An Approximate Approach to H^2 Optimal Model Reduction

Wei-Yong Yan and James Lam, *Senior Member, IEEE*

***Abstract*—This paper deals with the problem of computing an H_2 optimal reduced-order model for a given stable multivariable linear system. By way of orthogonal projection, the problem is formulated as that of minimizing the H_2 model-reduction cost over the Stiefel manifold so that the stability constraint on reduced-order models is automatically satisfied and thus totally avoided in the new problem formulation. The closed form expression for the gradient of the cost over the manifold is derived, from which a gradient flow results as an ordinary differential equation (ODE). A number of nice properties about such a flow are established. Furthermore, two explicit iterative convergent algorithms are developed from the flow; one has a constant step-size and the other has a varying step-size and is much more efficient. Both of them inherit the properties that the iterates remain on the manifold starting from any orthogonal initial point and that the model reduction cost is decreasing to minima along the iterates. A procedure for closing the gap between the original and modified problem is proposed. In the symmetric case, the two problems are shown to be equivalent. Numerical examples are presented to illustrate the effectiveness of the proposed algorithms as well as convergence.**

***Index Terms*—Linear systems, model reduction, optimization, system approximation.**

I. Introduction

A LOWER order approximation to a high-order system is often desirable and used in practice. Among many developed methods for order reduction are the balanced truncation method [1], [2] and the Hankel norm approximation method [3] with the former extended to the case of second-order form linear systems by Meyer and Srinivasan [4] most recently. Also quite recently, Zhou [5] has proposed a new method based on these two methods for L_∞ norm model reduction with some L_∞ error bounds derived. However, the problem of finding an optimal reduced-order model in the H_2 or H_∞ sense is still largely open.

The present paper is concerned with minimizing the H_2 (also called L_2) norm of the model mismatch between a given model and a reduced-order one. This minimization problem has received a great deal of attention over the past several decades. The importance of the problem stems from the fact that the H_2 norm of a system is the expected root-mean-square (rms) value of the output when the input is a unit variance white noise process.

However, rigorous and convergent algorithms have remained to be found in the general multi-input/multi-output (MIMO) case. So far, the most commonly taken approach to H_2-optimal model reduction problem is to work with first-order necessary conditions for optimality, which were developed and simplified in one way or another by Meier and Luenberger [6], Wilson [7], Hyland and Bernstein [8], Halevi [9], Bryson and Carrier [10], Baratchart *et al.* [11], and more recently Spanos *et al.* [12]. Accordingly, they proposed their respective algorithms to seek a solution satisfying the conditions expressed in terms of nonlinear matrix equations. Many of the algorithms lack the proof of convergence and mathematical rigor, and some of them may even become divergent for certain initial conditions or converge to a maximum. Though Baratchart *et al.* [11] and Spanos *et al.* [12] established the convergence of their respective algorithms under certain conditions, the algorithms are only applicable to the single-input/single-output (SISO) case.

So far, it is unclear whether the global minimum of the cost exists or not in the continuous-time MIMO case, though the answer to this question in the discrete-time case was positive according to Baratchart [13]. This issue inevitably sheds some doubt on the theoretic basis of the above approach. Moreover, as pointed out by Spanos *et al.* [12], there are two technical difficulties associated with the approach; one is the stability constraint on reduced-order models and the other is the unboundedness of the level sets of the H_2 cost functional. It goes without saying that the first one is fundamentally intricate to accommodate and thus represents a major obstacle to the effectiveness of any algorithm based on that approach. We believe that this difficulty is due to direct parameterizations of all the reduced-order models in one form or another.

In this paper, we take a different approach to the H_2-optimal model-reduction problem in the continuous-time case. The main idea is to treat the minimization problem over a subclass of stable reduced-order models parameterized by a projection matrix instead of the whole class of all the reduced-order models. The restriction to this subclass enables one to avoid the stability constraint entirely and leads to a tractable minimization problem over the Stiefel manifold, which is compact. In addition, the global minimum is guaranteed to exist over the subclass. Our main purpose is to develop both continuous and iterative convergent algorithms which are rigorous and universally applicable.

Manuscript received July 21, 1995; revised July 8, 1996, July 23, 1997, and August 28, 1998. Recommended by Associate Editor, A. Vicino. This work was supported in part by the Australian Research Council, Curtin R&D Office, and RGC Grant HKU 544/96E.

W.-Y. Yan is with the School of Electrical and Computer Engineering, Curtin University of Technology, Perth WA 6845, Australia (e-mail: wyy@wdc.ece.curtin.edu.au).

J. Lam is with the Department of Mechanical Engineering, University of Hong Kong, Hong Kong.

Publisher Item Identifier S 0018-9286(99)05460-4.

Even though it is theoretically unclear about the exact mismatch between the original problem and the approximation at this point, the problem approximation proposed in the paper appears to be appealing and promising due to several reasons in addition to tractability and the ability to allow algorithms with proven convergence. First, any solution to the modified problem has the property that it is most compatible with the original model in the sense that the commutative diagram formed by them is least incompatible, a property which seems desirable on system-theoretic grounds. Second, an extensive numerical investigation reveals that the global optimum associated with the modified problem is consistently close, if not identical, to that associated with the original problem for a variety of examples treated in other papers, which indicates that the problem simplification may not sacrifice the global optimality in any significant way. Moreover, any possible conservativeness due to the approximation is limited by an *a priori* error bound derived in the paper. Third, the original and approximate problems turn out to be equivalent in the symmetric case involving relaxation systems. Finally, a high level procedure can be proposed to bridge the gap between the two problems by converting a minimum of the approximate problem into a minimum of the original problem.

The paper is briefly outlined as follows. In the next section, we modify the H_2-optimal model-reduction problem as an unconstrained minimization problem over the Stiefel manifold. Section III centers on the development of the gradient flow of the model-reduction cost and establishment of its associated properties including convergence by using differential manifold techniques. In Section IV, we turn to derive two recursive algorithms with detailed convergence analysis. In Section V, an upper bound on a minimum of the problem is derived, the symmetric case is treated, and a complete algorithm is described. In Section VI, we test our algorithms on a number of well-known examples and compare our results with those obtained by other methods. The last section contains some conclusions.

II. Problem Formulation

Consider a linear time-invariant stable system $G(s)$ with the realization

$$\dot{x} = Ax + Bu \tag{1}$$

$$y = Cx \tag{2}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$, $C \in \mathbb{R}^{q\times n}$. An admissible reduced-order model $G_m(s)$ is defined to be of the form

$$\dot{x}_m = A_m x_m + B_m u \tag{3}$$

$$y_r = C_m x_m \tag{4}$$

where $A_m \in \mathbb{R}^{m\times m}$, $B_m \in \mathbb{R}^{m\times p}$, $C_m \in \mathbb{R}^{q\times m}$ with A_m a stable matrix. The mismatch between the full-order $G(s)$ and reduced-order $G_m(s)$ will be measured by the square of the H_2 norm of their difference $G_e(s)$, i.e.,

$$||G_e(s)||_2^2$$

which is often termed the quadratic model-reduction cost.

The so-called H_2 or quadratically optimal model-reduction problem is to minimize the above cost over all the admissible reduced-order models $G_m(s)$. Note that one realization (A_e, B_e, C_e) of the error model $G_e(s)$ is given by

$$(A_e, B_e, C_e) = \left(\begin{bmatrix} A & 0 \\ 0 & A_m \end{bmatrix}, \begin{bmatrix} B \\ B_m \end{bmatrix}, [C \quad -C_m]\right).$$

Then, it is a standard fact that the cost can be conveniently expressed in terms of the controllability gramian L_c and observability gramian L_o of this realization. Namely, there holds

$$\begin{aligned} ||G_e(s)||_2^2 &= J(A_m, B_m, C_m) \\ &\triangleq \operatorname{trace}\left(C_e L_c C_e^T\right) \\ &= \operatorname{trace}\left(B_e^T L_o B_e\right) \end{aligned} \tag{5}$$

with

$$A_e L_c + L_c A_e^T + B_e B_e^T = 0 \tag{6}$$

$$A_e^T L_o + L_o A_e + C_e^T C_e = 0. \tag{7}$$

Remark 2.1: For a certain purpose which will become clearer, we need to stress that $J(A_m, B_m, C_m)$ is not defined to be the model-reduction cost. Rather, it is defined in terms of the gramians L_c and L_o. One easily overlooked difference between $J(A_m, B_m, C_m)$ and $||G_e(s)||_2^2$ is that they do not share the same domain of definition though they are equal for any admissible reduced-order model. An implication of this in relation to the issue of stability will be explained shortly.

It is known from [7] and [8] that any minimizing solution (A_m, B_m, C_m) must be of the form

$$(A_m, B_m, C_m) = (TAV, TB, CV) \tag{8}$$

where $V \in \mathbb{R}^{n\times m}$ and $T \in \mathbb{R}^{m\times n}$ satisfy

$$TV = I. \tag{9}$$

Hence, the original model-reduction problem amounts to minimizing $J(TAV, TB, CV)$ with respect to $(T, V) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m}$ subject to the two constraints

$$1)\ TV = I \quad \text{and} \quad 2)\ TAV \text{ is stable.}$$

This is essentially a nonlinear optimization problem subject to both equality and inequality constraints as the stability constraint can be expressed in terms of inequalities by the Hurwitz criterion. Though it may be possible to use some constrained optimization techniques to find a local minimum, the computation involved could be formidable. To our best knowledge, no convergent and generally applicable algorithm for solving the problem has been found by now.

To formulate a more tractable problem, we observe that T given by

$$T = V^{\dagger} = \left(V^T V\right)^{-1} V^T \tag{10}$$

and V given by

$$V = T^{\dagger} = T^T \left(TT^T\right)^{-1} \tag{11}$$

satisfy constraint (9) for any V of full column rank and any T of full row rank, respectively. It is therefore interesting to consider the following modified problem:

$$\begin{aligned}\text{minimize}\quad & J(V) \triangleq J\left(V^\dagger AV,\, V^\dagger B,\, CV\right)\\ & \text{over } V \in \mathbb{R}^{n\times m}\\ \text{subject to stability of } & V^\dagger AV.\end{aligned}$$

Remark 2.2: Admittedly, the modified problem represents an approximation to the original problem as the new set of reduced-order models over which the model-reduction cost is minimized is a subset of the original set. Nevertheless, the numerical results to be given later on suggest that the approximation tends to be sufficient for the purpose of finding the global minimum. This does not seem surprising because any minimizing solution to the original problem falls into the new model set associated with a certain realization. In fact, the necessary condition (9) for the optimality apparently implies the existence of a similarity transformation Θ such that

$$T = [I \quad 0]\Theta \quad \text{and} \quad V = \Theta^{-1}\begin{bmatrix} I \\ 0 \end{bmatrix}.$$

Consequently, the minimizing solution $(A_m,\, B_m,\, C_m)$ becomes the truncation of the new realization $(\Theta A\Theta^{-1},\, \Theta B,\, C\Theta^{-1})$ for the original system. This fact will be exploited in Section V.

Also, the above modification can be motivated from a geometric point of view. To see this, we decompose the state space $\mathbb{R}^n$ into

$$\mathbb{R}^n = \text{range}(V) + \text{range}(V)^\perp.$$

That is, any state $x \in \mathbb{R}^n$ is expressed as

$$x = Vw + e$$

with $w \in \mathbb{R}^m$ and $e \in \text{range}(V)^\perp$. Here, Vw is the orthogonal projection of the state onto the subspace $\text{range}(V)$. By rewriting the state equation as

$$V\dot{w} - (AVw + Bu) = Ae - \dot{e}$$

and appealing to the fact that $\|Vz - s\|_2$ is minimized at $z = V^\dagger s$ for any given $s \in \mathbb{R}^n$, one sees that the best approximate to $\dot{w}$ given w is $V^\dagger(AVw + Bu)$ in the sense that e has the minimal effect. Removing e from the output equation naturally results in a reduced-order model

$$\dot{w} = V^\dagger AVw + V^\dagger Bu \tag{12}$$

$$y = CVw. \tag{13}$$

As such, the above-modified minimization problem may well be thought of as finding a dominant state subspace of dimension m, which is spanned by the columns of V. As a matter of fact, such a projection idea shares the same underlying principle with the method of aggregation [14]–[16].

In order to provide a system-theoretic interpretation of the replacement (11), let us consider an arbitrary state-space realization $(\overline{A},\, \overline{B},\, \overline{C})$ on the subspace $\overline{\mathfrak{X}}$ resulting from the projection T. Obviously, if T happens to be a similarity transformation, the triple $(\overline{A},\, \overline{B},\, \overline{C})$ is uniquely determined as follows by the commutative diagram in Fig. 1:

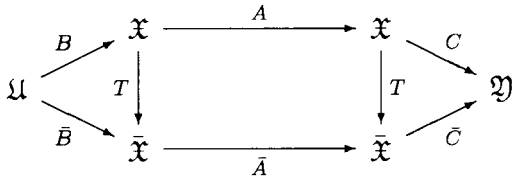


Fig. 1. Commutative diagram.

$$\overline{A} = TAT^{-1}, \qquad \overline{B} = TB, \qquad \overline{C} = CT^{-1}.$$

More precisely, commutativity of the diagram is equivalent to the requirement that the two minimal realizations represent the same system if they are of the same order. In the other case where the two realizations do not have the same order, the above diagram will no longer be commutative. Accordingly, the two realizations will not be compatible. Logically, the extent to which the diagram is not commutative reflects the degree of incompatibility between the two realizations. It obviously makes sense to measure the former by three matrix spectral norms

$$\|\overline{A}T - TA\|, \qquad \|\overline{B} - TB\|, \qquad \|\overline{C}T - C\|.$$

Since these norms are simultaneously minimized by the unique triple

$$(\overline{A},\, \overline{B},\, \overline{C}) = \left(TAT^\dagger,\, TB,\, CT^\dagger\right)$$

given the projection T, the corresponding realization could be regarded as least incompatible with the original one. Hence, (11) is nothing but to restrict the minimization to being over all the low-order models least incompatible with the full-order system in the above sense.

Perhaps it is also interesting and relevant to note that the model-reduction cost can be expressed as

$$\begin{aligned}&J(TAV,\, TB,\, CV)\\ &\quad = \left\|sC(sI - VTA)^{-1}(VT - I)(sI - A)^{-1}B\right\|_2^2.\end{aligned} \tag{14}$$

This implies that the cost only depends on the product VT and contains as one factor $VT - I$ whose norm is minimized at $T = V^\dagger$ as a function of T.

Another crucial implication of the above observation is that the modified problem is equivalent to the minimization problem over a much smaller set. To see this, note that

$$J\left(V^\dagger AV,\, V^\dagger B,\, CV\right) = J\left(U^T AU,\, U^T B,\, CU\right) \tag{15}$$

with $U = V(V^TV)^{-1/2}$ due to

$$UU^T = V\left(V^TV\right)^{-1}V^T = VV^\dagger.$$

Thus, the minimal model-reduction cost over the reduced-order model set

$$\begin{aligned}\Big\{(A_m,\, B_m,\, C_m) = \left(V^\dagger AV,\, V^\dagger B,\, CV\right) \;\Big|\; & V \in \mathbb{R}^{n\times m}\\ \text{and } V^\dagger AV \text{ is stable}\Big\}\end{aligned} \tag{16}$$

is exactly equal to that over

$$\{(A_m, B_m, C_m) = (U^TAU, U^TB, CU) | U \in St(m, n) \text{ and } U^TAU \text{ is stable}\} \quad (17)$$

where $St(m, n)$ is the so-called Stiefel manifold defined by

$$St(m, n) = \{U \in \mathbb{R}^{n\times m} | U^TU = I\}.$$

Since the latter model set is much smaller than the former one and is actually a compact set, the minimization over such a set is likely to lead to the global minimum more quickly and the associated computation may be less expensive.

Quite evidently, the stability constraint computationally hinders the search for local minima though narrowing the set to be searched. We notice that Spanos *et al.* [12] impose a certain line search condition on their algorithms in order to maintain the stability of the iterates.

To overcome this difficulty, we observe that the stability constraint becomes superfluous when the original realization (A, B, C) is such that $A + A^T$ is negative definite since $U^T(A + A^T)U$ remains negative definite for any U on the Stiefel manifold. Therefore, in this case one is led to the minimization problem over the following set:

$$\{(A_m, B_m, C_m) = (U^TAU, U^TB, CU) | U \in St(m, n)\}. \quad (18)$$

Moreover, this problem is guaranteed to have the global minimum because the set in (18) is compact, and it is equivalent to the minimization problem over the much larger reduced-order model set in (16). Furthermore, the choice of a realization with $A + A^T < 0$ turns out to be very simple. In fact, since A is stable, for any symmetric matrix $Q > 0$ there exists an infinite number of nonsingular matrix T such that

$$ATT^T + TT^TA^T = Q$$

i.e.,

$$T^{-1}AT + (T^{-1}AT)^T = T^{-1}Q(T^{-1})^T. \quad (19)$$

From this, it is plain that the use of any such T as a similarity transformation will result in a new realization with the required property.

Remark 2.3: Note that the property $A+A^T < 0$ is nothing but the strict dissipativity of the realization. In addition, a realization in modal form is also strictly dissipative.

Based on and motivated by the above discussion, we now formally pose the following approximate model-reduction problem.

Given a realization (1) and (2) with $A+A^T < 0$, minimize

$$\mathfrak{J}(U) \triangleq J(U^TAU, U^TB, CU) \quad (20)$$

over the Stiefel manifold $St(m, n)$.

III. Gradient Flow on Manifold

In this section, we aim to solve the approximate problem posed in the last section using the gradient flow approach. Recall that an optimal solution to this problem exists. So the question is really how to find one. Also, recall that there is no loss of generality in assuming that $A + A^T$ is negative definite for the original realization (1) and (2), which will be our standing assumption throughout. In addition, we adopt the convention that $\|\cdot\|$ means the spectral norm of a matrix, i.e., the maximum singular value while $\|\cdot\|_F$ means the Frobenius norm.

Let us first obtain a more explicit formula for $\mathfrak{J}(U)$. To do this, partition the solutions L_c and L_o to the Lyapunov equations (6) and (7) as

$$L_c = \begin{bmatrix} \Sigma_c & X \\ X^T & P \end{bmatrix} \quad \text{and} \quad L_o = \begin{bmatrix} \Sigma_o & Y \\ Y^T & Q \end{bmatrix}. \quad (21)$$

As a result, the Lyapunov equations (6) and (7) become equivalent to

$$A\Sigma_c + \Sigma_cA^T + BB^T = 0 \quad (22)$$
$$AX + XU^TA^TU + BB^TU = 0 \quad (23)$$
$$U^TAUP + PU^TA^TU + U^TBB^TU = 0 \quad (24)$$
$$A^T\Sigma_o + \Sigma_oA + C^TC = 0 \quad (25)$$
$$A^TY + YU^TAU - C^TCU = 0 \quad (26)$$
$$U^TA^TUQ + QU^TAU + U^TC^TCU = 0 \quad (27)$$

and the cost $\mathfrak{J}(U)$ can be rewritten as

$$\mathfrak{J}(U) = \text{trace}[C^TC(\Sigma_c + UPU^T - 2XU^T)] \quad (28)$$
$$= \text{trace}[BB^T(\Sigma_o + UQU^T + 2YU^T)]. \quad (29)$$

Quite obviously, $\mathfrak{J}(U)$ is a smooth function on the manifold $St(m, n)$. From [18] or [19], its tangent space at a given $U \in St(m, n)$ is known to be

$$T_USt(m, n) = \{\Pi \in \mathbb{R}^{n\times m} | \Pi^TU + U^T\Pi = 0\}.$$

By endowing $T_USt(m, n)$ with the inner product defined by

$$\langle \eta, \xi \rangle \triangleq 2\,\text{trace}(\eta^T\xi), \qquad \text{for } \eta, \xi \in T_USt(m, n).$$

$St(m, n)$ becomes a Riemannian manifold. Also, note that the derivative $D\mathfrak{J}_U$ of $\mathfrak{J}(U)$ at $U \in St(m, n)$ is a linear functional on the tangent space $T_USt(m, n)$ and that the gradient $\nabla\mathfrak{J}(U)$ of $\mathfrak{J}(U)$ at $U \in St(m, n)$ is a tangent vector in $T_USt(m, n)$ such that

$$D\mathfrak{J}_U(\Pi) = \langle \nabla\mathfrak{J}(U), \Pi \rangle, \qquad \forall\, \Pi \in T_USt(m, n).$$

The explicit expression of $\nabla\mathfrak{J}(U)$ is now given in the following lemma.

Lemma 3.1: For any $U \in St(m, n)$, there holds

$$\nabla\mathfrak{J}(U) = (I - UU^T)R$$

where

$$R \triangleq (-C^TC + A^TUY^T)X + (C^TCU + A^TUQ)P + (BB^T + AUX^T)Y + (BB^TU + AUP)Q. \quad (30)$$

Proof: See Appendix B.

At this point, it is worth pointing out that the above gradient is different from the gradient of $\mathfrak{J}(U)$ as a usual function defined on $\mathbb{R}^{n\times m}$.

As an immediate consequence of the above lemma, it follows from advanced calculus that any minimum point of $\mathfrak{J}(U)$ in $St(m, n)$ must satisfy

$$\big(I - UU^T\big)R = 0 \quad \text{and} \quad U^TU = I \tag{31}$$

since any solution in $St(m, n)$ is a critical point of $\mathfrak{J}(U)$. So (31) expresses a first-order necessary condition for a minimum point. However, solving such an equation does not seem to be a sensible or effective way to go about finding a minimum point as it may be very difficult to solve and may have multiple solutions.

Remark 3.1: It can be verified that U^TR is always a symmetric matrix for any $U \in St(m, n)$, which is instrumental in constructing iterative algorithms later. In fact, there holds

$$U^TR = Y^TAX + QU^TAUP + X^TA^TY + PU^TA^TUQ.$$

Therefore, the first equation of (31) can be expressed as

$$R = UR^TU.$$

Now with the formula for $\nabla\mathfrak{J}(U)$ available, we can form the following gradient flow:

$$\dot{U} = \big(UU^T - I\big)R \tag{32}$$

as a basis for solving the problem of minimizing the model-reduction cost. Regarding this ordinary differential equation, it is natural to inquire questions such as whether a solution to the ordinary differential equation (ODE) always exists and lies on the manifold $St(m, n)$ on the whole time interval for any given initial value in $St(m, n)$, how the model-reduction cost evolves along a solution, and whether the solution can converge to a critical point of $\mathfrak{J}(U)$ on $St(m, n)$. The answers to these questions are crucial in order for the ODE to be able to serve as a continuous-time algorithm for computing an optimal reduced-order model. We now address the raised issues by stating the following theorem, which summarizes the main features of the gradient flow.

Theorem 3.1: Let the initial condition of (32) be given by

$$U(0) = U_0 \in St(m, n).$$

Then, we have the following.

1) The ODE (32) has a unique solution $U(t)$ defined for all $t \geq 0$.
2) The solution $U(t)$ stays in $St(m, n)$ for all $t \geq 0$.
3) The cost $\mathfrak{J}(U)$ is nonincreasing along $U(t)$ with
$$\mathfrak{J}(U(s_2)) - \mathfrak{J}(U(s_1)) = -2\int_{s_1}^{s_2} \big\|\big(I - UU^T\big)R\big\|_F^2\, dt, \qquad \forall\, s_2 \geq s_1 \geq 0.$$
4) There holds
$$\lim_{t\to\infty} \dot{U}(t) = \lim_{t\to\infty} \big(UU^TR - R\big) = 0.$$
5) The solution $U(t)$ converges to a connected component of the set of critical points of $\mathfrak{J}(U)$.
6) There exists a time sequence $\{s_k\}$ with
$$s_k \geq 0 \quad \text{and} \quad \lim_{k\to\infty} s_k = \infty$$
such that the corresponding sequence $U(s_k)$ converges to a critical point of $\mathfrak{J}(U)$.

Proof: The first two statements follow from the compactness properties of the Stiefel manifold. In fact, it is straightforward to verify that the derivative of $U^T(t)U(t)$ is identically zero for all $t \geq 0$. Statement 3) is immediately obtained by noting that the derivative of $\mathfrak{J}(U(t))$ is equal to

$$\begin{aligned}\dot{\mathfrak{J}}(U(t)) &= \langle \nabla\mathfrak{J}(U), \dot{U}\rangle \\ &= -2\,\mathrm{trace}\big[R^T\big(I - UU^T\big)R\big] \\ &= -2\big\|\big(I - UU^T\big)R\big\|_F^2 \leq 0.\end{aligned}$$

Statement 4) is due to the two facts—finiteness of the integral $\int_0^\infty \|(I - UU^T)R\|_F^2\, dt$ and uniform continuity of $\dot{U}(t)$ on $[0, \infty)$. Finally, the last two statements are typical properties of a gradient flow on a Riemannian manifold.

The above summarized properties of the gradient flow (32) give us confidence in finding a minimum of $\mathfrak{J}(U)$ by integrating the differential equation, which can be done using any numerical ODE package, e.g., in Matlab. Since the model-reduction cost is getting smaller and smaller as the iteration goes on and no finite escape time will occur, one can keep on solving the ODE until a satisfactory suboptimal solution is reached. Moreover, the last two statements suggest that a minimum point could be found from the solution history. In particular, it is guaranteed that if the cost has only isolated minimum points, the solution $U(t)$ is bound to converge to one of them.

Remark 3.2: It should be pointed out that if the initial U_0 does not happen to be a critical point, then the cost $\mathfrak{J}(U)$ is actually strictly decreasing along the ODE solution $U(t)$, which is because of the uniqueness of solutions to an ODE.

Remark 3.3: Note that the assumption $A + A^T < 0$ has only been used to guarantee that the ODE (32) has no finite escape time. Without this assumption, the solution still exists for at least some finite time provided the initial condition U_0 is such that $U_0^TAU_0$ is stable.

IV. Iterative Gradient Flow

In this section, we will consider discretizing the gradient flow (32), which is necessary or desirable in order to take full advantage of digital computers as far as computation is concerned. In other words, we will seek iterative algorithms which can produce a sequence of iterates whose corresponding model-reduction costs are decreasing to its minimum. Recall that the projection matrix U is required to be orthogonal. This restriction makes it difficult if not impossible to apply common discretizing techniques such as Runge–Kutta methods to derive an efficient iterative algorithm.

In what follows, a general form of iterative algorithm will first be suggested which automatically guarantees that all the iterates generated evolve on the manifold $St(m, n)$ for an

arbitrary step-size. Two schemes for selecting the step-size will then be developed—one is constant and the other is varying and more effective.

We start by noting that the gradient flow can be rewritten as

$$\dot{U} = \Gamma U \tag{33}$$

because of Remark 3.1, where Γ is defined by

$$\Gamma = UR^T - RU^T. \tag{34}$$

In addition, it is trivial but vital to observe that Γ is skew-symmetric. As a result, the matrix exponential $e^{t\Gamma}$ is orthogonal for any real scalar t. With this observation and the special structure of the gradient flow, it seems natural to propose the algorithm of the following form:

$$U_{k+1} = e^{t_k \Gamma_k} U_k \tag{35}$$

where Γ_k is associated with U_k via (30) and (34), and t_k is the kth step-size to be determined. One nice thing about this algorithm is its ability to generate a sequence of orthogonal matrices from any starting orthogonal U_0 for any step-size, and another is its simplicity in form in spite of the involved calculation of the matrix exponential. Of course, for such an algorithm to work, it remains to develop a mechanism for selecting the step-size t_k so that the algorithm can converge to an orthogonal U at which the model-reduction cost is minimum. As will be determined, a certain constant step-size can be chosen for this purpose.

Understandably, a workable step-size should consistently reduce the model-reduction cost as the iteration goes on. With this in mind, we proceed by establishing the following auxiliary lemma before coming up with a scheme for choosing a constant step-size.

Lemma 4.1: Consider (22)–(27). Let $U \in St(m, n)$ be any differentiable function of t with the derivative U', and let R be defined by (30) accordingly. Then R and its derivative R' satisfy

$$\|R(t)\|_F \le \alpha_1 \tag{36}$$

$$\|R'(t)\|_F \le \alpha_2 \|U'(t)\|_F \tag{37}$$

where

$$\alpha_1 \triangleq \frac{4\sqrt{m}\|B\|^2\|C\|^2(\alpha + \|A\|)}{\alpha^2} \tag{38}$$

$$\alpha_2 \triangleq \frac{4\|B\|^2\|C\|^2(\alpha + 2\|A\|)(2\alpha + 3\|A\|)}{\alpha^3} \tag{39}$$

and α denotes the minimum eigenvalue of $-A - A^T$.

Proof: See Appendix C.

Theorem 4.1: Consider the iterative algorithm (35) with $U_0 \in St(m, n)$ and

$$0 < t_k < \frac{\sqrt{2}}{\alpha_1 + \sqrt{2}\alpha_2} \tag{40}$$

where α_1 and α_2 are defined as in Lemma 4.1. Then there holds

$$\mathfrak{J}(U_{k+1}) \le \mathfrak{J}(U_k), \qquad \forall\, k = 0, 1, 2, \cdots.$$

Moreover, the equality holds if and only if U_k becomes a critical point of $\mathfrak{J}(U)$.

Proof: Set

$$U(t) = e^{t\Gamma_k} U_k$$

and let $R(t)$ be the corresponding R defined via (30). Then it is clear that $U(0) = U_k$ and $R(0) = R_k$. By the Taylor expansion, there exists some θ between zero and t such that

$$\mathfrak{J}(U(t)) - \mathfrak{J}(U_k) = t\mathfrak{J}'(U(0)) + \frac{t^2}{2}\mathfrak{J}''(U(\theta)).$$

It is obvious from (B1) that

$$\begin{aligned}\mathfrak{J}'(U(t)) &= 2\,\mathrm{trace}\big[R^T(t)U'(t)\big] \\ &= 2\,\mathrm{trace}\big[R^T(t)\Gamma_k U(t)\big] \end{aligned} \tag{41}$$

$$\mathfrak{J}''(U(t)) = 2\,\mathrm{trace}\big[(R')^T(t)\Gamma_k U(t) + R^T(t)\Gamma_k^2 U(t)\big] \tag{42}$$

which imply that

$$\mathfrak{J}'(U(0)) = 2\,\mathrm{trace}\big(R_k^T\Gamma_k U_k\big) = -\mathrm{trace}\big(\Gamma_k^T\Gamma_k\big) \tag{43}$$

$$|\mathfrak{J}''(U(t))| \le 2\big(\|R'(t)\|_F\|\Gamma_k\|_F + \|R(t)\|_F\big\|\Gamma_k^2\big\|_F\big). \tag{44}$$

Furthermore, it follows by Lemma 4.1 that

$$\begin{aligned}|\mathfrak{J}''(U(t))| &\le 2\big(\alpha_2\|U'(t)\|_F\|\Gamma_k\|_F + \alpha_1\|\Gamma_k^2\|_F\big) \\ &\le 2\big(\alpha_2\|\Gamma_k\|_F^2 + \alpha_1\|\Gamma_k\|\,\|\Gamma_k\|_F\big).\end{aligned}$$

Consequently, there results

$$\begin{aligned}&\mathfrak{J}(U(t)) - \mathfrak{J}(U_k) \\ &\quad \le -t\|\Gamma_k\|_F^2 + t^2\big(\alpha_2\|\Gamma_k\|_F^2 + \alpha_1\|\Gamma_k\|\,\|\Gamma_k\|_F\big).\end{aligned} \tag{45}$$

As Γ_k is skew-symmetric, all its eigenvalues must be on the imaginary axis, and thus the multiplicity of every nonzero singular value is at least two, which implies that $\|\Gamma_k\| \le \|\Gamma_k\|_F/\sqrt{2}$. Therefore, it is true that $\mathfrak{J}(U(t)) \le \mathfrak{J}(U_k)$ for any t with

$$0 < t < \frac{\sqrt{2}}{\alpha_1 + \sqrt{2}\alpha_2}$$

and that the equality holds if and only if $\Gamma_k = 0$.

Two important remarks are in order.

Remark 4.1: Quite clearly, the model-reduction cost $\mathfrak{J}(U_k)$ is convergent as $k \to \infty$.

Remark 4.2: With the inequality (40), note from (45) that

$$\|\Gamma_k\|^2 \le \frac{\mathfrak{J}(U_k) - \mathfrak{J}(U_{k+1})}{2t_k - 2t_k^2\big(\alpha_2 + \alpha_1/\sqrt{2}\big)}.$$

As a result, when in addition t_k is chosen to be greater than a positive constant, there holds

$$\lim_{k\to\infty} \Gamma_k = 0$$

which implies that U_k generated by the algorithm will approach the critical points satisfying the first-order necessary conditions (31) for optimality as $k \to \infty$.

Since the step-size condition (40) is independent of the current iterates, the step-size tends to be small and conservative and the associated algorithm may have a poor convergence

rate. The remainder of this section will be devoted to developing a more effective step-size selection scheme which makes use of the information available at each iteration. To this end, it is useful to establish a local upper bound on the third derivative of the model-reduction cost. For notational convenience, we introduce the following Lie bracket operations:

$$\mathfrak{L}_1(X, Y) = \mathfrak{L}(X, Y) = XY - YX \tag{46}$$

$$\mathfrak{L}_n(X, Y) = \mathfrak{L}_{n-1}(X, Y)Y - Y\mathfrak{L}_{n-1}(X, Y) \tag{47}$$

and let X_k, $\overline{X}_k$, $\tilde{X}_k$, P_k, $\overline{P}_k$, $\tilde{P}_k$ be recursively defined by

$$AX_k + X_kU_k^TA^TU_k + BB^TU_k = 0 \tag{48}$$

$$A\overline{X}_k + \overline{X}_kU_k^TA^TU_k + BB^T\Gamma_kU_k + X_kU_k^T\mathfrak{L}_1(A^T, \Gamma_k)U_k = 0 \tag{49}$$

$$A\tilde{X}_k + \tilde{X}_kU_k^TA^TU_k + BB^T\Gamma_k^2U_k + X_kU_k^T\mathfrak{L}_2(A^T, \Gamma_k)U_k + 2\overline{X}_kU_k^T\mathfrak{L}_1(A^T, \Gamma_k)U_k = 0 \tag{50}$$

$$U_k^TAU_kP_k + P_kU_k^TA^TU_k + U_k^TBB^TU_k = 0 \tag{51}$$

$$U_k^TAU_k\overline{P}_k + \overline{P}_kU_k^TA^TU_k + U_k^T\mathfrak{L}_1(BB^T, \Gamma_k)U_k + U_k^T\mathfrak{L}_1(A, \Gamma_k)U_kP_k + P_kU_k^T\mathfrak{L}_1(A^T, \Gamma_k)U_k = 0 \tag{52}$$

$$U_k^TAU_k\tilde{P}_k + \tilde{P}_kU_k^TA^TU_k + U_k^T\mathfrak{L}_2(BB^T, \Gamma_k)U_k + U_k^T\mathfrak{L}_2(A, \Gamma_k)U_kP_k + P_kU_k^T\mathfrak{L}_2(A^T, \Gamma_k)U_k + 2U_k^T\mathfrak{L}_1(A, \Gamma_k)U_k\overline{P}_k + 2\overline{P}_kU_k^T\mathfrak{L}_1(A^T, \Gamma_k)U_k = 0. \tag{53}$$

Lemma 4.2: Let $U(t) = e^{t\Gamma_k}U_k$ where U_k is orthogonal and Γ_k is skew-symmetric, and let l_k denote the unique positive root of the polynomial

$$-2\|\mathfrak{L}_4(A, \Gamma_k)\|l^4 - 8\|\mathfrak{L}_3(A, \Gamma_k)\|l^3 - 12\|\mathfrak{L}_2(A, \Gamma_k)\|l^2 - 6\|\mathfrak{L}_1(A, \Gamma_k)\|l + \alpha$$

where α is defined as in Lemma 4.1. Then for any given τ_k with $0 < \tau_k < l_k$, the third derivative $\mathfrak{J}'''(U(t))$ of the model-reduction cost $\mathfrak{J}(U(t))$ with respect to t obeys

$$\max_{|t|\le\tau_k} |\mathfrak{J}'''(U(t))| \le \xi_k \tag{54}$$

as shown in (55) and (56) at the bottom of the page.

Proof: See Appendix D.

Remark 4.3: In the above lemma, l_k should be understood to be ∞ when Γ_k equals zero, in which case a local minimum is reached.

Remark 4.4: From the definition of l_k, it is easily seen that $l_k\|\Gamma_k\|$ must be greater than some positive constant, which implies that

$$\lim_{k\to\infty} \|\Gamma_k\| = 0 \quad \Longrightarrow \quad \lim_{k\to\infty} l_k = \infty.$$

If in particular Γ_k is defined by U_k through (34), then $\|\Gamma_k\|$ is bounded by a constant due to (36), and thus l_k is greater than some positive constant.

From the proof of Lemma 4.2, it can be seen that the smaller the τ_k, the tighter the upper bound ξ_k of $|\mathfrak{J}'''(U(t))|$ on the interval $[-\tau_k, \tau_k]$. On the other hand, from Remark 4.4 we know that τ_k can be allowed to be very large when U_k is close to a critical point. Therefore, it is natural to query whether the upper bound ξ_k will become too conservative as τ_k is large. The following lemma answers this question by giving a bound on ξ_k, which will be used to establish the convergence to critical points of the iterative algorithm with a varying step-size. This bound is not only uniform but converges to zero as fast as $\|\Gamma_k\|_F^3$.

Lemma 4.3: Adopt the same hypotheses and notation as in Lemma 4.2. Let

$$\tau_k = \rho l_k \tag{57}$$

$$\begin{aligned}
\xi_k \triangleq\ & \begin{bmatrix} \left\|\begin{bmatrix} 2\Gamma_k^3C^TC & \mathfrak{L}_3(C^TC, \Gamma_k)\end{bmatrix}\right\|_F \\ 3\left\|\begin{bmatrix} 2\Gamma_k^2C^TC & \mathfrak{L}_2(C^TC, \Gamma_k)\end{bmatrix}\right\|_F \\ 3\left\|\begin{bmatrix} 2\Gamma_kC^TC & \mathfrak{L}_1(C^TC, \Gamma_k)\end{bmatrix}\right\|_F \\ \sqrt{5}\|C^TC\|_F \end{bmatrix}^T \\
& \times \begin{bmatrix} 1 & -\tau_k & 0 & 0 \\ 0 & 1 & -\tau_k & 0 \\ 0 & 0 & 1 & -\tau_k \\ -2\tau_k\|\mathfrak{L}_4(A, \Gamma_k)\| & -8\tau_k\|\mathfrak{L}_3(A, \Gamma_k)\| & -12\tau_k\|\mathfrak{L}_2(A, \Gamma_k)\| & \alpha - 6\tau_k\|\mathfrak{L}_1(A, \Gamma_k)\| \end{bmatrix}^{-1} \\
& \times \begin{bmatrix} \left\|\begin{bmatrix} X_k^T & P_k\end{bmatrix}\right\|_F \\ \left\|\begin{bmatrix} \overline{X}_k^T & \overline{P}_k\end{bmatrix}\right\|_F \\ \left\|\begin{bmatrix} \tilde{X}_k^T & \tilde{P}_k\end{bmatrix}\right\|_F \\ \|\Omega_k\|_F + \tau_k\left\|\begin{bmatrix} \Gamma_k^4BB^T & \mathfrak{L}_4(BB^T, \Gamma_k)\end{bmatrix}\right\|_F \end{bmatrix}
\end{aligned} \tag{55}$$

$$\begin{aligned}
\Omega_k \triangleq\ & [-U_k^T\Gamma_k^3BB^T \quad U_k^T\mathfrak{L}_3(BB^T, \Gamma_k)U_k] \\
& + [U_k^T\mathfrak{L}_3(A, \Gamma_k)U_kX_k^T \quad U_k^T\mathfrak{L}_3(A, \Gamma_k)U_kP_k + P_kU_k^T\mathfrak{L}_3(A^T, \Gamma_k)U_k] \\
& + 3[U_k^T\mathfrak{L}_2(A, \Gamma_k)U_k\overline{X}_k^T \quad U_k^T\mathfrak{L}_2(A, \Gamma_k)U_k\overline{P}_k + \overline{P}_kU_k^T\mathfrak{L}_2(A^T, \Gamma_k)U_k] \\
& + 3[U_k^T\mathfrak{L}_1(A, \Gamma_k)U_k\tilde{X}_k^T \quad U_k^T\mathfrak{L}_1(A, \Gamma_k)U_k\tilde{P}_k + \tilde{P}_kU_k^T\mathfrak{L}_1(A^T, \Gamma_k)U_k]
\end{aligned} \tag{56}$$

where $0 < \rho < 1$. Then there exists a constant ψ independent of U_k such that

$$\xi_k \le \psi \|\Gamma_k\|_F^3. \tag{58}$$

Proof: See Appendix E.

With the above preparations, we are now in a position to come up with a scheme for choosing a varying step-size for the iterative algorithm (35).

Theorem 4.2: Let ξ_k and l_k be defined as in Lemma 4.2 with $0 < \tau_k < l_k$, and

$$\begin{aligned}\gamma_k \triangleq \operatorname{trace}\Big\{U_k^T\Big[&\mathfrak{L}_2\big(C^TC, \Gamma_k\big)U_kP_k + 2\mathfrak{L}_1\big(C^TC, \Gamma_k\big)\\ &\cdot U_k\overline{P}_k + C^TCU_k\tilde{P}_k - 2\Gamma_k^2C^TCX_k\\ &+ 4\Gamma_kC^TC\overline{X}_k - 2C^TC\tilde{X}_k\Big]\Big\}\end{aligned} \tag{59}$$

$$\phi_k \triangleq \frac{-3\gamma_k + \sqrt{9\gamma_k^2 + 24\xi_k\|\Gamma_k\|_F^2}}{2\xi_k}. \tag{60}$$

Then given any step-size t_k satisfying

$$t_k \le \min(\tau_k, \phi_k) \tag{61}$$

the iterative gradient flow (35) generates a sequence of matrices $\{U_k\}$ in the Stiefel manifold satisfying

$$\mathfrak{J}(U_{k+1}) \le \mathfrak{J}(U_k), \qquad \forall\, k = 0, 1, 2, \cdots \tag{62}$$

from any initial $U_0 \in St(m, n)$. Moreover, if

$$\tau_k = \rho_1 l_k \quad \text{and} \quad t_k = \min(\tau_k, \rho_2\phi_k) \tag{63}$$

then there holds

$$\lim_{k\to\infty} \Gamma_k = 0 \tag{64}$$

where ρ_1 and ρ_2 are any two fixed constants between zero and one.

Proof: Let $U(t)$ be defined as in Lemma 4.2. Then from the proofs of Theorems 4.1 and 4.2, it is seen that

$$\mathfrak{J}'(U(0)) = -\|\Gamma_k\|_F^2 \quad \text{and} \quad \mathfrak{J}''(U(0)) = \gamma_k.$$

By the Taylor expansion and Lemma 4.2, it follows that for any t with $0 \le t \le \tau_k$

$$\begin{aligned}&\mathfrak{J}(U(t)) - \mathfrak{J}(U(0))\\ &\quad\le \mathfrak{J}'(U(0))t + \mathfrak{J}''(U(0))\frac{t^2}{2} + \max_{0\le t\le\tau_k}|\mathfrak{J}'''(U(t))|\frac{t^3}{6}\\ &\quad\le t\left(-\|\Gamma_k\|_F^2 + \gamma_k\frac{t}{2} + \xi_k\frac{t^2}{6}\right).\end{aligned} \tag{65}$$

Hence, there holds $\mathfrak{J}(U(t)) \le \mathfrak{J}(U(0))$ for any t with

$$0 \le t \le t_k.$$

In particular, it follows that $\mathfrak{J}(U(t_k)) \le \mathfrak{J}(U(0))$, i.e., (62). Now with the selection (63), one has

$$\begin{aligned}&t_k\left(-\|\Gamma_k\|_F^2 + \gamma_k\frac{t_k}{2} + \xi_k\frac{t_k^2}{6}\right)\\ &\quad= \frac{\xi_k t_k}{6}\left(t_k + \frac{3\gamma_k + \sqrt{9\gamma_k^2 + 24\xi_k\|\Gamma_k\|_F^2}}{2\xi_k}\right)(t_k - \phi_k)\\ &\quad\le \frac{3\gamma_k + \sqrt{9\gamma_k^2 + 24\xi_k\|\Gamma_k\|_F^2}}{12}\,t_k(t_k - \phi_k)\\ &\quad\le \frac{3\gamma_k + \sqrt{9\gamma_k^2 + 24\xi_k\|\Gamma_k\|_F^2}}{12}\,(\rho_2 - 1)t_k\phi_k\\ &\quad\le (\rho_2 - 1)\|\Gamma_k\|_F^2 t_k.\end{aligned} \tag{66}$$

In view of Remark 4.4, there is a constant $c > 0$ such that $\tau_k > c$ for all $k = 1, 2, \cdots$. In addition, by Lemma 4.3 and the fact that $\{\gamma_k\}$ is a bounded sequence, there is a constant $\Delta > 0$ such that

$$\phi_k = \frac{12\|\Gamma_k\|_F^2}{3\gamma_k + \sqrt{9\gamma_k^2 + 24\xi_k\|\Gamma_k\|_F^2}} \ge \Delta\|\Gamma_k\|_F^2.$$

As a consequence, one obtains

$$t_k = \begin{cases}\tau_k \ge c, & \text{if } \tau_k \le \rho_2\phi_k,\\ \rho_2\phi_k \ge \rho_2\Delta\|\Gamma_k\|_F^2, & \text{else.}\end{cases}$$

This together with (65) and (66) implies that at least one of the two inequalities

$$\|\Gamma_k\|_F^2 \le \frac{\mathfrak{J}(U_k) - \mathfrak{J}(U_{k+1})}{c(1-\rho_2)}$$

and

$$\|\Gamma_k\|_F^4 \le \frac{\mathfrak{J}(U_k) - \mathfrak{J}(U_{k+1})}{\Delta\rho_2(1-\rho_2)}$$

must hold. In this way, (64) immediately follows from the convergence of $\{\mathfrak{J}(U_k)\}$.

Corollary 4.1: Adopt the same notation as in Theorem 4.2 and assume that the step-size scheme (63) is implemented. Then there holds

$$\lim_{k\to\infty} \xi_k = 0. \tag{67}$$

Corollary 4.2: Adopt the same notation as in Theorem 4.2 and assume that the step-size condition (61) is satisfied. Then there exists a constant $c > 0$ such that

$$\phi_k \ge c, \qquad \forall\, k = 1, 2, \cdots. \tag{68}$$

Proof: From the proof of Theorem 4.1, it is true that

$$|\gamma_k| = |\Im''(U_k)| \le c_1 \|\Gamma_k\|_F^2, \qquad \forall\, k = 1, 2, \cdots$$

for some constant $c_1 > 0$. This together with (61) results in

$$\begin{aligned}\phi_k &= \frac{12\|\Gamma_k\|_F^2}{3\gamma_k + \sqrt{9\gamma_k^2 + 24\xi_k\|\Gamma_k\|_F^2}} \\ &\ge \frac{2\|\Gamma_k\|_F^2}{|\gamma_k| + \sqrt{\xi_k}\|\Gamma_k\|_F} \\ &\ge \frac{2}{c_1 + \sqrt{\psi\|\Gamma_k\|_F}}\end{aligned}$$

which leads to (68) because of the boundedness of $\{\|\Gamma_k\|_F\}$.

Remark 4.5: Corollary 4.2 implies that the step-size given by (63) is always greater than some positive number.

Remark 4.6: The implementation of the step-size selection (63) involves solving eight Lyapunov equations (26), (27), and (48)–(53). These equations can be grouped into the following four subgroups:

$$\{(26)\}, \quad \{(27)\}, \quad \{(33), (49), (50)\} \quad \{(51)\text{–}(53)\}$$

which are obviously decoupled from each other.

V. Error Bound, Symmetric Case, and Complete Algorithm

The objective of this section is threefold. First, an upper bound on the global minimum of the approximate minimization over the Stiefel manifold will be derived in terms of the Hankel singular values of the original system. Second, it will be shown that the approximate problem becomes exactly equivalent to the original problem if the full-order system is symmetric and the symmetry constraint is imposed on reduced-order models. Finally, a complete algorithm will be proposed for bridging the gap between the original and approximate problems.

Our first result reveals an explicit way in which the last Hankel singular values affect the H_2 model-reduction error. This is reminiscent of two well-known H_∞ error bounds [3], [20].

Lemma 5.1: Consider an nth-order stable balanced realization (A, B, C). Let A, B, C, and the controllability gramian Σ be compatibly partitioned as

$$A = \begin{bmatrix} A_1 & A_{12} \\ A_{21} & A_2 \end{bmatrix}, \qquad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$
$$C = [C_1 \quad C_2], \qquad \Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$$

where $A_1 \in \mathbb{R}^{m\times m}$ with $m < n$. Assume that A_1 is stable. Then there holds

$$\begin{aligned}&\|C(sI - A)^{-1}B - C_1(sI - A_1)^{-1}B_1\|_2 \\ &\quad = \sqrt{\operatorname{trace}\big[(C_2^TC_2 + 2\Delta A_{12})\Sigma_2\big]}\end{aligned} \tag{69}$$

where Δ is the $(n-m)\times m$ lower submatrix of D, the unique $n \times m$ matrix solution to the equation

$$A^TD + DA_1 + C^TC_1 = 0. \tag{70}$$

Proof: Denote the right-hand side of (69) by E. Then it follows from (28) that

$$E^2 = \operatorname{trace}\big(C\Sigma C^T + C_1\Sigma_1C_1^T - 2CZC_1^T\big)$$

where Z is the solution to the equation

$$AZ + ZA_1^T + BB_1^T = 0.$$

Note that

$$A\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix} + \Sigma\begin{bmatrix}A_1^T \\ A_{12}^T\end{bmatrix} + BB_1^T = 0$$

because of $A\Sigma + \Sigma A^T + BB^T = 0$. Thus, by subtraction one obtains

$$A\left(Z - \begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\right) + \left(Z - \begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\right)A_1^T - \begin{bmatrix}0 \\ \Sigma_2\end{bmatrix}A_{12}^T = 0. \tag{71}$$

Since E^2 can be rewritten as

$$\begin{aligned}E^2 &= \operatorname{trace}\left[C\Sigma C^T + C_1\Sigma_1C_1^T - 2C\begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}C_1^T \right. \\ &\qquad \left. - 2C\left(Z - \begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\right)C_1^T\right] &(72)\\ &= \operatorname{trace}\left[C_2\Sigma_2C_2^T - 2C\left(Z - \begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\right)C_1^T\right] &(73)\\ &= \operatorname{trace}\left[C_2^TC_2\Sigma_2 - 2C_1^TC\left(Z - \begin{bmatrix}\Sigma_1 \\ 0\end{bmatrix}\right)\right] &(74)\end{aligned}$$

making use of Lemma A.1 yields

$$E^2 = \operatorname{trace}\big(C_2^TC_2\Sigma_2 + 2A_{12}[0 \quad \Sigma_2]D\big)$$

where D is the unique solution to (70). Quite obviously, this is equivalent to (69). The proof is completed.

The following result being a consequence of Lemma 5.1 gives an *a priori* upper bound on the global minimum of the proposed approximate problem. This error bound provides a simple way to predetermine a lower order so as to avoid incurring a large L_2 model-reduction error before invoking any optimization algorithm.

Theorem 5.1: Assume that (A, B, C) is a balanced realization with distinct Hankel singular values $\sigma_1, \sigma_2, \cdots, \sigma_n$. Let $\Im(U)$ be the model-reduction cost as defined in (20) and let $\mathbb{K}$ denote the set of all the ordered subsets of the set $N \triangleq \{1, 2, \cdots, n\}$ with cardinality $n - m$. Then there holds

$$\min_{U\in St(m,n)} \Im(U) \le \min_{\kappa\in\mathbb{K}} \operatorname{trace}\big[\big(C_\kappa^TC_\kappa + 2\Delta_\kappa A_{N\setminus\kappa,\kappa}\big)\Sigma_\kappa\big]$$

where Σ_κ is the diagonal matrix with diagonal elements σ_{k_1}, $\sigma_{k_2}, \cdots, \sigma_{k_{n-m}}$, C_κ is the matrix consisting of the k_1th, k_2th, $\cdots$, k_{n-m}th columns of C, A_{κ_1,κ_2} is the submatrix of A resulting from deleting those columns whose index is not κ_1 and those rows whose index is not in κ_2, and Δ_κ is the $(n-m)\times m$ lower submatrix of D_κ, the unique matrix solution to the equation

$$\begin{bmatrix}A_{N\setminus\kappa,N\setminus\kappa} & A_{N\setminus\kappa,\kappa} \\ A_{\kappa,N\setminus\kappa} & A_{\kappa,\kappa}\end{bmatrix}^T D_\kappa + D_\kappa A_{N\setminus\kappa,N\setminus\kappa} + [C_{N\setminus\kappa} \quad C_\kappa]^TC_{N\setminus\kappa} = 0$$

when $\kappa = \{k_1, k_2, \cdots, k_{n-m}\}$.

Proof: The proof follows from Lemma 5.1 and the fact that each model truncation associated with $\kappa \in \mathbb{K}$ can be realized with a corresponding $U \in St(m, n)$.

Remark 5.1: Interestingly, Glover *et al.* [21] gave a different L_2 error bound on the truncation error for infinite-dimensional systems of nuclear type with an output normal realization. An example there shows that the bound decreases quite slowly as the reduced-order increases.

Now we come to address the question as to whether there is any special case in which the original problem can be exactly reduced to the approximate problem. As is identified below, the symmetric case is one such case.

Theorem 5.2: Let (A, B, B^T) be a given nth-order realization with $A = A^T < 0$ and $B \in \mathbb{R}^{n\times p}$. There holds

$$\min_{(\mathbf{A},\mathbf{B})\in\mathfrak{M}} S(\mathbf{A}, \mathbf{B}) = \min_{U\in St(m,n)} S(U^TAU, U^TB) \tag{75}$$

if the global minimum over $\mathfrak{M}$ exists, where

$$S(\mathbf{A}, \mathbf{B}) \triangleq \left\|B^T(sI - A)^{-1}B - \mathbf{B}^T(sI - \mathbf{A})^{-1}\mathbf{B}\right\|_2^2, \quad (\mathbf{A}, \mathbf{B}) \in \mathfrak{M}$$

$$\mathfrak{M} \triangleq \mathfrak{A} \times \mathbb{R}^{m\times p}$$

$$\mathfrak{A} \triangleq \{\mathbf{A} \in \mathbb{R}^{m\times m} | \mathbf{A} = \mathbf{A}^T < 0\}.$$

Proof: Let the cost function $S(\mathbf{A}, \mathbf{B})$ attain the global minimum over $\mathfrak{M}$ at $(A_m, B_m) \in \mathfrak{M}$. Then the gradient of $S(\mathbf{A}, \mathbf{B})$ over $\mathfrak{A}$ and the gradient over $\mathbb{R}^{m\times p}$ both must vanish at $(\mathbf{A}, \mathbf{B}) = (A_m, B_m)$. Denote the Fréchet derivative of $S(\mathbf{A}, \mathbf{B})$ with respect to $\mathbf{A} \in \mathbb{R}^{m\times m}$ and the Fréchet derivative with respect to $\mathbf{B} \in \mathbb{R}^{m\times p}$ at (A_m, B_m) by DS_{A_m} and DS_{B_m}, respectively. Then it is a routine exercise to find them as follows:

$$DS_{A_m}(\eta) = 2\,\mathrm{trace}\left[\eta^T\left(P_{22}^2 - P_{12}^TP_{12}\right)\right], \qquad \eta \in \mathbb{R}^{m\times m}$$

$$DS_{B_m}(\xi) = 4\,\mathrm{trace}\left[\xi^T\left(P_{22}B_m - P_{12}^TB\right)\right], \qquad \xi \in \mathbb{R}^{m\times p}$$

where $\left[\begin{smallmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22}\end{smallmatrix}\right]$ is the controllability gramian of the error system

$$\left(\begin{bmatrix} A & 0 \\ 0 & A_m\end{bmatrix}, \begin{bmatrix} B \\ B_m\end{bmatrix}, \begin{bmatrix} B \\ -B_m\end{bmatrix}^T\right).$$

The derivation of the above formulas has made use of the fact that $\left[\begin{smallmatrix} P_{11} & -P_{12} \\ -P_{12}^T & P_{22}\end{smallmatrix}\right]$ is the observability gramian of the same error system, which is in turn due to the assumption that both the full and reduced-order realizations are symmetric. Since the tangent space of $\mathfrak{A}$ is the set of all $m \times m$ symmetric matrices, it follows from the necessary conditions for optimality that

$$P_{22}^2 - P_{12}^TP_{12} = 0 \tag{76}$$

$$P_{22}B_m - P_{12}^TB = 0 \tag{77}$$

implying

$$B_m = U^TB \quad \text{and} \quad U \triangleq P_{12}P_{22}^{-1} \in St(m, n). \tag{78}$$

Since P_{12} and P_{22} satisfy

$$AP_{12} + P_{12}A_m + BB_m^T = 0 \tag{79}$$

$$A_mP_{22} + P_{22}A_m + B_mB_m^T = 0 \tag{80}$$

premultiplying (79) by U^T and subtracting leads to

$$U^TAP_{12} - A_mP_{22} = 0 \quad \text{i.e.,} \quad A_m = U^TAU. \tag{81}$$

From this and (78), (75) is concluded.

Remark 5.2: In [22], physical systems with a symmetric realization and without poles in the open right-half plane are called relaxation systems, of which RC or RL electrical networks and chemical reactions are typical examples.

Since the gradient flow algorithm or its iterative version for solving the approximate problem may sometimes lead to local minima, a more complete algorithm is needed in order to overcome or alleviate this problem. But first, let us describe the construction of a new full-order realization from any given full-order realization (A, B, C) and any given mth-order realization (A_m, B_m, C_m). To this end, assume that L_c and L_o are the controllability and observability gramians of the model error and partitioned as in (21). Set

$$\hat{P} \triangleq XP^{-1}X^T \quad \text{and} \quad \hat{Q} \triangleq YQ^{-1}Y^T.$$

It is known from [23, Th. 6.2.5] that an invertible $\Phi \in \mathbb{R}^{n\times n}$ can be constructed so that

$$\hat{P} = \Phi^{-1}\begin{bmatrix}\Lambda_{\hat{P}} & 0 \\ 0 & 0\end{bmatrix}\Phi^{-T} \quad \text{and} \quad \hat{Q} = \Phi^T\begin{bmatrix}\Lambda_{\hat{Q}} & 0 \\ 0 & 0\end{bmatrix}\Phi.$$

With Φ as a similarity transformation, we obtain the new full-order realization

$$(\hat{A}, \hat{B}, \hat{C}) \triangleq \left(\Phi A\Phi^{-1}, \Phi B, C\Phi^{-1}\right). \tag{82}$$

For ease of reference, this realization will be called an induced realization from (A, B, C) and (A_m, B_m, C_m). An important fact about the induced realization $(\hat{A}, \hat{B}, \hat{C})$ is that if (A_m, B_m, C_m) is an optimal reduced-order model, then it coincides with the mth-order truncation of $(\hat{A}, \hat{B}, \hat{C})$; see [8]. Put another way, any optimal reduced-order model is the direct truncation of an induced realization from the full-order model and itself.

We are now in a position to propose a complete algorithm for circumventing the case where the balanced realization and truncation fails to lead to the global minimum via the gradient flow alone. This algorithm employs the gradient flow as a core ingredient. The underlying idea is to switch to an induced full-order realization and its truncation based on the current full-order realization and the obtained locally optimal reduced-order model. As will be illustrated through simulation, this idea turns out to work very well in getting out of a local minimum toward the global minimum.

Algorithm for Computing L_2 Optimal Reduced-Order Models:

Step 1: Choose a balanced realization of the full-order model and an initial projection matrix U_0.

Step 2: Solve the ODE (32) or the recursive equation (35) with U_0 as the starting point to get a suboptimal reduced-order model.

Step 3: Construct an induced realization from the current full-order realization and the reduced-order realization.

TABLE I
THE COMPARISON OF RELATIVE ERRORS AMONG FIVE METHODS

Model	Lower Order	GF	OP	BT	SMM	LPMV	UB
1	1	0.17658	0.17658	0.52487	-	-	0.38493
2	1	9.7533e-2	9.7533e-2	0.99504	-	-	9.7533e-2
3	3	1.3107e-3	1.3047e-3	1.3107e-3	-	-	1.3107e-3
	2	3.9299e-2	3.9290e-2	3.9378e-2	-	-	3.9378e-2
	1	0.42709	0.42683	0.43212	-	-	0.43221
4	5	0.21431	Divergent	0.24037	-	-	0.24037
	4	0.21454	0.21454	0.22018		-	0.22018
	3	0.51632	Divergent	0.58552	-	-	0.53010
	2	0.51670	0.51670	0.51674	-	-	0.51674
	1	0.97080*	Divergent	1.14716	-	-	0.99577
5	6	5.817e-5	5.817e-5	5.822e-5	-	2.864e-4	5.822e-5
	5	2.132e-3*	Divergent	2.452e-3	-	2.132e-3	2.452e-3
	4	8.199e-3	8.199e-3	8.226e-3	-	8.199e-3	8.226e-3
	3	0.1171*	Divergent	0.2384	-	0.1171	0.1440
6	3	0.0598	0.0574	0.0599	0.0574	-	0.0599
	2	0.2443*	Divergent	0.3332	0.2443	-	0.3332
	1	0.4818	0.4818	0.4848	0.4818	-	0.4848
7	1	0.0985	0.0985	0.9949	0.0985	-	0.0985
8	4	0.4005	Divergent	0.4175	0.4005	-	0.4175
	2	0.6929	Divergent	0.8517	0.6929	-	0.6973

Step 4: If the direct truncation of the induced realization achieves the same cost as the reduced-order model or is unstable, stop; otherwise, go back to Step 2 with the induced realization and $U_0 = [I \;\; 0]^T$.

VI. NUMERICAL INVESTIGATION

In this section, we shall discuss a number of examples for illustrating the effectiveness and power of our approach to solving the optimal H_2 model-reduction problem. In particular, the following three issues will be looked at in relation to the proposed technique:

- overall performance;
- applicability to the multivariable case;
- possible conservativeness.

A. Overall Performance

For a comprehensive comparison, we consider the following well-known examples, in all of which no single method has reportedly been tested previously:

Model 1: the second-order model in [8, Example 6.3] as well as from [24];
Model 2: the second-order model in [8, Example 6.2];
Model 3: the fourth-order model in [8, Example 6.1];
Model 4: the sixth-order four-disc model from [25];
Model 5: the seventh-order model from [26];
Model 6: the fourth-order model in [12, Example 1];
Model 7: the second-order model in [12, Example 2];
Model 8: the sixth-order model of the flexible structure in [12, Example 3].

The focus of our comparison is on the five methods:

- the currently proposed gradient flow (GF) method;
- the orthogonal projection (OP) method proposed in [8];
- the balanced truncation (BT) method;
- the method proposed in [12] (SMM);
- the method proposed in [26] (LPMV).

We now summarize our obtained relative errors (i.e., $\|G_e\|_2/\|G\|_2$) in Table I as well as their upper bounds (UB) calculated by using the formula in Theorem 5.1. Those obtained by using the other methods are also included where available. Examining the table manifests the consistent success of the proposed technique in solving the H_2 optimal model-reduction and the tightness of the derived error bound.

Remark 6.1: The results shown in Table I obtained by the gradient flow algorithm are based on initial guesses corresponding to balanced truncations. An alternative way to initialize the proposed algorithm is to use a truncation of a realization in modal form.

Remark 6.2: The superscript $*$ in the table means that the complete algorithm described in the previous section is invoked. Without appealing to this procedure, the relative errors will be 0.998 44, 2.448e-3, 0.1229, and 0.2709 as opposed to 0.970 80, 2.132e-3, 0.1171, and 0.2443, which shows the ability of the algorithm to avoid getting stuck at a local minimum close to the global minimum. For example, in the case of Model 6, the convergence of the relative error to the global minimum is depicted in Fig. 2, where it is seen that the gradient flow is solved three times with the starting points approximately corresponding to the respective iteration numbers 1, 28, and 49.

B. Multivariable Case

To verify the applicability of our algorithm to the multivariable case, let us consider the automobile gas turbine model with 2 inputs, 2 outputs, and 12 states from [3] and [27]. With the gradient flow algorithm initialized with a balanced truncation, the relative errors in the respective cases of fifth- and sixth-order reductions are depicted in Fig. 3 along time t. The final results of the relative errors are shown in Table II against their respective initial relative errors resulting from balanced truncation. Once again, this illustrates that the H_2 model-reduction error can be substantially reduced by starting

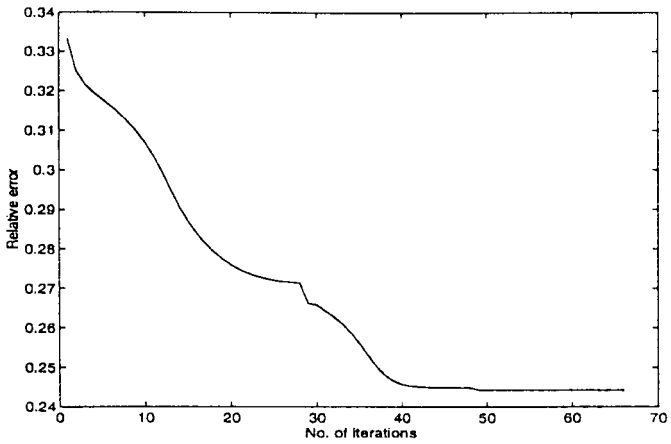


Fig. 2. Evolution of relative error.

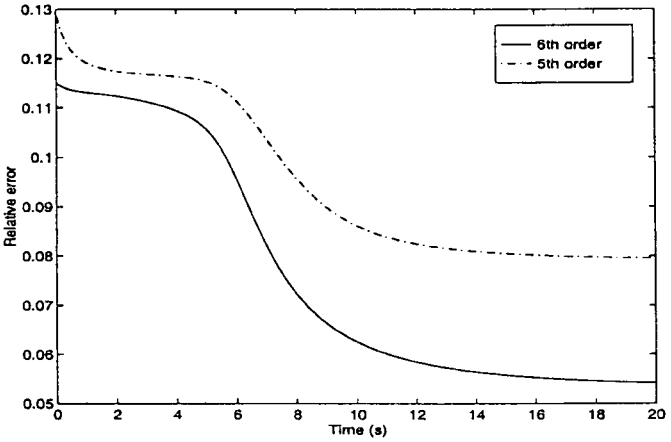


Fig. 3. Variation of relative errors over time.

TABLE II
FINAL RELATIVE ERRORS AGAINST INITIAL ONES

Reduced Order	GF	BT
4	0.1354	0.3687
5	0.0795	0.1295
6	0.0541	0.1151

with a balanced truncation. We note that the optimal projection method of Hyland and Bernstein fails to give a converging solution for this MIMO example.

C. On Possible Conservativeness

Recall that the proposed algorithm is guaranteed to produce a local minimum for any initial condition. Obviously, how close the resulting local minimum is to the global minimum depends on the choice of an initial condition. In other words, the performance of the algorithm may be influenced by the chosen initial condition. In this subsection, we examine if the majority of a large number of given starting points can lead to the global minimum with the algorithm. For this purpose, it is convenient to consider the first-order reduction of a second-order SISO system and of a third-order SISO system. In both cases, the H_2 model-reduction cost is a function of two scalar variables which uniquely determine a first-order model, and thus it is possible to find the global minimum through an extensive search over the set of all stable reduced-order models together with simple necessary conditions for extremality. That is why the two cases are taken.

As a matter of fact, it can be established in the SISO case that

$$f(a,b) \triangleq \left\|C(sI-A)^{-1}B - b(s-a)^{-1}\right\|_2 = \sqrt{\|C(sI-A)^{-1}B\|_2^2 + 2bC(A+aI)^{-1}B - \frac{b^2}{2a}}$$

for a full-order system $C(sI-A)^{-1}B$ and a reduced-order model $b(s-a)^{-1}$. Clearly, finding a globally optimal stable reduced-order model amounts to finding a global minimum point of the cost function $f(a,b)$ over the region $(0,\infty)\times(-\infty,\infty)$. Further, the necessary conditions for extremality can be derived as

$$b = 2aC(aI+A)^{-1}B \tag{83}$$

$$0 = C(aI+A)^{-2}(aI-A)B \tag{84}$$

which will be used to determine the global minimum for the following examples.

Example 6.1: Let a full-order system be given by

$$T(s) = \frac{0.5129s + 0.4605}{s^2+3s+2}$$

with randomly generated numerator coefficients. Then the globally optimal first-order model can be found to be $b_0/(s-a_0)$ with

$$(a_0, b_0) = (-2.1904,\ 0.5190) \tag{85}$$

which gives the optimal H_2 cost of 0.0046.

Note that in this case the Stiefel manifold reduces to the unit circle and that any point on it can be parameterized as follows:

$$U = [\cos(\theta) \quad \sin(\theta)]^T, \qquad \theta \in [0, 2\pi]$$

and that the cost apparently assumes the same global minimum over the upper half circle as over the whole circle. Then with each of the 40 regularly spaced points on the upper half circle corresponding to

$$\theta_i = \frac{i}{40}\pi, \qquad i = 1, \cdots, 40$$

as an initial condition, the algorithm arrives at 40 suboptimal model-reduction costs. The relative deviations of the obtained costs from the global minimum of 0.0046 are depicted in Fig. 4, which verifies the closeness between the obtained costs and the optimal one. It takes about 3 min to complete the 40 simulations on an HP workstation, which means that on average the algorithm takes less than 5 s to get to the near global minimum for each of the specified initial conditions.

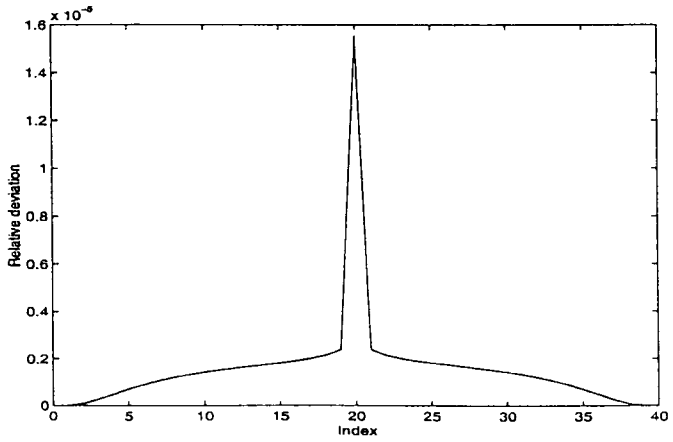


Fig. 4. Relative deviations between achieved costs and the optimum for Example 6.1.

VII. Conclusions

The H_2 optimal model-reduction problem has approximately been formulated as an unconstrained minimization problem over the Stiefel manifold. The two problems have been proved to be equivalent in the symmetric case. Using the differential techniques, we have derived explicit formulas for the gradient of the model-reduction cost function over the manifold. Several convergent algorithms have been proposed. The first one is given in terms of an ordinary differential equation formed by the gradient flow, and concerning this algorithm a number of nice theoretical properties are obtained. For example, the cost is always decreasing along the solution to the ODE evolving on the Stiefel manifold until a minimum point is reached. Based on this gradient flow algorithm, an iterative algorithm in closed form has been generated, for which it has been shown that a fixed step-size is adequate to ensure that the cost is decreasing to a minimum. However, an adaptive scheme derived for choosing the step-size tends to achieve a greater convergence rate. All the proposed algorithms are well applicable to the MIMO case. Numerical tests have indicated the reliability of the algorithms as well as the convergence to a minimum.

Developed in this paper, the techniques have since been applied to solving several related problems such as frequency-weighted model reduction as well as filter reduction [28], [29].

Appendix A
Auxiliary Lemmas

Lemma A.1: If P and Q satisfy

$$AP+PB+X=0 \quad \text{and} \quad A^TQ+QB^T+Y=0$$

then there holds

$$\operatorname{trace}\big(Y^TP\big)=\operatorname{trace}\big(X^TQ\big). \tag{A1}$$

Proof: The lemma follows directly from the respective substitution of $-(AP+PB)$ and $-(A^TQ+QB^T)$ for X and Y into the both sides of (A1).

Lemma A.2: Suppose two matrices $A_1 \in \mathbb{R}^{n\times n}$ and $A_2 \in \mathbb{R}^{m\times m}$ are given with μ_i denoting the maximum eigenvalue of $(A_i+A_i^T)/2$ for $i=1,2$. Assume that P is the unique solution to the Lyapunov equation

$$A_1P+PA_2+Z=0$$

with $Z \in \mathbb{R}^{n\times m}$. If $\mu_1+\mu_2<0$, then there hold

$$\|P\|_F \le -\frac{\|Z\|_F}{\mu_1+\mu_2} \quad \text{and} \quad \|P\| \le -\frac{\|Z\|}{\mu_1+\mu_2}.$$

Proof: It is not difficult to establish that

$$\left\|e^{A_it}\right\| \le e^{\mu_it}, \qquad i=1,2$$

from which it follows that

$$\left\|e^{A_1t}Ze^{A_2t}\right\|_F \le \|Z\|_F e^{(\mu_1+\mu_2)t}$$

and

$$\left\|e^{A_1t}Ze^{A_2t}\right\| \le \|Z\| e^{(\mu_1+\mu_2)t}. \tag{A2}$$

Since $\mu_1+\mu_2<0$, the above in turn implies that

$$P=\int_0^\infty e^{A_1t}Ze^{A_2t}\,dt.$$

Again from (A2), the lemma is concluded.

Lemma A.3: Let $F(t)$ be a differentiable $p\times q$ matrix function of t on the interval $[a,b]$ containing zero inside. Then there holds

$$\|F(t)\|_F \le \|F(0)\|_F + |t| \max_{a\le t\le b}\|F'(t)\|_F, \qquad \forall\, t\in[a,b]. \tag{A3}$$

Proof: Let ϵ be an arbitrary positive scalar. Then

$$f(t) \triangleq \sqrt{\operatorname{trace}[F^T(t)F(t)]+\epsilon}$$

is a differentiable scalar function of t on $[a,b]$. Therefore, $f(t)$ has the following Taylor expansion

$$f(t)=f(0)+tf'(\theta)$$

where θ is between zero and t. This implies that

$$|f(t)| \le |f(0)| + |t|\,|f'(\theta)|.$$

But, one has

$$\begin{aligned}|f'(\theta)| &= \left[\|F(\theta)\|_F^2+\epsilon\right]^{-(1/2)}\left|\operatorname{trace}[(F'(\theta))^TF(\theta)]\right| \\ &\le \|F'(\theta)\|_F \\ f(0) &\le \|F(0)\|_F+\sqrt{\epsilon}.\end{aligned}$$

Hence, it follows that

$$|f(t)| \le \|F(0)\|_F + |t| \max_{a\le t\le b}\|F'(t)\|_F + \sqrt{\epsilon}.$$

As ϵ is arbitrary, and this immediately results in (A3).

Lemma A.4: Let $A\in\mathbb{R}^{n\times n}$ be given. Assume that $A+A^T$ is positive definite. Then, there exists some constant κ such that for any skew-symmetric $\Gamma\in\mathbb{R}^{n\times n}$ there holds

$$\|\Gamma\|_F \le \kappa\|\mathfrak{L}(A,\Gamma)\|_F.$$

Proof: Choose an orthogonal $V \in \mathbb{R}^{n\times n}$ so that

$$V^T(A+A^T)V = \operatorname{diag}\{l_1, \cdots, l_n\}.$$

Note that $V^T\Gamma V$ is still skew-symmetric and

$$\mathfrak{L}\big(\operatorname{diag}\{l_1, \cdots, l_n\}, V^T\Gamma V\big) = V^T\mathfrak{L}(A+A^T, \Gamma)V$$

which means that the (i, j)-element of $V^T\Gamma V$ is equal to that of $V^T\mathfrak{L}(A+A^T, \Gamma)V$ times $1/(l_i+l_j)$. Thus, there results

$$\begin{aligned}\|\Gamma\|_F &\le \max_{1\le i<j\le n}\left(\frac{1}{l_i+l_j}\right)\big\|\mathfrak{L}(A+A^T, \Gamma)\big\|_F \\ &\le 2\max_{1\le i<j\le n}\left(\frac{1}{l_i+l_j}\right)\|\mathfrak{L}(A, \Gamma)\|_F.\end{aligned}$$

Setting

$$\kappa = 2\max_{1\le i<j\le n}\left(\frac{1}{l_i+l_j}\right)$$

completes the proof.

Appendix B
Proof of Lemma 3.1

First it is straightforward to compute the Fréchet derivative $D\mathfrak{J}_U$ of $\mathfrak{J}(U)$ as follows:

$$\begin{aligned}D\mathfrak{J}_U(\xi) &= \operatorname{trace}\big\{C^TC\big(\xi PU^T + UP\xi^T - 2X\xi^T \\ &\qquad + U[DP_U(\xi)]U^T - 2[DX_U(\xi)]U^T\big)\big\} \\ &= \operatorname{trace}\big\{2\big(PU^TC^TC - X^TC^TC\big)\xi \\ &\qquad + U^TC^TCU[DP_U(\xi)] - 2U^TC^TC[DX_U(\xi)]\big\}.\end{aligned}$$

By differentiating the both sides of (23) and (24), it follows that $DX_U(\xi)$ and $DP_U(\xi)$ satisfy

$$\begin{aligned}&A[DX_U(\xi)] + [DX_U(\xi)]U^TA^TU + X\xi^TA^TU \\ &\quad + XU^TA^T\xi + BB^T\xi = 0 \\ &U^TAU[DP_U(\xi)] + [DP_U(\xi)]U^TA^TU + Z + Z^T = 0\end{aligned}$$

with

$$Z = PU^TA^T\xi + P\xi^TA^TU + U^TBB^T\xi.$$

Hence by Lemma A.1, one obtains

$$\begin{aligned}D\mathfrak{J}_U(\xi) &= \operatorname{trace}\Big[2\Big(PU^TC^TC - X^TC^TC\Big)\xi + \big(Z+Z^T\big)Q \\ &\qquad + 2\big(X\xi^TA^TU + XU^TA^T\xi + BB^T\xi\big)^TY\Big] \\ &= 2\operatorname{trace}\big[\big(PU^TC^TC - X^TC^TC\big)\xi \\ &\qquad + Y^T\big(X\xi^TA^TU + XU^TA^T\xi + BB^T\xi\big) + ZQ\big] \\ &= 2\operatorname{trace}\big[\big(PU^TC^TC - X^TC^TC + X^TYU^TA \\ &\qquad + Y^TXU^TA^T + Y^TBB^T\big)\xi + ZQ\big] \\ &= 2\operatorname{trace}\big[\big(PU^TC^TC - X^TC^TC + X^TYU^TA \\ &\qquad + Y^TXU^TA^T + Y^TBB^T\big)\xi \\ &\qquad + \big(QPU^TA^T + PQU^TA + QU^TBB^T\big)\xi\big] \\ &= 2\operatorname{trace}\big(R^T\xi\big). \qquad \text{(B1)}\end{aligned}$$

The gradient $\nabla\mathfrak{J}$ is the uniquely determined vector field on $St(m, n)$ which satisfies the two conditions:

1) $\nabla\mathfrak{J}(U) \in T_USt(m, n), \qquad \forall U \in St(m, n)$

2) $D\mathfrak{J}_U(\xi) = \langle\nabla\mathfrak{J}(U), \xi\rangle, \qquad \forall \xi \in T_USt(m, n).$

Due to (B1), condition 2) is equivalent to

$$(\nabla\mathfrak{J}(U) - R)^T\xi = 0, \qquad \forall \xi \in T_USt(m, n). \qquad \text{(B2)}$$

Since

$$T_USt(m, n)^\perp = \Big\{U\Lambda | \Lambda = \Lambda^T \in \mathbb{R}^{k\times k}\Big\} \qquad \text{(B3)}$$

(B2) together with condition 1) gives

$$\nabla\mathfrak{J}(U) = \big(I - UU^T\big)R. \qquad \text{(B4)}$$

□

Appendix C
Proof of Lemma 4.1

By Lemma A.2, one has

$$\max(\|X\|_F, \|P\|_F) \le \frac{\big\|BB^T\big\|_F}{\alpha} \qquad \text{(C1)}$$

$$\max(\|Y\|_F, \|Q\|_F) \le \frac{\big\|C^TC\big\|_F}{\alpha} \qquad \text{(C2)}$$

$$\max(\|X\|, \|P\|) \le \frac{\|B\|^2}{\alpha} \qquad \text{(C3)}$$

$$\max(\|Y\|, \|Q\|) \le \frac{\|C\|^2}{\alpha}. \qquad \text{(C4)}$$

From the first two inequalities, (36) follows immediately by recalling the definition of R in Lemma 3.1. To prove (37), differentiate the both sides of (23) to yield

$$\begin{aligned}&AX' + X'U^TA^TU \\ &\quad + \big\{BB^TU' + X\big[(U')^TA^TU + U^TA^TU'\big]\big\} = 0.\end{aligned}$$

Again by Lemma A.2, one obtains

$$\begin{aligned}\|X'\|_F &\le \frac{\big\|BB^TU' + X\big[(U')^TA^TU + U^TA^TU'\big]\big\|_F}{\alpha} \\ &\le \frac{\big(\|B\|^2 + 2\|X\|\,\|A\|\big)\|U'\|_F}{\alpha} \\ &\le \frac{\|B\|^2(\alpha + 2\|A\|)\|U'\|_F}{\alpha^2}. \qquad \text{(C5)}\end{aligned}$$

In the same way, it can be established that

$$\|P'\|_F \le \frac{2\|B\|^2(\alpha + 2\|A\|)\|U'\|_F}{\alpha^2} \qquad \text{(C6)}$$

$$\|Y'\|_F \le \frac{\|C\|^2(\alpha + 2\|A\|)\|U'\|_F}{\alpha^2} \qquad \text{(C7)}$$

$$\|Q'\|_F \le \frac{2\|C\|^2(\alpha + 2\|A\|)\|U'\|_F}{\alpha^2}. \qquad \text{(C8)}$$

Since

$$\begin{aligned}R' &= \left[(-C^TC + A^TUY^T)X' + AU(X')^TY\right] \\ &\quad + \left[(C^TCU + A^TUQ)P' + AUP'Q\right] \\ &\quad + \left[(BB^T + AUX^T)Y' + A^TU(Y')^TX\right] \\ &\quad + \left[(BB^TU + AUP)Q' + A^TUQ'P\right] \\ &\quad + \left[(C^TCU' + A^TU'Q)P + (BB^TU' + AU'P)Q\right] \\ &\quad + A^TU'Y^TX + AU'X^TY\end{aligned}$$

it follows that

$$\begin{aligned}\|R'\|_F &\le \left(\|C\|^2 + 2\|A\|\,\|Y\|\right)\|X'\|_F \\ &\quad + \left(\|C\|^2 + 2\|A\|\,\|Q\|\right)\|P'\|_F \\ &\quad + \left(\|B\|^2 + 2\|A\|\,\|X\|\right)\|Y'\|_F \\ &\quad + \left(\|B\|^2 + 2\|A\|\,\|P\|\right)\|Q'\|_F \\ &\quad + \left[\|C\|^2\|P\| + \|B\|^2\|Q\| \right. \\ &\quad \left. + 2\|A\|(\|X\|\,\|Y\| + \|P\|\,\|Q\|)\right]\|U'\|_F \\ &\le \frac{6\|B\|^2\|C\|^2(\alpha + 2\|A\|)^2}{\alpha^3}\|U'\|_F \\ &\quad + \left(\frac{2\|B\|^2\|C\|^2}{\alpha} + \frac{4\|A\|\,\|B\|^2\|C\|^2}{\alpha^2}\right)\|U'\|_F \\ &= \frac{4\|B\|^2\|C\|^2(\alpha + 2\|A\|)(2\alpha + 3\|A\|)}{\alpha^3}\|U'\|_F\end{aligned}$$

as required. □

Appendix D
Proof of Lemma 4.2

Set

$$\Phi(t) \triangleq \begin{bmatrix} 0 & X(t) \\ X^T(t) & P(t)\end{bmatrix}$$

where $X(t)$ and $P(t)$ are defined by (23) and (24) with $U = U(t)$. Then, Φ is obviously a smooth function of t and satisfies

$$A_U\Phi + \Phi A_U^T + \begin{bmatrix} 0 & BB^TU \\ U^TBB^T & U^TBB^TU\end{bmatrix} = 0 \qquad \text{(D1)}$$

where

$$A_U \triangleq \begin{bmatrix} A & 0 \\ 0 & U^TAU\end{bmatrix}.$$

Moreover, by successive differentiation of the above, one can reach the relations

$$\begin{aligned}&A_U\Phi' + \Phi'A_U^T + \begin{bmatrix} 0 & BB^T\Gamma_kU \\ -U^T\Gamma_kBB^T & U^T\mathfrak{L}_1(BB^T, \Gamma_k)U\end{bmatrix} \\ &\quad + \begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_1(A, \Gamma_k)U\end{bmatrix}\Phi \\ &\quad + \Phi\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_1(A^T, \Gamma_k)U\end{bmatrix} = 0 \qquad \text{(D2)}\end{aligned}$$

$$\begin{aligned}&A_U\Phi'' + \Phi''A_U^T + \begin{bmatrix} 0 & BB^T\Gamma_k^2U \\ U^T\Gamma_k^2BB^T & U^T\mathfrak{L}_2(BB^T, \Gamma_k)U\end{bmatrix} \\ &\quad + \begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_2(A, \Gamma_k)U\end{bmatrix}\Phi + \Phi\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_2(A^T, \Gamma_k)U\end{bmatrix} \\ &\quad + 2\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_1(A, \Gamma_k)U\end{bmatrix}\Phi' \\ &\quad + 2\Phi'\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_1(A^T, \Gamma_k)U\end{bmatrix} = 0 \qquad \text{(D3)}\end{aligned}$$

$$\begin{aligned}&A_U\Phi''' + \Phi'''A_U^T + \begin{bmatrix} 0 & BB^T\Gamma_k^3U \\ -U^T\Gamma_k^3BB^T & U^T\mathfrak{L}_3(BB^T, \Gamma_k)U\end{bmatrix} \\ &\quad + \begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_3(A, \Gamma_k)U\end{bmatrix}\Phi + \Phi\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_3(A^T, \Gamma_k)U\end{bmatrix} \\ &\quad + 3\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_2(A, \Gamma_k)U\end{bmatrix}\Phi' \\ &\quad + 3\Phi'\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_2(A^T, \Gamma_k)U\end{bmatrix} \\ &\quad + 3\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_1(A, \Gamma_k)U\end{bmatrix}\Phi'' \\ &\quad + 3\Phi''\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_1(A^T, \Gamma_k)U\end{bmatrix} = 0. \qquad \text{(D4)}\end{aligned}$$

In particular, with

$$W(t) \triangleq \begin{bmatrix} X^T(t) & P(t)\end{bmatrix}$$

it follows that

$$\begin{aligned}&U^TAUW''' + W'''A_U^T \\ &\quad + \begin{bmatrix} -U^T\Gamma_k^3BB^T & U^T\mathfrak{L}_3(BB^T, \Gamma_k)U\end{bmatrix} \\ &\quad + U^T\mathfrak{L}_3(A, \Gamma_k)UW + W\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_3(A^T, \Gamma_k)U\end{bmatrix} \\ &\quad + 3U^T\mathfrak{L}_2(A, \Gamma_k)UW' + 3W'\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_2(A^T, \Gamma_k)U\end{bmatrix} \\ &\quad + 3U^T\mathfrak{L}_1(A, \Gamma_k)UW'' \\ &\quad + 3W''\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_1(A^T, \Gamma_k)U\end{bmatrix} = 0.\end{aligned}$$

Therefore, making use of Lemma A.2 yields

$$\begin{aligned}&\alpha\|W'''(t)\|_F \\ &\quad \le \left\| \left[-U^T\Gamma_k^3BB^T \quad U^T\mathfrak{L}_3(BB^T, \Gamma_k)U\right] \right. \\ &\qquad + U^T\mathfrak{L}_3(A, \Gamma_k)UW + W\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_3(A^T, \Gamma_k)U\end{bmatrix} \\ &\qquad + 3U^T\mathfrak{L}_2(A, \Gamma_k)UW' \\ &\qquad + 3W'\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_2(A^T, \Gamma_k)U\end{bmatrix} \\ &\qquad + 3U^T\mathfrak{L}_1(A, \Gamma_k)UW'' \\ &\qquad \left. + 3W''\begin{bmatrix} 0 & 0 \\ 0 & U^T\mathfrak{L}_1(A^T, \Gamma_k)U\end{bmatrix}\right\|_F, \qquad \forall |t| \le \tau_k. \qquad \text{(D5)}\end{aligned}$$

Since the matrix on the right-hand side equals $\emptyset_k$ at $t=0$ and has the following derivative:

$$\begin{aligned}&[U^T\Gamma_k^4BB^T \quad U^T\mathfrak{L}_4(BB^T,\Gamma_k)U]\\&\quad+U^T\mathfrak{L}_4(A,\Gamma_k)UW+W\begin{bmatrix}0&0\\0&U^T\mathfrak{L}_4(A^T,\Gamma_k)U\end{bmatrix}\\&\quad+4U^T\mathfrak{L}_3(A,\Gamma_k)UW'+4W'\begin{bmatrix}0&0\\0&U^T\mathfrak{L}_3(A^T,\Gamma_k)U\end{bmatrix}\\&\quad+6U^T\mathfrak{L}_2(A,\Gamma_k)UW''+6W''\begin{bmatrix}0&0\\0&U^T\mathfrak{L}_2(A^T,\Gamma_k)U\end{bmatrix}\\&\quad+3U^T\mathfrak{L}_1(A,\Gamma_k)UW'''+3W'''\begin{bmatrix}0&0\\0&U^T\mathfrak{L}_1(A^T,\Gamma_k)U\end{bmatrix}\end{aligned}$$

applying Lemma A.3 to the right-hand side of (D5) gives rise to

$$\begin{aligned}&\alpha\max_{|t|\le\tau_k}\|W'''(t)\|_F\\&\quad\le\|\Omega_k\|_F+\tau_k\|[\Gamma_k^4BB^T\quad\mathfrak{L}_4(BB^T,\Gamma_k)]\|_F\\&\qquad+2\tau_k\Big\{\|\mathfrak{L}_4(A,\Gamma_k)\|\max_{|t|\le\tau_k}\|W\|_F\\&\qquad\quad+4\|\mathfrak{L}_3(A,\Gamma_k)\|\max_{|t|\le\tau_k}\|W'\|_F\\&\qquad\quad+6\|\mathfrak{L}_2(A,\Gamma_k)\|\max_{|t|\le\tau_k}\|W''\|_F\\&\qquad\quad+3\|\mathfrak{L}_1(A,\Gamma_k)\|\max_{|t|\le\tau_k}\|W'''\|_F\Big\}.\end{aligned}\tag{D6}$$

In the meantime, by Lemma A.3 one has

$$\max_{|t|\le\tau_k}\|W(t)\|_F\le\|W(0)\|_F+\tau_k\max_{|t|\le\tau_k}\|W'(t)\|_F\tag{D7}$$

$$\max_{|t|\le\tau_k}\|W'(t)\|_F\le\|W'(0)\|_F+\tau_k\max_{|t|\le\tau_k}\|W''(t)\|_F\tag{D8}$$

$$\max_{|t|\le\tau_k}\|W''(t)\|_F\le\|W''(0)\|_F+\tau_k\max_{|t|\le\tau_k}\|W'''(t)\|_F.\tag{D9}$$

Due to

$$\begin{aligned}\alpha>2\tau_k\big(&\|\mathfrak{L}_4(A,\Gamma_k)\|\tau_k^3+4\|\mathfrak{L}_3(A,\Gamma_k)\|\tau_k^2\\&+6\|\mathfrak{L}_2(A,\Gamma_k)\|\tau_k+3\|\mathfrak{L}_1(A,\Gamma_k)\|\big)\end{aligned}$$

it is not difficult to see that the inequalities (D6)–(D9) can be combined into the compact form as shown in (D10) at the bottom of the page.

On the other hand, recall that $\mathfrak{J}(U(t))$ can be expressed as

$$\mathfrak{J}(U(t))=\operatorname{trace}\left(C\Sigma_cC^T+[C\quad -CU]\Phi\begin{bmatrix}C^T\\-U^TC^T\end{bmatrix}\right)$$

where Σ_c is the controllability gramian of the full-order system, i.e., the solution to (22). Then it is routine to compute the following derivatives with respect to t:

$$\begin{aligned}\mathfrak{J}'(U(t))=\operatorname{trace}\Bigg(&\begin{bmatrix}0&-C^TC\Gamma_kU\\U^T\Gamma_kC^TC&U^T\mathfrak{L}_1(C^TC,\Gamma_k)U\end{bmatrix}\Phi\\&+\begin{bmatrix}0&-C^TCU\\-U^TC^TC&U^TC^TCU\end{bmatrix}\Phi'\Bigg)\\\mathfrak{J}''(U(t))=\operatorname{trace}\Bigg(&\begin{bmatrix}0&-C^TC\Gamma_k^2U\\-U^T\Gamma_k^2C^TC&U^T\mathfrak{L}_2(C^TC,\Gamma_k)U\end{bmatrix}\Phi\\&+2\begin{bmatrix}0&-C^TC\Gamma_kU\\U^T\Gamma_kC^TC&U^T\mathfrak{L}_1(C^TC,\Gamma_k)U\end{bmatrix}\Phi'\\&+\begin{bmatrix}0&-C^TCU\\-U^TC^TC&U^TC^TCU\end{bmatrix}\Phi''\Bigg)\\\mathfrak{J}'''(U(t))=\operatorname{trace}\Bigg(&\begin{bmatrix}0&-C^TC\Gamma_k^3U\\U^T\Gamma_k^3C^TC&U^T\mathfrak{L}_3(C^TC,\Gamma_k)U\end{bmatrix}\Phi\\&+3\begin{bmatrix}0&-C^TC\Gamma_k^2U\\-U^T\Gamma_k^2C^TC&U^T\mathfrak{L}_2(C^TC,\Gamma_k)U\end{bmatrix}\Phi'\\&+3\begin{bmatrix}0&-C^TC\Gamma_kU\\U^T\Gamma_kC^TC&U^T\mathfrak{L}_1(C^TC,\Gamma_k)U\end{bmatrix}\Phi''\\&+\begin{bmatrix}0&-C^TCU\\-U^TC^TC&U^TC^TCU\end{bmatrix}\Phi'''\Bigg)\\=\operatorname{trace}\big\{&2U^T\Gamma_k^3C^TCX(t)\\&+U^T\mathfrak{L}_3(C^TC,\Gamma_k)UP(t)\\&+3[-2U^T\Gamma_k^2C^TCX'(t)\\&+U^T\mathfrak{L}_2(C^TC,\Gamma_k)UP'(t)]\\&+3[2U^T\Gamma_kC^TCX''(t)\\&+U^T\mathfrak{L}_1(C^TC,\Gamma_k)UP''(t)]\\&-2U^TC^TCX'''(t)+U^TC^TCUP'''(t)\big\}.\end{aligned}$$

$$\begin{bmatrix}\max_{|t|\le\tau_k}\|W(t)\|\\\max_{|t|\le\tau_k}\|W'(t)\|\\\max_{|t|\le\tau_k}\|W''(t)\|\\\max_{|t|\le\tau_k}\|W'''(t)\|\end{bmatrix}\le\Psi^{-1}\times\begin{bmatrix}\|[X_k^T\quad P_k]\|_F\\\|[\overline{X}_k^T\quad\overline{P}_k]\|_F\\\|[\hat{X}_k^T\quad\hat{P}_k]\|_F\\\|\Omega_k\|_F+\tau_k\|[\Gamma_k^4BB^T\quad\mathfrak{L}_4(BB^T,\Gamma_k)]\|_F\end{bmatrix}\tag{D10}$$

$$\Psi=\begin{bmatrix}1&-\tau_k&0&0\\0&1&-\tau_k&0\\0&0&1&-\tau_k\\-2\tau_k\|\mathfrak{L}_4(A,\Gamma_k)\|&-8\tau_k\|\mathfrak{L}_3(A,\Gamma_k)\|&-12\tau_k\|\mathfrak{L}_2(A,\Gamma_k)\|&\alpha-6\tau_k\|\mathfrak{L}_1(A,\Gamma_k)\|\end{bmatrix}^{-1}\tag{D11}$$

$$\xi_k \le \frac{\kappa_1\kappa_3\|\Gamma_k\|_F^3 \sum_{1\le i,j\le 4}\varphi_{ij}\|\Gamma_k\|_F^{j-i}}{\alpha - (2\|\mathfrak{L}_4(A,\Gamma_k)\|\tau_k^4 + 8\|\mathfrak{L}_3(A,\Gamma_k)\|\tau_k^3 + 12\|\mathfrak{L}_2(A,\Gamma_k)\|\tau_k^2 + 6\|\mathfrak{L}_1(A,\Gamma_k)\|\tau_k)} \tag{E4}$$

$$\begin{aligned}\xi_k &\le \frac{16\kappa_1\kappa_2\kappa_3\kappa_4\|\Gamma_k\|_F^3}{\alpha - (2\|\mathfrak{L}_4(A,\Gamma_k)\|\tau_k^4 + 8\|\mathfrak{L}_3(A,\Gamma_k)\|\tau_k^3 + 12\|\mathfrak{L}_2(A,\Gamma_k)\|\tau_k^2 + 6\|\mathfrak{L}_1(A,\Gamma_k)\|\tau_k)} \\ &= \frac{16\kappa_1\kappa_2\kappa_3\kappa_4}{(1-\rho)\alpha}\|\Gamma_k\|_F^3\end{aligned} \tag{E5}$$

By noting that $\|U(t)\| = 1$, it is easily deduced that

$$\begin{aligned}&|\Im'''(U(t))| \\ &\quad\le [\|W(t)\|_F \;\; \|W'(t)\|_F \;\; \|W''(t)\|_F \;\; \|W'''(t)\|_F] \\ &\qquad\times \begin{bmatrix} \|[2\Gamma_k^3C^TC \quad \mathfrak{L}_3(C^TC,\Gamma_k)]\|_F \\ 3\|[2\Gamma_k^2C^TC \quad \mathfrak{L}_2(C^TC,\Gamma_k)]\|_F \\ 3\|[2\Gamma_kC^TC \quad \mathfrak{L}_1(C^TC,\Gamma_k)]\|_F \\ \sqrt{5}\|C^TC\|_F \end{bmatrix}.\end{aligned}$$

This together with (D10) leads to (54). □

APPENDIX E
PROOF OF LEMMA 4.3

First, it is easy to see that

$$\begin{bmatrix} \|[2\Gamma_k^3C^TC \quad \mathfrak{L}_3(C^TC,\Gamma_k)]\|_F \\ 3\|[2\Gamma_k^2C^TC \quad \mathfrak{L}_2(C^TC,\Gamma_k)]\|_F \\ 3\|[2\Gamma_kC^TC \quad \mathfrak{L}_1(C^TC,\Gamma_k)]\|_F \\ \sqrt{5}\|C^TC\|_F \end{bmatrix} \le \kappa_1 \begin{bmatrix} \|\Gamma_k\|_F^3 \\ \|\Gamma_k\|_F^2 \\ \|\Gamma_k\|_F \\ 1 \end{bmatrix} \tag{E1}$$

where κ_1 is a constant independent of U_k. Because of $\tau_k < l_k$, it is true that $6\tau_k\|\mathfrak{L}_1(A,\Gamma_k)\| \le \alpha$. By Lemma A.4, this leads to

$$\tau_k\|\Gamma_k\|_F \le \kappa_2 \tag{E2}$$

where $\kappa_2 > 0$ is a constant independent of U_k. Hence, by successively applying Lemma A.2 to (D1)–(D3), it is not difficult to establish the following inequality:

$$\begin{aligned}&\begin{bmatrix} \|[X_k^T \quad P_k]\|_F \\ \|[\overline{X}_k^T \quad \overline{P}_k]\|_F \\ \|[\tilde{X}_k^T \quad \tilde{P}_k]\|_F \\ \|\Omega_k\|_F + \tau_k\|[\Gamma_k^4BB^T \quad \mathfrak{L}_4(BB^T,\Gamma_k)]\|_F \end{bmatrix} \\ &\quad\le \kappa_3[1 \quad \|\Gamma_k\|_F \quad \|\Gamma_k\|_F^2 \quad \|\Gamma_k\|_F^3]^T\end{aligned} \tag{E3}$$

for some constant κ_3 independent of U_k. Since all the elements of the inverse in (55) are nonnegative, combining (E1) with (E3) yields (E4), as shown at the top of the page, where φ_{ij} denotes the (i,j)-element of the adjoint of Ψ as shown in (D11), at the bottom of the previous page. By close inspection, it follows from (E2) that there exists a positive constant κ_4 independent of U_k and τ_k such that

$$\varphi_{ij} \le \begin{cases} \kappa_4\tau_k^{j-i}, & \text{if } i \le j \\ \kappa_4\|\Gamma_k\|^{i-j}, & \text{else.} \end{cases}$$

Consequently, again from (E2) we have (E5), shown at the top of the page. The proof is thus completed.

REFERENCES

[1] B. C. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Automat. Contr.*, vol. AC-26, pp. 17–32, 1981.

[2] L. Pernebo and L. M. Silverman, "Model reduction via balanced state space representations," *IEEE Trans. Automat. Contr.*, vol. AC-27, pp. 382–387, 1982.

[3] K. Glover, "All optimal Hankel-norm approximations of linear multivariable systems and their L_∞-error bounds," *Int. J. Contr.*, pp. 1115–1193, 1984.

[4] D. G. Meyer and S. Srinivasan, "Balancing and model reduction for second-order form linear systems," *IEEE Trans. Automat. Contr.*, vol. 41, pp. 1632–1644, 1996.

[5] K. Zhou, "Frequency-weighted L_∞ norm and optimal Hankel norm model reduction," *IEEE Trans. Automat. Contr.*, vol. 40, pp. 1687–1699, 1995.

[6] L. Meier and D. G. Luenberger, "Approximation of linear constant systems," *IEEE Trans. Automat. Contr.*, vol. AC-12, pp. 585–588, 1967.

[7] D. A. Wilson, "Optimum solution of model-reduction problem," in *Proc. Inst. Elec. Eng.*, pp. 1161–1165, 1970.

[8] D. C. Hyland and D. S. Bernstein, "The optimal projection equations for model reduction and the relationships among the methods of Wilson, Skelton and Moore," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 1201–1211, 1985.

[9] Y. Halevi, "Frequency weighted model reduction via optimal projection," in *Proc. IEEE Conf. Decision and Control*, 1990, pp. 2906–2911.

[10] A. E. Bryson and A. Carrier, "Second-order algorithm for optimal model order reduction," *J. Guidance Contr. Dynam.*, pp. 887–892, 1990.

[11] L. Baratchart, M. Cardelli, and M. Olivi, "Identification and rational L_2 approximation: A gradient algorithm," *Automatica*, pp. 413–417, 1991.

[12] J. T. Spanos, M. H. Milman, and D. L. Mingori, "A new algorithm for L_2 optimal model reduction," *Automatica*, pp. 897–909, 1992.

[13] L. Baratchart, "Recent and new results in rational L_2 approximation," in *Modeling, Robustness and Sensitivity Reduction in Control Systems*, R. F. Curtin, Ed. Berlin, Germany: Springer-Verlag, 1987, pp. 119–126.

[14] M. Aoki, "Control of large-scale dynamic systems by aggregation," *IEEE Trans. Automat. Contr.*, vol. AC-13, pp. 246–253, 1968.

[15] ——, "Some approximation methods for estimation and control of large scale systems," *IEEE Trans. Automat. Contr.*, pp. 173–182, 1978.

[16] J. Hickin and N. K. Sinha, "Canonical forms for aggregated models," *Int. J. Contr.*, pp. 473–485, 1978.

[17] J. Lam and Y. S. Hung, "On the stability of projections of balanced realizations," *Linear Algebra and Its Appl.*, vol. 257, pp. 163–182, 1997.

[18] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems*. London, U.K.: Springer-Verlag, 1994.

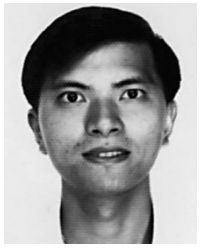
[19] I. M. James, *The Topology of Stiefel Manifolds*. Cambridge, U.K.: Cambridge Univ. Press, 1976.

[20] D. Enns, "Model reduction with balance realizations: An error bound and a frequency weighted generalizations," in *Proc. IEEE Conf. Decision and Control*, 1984, pp. 127–132.

[21] K. Glover, R. F. Curtain, and J. R. Partington, "Realization and approximation of linear infinite-dimensional systems with error bound," *SIAM J. Contr. Optim.*, vol. 26, pp. 863–898, 1988.

[22] J. C. Willems, "Realization of systems with internal passivity and symmetrical constraints," *J. Franklin Inst.*, vol. 301, pp. 605–621, 1976.

[23] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications*. New York: Wiley, 1971.
[24] P. T. Kabamba, "Balanced gains and their significance for L_2 model reduction," *IEEE Trans. Automat. Contr.*, vol. AC-30, pp. 690–693, 1985.
[25] D. Enns, "Model reduction for control system design," Ph.D. dissertation, Dept. of Aeronautics and Astronautics, Stanford Univ., 1984.
[26] A. Lepschy, G. A. Mian, G. Pinato, and U. Viaro, "Rational L_2 approximation: A nongradient algorithm," in *Proc. IEEE Conf. Decision and Control*, 1991, pp. 2321–2323.
[27] Y. Hung and A. MacFarlane, "Multivariable feedback: A quasiclassical approach," in *Lecture Notes in Control and Information Sciences.* New York: Springer-Verlag, 1982.
[28] W.-Y. Yan, L. Xie, and J. Lam, "Convergent algorithms for frequency weighted L_2 model reduction," *Syst. Contr. Lett.*, vol. 31, pp. 11–20, 1997.
[29] L. Xie, W.-Y. Yan, and Y. C. Soh, "L_2 optimal filter reduction: A closed-loop approach," *IEEE Trans. Signal Process.*, vol. 46, pp. 11–20, 1998.
[30] P. R. Aigrain and E. M. Williams, "Synthesis of n-reactance networks for desired transient response," *J. Appl. Phys.*, pp. 587–600, 1949.
[31] R. E. Skelton, "Cost decomposition of linear systems with application to model reduction," *Int. J. Contr.*, vol. 32, pp. 1031–1055, 1980.
[32] R. E. Skelton and A. Yousuff, "Component cost analysis of large scale systems," *Int. J. Contr.*, vol. 37, pp. 285–304, 1983.



Wei-Yong Yan received the B.S. degree in mathematics from Nankai University, Tianjin, China, in 1983, the M.S. degree in systems science from Academia Sinica, Beijing, China, in 1986, and the Ph.D. degree in systems engineering from the Australian National University, Canberra, in 1990.

From 1990 to 1992, he worked as a Research Fellow in the Department of Systems Engineering, Australian National University. He was a Lecturer in Applied Mathematics at the University of Western Australia from 1993 to 1994, prior to joining the Nanyang Technological University in Singapore first as a Lee Kuan Yew Fellow and later as a Senior Lecturer in the School of Electrical and Electronic Engineering. Since 1998, he has been a Senior Lecturer in the School of Electrical and Computer Engineering at Curtin University of Technology, Perth, Australia. His current research interests include the areas of control and signal processing.

James Lam (S'87–M'88–SM'99) received the first class B.Sc. degree in mechanical engineering from the University of Manchester in 1983. He then obtained the M.Phil. and Ph.D. degrees in the area of control engineering from the University of Cambridge in 1985 and 1988, respectively.

He has held faculty positions at the City University of Hong Kong and the University of Melbourne. He is now an Associate Professor in the Department of Mechanical Engineering, the University of Hong Kong, and is holding a concurrent Professorship at the Northeastern University, China. His research interests include model reduction, robust and fault tolerant control, delay systems, and generalized systems.

Dr. Lam is a Chartered Mathematician and a Fellow of the Institute of Mathematics and Its Applications (U.K.).