(A^\top, \Gamma_k) V \end{pmatrix} \\
&+ \sum_{j=1}^{m} \left(\begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(N_j, \Gamma_k) V \end{pmatrix} \Phi N_{ej}^\top + N_{ej} \Phi \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(N_j^\top, \Gamma_k) V \end{pmatrix} \right) \\
&+ 2 \sum_{j=1}^{m} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(N_j, \Gamma_k) V \end{pmatrix} \Phi \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(N_j^\top, \Gamma_k) V \end{pmatrix}, \qquad \text{(B.9)}
\end{aligned}
$$

$$
\begin{aligned}
\Psi_3 = &\begin{pmatrix} 0 & \mathfrak{B}\Gamma_k^3 V \\ -V^\top \Gamma_k^3 \mathfrak{B} & V^\top \mathcal{L}_2(\mathfrak{B}, \Gamma_k) V \end{pmatrix} + 3 \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(A, \Gamma_k) V \end{pmatrix} \ddot{\Phi} \\
&+ 3 \ddot{\Phi} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(A^\top, \Gamma_k) V \end{pmatrix} \\
&+ 3 \sum_{j=1}^{m} \left(\begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(N_j, \Gamma_k) V \end{pmatrix} \ddot{\Phi} N_{ej}^\top + N_{ej} \ddot{\Phi} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(N_j^\top, \Gamma_k) V \end{pmatrix} \right) \\
&+ 3 \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(A, \Gamma_k) V \end{pmatrix} \dot{\Phi} + 3 \dot{\Phi} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(A^\top, \Gamma_k) V \end{pmatrix} \\
&+ 3 \sum_{j=1}^{m} \left(\begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(N_j, \Gamma_k) V \end{pmatrix} \dot{\Phi} N_{ej}^\top + N_{ej} \dot{\Phi} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(N_j^\top, \Gamma_k) V \end{pmatrix} \right) \\
&+ 6 \sum_{j=1}^{m} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(N_j, \Gamma_k) V \end{pmatrix} \dot{\Phi} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(N_j^\top, \Gamma_k) V \end{pmatrix} \\
&+ \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_3(A, \Gamma_k) V \end{pmatrix} \Phi + \Phi \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_3(A^\top, \Gamma_k) V \end{pmatrix} \\
&+ \sum_{j=1}^{m} \left(\begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_3(N_j, \Gamma_k) V \end{pmatrix} \Phi N_{ej}^\top + N_{ej} \Phi \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_3(N_j^\top, \Gamma_k) V \end{pmatrix} \right) \\
&+ 3 \sum_{j=1}^{m} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(N_j, \Gamma_k) V \end{pmatrix} \Phi \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(N_j^\top, \Gamma_k) V \end{pmatrix} \\
&+ 3 \sum_{j=1}^{m} \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_1(N_j, \Gamma_k) V \end{pmatrix} \Phi \begin{pmatrix} 0 & 0 \\ 0 & V^\top \mathcal{L}_2(N_j^\top, \Gamma_k) V \end{pmatrix}. \qquad \text{(B.10)}
\end{aligned}
$$

Denote the second row of $\Psi_i,\ i = 0, 1, 2, 3$, as $Z_{\Phi_i},\ i = 0, 1, 2, 3$. According to Theorem 2, the bound of $\|W^{(3)}(\alpha)\|_F$ satisfies

$$(\mu - \eta^2 c^2) \max_{0\le\alpha\le\tau_k} \|W^{(3)}(\alpha)\|_F \le c^2 \max_{0\le\alpha\le\tau_k} \|Z_{\Phi_3}(\alpha)\|_F.$$

Now apply *Lemma A*.3 in [27] and let Ω_k denote $Z_{\Phi_3}(0)$, then we obtain the following:

$$\begin{aligned}
&(\mu - \eta^2 c^2) \max_{0\le\alpha\le\tau_k} \|W^{(3)}(\alpha)\|_F \\
\le\ & c^2\|\Omega_k\|_F + c^2\tau_k \| \left(\Gamma_k^4 \mathfrak{B}\ \mathcal{L}_4(\mathfrak{B}, \Gamma_k)\right) \|_F \\
&+2c^2\tau_k \Bigg(\beta_0 \max_{0\le\alpha\le\tau_k} \|W(\alpha)\|_F + \beta_1 \max_{0\le\alpha\le\tau_k} \|\dot{W}(\alpha)\|_F \\
&+ \beta_2 \max_{0\le\alpha\le\tau_k} \|\ddot{W}(\alpha)\|_F + \beta_3 \max_{0\le\alpha\le\tau_k} \|W^{(3)}(\alpha)\|_F \Bigg),
\end{aligned} \tag{B.11}$$

where $\beta_i, i = 0, 1, 2, 3$ are given by (3.16) to (3.20). Again, repetitively applying *Lemma A*.3 in [27], we derive that

$$\begin{aligned}
\max_{0\le\alpha\le\tau_k} \|W(\alpha)\|_F &\le \|W(0)\|_F + \tau_k \max_{0\le\alpha\le\tau_k} \|\dot{W}(\alpha)\|_F, \\
\max_{0\le\alpha\le\tau_k} \|\dot{W}(\alpha)\|_F &\le \|W(0)\|_F + \tau_k \max_{0\le\alpha\le\tau_k} \|\ddot{W}(\alpha)\|_F, \\
\max_{0\le\alpha\le\tau_k} \|\ddot{W}(\alpha)\|_F &\le \|W(0)\|_F + \tau_k \max_{0\le\alpha\le\tau_k} \|W^{(3)}(\alpha)\|_F.
\end{aligned}$$

Substituting the above inequalities to (B.11), it can be obtained that

$$\begin{aligned}
&(\mu - \eta^2 c^2 - 2\beta_0\tau_k^4 c^2 - 2\beta_1\tau_k^3 c^2 - 2\beta_2\tau_k^2 c^2 - 2\beta_3\tau_k c^2) \max_{0\le\alpha\le\tau_k} \|W^{(3)}(\alpha)\|_F \\
\le\ & a\ positive\ number.
\end{aligned}$$

To make sure that $\max_{0\le\alpha\le\tau_k} \|W^{(3)}(\alpha)\|_F$ is bounded, it must hold that

$$\mu - \eta^2 c^2 > 2\beta_0\tau_k^4 c^2 + 2\beta_1\tau_k^3 c^2 + 2\beta_2\tau_k^2 c^2 + 2\beta_3\tau_k c^2.$$

It is also not difficult to show that the upper bound of $W(\alpha)$ and its first three derivatives satisfy a system of linear inequalities as follows:

$$\begin{pmatrix} 1 & -\tau_k & 0 & 0 \\ 0 & 1 & -\tau_k & 0 \\ 0 & 0 & 1 & -\tau_k \\ -2\beta_0\tau_k c^2 & -2\beta_1\tau_k c^2 & -2\beta_2\tau_k c^2 & \mu - \eta^2 c^2 - 2\beta_3\tau_k c^2 \end{pmatrix} \begin{pmatrix} \max_{0\le\alpha\le\tau_k} \|W(\alpha)\|_F \\ \max_{0\le\alpha\le\tau_k} \|\dot{W}(\alpha)\|_F \\ \max_{0\le\alpha\le\tau_k} \|\ddot{W}(\alpha)\|_F \\ \max_{0\le\alpha\le\tau_k} \|W^{(3)}(\alpha)\|_F \end{pmatrix} \le \begin{pmatrix} \|W(0)\|_F \\ \|\dot{W}(0)\|_F \\ \|\ddot{W}(0)\|_F \\ \|\Omega_k\|_F + \| \left(\Gamma_k^4 \mathfrak{B} \;\; \mathcal{L}_4(\mathfrak{B}, \Gamma_k)\right) \|_F \end{pmatrix}. \tag{B.12}$$

By following the proofs of *Lemma* 4.2 and 4.3 in [27], the proof can be completed.

References

1. Absil, P.A., Mahony, R., Sepulchre R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2009)
2. Baars, S., Viebahn, J., Mulder, T., Kuehn, C., Wubs, F.W., Dijkstra, H.A.: Continuation of probability density functions using a generalized Lyapunov approach. J. Comput. Phys. **336**, 627–643 (2017)
3. Benner, P., Breiten, T.: On $\mathcal{H}_2$-model reduction of linear parameter-varying systems. PAMM **11**(1), 805–806 (2011)
4. Benner, P., Breiten, T.: Interpolation-based $\mathcal{H}_2$-model reduction of bilinear control systems. SIAM J. Matrix Anal. Appl. **33**(3), 859–885 (2012)
5. Benner, P., Damm, T.: Lyapunov equations, energy functionals, and model order reduction of bilinear and stochastic systems. SIAM J. Control. Optim. **49**(2), 686–711 (2011)
6. Benner, P., Gugercin, S., Willcox, K.: A survey of projection-based model reduction methods for parametric dynamical systems. SIAM Rev. **57**(4), 483–531 (2015)
7. Breiten, T., Damm, T.: Krylov subspace methods for model order reduction of bilinear control systems. Syst. Control Lett. **59**(8), 443–450 (2010)
8. Bruns, A., Benner, P.: Parametric model order reduction of thermal models using the bilinear interpolatory rational Krylov algorithm. Math. Comput. Model. Dyn. Syst. **21**(2), 103–129 (2015)
9. Bruns, A.S.: Bilinear $\mathcal{H}_2$-Optimal Model Order Reduction with Applications to Thermal Parametric Systems. PhD thesis, Otto-von-Guericke Universität Magdeburg (2015)
10. Castañé Selga, R.: The Matrix Measure Framework for Projection-Based Model Order Reduction. PhD thesis, Technische Universität München (2011)
11. D'Alessandro, P., Isidori, A., Ruberti, A.: Realization and structure theory of bilinear dynamical systems. SIAM J. Control. **12**(3), 517–535 (1974)
12. Damm, T.: Direct methods and ADI-preconditioned Krylov subspace methods for generalized Lyapunov equations. Numerical Linear Algebra with Applications **15**(9), 853–871 (2008)
13. Dorissen, H.: Canonical forms for bilinear systems. Syst. Control Lett. **13**(2), 153–160 (1989)
14. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**(2), 303–353 (1998)
15. Flagg, G., Gugercin, S.: Multipoint Volterra series interpolation and $\mathcal{H}_2$ optimal model reduction of bilinear systems. SIAM J. Matrix Anal. Appl. **36**(2), 549–579 (2015)
16. Jost, J.: Riemannian Geometry and Geometric Analysis, vol. 42005. Springer, Berlin (2008)
17. Mohler, R.R.: Nonlinear Systems (Vol 2): Applications to Bilinear Control. Prentice-Hall, Inc (1991)

18. MORwiki-Community: MORwiki - Model Order Reduction Wiki. http://modelreduction.org (2018)
19. Negri, F., Manzoni, A., Amsallem, D.: Efficient model reduction of parametrized systems by matrix discrete empirical interpolation. J. Comput. Phys. **303**, 431–454 (2015)
20. Nijmeijer, H., Van der Schaft, A.: Nonlinear Dynamical Control Systems, vol. 175. Springer, Berlin (1990)
21. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, USA (1999)
22. Sato, H., Sato, K.: Riemannian trust-region methods for H^2 optimal model reduction. In: 2015 IEEE 54th Annual Conference on Decision and Control (CDC), pp 4648–4655, IEEE (2015)
23. Sato, H., Sato, K.: A New H^2 optimal model reduction method based on riemannian conjugate gradient method. In: 2016 IEEE 55th Annual Conference on Decision and Control (CDC), pp. 5762–5768, IEEE (2016)
24. Shank, S.D., Simoncini, V., Szyld, D.B.: Efficient low-rank solution of generalized Lyapunov equations. Numer. Math. **134**(2), 327–342 (2016)
25. Stykel, T., Simoncini, V.: Krylov subspace methods for projected Lyapunov equations. Appl. Numer. Math. **62**(1), 35–50 (2012)
26. Xu, Y., Zeng, T.: Fast optimal $\mathcal{H}_2$ model reduction algorithms based on Grassmann manifold optimization. Int. J. Numer. Anal. Model. **10**, 972–991 (2013)
27. Yan, W.Y., Lam, J.: An approximate approach to H^2 optimal model reduction. IEEE Trans. Autom. Control **44**(7), 1341–1358 (1999)
28. Zhang, L., Lam, J.: On H_2 model reduction of bilinear systems. Automatica **38**(2), 205–216 (2002)