The Sylvester equation and approximate balanced reduction*

D.C. Sorensen
Department of Computational and Applied Mathematics
Rice University
Houston, Texas 77251
sorensen@rice.edu
http://www.caam.rice.edu/~sorensen
and
A.C. Antoulas[1]
Department of Electrical and Computer Engineering
Rice University
Houston, Texas 77251
aca@rice.edu
http://www.ece.rice.edu/~aca

January 14, 2002

Abstract

The purpose of this paper is to investigate the problem of iterative computation of approximately balanced reduced order systems. The resulting approach is completely automatic once an error tolerance is specified and also yields an error bound. This is to be contrasted with existing projection methods, namely PVL (Padé via Lanczos) and rational Krylov, which do not satisfy these properties. Our approach is based on the computation and approximation of the *cross gramian* of the system. The cross gramian is the solution of a Sylvester equation and therefore some effort is dedicated to the study of this equation leading to some new insights. Our method produces a low rank approximation to this gramian in factored form and thus directly provides a reduced order model and a reduced basis for the orignal system. It is well suited to large scale problems because there are no matrix factorizations of the large (sparse) system matrix. Only matrix-vector products are required.

Key words: model reduction, projection methods, balancing, gramians, Sylvester equations.

Contents

*This work was supported in part by the NSF through Grants DMS-9972591, CCR-9988393, and ACI-0082645.
[1] Corresponding author.

1 Introduction

Model reduction has a long history in the area of systems and control. In fact, there is a vast literature on the general topic of dimension reduction in dynamical systems. Recently, there has been renewed interest in projection methods for model reduction. Three leading efforts in this area are Padé via Lanczos (PVL) [10], multipoint rational interpolation [14], and implicitly restarted dual Arnoldi [23].

The PVL approach exploits the deep connection between the (nonsymmetric) Lanczos process and classic moment matching techniques [12], [10]; for an overview see [37]. The multipoint rational interpolation approach utilizes the rational Krylov method of Ruhe [26] to provide moment matching of the transfer function at selected frequencies and therefore to obtain enhanced approximation of the transfer function over a broad frequency range. These techniques have proven to be very effective. PVL has enjoyed considerable success in circuit simulation applications. Rational interpolation achieves remarkable approximation of the transfer function with very low order models. Nevertheless, there are shortcomings to both approaches. In particular, since the methods are local in nature, it is difficult to establish rigorous error bounds. Heuristics have been developed that appear to work, but no global results exist. Secondly, the rational interpolation method requires selection of interpolation points. At present, this is not an automated process and relies on ad-hoc specification by the user.

The approach that we are proposing here is more closely related to the implicitly restarted dual Arnoldi approach developed in [23]. The dual Arnoldi method runs two separate Arnoldi processes, one for the reachability subspace, and the other for the observability subspace and then constructs an oblique projection from the two orthogonal Arnoldi basis sets. The basis sets and the reduced model are updated using a generalized notion of implicit restarting. The updating process is designed to iteratively improve the approximation properties of the model. Essentially, the reduced model is reduced further, keeping the best features, and then expanded via the dual Arnoldi processes to include new information. The goal is to achieve approximation properties related to balanced realizations. Other related approaches [8, 35, 36, 33] work directly with projected forms of the two Lyapunov equations (2) to obtain low rank approximations to the system gramians. An overview of similar model reduction methods can be found in [38].

One problem with working with the two Lyapunov equations separately and then applying dense methods to the reduced equations is consistency. One cannot be certain that the two separate basis sets are the ones that would have been selected if the full system gramians had been available.

This difficulty has led us to consider a different approach based upon the notion of a cross gramian that was introduced in [11].

The systems $\mathbf{\Sigma}$ that we will deal with are linear and time invariant of the form

$$\mathbf{\Sigma}: \begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases} \quad \text{denoted as} \quad \mathbf{\Sigma} := \left(\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right) \tag{1}$$

Here A, B, C, D are real $n \times n$, $n \times m$, $p \times n$, and $p \times m$ matrices, while u, y, x are vector valued functions of time. We assume throughout that A is stable (eigenvalues are in the open left half-plane) and minimal, that is, reachable and observable. In the sequel, since D is not important for model reduction, it will not be considered ($D = 0$).

Large systems of this form arise in circuit simulation; they also arise through spatial discretization of certain time dependent PDE control systems such as parabolic equations subject to boundary control.

We present a new computational approach to model reduction that addresses some fundamental difficulties with existing dimension reduction techniques. These issues are central to the development of robust and widely applicable software. In this paper we develop a projection method for obtaining an approximate reduced order balancing transformation directly. Exact balancing is a very attractive form of model reduction since:

- There is a bound on the error of the response of the reduced system.
- Stability is preserved.
- It is fully automatic once a desired error tolerance is specified.

Exact balancing however, requires complete determination of either the cross gramian or of both the reachability and observability gramians. The approach which is proposed in the sequel, preserves the third property, but because of approximation error it only approximately satisfies the first; moreover, it may need a slight correction to satisfy the second property. In practice, all three properties seem to be achieved regularly.

Our primary goal is to develop an *implicit restarting method* that iteratively produces an approximation of specified rank k to the *cross-gramian* that is nearly best approximation of rank k to the full gramian.

In section 6, we provide simple examples which illustrate how the proposed model reduction method is applied to linear systems arising in various situations. We observe that the Hankel singular values of these systems decay extremely rapidly. Hence very low rank approximations to the system gramians are possible resulting in accurate low order reduced models. A discussion of this phenomenon is available in [4].

The remainder of the paper is organized as follows. The next section reviews some known results and introduces the cross gramian. Section 3 revisits the Sylvester equation and offers some insights into its solution; a relation with rational interpolation is also noted. Furthermore an expression for the error of the approximants in the $\mathcal{H}_2$ norm is derived. Section 4 presents the proposed iterative model reduction algorithm. The issue of stability and the issue of extensions to MIMO (multi-input and multi-output) systems are discussed in section 5. The paper concludes with experimental results and a short summary.

2 Review of some known facts and some new ones

Suppose that $\mathbf{\Sigma}$ defined in (1) is reachable, observable, and stable. It is well known [1] that there are unique symmetric positive definite matrices $\mathcal{P}$ and $\mathcal{Q}$ which are solutions to the following Lyapunov equations:

$$A\mathcal{P} + \mathcal{P}A^{^T} + BB^{^T} = 0, \quad A^{^T}\mathcal{Q} + \mathcal{Q}A + C^{^T}C = 0 \tag{2}$$

$\mathcal{P}$ is referred to as the *reachability* gramian and $\mathcal{Q}$ is referred to as the *observability* gramian. There are well known numerically stable direct methods to solve these equations [21, 24, 17].

The matrices $\mathcal{P}$ and $\mathcal{Q}$ are indeed gramians in the following sense: Recall that the impulse response of a system $\mathbf{\Sigma}$ is: $h(t) = Ce^{At}B,\ t \geq 0$. Now, consider the following two maps:

the *input-to-state map*: $\xi(t) = e^{At}B$ and

the *state-to-output map*: $\eta(t) = Ce^{At}$.

If the input to the system is the impulse $\delta(t)$, the resulting state is $\xi(t)$; moreover, if the initial condition of the system is $x(0)$, in the absence of a forcing function u, the resulting output is $y(t) = \eta(t)x(0)$. The *gramians* corresponding to $\xi(t)$ and $\eta(t)$ are:

$$\mathcal{P} = \int_0^\infty \xi(t)\xi(t)^{^T}\,dt = \int_0^\infty e^{At}BB^{^T}e^{A^{^T}t}\,dt \ \text{ and } \ \mathcal{Q} = \int_0^\infty \eta(t)^{^T}\eta(t)\,dt = \int_0^\infty e^{A^{^T}t}C^{^T}Ce^{At}\,dt,$$

which are well known to be solutions to the Lyapunov equations (2). The *Hankel singular values* $\sigma_i(\mathbf{\Sigma})$ of the system $\mathbf{\Sigma}$, are an important set of parameters for model reduction. These are defined with respect to the Hankel operator.

One can show [1] that the Hankel singular values are the square roots of the eigenvalues of the product $\mathcal{PQ}$: $\sigma_i(\mathbf{\Sigma}) = \sqrt{\lambda_i(\mathcal{PQ})}$. These Hankel singular values are easily seen to be invariant under state space transformations. If T is any non-singular matrix of order n, the transformed system $\mathbf{\Sigma} = \left(\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & \end{array}\right)$, where $\hat{A} = T^{-1}AT$, $\hat{B} = T^{-1}B$, $\hat{C} = CT$, has gramians $\hat{\mathcal{P}} = T^{-1}\mathcal{P}T^{-^T}$ and $\hat{\mathcal{Q}} = T^T\mathcal{Q}T$. Thus $\hat{\mathcal{P}}\hat{\mathcal{Q}} = T^{-1}\mathcal{P}\mathcal{Q}T$, and the eigenvalues of the product are preserved. The transformed system is in *balanced* form if $\hat{\mathcal{P}} = \hat{\mathcal{Q}}$; it is called *principal axis balanced* if

$$\hat{\mathcal{P}} = \hat{\mathcal{Q}} = \Sigma := \mathrm{diag}\,(\sigma_1, \sigma_2, \cdots, \sigma_n);$$

that is, both gramians are equal *and* diagonal. In this case T is called a balancing transformation. For details on balancing we refer to the original source [25]. We shall assume without loss of generality that the gramians of a principal axis balanced system have positive diagonal entries σ_i arranged in decreasing order $\sigma_i \geq \sigma_{i+1}$.

Model reduction of a balanced system is achieved through simple truncation. For any positive integer k, one may specify a reduced system $\mathbf{\Sigma}_k = \left(\begin{array}{c|c} A_k & B_k \\ \hline C_k & \end{array}\right)$, $\mathcal{P}_k$, $\mathcal{Q}_k$, obtained by taking the $k \times k$, $k \times m$, $p \times k$ leading blocks of $\hat{A}$, $\hat{B}$, $\hat{C}$, and the $k \times k$ leading blocks of the gramians, respectively. Each such reduced system has system gramians that are diagonal and equal since each of the leading $k \times k$ blocks of the diagonal gramians each satisfies the corresponding truncated k^{th} order Lyapunov equations. Each k^{th} order truncation provides a balanced reduced order system. Moreover, these reduced models have two important properties: the first is stability, and the second is the existence of an error bound. In order to state this error bound we need a definition. The $\mathcal{H}_\infty$ norm of a stable system $\mathbf{\Sigma}$ is the 2-induced norm of the associated input-output (convolution) operator:

$$\|\mathbf{\Sigma}\|_{\mathcal{H}_\infty} = \sup_{u \neq 0} \frac{\|y\|_2}{\|u\|_2} \quad \text{where} \quad y = h * u = \int_{-\infty}^{t} h(t-\tau)u(\tau), \;\; h(t) = Ce^{At}B, \;\; t \geq 0$$

It turns out that the $\mathcal{H}_\infty$ norm can be computed as the supremum of the largest singular value of the transfer function on the imaginary axis:

$$\|\mathbf{\Sigma}\|_{\mathcal{H}_\infty} = \sup_{\omega} \left(\sigma_{\max}\left[G(j\omega)\right]\right), \quad G(s) = C(sI - A)^{-1}B$$

The $\mathcal{H}_\infty$ norm can be interpreted as follows. Let the input u produce the output y; if u has unit 2-norm (i.e. unit energy), the 2-norm (energy) of the corresponding output y is bounded from above by the $\mathcal{H}_\infty$ norm of the system: $\|u\|_2 = 1 \Rightarrow \|y\|_2 \leq \|\mathbf{\Sigma}\|_{\mathcal{H}_\infty}$.

We can now state the two properties satisfied by the reduced systems obtain by means of balanced truncation.

- **Error bound**. The $\mathcal{H}_\infty$-norm of the error system has the following upper bound[1]:

$$\|\mathbf{\Sigma} - \mathbf{\Sigma}_k\|_{\mathcal{H}_\infty} \leq 2(\sigma_{k+1} + \sigma_{k+2} + \ldots + \sigma_n). \tag{3}$$

- **Stability**. The reduced system $\mathbf{\Sigma}_k$ is stable.

The error bound follows from the work of Glover [16] and Enns [9]. In terms of responses of the two systems (the original and the reduced) given the same input of unit norm, inequality (3) implies

$$\|y - \hat{y}\|_2 \;\leq\; 2(\sigma_{k+1} + \sigma_{k+2} + \ldots + \sigma_n),$$

where y is the response of the original system and $\hat{y}$ is the response of the reduced system corresponding to the same input u of unit 2-norm [16].

The stability result follows from the Lyapunov inertia relation. Given a square matrix A, let the number of eigenvalues in the left half-plane, on the imaginary axis, in the right half-plane be denoted by $in_-(A)$, $in_0(A)$, $in_+(A)$ respectively. The triple $(in_-(A), in_0(A), in_+(A))$ is called the *inertia* of A and is denoted by $\mathrm{in}(A)$.

Lemma 2.1 *Let (A, B) be a reachable pair and let the eigenvalues of A satisfy $\lambda_i(A) + \lambda_j(A) \neq 0$, for all i, j. The Lyapunov equation $A\mathcal{P} + \mathcal{P}A^T + BB^T = 0$, has a unique symmetric solution, and moreover* $\mathrm{in}(-A) = \mathrm{in}(\mathcal{P})$.

[1] In this error bound multiple singular values are counted only once.

In particular, this result implies that A is stable if and only if $\mathcal{P}$ is positive definite. For a new proof of this result see [2]. The significance of these results is to provide rigorous justification for using the reduced order model to predict behavior and enact control of the full system.

In the sequel, we will also make use of a different way of measuring the discrepancy between the original and the reduced systems, namely, the $\mathcal{H}_2$-norm. The $\mathcal{H}_2$ norm of a stable system $\mathbf{\Sigma}$, is defined as follows:

$$\|\mathbf{\Sigma}\|_{\mathcal{H}_2}^2 = \text{trace}\left[\int_0^\infty h^T(t)\, h(t)\, dt\right]$$

It follows readily that this can be expressed in terms of the gramians:

$$\|\mathbf{\Sigma}\|_{\mathcal{H}_2}^2 = \text{trace}\left[C\mathcal{P}C^T\right] = \text{trace}\left[B^T\mathcal{Q}B\right] \tag{4}$$

In section 3.2, we devote our attention to the derivation of an expression for the $\mathcal{H}_2$ norm of the error system.

Computational techniques for producing balancing transformations T for small to medium scale problems are well known [21, 24]. Such methods rely upon an initial Schur decomposition of A followed by additional factorization schemes of dense linear algebra. The computational complexity involves $\mathcal{O}(n^3)$ arithmetic operations and the storage of several dense matrices of order n; i.e. $\mathcal{O}(n^2)$ storage. For large state space systems, such as those arising as a discretization of the spatial operator in a time dependent PDEs, this approach to obtain a reduced model is clearly intractable. Yet, computational experiments indicate that such systems are representable with very low order models. This provides the primary motivation for seeking methods to construct projections of low order.

Our approach to model reduction as reported in section 4, consists in constructing low rank k approximations of a gramian which is different from the reachability and observability gramians. This is the **cross gramian** X which is introduced in the next section. In particular we set $X = V\hat{X}W^T$ with $W^TV = I_k$ and then project using V together with W (see (25)). We have developed an implicit restart mechanism that allows us to compute an approximation to the best rank k approximation to X. Furthermore, a reduced basis constructed from this procedure has an error estimate in the SISO case.

2.1 The cross gramian and balanced model reduction

The *cross gramian* X for square systems $\mathbf{\Sigma}$ $(m = p)$, is defined as the solution to the Sylvester equation

$$AX + XA + BC = 0. \tag{5}$$

If A is stable, it is well known that the solution of this equation can be written as

$$X = \int_0^\infty e^{At}BCe^{At}\, dt \tag{6}$$

The next three lemmas summarize the important properties of the cross gramian.

Lemma 2.2 *For square systems $\mathbf{\Sigma}$ the non-zero eigenvalues of the Hankel operator $\mathcal{H}$ are equal to the eigenvalues of the cross gramian X.*

Proof. Recall that the Hankel operator maps past inputs into future outputs, namely

$$\mathcal{H}:\ u_- \to y_+,\quad y_+(t) = \mathcal{H}(u_-)(t) = \int_{-\infty}^0 h(t-\tau)u(\tau)d\tau,\ t \geq 0$$

where $h(t) = Ce^{At}B$, $t \geq 0$ is the impulse response of $\mathbf{\Sigma}$. The eigenvalue problem of $\mathcal{H}$ for square systems is $\mathcal{H}(u_-) = \lambda y_+$, where $y_+(t) = u_-(-t)$. Let the function u_- be an eigenfunction of $\mathcal{H}$. Then

$$\int_{-\infty}^0 Ce^{A(t-\tau)}Bu_-(\tau)d\tau = \lambda u_-(-t)\ \Rightarrow\ Ce^{At}\int_{-\infty}^0 e^{-A\tau}Bu_-(\tau)d\tau = \lambda u_-(-t)$$

$$\int_0^\infty e^{At}BCe^{At}dt \int_{-\infty}^0 e^{-A\tau}Bu_-(\tau)d\tau = \lambda \int_0^\infty e^{At}Bu_-(-t)dt$$

The first integral is equal to the cross gramian X, the second and the third are equal to the same constant vector, say, $v \in \mathbb{R}^n$. We thus have $Xv = \lambda v$, which shows that if λ is a non-zero eigenvalue of $\mathcal{H}$ it is also an eigenvalue of X. Conversely, let (λ, v) be an eigenpair of X, i.e. $Xv = \lambda v$:

$$\left[\int_0^\infty e^{A\tau}BCe^{A\tau}d\tau\right] v = \lambda v \;\Rightarrow\; Ce^{At}\int_{-\infty}^0 \left[e^{-A\tau}BCe^{-A\tau}\right] v \, d\tau = \lambda Ce^{At} v$$

$$\Rightarrow \int_{-\infty}^0 Ce^{A(t-\tau)}B\underbrace{Ce^{-A\tau}v}_{\tilde{u}(\tau)} d\tau = \lambda Ce^{At}v = \lambda\tilde{u}(-t) \;\Rightarrow\; \mathcal{H}(\tilde{u})(t) = \lambda\tilde{u}(-t),\ t \geq 0$$

Therefore $\tilde{u}$ is an eigenfunction of $\mathcal{H}$. The proof is thus complete. ■

Remark 2.1 If the system is not symmetric, the Hankel singular values σ_i and the singular values π_i of X satisfy the following majorization inequalities: $\sum_{i=1}^k \sigma_i \geq \sum_{i=1}^k \pi_i$ and $\sum_{i=1}^k \pi_{n-i+1} \geq \sum_{i=1}^k \sigma_{n-i+1},\ i = 1, \cdots, n$. ■

Lemma 2.3 *Let* $\left(\begin{array}{c|c} A & B \\ \hline C & \end{array}\right)$ *be a stable SISO system that is reachable and observable. There is a nonsingular symmetric matrix* J *such that* $AJ = JA^T$, *and* $CJ = B^T$.

Proof. Let $\mathcal{K}_b := [B, AB, A^2B, \ldots, A^{n-1}B]$, $\mathcal{K}_c := [C^T, A^TC^T, (A^T)^2C^T, \ldots, (A^T)^{n-1}C^T]^T$, and define

$$\mathcal{H}_k := \mathcal{K}_c A^k \mathcal{K}_b.$$

The hypothesis implies both $\mathcal{K}_b$ and $\mathcal{K}_c$ are nonsingular, and it is easily shown that the Hankel matrix $\mathcal{H}_k$ is symmetric. Define $J := \mathcal{K}_b\mathcal{K}_c^{-T}$. Note $J = \mathcal{K}_c^{-1}\mathcal{H}_o\mathcal{K}_c^{-T}$ so that $J = J^T$. Moreover,

$$CJ = e_1^T\mathcal{K}_cJ^T = e_1^T\mathcal{K}_c\mathcal{K}_c^{-1}\mathcal{K}_b^T = e_1^T\mathcal{K}_b^T = B^T.$$

To complete the proof, we note that $AJ = A\mathcal{K}_b\mathcal{K}_c^{-T} = \mathcal{K}_c^{-1}\mathcal{H}_1\mathcal{K}_c^{-T}$, and hence $AJ = (AJ)^T = JA^T$. ■

As mentioned earlier, construction of a balancing transformation has typically relied upon solving the reachability and observability gramians $\mathcal{P}, \mathcal{Q}$ and developing a balancing transformation from the EVD of the product $\mathcal{PQ}$. However, it turns out that in the SISO case and in the case of symmetric MIMO systems, a balancing transformation can be obtained directly from the eigenvector basis for the cross gramian. Recall that a system $\mathbf{\Sigma}$ is called *symmetric* if its Markov parameters are symmetric, that is $CA^kB = (CA^kB)^T$, for all $k \geq 0$.

Lemma 2.4 *Let* $\left(\begin{array}{c|c} A & B \\ \hline C & \end{array}\right)$ *be a stable SISO system that is reachable and observable. Suppose that* (5) *is satisfied. Then* $X^2 = \mathcal{PQ}$ *which implies that* X *is diagonalizable with* $XZ = ZD$ *where* Z *is nonsingular and* D *is real and diagonal. Moreover, up to a diagonal scaling of its columns,* Z *is a balancing transformation for the system.*

The following corollary is stated without proof.

Corollary 2.1 *The statements of lemmas 2.3 and 2.4 hold for symmetric MIMO systems.*

Proof of lemma 2.4. First we note that the proof of Lemma 2.3 can be easily extended to hold for symmetric MIMO systems; consequently, there is a nonsingular matrix J such that $AJ = JA^T$, $CJ = B^T$, $J = J^T$. This may be used to obtain

$$A(XJ) + (XJ)A^T + BB^T = 0 \;\Rightarrow\; \boxed{\mathcal{P} = XJ}$$

From Lemma 2.1 it follows that $\mathcal{P} = XJ$ and $(XJ) = (XJ)^T$ is symmetric positive definite. Therefore, XJ has a Cholesky factorization $XJ = LL^T$ with L lower triangular and nonsingular. This implies $L^{-1}XL = L^TJ^{-1}L = QDQ^T$, where Q is orthogonal and D is real and diagonal since $L^TJ^{-1}L$ is symmetric. Thus,

$$X = ZDZ^{-1}, \quad \text{and} \quad J = ZD^{-1}Z^T, \quad \text{with} \quad Z := LQ.$$

Since $|D|^{1/2}D = D|D|^{1/2}$ we may replace $Z \leftarrow Z|D|^{-1/2}$ to obtain

$$J = ZD_JZ^T, \text{ and } X = ZDZ^{-1}, \text{ with } D_J = |D|D^{-1} = diag(\pm 1).$$

It follows that $XJ = Z(DD_J)Z^T$, and $J^{-1}X = Z^{-T}(D_JD)Z^{-1}$, since $D_J = diag(\pm 1)$ implies $D_J = D_J^{-1}$. If we put $S := DD_J = |D|$, we note S is a diagonal matrix with positive diagonal elements, and the above discussion together with Equation 5 gives

$$(Z^{-1}AZ)S + S(Z^TA^TZ^{-T}) + Z^{-1}BB^TZ^{-T} = 0.$$

Multiplying the cross gramian equation on the left by J^{-1} we obtain

$$A^TJ^{-1}X + J^{-1}XA + C^TC = 0 \Rightarrow \boxed{\mathcal{Q} = J^{-1}X}$$

This equation in turn implies that $\mathcal{Q} = (J^{-1}X) = (J^{-1}X)^T$. and after some straightforward manipulations, we find that $(Z^{-1}AZ)^TS + S(Z^{-1}AZ) + Z^TC^TCZ = 0$. To conclude the proof, we note that $CJ = B^T$ implies $CZD_J = B^TZ^{-T}$, and hence the system transformed by Z is indeed balanced. ■

Corollary 2.2 *If the cross gramian is diagonal then the system is essentially balanced, that is, a diagonal linear transformation will balance it. In particular Z can be taken to be a diagonal matrix such that $CZe_j = \pm B^TZ^{-T}e_j$. This diagonal transformation is constructed trivially from the entries of B and C.*

As mentioned earlier, in the SISO case the absolute values of the diagonal entries of D are the Hankel singular values of the system. If we assume that the diagonal entries of D have been ordered in decreasing order of magnitude then the $n \times k$ matrix Z_k consisting of the leading k columns of Z provide a truncated balanced realization with all of the desired stability and error properties.

The question then is how to compute a reasonable approximation to Z_k directly. We wish to avoid computing **all** *of Z and then* **truncating**, *especially in the large-scale setting.*

3 A closer look at the Sylvester equation

Since the *Sylvester equation* underlies the definition and hence construction of the *cross gramian*, this section is devoted to a re-examination of this equation. Below, an expression for the solution of the Sylvester equation (7), is derived. This solution is in terms of the eigenvalues and the eigenvectors of the matrices involved. This result, although trivially derived, has important consequences. The form of the solution establishes namely a connection with the so-called Löwner matrix and consequently with rational interpolation.

In section 3.2, we compute explicitly the $\mathcal{H}_2$ norm of the error system (see (20)). Notice that this is a *computable upper bound* for the the $\mathcal{H}_2$ norm of the error system (23). Special cases, like (21) and an upper bound (23) are also discussed.

3.1 A solution of the Sylvester equation

Consider the Sylvester equation:

$$AX + XH = M,\ A \in \mathbb{R}^{n\times n},\ H \in \mathbb{R}^{k\times k},\ M \in \mathbb{R}^{n\times k} \Rightarrow X \in \mathbb{R}^{n\times k} \tag{7}$$

We will assume for simplicity that A and H are diagonalizable matrices; let their EVD be

$$AV = V\Lambda,\ W^*H = \Omega W^*$$

where $V := [v_1\ \cdots\ v_n]$, $\Lambda := \text{diag}\,(\lambda_i)$, $W := [w_1\ \cdots\ w_k]$, $\Omega := \text{diag}\,(\mu_j)$; let also $\hat{W} := W^{-*} = [\hat{w}_1 \cdots \hat{w}_k]$.

Theorem 3.1 *Using the above data, the solution of the Sylvester equation (7), can be written as the* **sum** *of* **rank-one** *matrices:*

$$\boxed{X = \sum_{i=1}^{n} v_i \hat{v}_i^* \, M \, (\lambda_i I + H)^{-1} = \sum_{j=1}^{k} (\mu_j I + A)^{-1} \, M \, \hat{w}_j w_j^*} \tag{8}$$

These expressions can also be written more compactly as:

$$\boxed{X = V \begin{bmatrix} \hat{v}_1^* \, M \, (\lambda_1 I + H)^{-1} \\ \vdots \\ \hat{v}_n^* \, M \, (\lambda_n I + H)^{-1} \end{bmatrix} = \begin{bmatrix} (\mu_1 I + A)^{-1} \, M \, \hat{w}_1 & \cdots & (\mu_k I + A)^{-1} \, M \, \hat{w}_k \end{bmatrix} W^*} \tag{9}$$

Proof. From $AX + XH = M \Rightarrow (A - \lambda_i I)X + X(H + \lambda_i I) = M$, where λ_i is an eigenvalue of A, with corresponding left/right eigenvectors $\hat{v}_i$, v_i; hence $\hat{v}_i^*(A - \lambda_i I) = 0$, which leads to $\hat{v}_i^* X = \hat{v}_i^* M (\lambda_i I + H)^{-1} \Rightarrow v_i \hat{v}_i^* X = v_i \hat{v}_i^* M (\lambda_i I + H)^{-1}$. Due to $\sum_{i=1}^{n} v_i \hat{v}_i^* = I_n$ we have $\sum_{i=1}^{n} v_i \hat{v}_i^* X = X$. This proves the formula on the left-hand side. The proof of the right-hand side formula follows similarly. ■

Remark 3.1 **(a)** To use one of the above formulas to compute X will require either computing the full eigensystem of the large (sparse) matrix A followed by n inversions (factorizations) of $k \times k$ matrices $(\lambda_i I + H)$, or computing the full eigensystem of the small matrix H followed by k inversions (factorizations) of $n \times n$ matrices $(\mu_i I + A)$. If $n \gg k$, and A is sparse, the latter is clearly preferable.
(b) The evaluation of the norm of such an expression (which will be required in section 3.2) depends on the norm of $v_i \hat{v}_i^*$ or of the norm of $\hat{w}_j w_j^*$; these quantities are encoutered frequently in calculations involving the resultant [13].
(c) The most striking aspect of the above solution however is the explicit connection that it provides to rational interpolation. This connection is the subject of section 3.1.1 which follows. ■

3.1.1 The Sylvester equation and the Löwner matrix

This section is devoted to the connection between the solution to the Sylvester equation and the Löwner matrix, and consequently rational interpolation. Several discussions of the second author with Paul van Dooren in July-August 2001, contributed to some of the ideas below.

Some results on the Löwner matrix

A few basic results explored in [5] and [6] will now be summarized. We are given two arrays of complex numbers:

$$C = \{(x_i, y_i) : i = 1, \cdots, r\}, \; R = \{(\hat{x}_i, \hat{y}_i) : i = 1, \cdots, p\}$$

Assuming that $x_i \neq \hat{x}_j$, for all i, j, we can form the following matrix

$$\mathcal{L} := \begin{bmatrix} \frac{\hat{y}_1 - y_1}{\hat{x}_1 - x_1} & \cdots & \frac{\hat{y}_1 - y_r}{\hat{x}_1 - x_r} \\ \vdots & & \vdots \\ \frac{\hat{y}_p - y_1}{\hat{x}_p - x_1} & \cdots & \frac{\hat{y}_p - y_r}{\hat{x}_p - x_r} \end{bmatrix} \in \mathbb{R}^{p \times r} \tag{10}$$

$\mathcal{L}$ is called the *Löwner matrix* defined by means of the *row array* R and the *column array* C. As it turns out $\mathcal{L}$ is an important tool for solving the rational interpolation problem, which can be formulated as follows. We seek all rational functions

$$y(x) = \frac{n(x)}{d(x)}, \quad \text{with} \quad (n, d) \quad \text{coprime}$$

which *interpolate* the points of the arrays C and R, that is

$$y(x_i) = y_i, \; i = 1, \cdots, r \quad \text{and} \quad y(\hat{x}_i) = \hat{y}_i, \; i = 1, \cdots, p$$

The degree of y is defined as the largest between the degrees of the numerator and denominator: $\deg y = \max\{\deg n, \deg d\}$.

Let us now assume that such an interpolant $y(x)$ of degree n exists and is proper rational. Then according to realization theory, it can be expressed in terms of four constant matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^n$, $C \in \mathbb{R}^{1\times n}$, $D \in \mathbb{R}$:

$$y(x) = C(xI - A)^{-1}B + D$$

This expression immediately implies

$$[\mathcal{L}]_{i,j} := \frac{\hat{y}_i - y_j}{\hat{x}_i - x_j} = -C(\hat{x}_i I - A)^{-1}(x_j I - A)^{-1}B$$

Consequently, $\mathcal{L}$ can be factorized as follows:

$$\mathcal{L} = -\mathcal{O}_p(A, B)\mathcal{R}_r(C, A)$$

where

$$\mathcal{R}_r(A, B) := [(x_1 I - A)^{-1}B \ \cdots \ (x_r I - A)^{-1}B] \in \mathbb{R}^{n\times r} \tag{11}$$

and

$$\mathcal{O}_p(C, A) := \begin{bmatrix} C(\hat{x}_1 I - A)^{-1} \\ \vdots \\ C(\hat{x}_p I - A)^{-1} \end{bmatrix} \in \mathbb{R}^{p\times n} \tag{12}$$

It can be shown that *realization* can be viewed as *rational interpolation* with interpolating points at *infinity* [7]. In this case the Hankel matrix factors in a product of an observability times a reachability matrix. In analogy, as shown in [7], if the interpolation points lie in finite locations in the complex plane, the Löwner matrix replaces the Hankel matrix as main tool. Furthermore, $\mathcal{L}$ factors in a product as shown above. In analogy with realization, we will call $\mathcal{O}_p$ the *generalized observability matrix* and $\mathcal{R}_r$ the *generalized reachability matrix* associated with the arrays C, R of the underlying interpolation problem.

Connection with the Sylvester equation

Let us now assume for argument's sake that M is rank one; it can then be written as $M = m_1 m_2^T$, $m_1 \in \mathbb{R}^n$, $m_2 \in \mathbb{R}^k$. Thus (9) can be expressed as follows:

$$X = \underbrace{V \begin{bmatrix} \hat{v}_1^* m_1 & & \\ & \ddots & \\ & & \hat{v}_n^* m_1 \end{bmatrix}}_{\tilde{V}} \underbrace{\begin{bmatrix} m_2^*(\lambda_1 I + H)^{-1} \\ \vdots \\ m_2^*(\lambda_n I + H)^{-1} \end{bmatrix}}_{\mathcal{O}_n(m_2^*, H)} = \underbrace{\begin{bmatrix} (\mu_1 I + A)^{-1}m_1 \ \cdots \ (\mu_k I + A)^{-1}m_1 \end{bmatrix}}_{\mathcal{R}_k(A, m_1)} \underbrace{\begin{bmatrix} m_2^*\hat{w}_1 & & \\ & \ddots & \\ & & m_2^*\hat{w}_k \end{bmatrix} W^*}_{\tilde{W}^*} \tag{13}$$

In the above expression $\tilde{V}$ is the matrix of scaled right eigenvectors of A while $\tilde{W}$ is the matrix of scaled left eigenvectors of H. It is interesting to notice that the remaining matrices $\mathcal{O}_n$ and $\mathcal{R}_k$ are directly related with rational interpolation. In order to establish this relationship we refer to the results of [5] and [6]. There it is shown that the Löwner matrix can be factorized as a product of *generalized* reachability and observability matrices. The matrices defined above, namely $\mathcal{R}_k(A, m_1)$ and $\mathcal{O}_n(m_2^*, H)$ are precisely these generalized matrices:

$$\boxed{X = \tilde{V}\mathcal{O}_n(m_2^*, H) = \mathcal{R}_k(A, m_1)\tilde{W}^*} \tag{14}$$

Let us now examine the special case where $H = A \in \mathbb{R}^{n\times n}$ and $m_1 = B \in \mathbb{R}^n$ and $m_2^* = C \in \mathbb{R}^{1\times n}$, we have $VW^* = I$ and therefore upto scaling X^2 is similar to the associated Löwner matrix

$$\boxed{X^2 = \tilde{V}\mathcal{O}_n(C, A)\mathcal{R}_k(A, B)\tilde{W}^*, \ \mathcal{L} = -\mathcal{O}_n(C, A)\mathcal{R}_k(A, B)} \tag{15}$$

3.2 The $\mathcal{H}_2$ norm of the error system in model reduction

Consider the system $\mathbf{\Sigma} = \left(\begin{array}{c|c} A & B \\ \hline C & \end{array}\right)$, which is stable and minimal. In projection methods there always exists a basis in which the reduced order system is obtained by simple truncation. The balanced basis and the associated balanced truncation are a special case. We will thus assume that in this special basis A, B, and C are partitioned accordingly: $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, $B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}$, $C = \begin{pmatrix} C_1 & C_2 \end{pmatrix}$; the reduced order system is $\hat{\mathbf{\Sigma}} = \left(\begin{array}{c|c} A_{11} & B_1 \\ \hline C_1 & \end{array}\right)$. Let the three gramians, that is, the reachability, observability, and cross gramians, be $\mathcal{P}$, $\mathcal{Q}$, and X, respectively. As mentioned earlier (see (4)) the $\mathcal{H}_2$ norm of $\mathbf{\Sigma}$ can be expressed in terms of the reachability and the observability gramians. Consequently by (6), for square symmetric systems it can also be expressed in terms of the cross gramian:

$$\|\mathbf{\Sigma}\|_{\mathcal{H}_2}^2 = \mathrm{trace}\,[C\mathcal{P}C^*] = \mathrm{trace}\,[B^*\mathcal{Q}B] = \mathrm{trace}\,[CXB]$$

In order to compute the $\mathcal{H}_2$ norm of the error system, in addition to X, we need the quantities Y, Z, and $\hat{X}$, which satisfy the Sylvester equations below. In partitioned form, $X = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}$, $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, $Z = \begin{pmatrix} Z_1 & Z_2 \end{pmatrix}$, and $\hat{X}$ satisfy

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix} + \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} + \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}\begin{pmatrix} C_1 & C_2 \end{pmatrix} = 0 \tag{16}$$

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} + \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} A_{11} + \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} C_1 = 0 \tag{17}$$

$$A_{11}\begin{pmatrix} Z_1 & Z_2 \end{pmatrix} + \begin{pmatrix} Z_1 & Z_2 \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} + B_1\begin{pmatrix} C_1 & C_2 \end{pmatrix} = 0 \tag{18}$$

$$A_{11}\hat{X} + \hat{X}A_{11} + B_1C_1 = 0 \tag{19}$$

Thus the $\mathcal{H}_2$ norm of the error system resulting from model reduction is the square root of the following expression:

$$\|\,\mathbf{\Sigma}_{\mathrm{e}}\,\|_{\mathcal{H}_2}^2 = \mathrm{trace}\left[\begin{pmatrix} C_1 & C_2 & C_1 \end{pmatrix}\begin{pmatrix} X_{11} & X_{12} & -Y_1 \\ X_{21} & X_{22} & -Y_2 \\ -Z_1 & -Z_2 & \hat{X} \end{pmatrix}\begin{pmatrix} B_1 \\ B_2 \\ B_1 \end{pmatrix}\right]$$

Theorem 3.2 *In terms of the quantities defined above, the $\mathcal{H}_2$ norm of the error system is*

$$\|\,\mathbf{\Sigma}_{\mathrm{e}}\,\|_{\mathcal{H}_2}^2 = \mathrm{trace}\,[\,C_2X_{22}B_2\,] + \mathrm{trace}\,[\,C_1(\hat{X} - X_{11})B_1\,] + \mathrm{trace}\,[\,A_{12}(X_{2,:}\,Y) + (Z\,X_{:,2})A_{21}\,] \tag{20}$$

The proof of this result is given in the Appendix.

Remark 3.2 The first term in the above expression is the $\mathcal{H}_2$-norm of the neglected subsystem of the original system; the second term, is the difference between the $\mathcal{H}_2$-norms of the reduced-order system and the dominant subsystem of the original system; finally the third term is the sum of the inner product of the second block row of the cross gramian with Y and that of Z with the second block column of the cross gramian (each term weighted by the block off-diagonal terms of A). Finally, $\hat{X} - X_{11}$ satisfies the Sylvester equation:

$$A_{11}(\hat{X} - X_{11}) + (\hat{X} - X_{11})A_{11} = A_{12}X_{21} + X_{12}A_{21}$$

This implies that if either the cross gramian or A have small (zero) off-diagonal elements, then $\hat{X} - X_{11}$ will be small (zero). ■

It turns out that the above formula becomes really useful when the cross gramian is block diagonal, i.e. $X_{12} = 0$ and $X_{21} = 0$. In particular, when the system is balanced, X is diagonal. The first consequence of this assumption is that $\hat{X} = X_{11}$, hence the second term in (20) vanishes. As for the last term it becomes $A_{12}X_{22}Y_2 + Z_2X_{22}A_{21}$; a moment's reflection shows that the trace of these two terms is the same. Hence there holds

Corollary 3.1 *Under the assumption of theorem 3.2, if $X_{12} = 0$ and $X_{21} = 0$, the $\mathcal{H}_2$ norm of the error system reduces to*

$$\boxed{\| \mathbf{\Sigma}_e \|_{\mathcal{H}_2}^2 = \text{trace}\,[\, C_2 X_{22} B_2 \,] + 2\,\text{trace}\,[\, A_{12} X_{22}\, Y_2 \,]} \tag{21}$$

To further simplify this expression we notice that $Y - X_{:,1}$ satisfies the Sylvester equation

$$A \begin{pmatrix} Y_1 - X_{11} \\ Y_2 \end{pmatrix} + \begin{pmatrix} Y_1 - X_{11} \\ Y_2 \end{pmatrix} A_{11} = \begin{pmatrix} 0 \\ X_{22} \end{pmatrix} A_{21}$$

Thus, using the expression (8) derived for the solution of the Sylvester equation, we obtain

$$A_{12} X_{22} Y_2 = \sum_{i=1}^{k} w_i \bar{w}_i^* \begin{pmatrix} 0 & A_{12} X_{22} \end{pmatrix} (\mu_i I + A)^{-1} \begin{pmatrix} 0 \\ X_{22} A_{21} \end{pmatrix} \tag{22}$$

where μ_i, w_i, $\bar{w}_i$ are the eigenvalues, right-, left-eigenvectors of the A-matrix of the reduced-order system: A_{11}. Thus we can derive an upper bound for the $\mathcal{H}_2$ norm of the error in terms of the $\mathcal{H}_\infty$ norm of the auxiliary system

$$\mathbf{\Sigma}_{\text{aux}} = \left(\begin{array}{cc|c} A_{11} & A_{12} & 0 \\ A_{21} & A_{22} & X_{22} A_{21} \\ \hline 0 & A_{12} X_{22} & 0 \end{array} \right)$$

Notice that the transfer function of $\mathbf{\Sigma}_{\text{aux}}$ can be written as

$$G_{\text{aux}}(s) = A_{12} X_{22} \left[sI - A_{22} - A_{21} (sI - A_{11})^{-1} A_{12} \right]^{-1} X_{22} A_{21}$$

It is worth noting that this expression is *quadratic* in the neglected part X_{22}, of the cross gramian. We thus obtain the following upper bound for the $\mathcal{H}_2$ norm

$$\boxed{\| \mathbf{\Sigma}_e \|_{\mathcal{H}_2}^2 \;\leq\; \text{trace}\,[\, C_2 X_{22} B_2 \,] + 2\, k\, \rho\, \| \mathbf{\Sigma}_{\text{aux}} \|_{\mathcal{H}_\infty}} \tag{23}$$

where $\rho = \max_i \{ \| w_i \bar{w}_i^* \| \;:\; i = 1, \cdots, k \}$, is related to the condition number of the eigenvectors of A_{11}. If we are dealing with a balanced realization, since this realization is sign symmetric, the corresponding left and right eigenvectors are sign orthogonal and therefore $\rho = 1$.

Remark 3.3 **(a)** It should be stressed that while the first term in (23) is linear in X_{22}, the second is quadratic. Therefore if the system is well conditioned and X_{22} contains all the small Hankel singular values, the second term can be neglected and the expression becomes: $\| \mathbf{\Sigma}_e \|_{\mathcal{H}_2}^2 \approx \text{trace}\,[\, C_2 X_{22} B_2 \,]$.
(b) If $\mathbf{\Sigma}$ is in balanced form then the cross gramian is not only block diagonal, put simply diagonal. Therefore, taking into account the sign symmetry of balanced realizations, we have: $\| \mathbf{\Sigma}_e \|_{\mathcal{H}_2}^2 \approx \sigma_{k+1} B_{k+1}^2 + \cdots + \sigma_n B_n^2$, where k is the size of A_{11}, σ_i are as usual the Hankel singular values and B_i the i$^{\text{th}}$ entry of B (which, up to sign, happens to be equal to the i$^{\text{th}}$ entry of C). Recall that in the balanced case the $\mathcal{H}_2$ norm of the original system is $\sigma_1 B_1^2 + \cdots + \sigma_n B_n^2$, which is a weighted sum of the Hankel singular values. ■

4 Approximate balancing through low rank approximation of the cross gramian

4.1 Computing a rank k approximation to the cross gramian

We now consider the problem of computing the best rank k approximation X_k to the cross gramian X. We seek a restarting mechanism analogous to implicit restarting for eigenvalue computations that will enable us to compute X_k directly instead of computing all of X and truncating.

Suppose $X = USV^T$ is the SVD of X. Let $X = U_k S_k V_k^T + \hat{U}_k \hat{S}_k \hat{V}_k^T$ where $U = [U_k, \hat{U}_k]$, and $V = [V_k, \hat{V}_k]$. Then

$$(AX + BC)V_k = -XAV_k$$

and it follows that

$$AU_k S_k + BCV_k = -U_k S_k (V_k^T AV_k) + E_k$$

where $E_k = -\hat{U}_k \hat{S}_k \hat{V}_k^T AV_k$. Observe that $\|E_k\|_2 = \mathcal{O}(\sigma_{k+1})$ where σ_{k+1} is the first neglected singular value, that is, the (k+1)-st singular value of X.

This suggests the following type of iteration; First obtain a projection of A

$$-AV = VH + F, \quad \text{with} \quad V^T V = I, \quad V^T F = 0,$$

where V is $n \times m$ and H is $m \times m$ with $k < m \ll n$. We require that this projection yield a stable matrix H. If it does not do so, then V must be modified. As developed in [15] , we can use the technique of implicit restarting to cast out the unstable eigenvalues. This would result in a V and an H of smaller dimension. As long as this dimension remains greater than k, we can execute the remaining steps described below. Should the dimension fall below k, then some method must be used to expand the basis set. As an example, one could use a block form of Arnoldi with the remaining V as a starting block of vectors.

Once V and H have been computed, solve the Sylvester equation

$$AW - WH = -BCV$$

for W. Take the SVD of W. If Y_k denotes the matrix of right singular vectors for W corresponding to the largest k singular values then we have

$$\begin{aligned} AWY_k + BCVY_k = WY_k(Y_k^T HY_k) + \hat{E}_k \quad &\Rightarrow \quad AU_k S_k + BCV_k = U_k S_k H_k + \hat{E}_k \\ -AVY_k = VY_k(Y_k^T HY_k) + \hat{F}_k \quad &\Rightarrow \quad -AV_k = V_k H_k + \hat{F}_k. \end{aligned}$$

We must now have a mechanism to expand the subspace generated by $V_k \equiv VY_k$ so that the above process can be repeated. One possibility is to form

$$E = A^T V_k S_k + V_k S_k H_y^T + C^T B^T U_k, \quad \text{where} \quad H_k = Y_k^T HY_k,$$

and then solve $A^T Z + ZH_k^T = -E$. This represents a residual correction to the solution of the Sylvester equation $AX + XA + BC = 0$ when projected on the left through multiplication by V_k^T. However, instead of adding this correction to the basis set V_k, we simply adjoin the columns of Z to the subspace spanned by the columns of V_k and project $-A$ onto this space. This portion of the iteration is analogous to the "Davidson part" of the Jacobi-Davidson algorithm for eigenvalue calculation proposed by Sleijpen and van der Vorst [32]. These ideas are summarized in the algorithm sketched in Figure 1 for the complete iteration.

There have been many ideas for the numerical solution of Lyapunov and Sylvester equations [18, 34, 22, 27, 28, 33, 35, 36]. This approach nearly gives the best rank k approximation directly. It is prevented from giving the best such approximation since $\|\hat{E}_k\| = \mathcal{O}(\sigma_{k+1})$. In fact, the best rank k approximation is not a fixed point of this iteration. A slightly modified correction equation is needed to achieve this; for details we refer the reader to [30]. Note the iteration of Hodel, et. al. [18] for the Lyapunov equation also suffers from this difficulty.

We take the QR factorization of W at step 2.1 and then compute the SVD of the very low dimensional matrix R to find the truncated basis. The single matrix multiplication $U \leftarrow UQ_k$ at step 2.2 is much more effecient than accumulating the Givens transformatons required to compute the SVD against an $n \times m$ matrix, and then discarding the last $m - k$ columns.

The equation $AW + BCV = WH$ will introduce directions in the Krylov space $\mathcal{K}(A, B)$ while the correction equation $A^T Z + ZH_w = -E$ will introduce directions from the Krylov space $\mathcal{K}(A^T, C^T)$ although this is not quite as obvious.

(Implicit) Restarting Algorithm

1. $(-\mathbf{A})\mathbf{V} = \mathbf{V}\mathbf{H} + \mathbf{F}$ with $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ and $\mathbf{V}^T\mathbf{F} = \mathbf{0}$
2. while (*not_converged*),

2.1 Solve Projected Sylvester Eqn. in $\mathcal{K}(\mathbf{A}, \mathbf{B})$:

Solve $\mathbf{A}\mathbf{W} + \mathbf{B}\mathbf{C}\mathbf{V} = \mathbf{W}\mathbf{H}$
$[\mathbf{U}, \mathbf{R}] = qr(\mathbf{W})$; $[\mathbf{Q}, \mathbf{S}, \mathbf{Y}] = svd(\mathbf{R})$;

2.2 Contract the Space (Keep Largest Singular Values):

$\mathbf{S}_k = \mathbf{S}(1:k, 1:k)$;
$\mathbf{U} \leftarrow \mathbf{U}\mathbf{Q}_k$; $\mathbf{V} \leftarrow \mathbf{V}\mathbf{Y}_k$;
$\mathbf{H} \leftarrow \mathbf{Y}_k^T\mathbf{H}\mathbf{Y}_k$;

2.3 Correct Projected Sylvester Eqn. in $\mathcal{K}(\mathbf{A}^T, \mathbf{C}^T)$:

Form $\mathbf{E} = \mathbf{A}^T\mathbf{V}\mathbf{S}_k + \mathbf{V}\mathbf{S}_k\mathbf{H}^T + \mathbf{C}^T\mathbf{B}^T\mathbf{U}$
Solve $\mathbf{A}^T\mathbf{Z} + \mathbf{Z}\mathbf{H}^T = -\mathbf{E}$

2.4 Expand the Space – Adjoin Correction and Project:

Factor $[\mathbf{V}, \mathbf{R}] = qr([\mathbf{V}, \mathbf{Z}])$
$\mathbf{F} = -\mathbf{A}\mathbf{V}$; $\mathbf{H} \leftarrow \mathbf{V}^T\mathbf{F}$;
$\mathbf{F} \leftarrow \mathbf{F} - \mathbf{V}\mathbf{H} = (-\mathbf{A})\mathbf{V} - \mathbf{V}\mathbf{H}$;

Figure 1: An Implicitly Restarted Method.

4.2 A special Sylvester equation

Efficient solution of a special Sylvester equation provide the key to Steps 2.1 and 2.3 in Algorithm 1. Both of these steps result in a special Sylvester equation of the form

$$AZ + ZH + M = 0$$

where H is a $k \times k$ stable matrix, and M is an $n \times k$ matrix with $k \ll n$. (Note that step 2.3 involves A^T but the same method applies.) The only thing special about this Sylvester equation is that A is very much larger in dimension than H.

The method we propose stems from the following observation. Suppose that a partial real Schur decomposition of the form

$$\begin{bmatrix} A & M \\ 0 & -H \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} R \tag{24}$$

has been obtained, where R is a real quasi-upper triangular $k \times k$ matrix, and $V_1^T V_1 + V_2^T V_2 = I_k$. This decomposition provides the solution Z if V_2 is nonsingular. Indeed, if V_2 is nonsingular, put $Z = V_1 V_2^{-1}$ to obtain

$$AZ + M = Z(V_2 R V_2^{-1}) = -ZH.$$

Moreover, it turns out that V_2 is nonsingular if and only if the eigenvalues of R are the eigenvalues of $-H$.

Since the eigenvalues of A are in the open left half-plane and the eigenvalues of $-H$ are in the open right half-plane, the k eigenvalues of largest real part for the block upper triangular matrix in (24) are the desired eigenvalues. When k is small, it is possible to compute the eigenvalues of H in advance of the computation of the partial Schur decomposition in (24). Within this framework, the implicitly restarted Arnoldi method [29] (implemented in ARPACK [31]) can be used effectively to compute this partial Schur decomposition. If there is a reasonable gap between the eigenvalues of H and the imaginary axis, then IRA will be successful in computing the k eigenvalues

of largest real part using only matrix vector products. In any case, exact knowledge of the desired eigenvalues provides several opportunities to enhance convergence. One possibility is to use a single Cayley transformation (costing one sparse direct factorization) to map the eigenvalues of A to the interior and the eigenvalues of H to the exterior of the unit disk. Further justification of this approach including a proof of the equivalence and a more thorough discussion of possibilities for acceleration are given in [30].

An alternative would be to construct a Schur decomposition of H and transform the equation to

$$A\hat{Z} + \hat{Z}R + \hat{M} = 0$$

and then solve for the columns of $\hat{Z}$ from left to right via

$$(A - \rho_{jj}I)z_j = -m_j - \sum_{i=1}^{j-1} z_i \rho_{ij}, \quad \text{for} \quad j = 1, 2, ..., k,$$

where ρ_{ij} are the elements of the upper triangular matrix R, and z_j, m_j are the columns of $\hat{Z}, \hat{M}$. This would require a separate sparse direct factorization of a large $n \times n$ complex matrix at each step j. There would be k such factorizations and each of these would have a potentially different sparsity pattern for L and U due to pivoting for stability. Staying in real arithmetic would require working with quadratic factors involving A and hence would destroy sparsity.

4.3 Approximate balancing transformation from X_k

The above procedure provides $X_k = U_k S_k V_k^T$. A closer examination of the iteration shows that U_k is an ortho-normal basis for a subspace of the reachability subspace $\mathcal{K}(A, B)$, while V_k provides an ortho-normal basis for a subspace of the observability subspace $\mathcal{K}(A^T, C^T)$. It is tempting to just take V_k as a reduced basis since this amounts to orthogonal projection. However, it is clear that this choice neglects important information. Therefore we are lead to the idea of obtaining an *approximate* balancing transformation through the eigenvectors of X_k corresponding to non-zero eigenvalues.

In the SISO case, we know that X has real eigenvalues. If X_k were obtained as $X_k = Z_k D_k W_k^T$ where the diagonal elements of D_k are the eigenvalues of largest magnitude (i.e. the dominant Hankel singular values) and $W_k^T Z_k = I_k$ then, as discussed previously, Z_k would provide a balancing transformation. Instead, we have the best rank k approximation to X in X_k. Therefore, we shall attempt to approximate the relevant eigenvector basis for X with an eigenvector basis for X_k. It is easily seen that any eigenvector of X_k corresponding to a nonzero eigenvalue must be in the range of U_k. In fact, we see that

$$X_k U_k S_k^{1/2} = U_k S_k^{1/2} G$$

where $G = S_k^{1/2} V_k^T U_k S_k^{1/2}$. If $GZ = ZD_k$ with D_k real and diagonal, then taking

$$Z_k = U_k S_k^{1/2} Z |D_k|^{-1/2} \text{sign}\,(D_k) \quad \text{and} \quad W_k = V_k S_k^{1/2} Z^{-T} |D_k|^{-1/2}$$

provides $X_k = Z_k D_k W_k^T$ with $W_k^T Z_k = I_k$, where $\text{sign}\,(D_k) = D_k |D_k|^{-1}$. Note also that

$$X Z_k = X_k Z_k + (X - X_k) Z_k = Z_k D_k + \mathcal{O}(\sigma_{k+1})$$

and thus we have an approximate balancing transformation represented by this Z_k. Our reduced model is (A_k, B_k, C_k), with

$$\boxed{A_k = W_k^T A Z_k, \quad B_k = W_k^T B, \quad C_k = C Z_k.} \tag{25}$$

This seems to be a natural construction, but it does not lead to a projected equation. If we define E_k as the residual via the equation

$$AX_k + X_k A + BC = E_k$$

and multiply on the left by W_k^T and on the right by Z_k, we obtain

$$A_k D_k + D_k A_k + B_k C_k^T = \hat{E}_k,$$

where $\hat{E}_k = W_k^T E_k Z_k$. If $\hat{E}_k = 0$ were true, we would be able to conclude that the reduced system is balanced and stable. However, we only have $\|\hat{E}\| = \|A\|\mathcal{O}(\sigma_{k+1})$, and therefore stability and/or balancing of the reduced model do not automatically follow.

Remark 4.1 The preceding discussion was primarily concerned with the SISO case. For MIMO systems we propose the idea of *embedding* the system in a symmetric system; this is explored in section 5.2. For square MIMO systems ($m = p$) it is also possible to proceed directly by considering the solution X of $AX + XA + BC = 0$. In this case X can have complex eigenvalues and the real Schur form is a more appropriate eigen-decomposition. Let

$$S_k^{1/2}(V_k^T U_k)S_k^{1/2}\hat{Z} = \hat{Z}R_k$$

be a real Schur form, i.e. $\hat{Z}^T\hat{Z} = I_k$ and R_k is real and quasi-upper triangular. We define

$$Z_k := U_k S_k^{1/2}\hat{Z}R_k^{-1} \quad \text{and} \quad W_k := V_k S^{1/2}\hat{Z}.$$

Then $W_k^T Z_k = I_k$, $X_k Z_k = Z_k R_k$, and $W_k^T X_k = R_k W_k^T$. Proceeding exactly as before with the reduced model, we arrive at the reduced equation

$$A_k R_k + R_k A_k + B_k C_k^T = \hat{E}_k$$

Again, we only have that $\|\hat{E}\| = \|A\|\mathcal{O}(\sigma_{k+1})$. This reduction appears to work quite well in practice, but further analysis is necessary to obtain a proof of convergence. ■

5 Further issues

5.1 Stability and balancing of the reduced model

Our methods pertain to stable systems and therefore, it is important in many applications that the reduced model be stable as well. We must have the eigenvalues of the reduced matrix A_k in the left half-plane. The reduced model obtained through the algorithm shown in Figure 1, is almost always stable in practice, but occasionally it might be unstable.

One approach to achieving a stable reduced model is to apply the techniques developed in Grimme, van Dooren, Sorensen [15] to the projected quantities. That would amount to applying the implicit restarting mechanism to rid the projected matrix of unstable modes.

Another important question is whether the reduced model is balanced. Since we are only *approximating* a balancing transformation, the reduced model obtained may not be balanced. Thus additional conditions would have to be worked out to assure this. For stability, these conditions are tied to the truncation tolerance.

5.2 Extension to MIMO systems

In the large scale setting, there is a clear advantage to working with the cross gramian instead of the working with the two gramians related to reachability and observability. In addition to the fact that only one Sylvester equation need be solved, there is the question of compatibility that arises when working with the pair of gramians. Since two separate projections must be made, one cannot be certain that the two subspaces are the same as the ones that would have been achieved through computing the full gramians and then truncating.

The crucial property of the three gramians in the SISO case is $X^2 = \mathcal{PQ}$. It is easy to see that this relationship holds true for MIMO systems which are *symmetric*, i.e. the transfer function is symmetric. Of course, this is not generally the case. In order to make use of this property, we propose to embed the given system into a symmetric

system with more inputs and outputs but the same number of state variables. Given the m-input, p-output system $\mathbf{\Sigma} = \left(\begin{array}{c|c} A & B \\ \hline C & \end{array}\right)$, we seek $\tilde{B} \in \mathbb{R}^{n\times p}$ and $\tilde{C} \in \mathbb{R}^{m\times n}$ such that the augmented system

$$\hat{\mathbf{\Sigma}} = \left(\begin{array}{c|c} \hat{A} & \hat{B} \\ \hline \hat{C} & \end{array}\right) = \left(\begin{array}{c|cc} A & \tilde{B} & B \\ \hline C & & \\ \tilde{C} & & \end{array}\right)$$

is square and symmetric, i.e. the Markov parameters $\hat{C}\hat{A}^{\ell}\hat{B} \in \mathbb{R}^{(m+p)\times(m+p)}$, are symmetric for all $\ell \geq 0$. That this can be done follows readily from properties of system realizations. The important aspect of this embedding is that the complexity (McMillan degree, number of states) of the system has not increased. Therefore the norm ($\mathcal{H}_2$ or $\mathcal{H}_\infty$) of the original system is upper-bounded by that of the augmented system.

5.2.1 MIMO systems and the symmetrizer

A basic result in linear algebra states that any (square) matrix has a *symmetrizer*, i.e. a symmetric matrix which commutes with its transpose. Let $J = J^{T}$ be a symmetrizer for A, i.e. $AJ = JA^{T}$. The following quantities are defined

$$\tilde{B} = JC^{T} \;\text{ and }\; \tilde{C} = B^{T}J^{-1} \;\;\Rightarrow\;\; \hat{B} = \begin{bmatrix}\tilde{B} & B\end{bmatrix},\; \hat{C} = \begin{bmatrix} C \\ \tilde{C}\end{bmatrix} \;\Rightarrow\; \hat{B} = J\hat{C}^{T}$$

The *augmented system* is

$$\left(\begin{array}{c|c} A & \hat{B} \\ \hline \hat{C} & \end{array}\right) \in \mathbb{R}^{(n+m+p)\times(n+p+m)}$$

This system has the property that its Hankel operator is symmetric. Therefore, using the same tools as above we can show that

$$\hat{\mathcal{P}} = \hat{X}J^{-1}, \hat{\mathcal{Q}} = J\hat{X} \;\Rightarrow\; \hat{X}^2 = \hat{\mathcal{P}}\hat{\mathcal{Q}}$$

The three equations satisfied by the augmented system are

$$\begin{aligned} A\hat{\mathcal{P}} + \hat{\mathcal{P}}A^{T} + \hat{B}\hat{B}^{T} = 0 \Rightarrow\; A\hat{\mathcal{P}} + \hat{\mathcal{P}}A^{T} + JC^{T}CJ + BB^{T} = 0 \;&\Rightarrow\; \boxed{\hat{\mathcal{P}} = \mathcal{P} + J\mathcal{Q}J} \\ A^{T}\hat{\mathcal{Q}} + \hat{\mathcal{Q}}A + \hat{C}^{T}\hat{C} = 0 \Rightarrow\; A^{T}\hat{\mathcal{Q}} + \hat{\mathcal{Q}}A + C^{T}C + J^{-1}BB^{T}J^{-1} = 0 \;&\Rightarrow\; \boxed{\hat{\mathcal{Q}} = \mathcal{Q} + J^{-1}\mathcal{P}J^{-1}} \\ A\hat{X} + \hat{X}A + \hat{B}\hat{C} = 0 \;&\Rightarrow\; \boxed{\hat{X} = J\hat{\mathcal{Q}} = \hat{\mathcal{P}}J^{-1}} \end{aligned}$$

5.2.2 The choice of symmetrizer

At this stage, the symmetrizer is any symmetric matrix which satisfies $AJ = JA^{T}$. For simplicity, let us assume that A is *diagonalizable* and has been transformed to the basis where A is diagonal. In this case the symmetrizer J is an arbitrary diagonal matrix. The question arises, of how to best choose the diagonal entries of J.

The criterion which will be chosen is that the *Hankel operator of the augmented system be* **close** *to that of the original system.* In order to address this issue we will make use of the variational characterization of balancing as developed in a series of papers by Helmke and Moore [19, 20].

Let T be a basis change in the state space. Then as already mentioned the gramians are transformed as follows: $\mathcal{P} \to T\mathcal{P}T^{T}$, $\mathcal{Q} \to T^{-T}\mathcal{P}T^{-1}$. Consider the following criterion:

$$\mathcal{J}(T) = \text{trace}\,(T\mathcal{P}T^{T}) + \text{trace}\,(T^{-T}\mathcal{Q}T^{-1}) \tag{26}$$

For a fixed state space basis, the above quantity is equal to the sum of the eigenvalues of the reachability and of the observability gramians. The question which arises is to find the minimum of $\mathcal{J}(T)$ as a function of all non-singular transformations T. First notice that $\mathcal{J}$ can be expressed in terms of the positive definite matrix $\Phi = T^{T}T$:

$$\mathcal{J} = \text{trace}\,\left[\mathcal{P}\Phi + \mathcal{Q}\Phi^{-1}\right],\; \Phi = T^{T}T > 0 \tag{27}$$

The following result is due to Helmke and Moore.

Proposition 5.1 *The minimum of $\mathcal{J}$ is $\mathcal{J}_* = \min_{\Phi>0} \mathcal{J} = 2\sum_{k=1}^{n} \sigma_k$, and the minimizer is* $\Phi_* = \mathcal{P}^{-1/2}\left(\mathcal{P}^{1/2}\mathcal{Q}\mathcal{P}^{1/2}\right)^{1/2}\mathcal{P}^{-1/2}$.

It readily follows that with the eigenvalue decomposition $\mathcal{P}^{1/2}\mathcal{Q}\mathcal{P}^{1/2} = U\Sigma^2 U^T$, a resulting balancing transformation is $T_b = \Sigma^{1/2}U^T\mathcal{P}^{-1/2}$. In other words

$$T_b \mathcal{P} T_b^T = T_b^{-T}\mathcal{Q}T_b^{-1} = P$$

The transformation T_b is unique up to orthogonal similarity (P need not be diagonal). In our case, we wish to compute an appropriate symmetrizer. The criterion (the sum of the traces of the two gramians) for the augmented system is as follows

$$\mathcal{J}(J) = \mathrm{trace}\,(\mathcal{P} + J\mathcal{Q}J) + \mathrm{trace}\,(\mathcal{Q} + J^{-1}\mathcal{P}J^{-1}) = \mathrm{trace}\,(\mathcal{P}+\mathcal{Q}) + \underbrace{\mathrm{trace}\,(J\mathcal{Q}J + J^{-1}\mathcal{P}J^{-1})}_{\mathcal{J}_1(J)}$$

We will compute the diagonal $J = \mathrm{diag}\,(j_1, \cdots, j_n)$, so that the above trace is minimized. The first summand does not depend on J. The second is

$$\mathcal{J}_1(J) = \sum_{i=1}^{n}\left[p_{ii}j_i^2 + q_{ii}\frac{1}{j_i^2}\right]$$

The minimum of $\mathcal{J}_1$ is achieved for $j_i^2 = \sqrt{\frac{q_{ii}}{p_{ii}}}$:

$$\min \mathcal{J}_1 = 2\sum_{i=1}^{n}\sqrt{p_{ii}q_{ii}} \quad \Rightarrow \quad \min \mathcal{J} = \left(\sqrt{p_{ii}} + \sqrt{q_{ii}}\right)^2 \tag{28}$$

This should be compared with twice the sum of the trace of the two gramians, namely $2\sum_{i=1}^{n}(p_{ii}+q_{ii})$. The difference of the two traces is $\sum_{i=1}^{n}(\sqrt{p_{ii}} - \sqrt{q_{ii}})^2$.

The above computation was carried through under the assumption that A is diagonal. Let in particular,

$$A = \mathrm{diag}\,(-\lambda_1, \cdots, -\lambda_n),\ B = \begin{pmatrix} b_1^* \\ \vdots \\ b_n^* \end{pmatrix},\ C = [c_1\ \cdots c_n] \ \text{ where } \ b_i \in \mathbb{R}^m,\ c_i \in \mathbb{R}^p$$

In this representation the diagonal entries of the two gramians are:

$$p_{ii} = \frac{b_i^* b_i}{\lambda_i + \lambda_i^*},\ q_{ii} = \frac{c_i c_i^*}{\lambda_i + \lambda_i^*}$$

Furthermore by applying the state space transformation T, which is diagonal with entries $t_{ii} = \sqrt{\frac{c_i c_i^*}{b_i^* b_i}}$, we can assure that $p_{ii} = q_{ii}$.

Conclusion. If A is diagonalizable there exists a symmetrizer which guarantees that the sum of the singular values of the augmented system is *exactly* twice that of the original. This is the best one can do in this case.

5.3 Computational Efficiency

The method we have proposed appears to be quite expensive computationally. However, it seems to converge quite rapidly. In fact, our test examples indicate that for medium scale problems ($n \approx 400$) it is already competitive with existing dense methods. Moreover, it can provide balanced realizations where most existing methods fail.

In the large scale setting, it is not clear that this approach will compete with those of (PVL) [10], or multipoint rational interpolation [14]. The error bounds and the stability of the reduced order model come at a price. The proposed implicit restarting approach involves a great deal of work associated with solving the required special Sylvester equations. However, the iterative method is based upon adjoining residual corrections to the current approximate solutions to the cross gramian equations. We observe very rapid convergence in practice. Usually three major iterations are sufficient. Nevertheless, there is no proof of convergence and this seems to be a general difficulty with a projection approach of this type (see e.g., [18]).

6 Experimental results

Matlab codes implementing the model reduction methods discussed above have been developed and tested considerably. We have experimented with a number of generated problems (randomly chosen eigenvalues in the left half-plane with a specified number of these forced to be near the imaginary axis). The largest systems we have dealt with successfully, have dimension $n = 10,000$, and have been reduced approximately to dimension $k = 20 - 30$. Here we will report numerical experiments performed on low order systems. The following 6 cases are described in detail in [3]; the last column shows the degree of the reduced system obtained for a tolerance of $\tau = 10^{-5}$:

	n	m	p	k
Structural Model	270	3	3	37
Building Model	48	1	1	31
Heat Model	197	2	2	5
CD Player	120	1	1	12
Clamped Beam	348	1	1	13
Butterworth Filter	100	1	1	35

Most of the problems we have seen exhibit the behavior that is indicative of the success for this approach. Namely, both the singular values and the eigenvalues of X decay very rapidly indicating that X is typically approximated well with an X_k of very low rank. In the examples above we compute X_k where the value k is determined by a cut-off, namely taking the first positive integer k such that

$$\sigma_{k+1} < \tau\sigma_1 .$$

A typical value of τ is $\tau = \sqrt{\text{machine precision}}$. This is a *user specified tolerance* that determines the user's desired bound on the $\mathcal{H}_\infty$ norm error estimate given by the neglected Hankel singular values. The important point is that the specification of this single tolerance determines everything. The above cut-off criterion determines the value of k required to meet the users error specification. The order k of the reduced model and the eigenvalue placement are completely automatic. Moreover, the value of τ directly determines the error of approximation for $\|y - y_k\|$ over the entire frequency range.

In figure 2, we show results for the 5th system in the list above, namely displacement and stress in a clamped beam with Rayleigh (or proportional) damping. The input is a force at the unclamped end, while the output is displacement. The figure shows a Bode plot of the response of the original system which is of order 348 (solid line) and response of the reduced order model of order 48 (dashed line). Again, we emphasize that the order of the reduced model is automatically determined once the error requirement is specified. The results for the remaining systems as well as comparisons and calculation of the $\mathcal{H}_2$ and $\mathcal{H}_\infty$ norms of the error systems can be found in [3]. It is worth noting that for the above systems the traditional balanced truncation method and the *approximate* balanced truncation method proposed in this paper, give practically the same results.

7 Conclusions

In this paper we propose an iterative model reduction method based on the cross gramian. This quantity satisfies a Sylvester equation and thefore we first devote our attention to a re-examination of various properties of this equation.

Subsequently the approach for computing approximate partially balanced realizations of linear systems is derived. We have demonstrated computationally that this approach performs well. Moreover, we have established error bounds for the response approximation in the SISO case. The approach enjoys the highly desirable quality of being completely automatic once an error tolerance has been specified. This is to be contrasted with existing projection methods. There are no global error bounds available for PVL or for rational interpolation methods. Moreover, the rational interpolation schemes all require the user to select interpolation points where moments of the transfer function are to be matched. Our approach selects interpolation points automatically according to the specified error tolerance.

Despite the obviously desirable features of the cross gramian approach proposed here, many open questions remain. There are a number of refinements with respect to performance, robustness, and accuracy that are worth

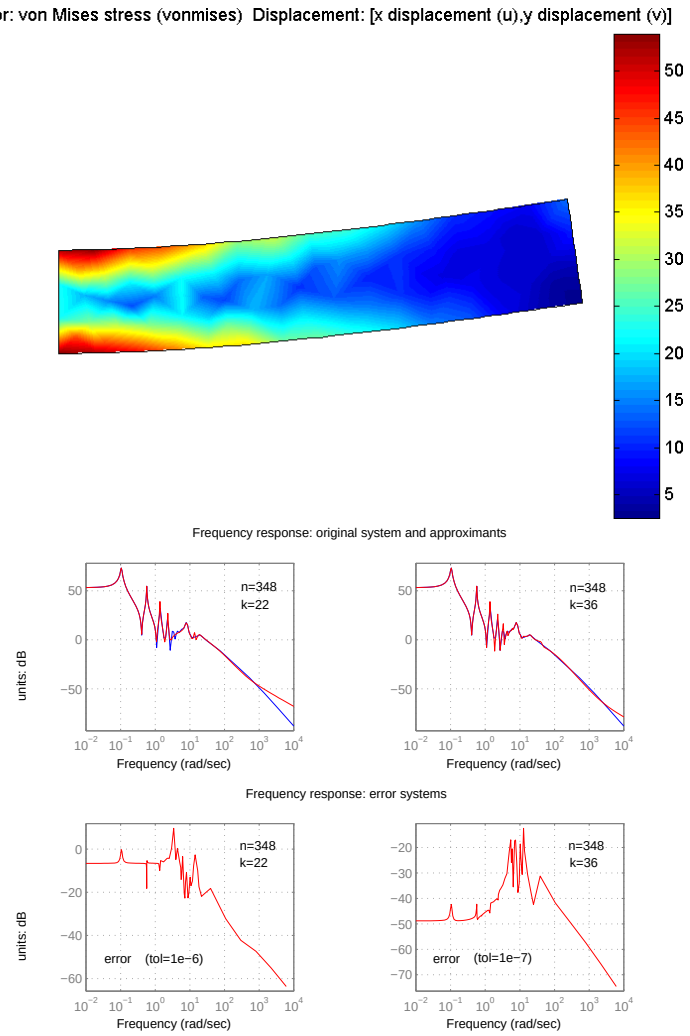

Figure 2: *Clamped Beam.* Upper plots: comparison of frequency response of original and reduced systems for tolerance 10^{-4} and 10^{-5}. Lower plots: frequency response of error systems for the given tolerances.

pursuing. There are also several areas related to MIMO systems that require both theoretical and algorithmic developments.

References

[1] A.C. Antoulas, *Lectures on the approximation of large-scale dynamical systems*, to be published by SIAM, Draft (2001).

[2] A.C. Antoulas and D.C. Sorensen, *Lyapunov, Lanczos, and Inertia*, Linear Algebra and Its Applications, **326**: 137-150 (2001).

[3] A.C. Antoulas, D.C. Sorensen, and S. Gugercin, *A survey of model reduction methods for large-scale systems*, Contemporary Mathematics, **280**: 193-219 (2001).

[4] A.C. Antoulas, D.C. Sorensen, and Y. Zhou *On the decay rate of the Hankel singular values and related issues*, Technical Report, Dept. of Computational and Applied Math., CAAM TR-01-09, Rice University, May 2001.

[5] A.C. Antoulas and B.D.O. Anderson, *State-space and polynomial approaches to rational interpolation*, in Realization and Modelling in System Theory, Proc. MTNS-89, edited by M.A. Kaashoek, J.H. van Schuppen, and A.C.M. Ran, Volume I, pages 73-81 (1990).

[6] B.D.O. Anderson and A.C. Antoulas, *Rational interpolation and state-variable realizations*, Linear Algebra and Its Applications, Special Issue on Matrix problems, **137/138**: 479-509 (1990).

[7] A.C. Antoulas and B.D.O. Anderson, *On the scalar rational interpolation problem*, IMA J. of Mathematical Control and Information, Special Issue on Parametrization problems, edited by D. Hinrichsen and J.C. Willems, **3**: 61-88 (1986).

[8] D.L. Boley, *Krylov space methods on state-space control models*, Circuits, Systems, and Signal Processing, **13**: 733-758 (1994).

[9] D.F. Enns, *Model reduction with balanced realizations: An error bound and frequency weighted generalization*, Proc. of the IEEE Conference on Decision and Control, 127-132 (1984).

[10] P. Feldman and R.W. Freund, *Efficient linear circuit analysis by Padé approximation via a Lanczos method*, IEEE Trans. Computer-Aided Design, **14**, 639-649, (1995).

[11] K.V. Fernando and H. Nicholson, *On the structure of balanced and other principal representations of SISO systems*, IEEE Trans. Automatic Control, **AC-28**: 228-231 (1983).

[12] W.B. Gragg and A. Lindquist, *On the partial realization problem*, Linear Algebra and Its Applications, Special Issue on Linear Systems and Control, **50**: 277-319 (1983).

[13] A. Greenbaum, *Using the Cauchy integral formula and partial fractions decomposition of the resolvent to estimate* $\|f(A)\|$, Technical Report, Dept. of Mathematics, University of Washington, Seattle, 17 pages (2000).

[14] E.J. Grimme, *Krylov Projection Methods for Model Reduction*, Ph.D. Thesis, ECE Dept., U. of Illinois, Urbana-Champaign,(1997).

[15] E.J. Grimme, D.C. Sorensen, and P. Van Dooren, *Model reduction of state space systems via an implicitly restarted Lanczos method*, Numerical Algorithms, **12**: 1-31 (1995).

[16] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their* L^∞*-error bounds*, International Journal of Control, **39**: 1115-1193 (1984).

[17] S.J. Hammarling, *Numerical solution of the stable, non-negative definite Lyapunov equation*, IMA J. Numerical Analysis, **2**: 303-323 (1982).

[18] A.S. Hodel, K. Poola, and B. Tenison, *Numerical solution of the Lyapunov equation by approximate power iteration*, Linear Algebra and Its Applications, **236**: 205-230 (1996).

[19] J.E. Perkins, U. Helmke, and J.B. Moore, *Balanced realizations via gradient flows*, Systems and Control Letters, **14**: 369-380 (1990).

[20] J.B. Moore, R.E. Mahoney, and U. Helmke, *Numerical gradient algorithms for eigenvalue and singular value calculations*, SIAM J. Matrix Anal. Applic., **SIMAX** (1994).

[21] M.T. Heath, A.J. Laub, C.H. Paige, and R.C. Ward, *Computing the singular value decomposition of a product of two matrices*, SIAM J. Sci. Stat. Computing, **7**: 1147-1159 (1986).

[22] D. Hu and L. Reichel, *Krylov-subspace methods for the Sylvester equation*, Linear Alg. and Appl., **172**, 283-313, (1992).

[23] I.M. Jaimoukha, E.M. Kasenally, *Implicitly restarted Krylov subspace methods for stable partial realizations*, SIAM J. Matrix Anal. Appl., **18**: 633-652 (1997).

[24] A.J. Laub, M.T. Heath, C.C. Paige, and R.C. Ward, *Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms*, IEEE Trans. Automatic Control, **AC-32**: 115-121 (1987).

[25] B.C. Moore, *Principal component analysis in linear systems: controllability, observability, and model reduction*, IEEE Transactions on Automatic Control, **AC-26**: 17-32 (1981).

[26] A. Ruhe, *Rational Krylov algorithms for nonsymmetric eigenvalue problems II: matrix pairs*, Linear Alg. Appl., **197**:283-295, (1984).

[27] Y. Saad, *Numerical solution of large Lyapunov equations*, in Signal Processing, Scattering and Operator Theory, and Numerical Methods, Proc. MTNS-89, **3**: 503-511, Birkhäuser (1990).

[28] V. Simoncini, *On the numerical solution of* $AX - XB = C$, BIT, **36**: 814-830 (1996).

[29] D.C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Applic., **13**: 357-385 (1992).

[30] D.C. Sorensen, *Low Rank Approximate Solutions to Lyapunov Equations*, Tech. Report TR01-05, CAAM Deptartment, Rice University, in preparation (2001).

[31] R. Lehoucq, D.C. Sorensen, and C. Yang, *ARPACK Users Guide: Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi methods*, SIAM Publications, Philadelphia, (1998). (software available at http://www.caam.rice.edu/software/ARPACK)

[32] G.L.G. Sleijpen and H. van der Vorst, *A Jacobi-Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl. **17**, 401-425,(1996).

[33] T. Penzl, *A cyclic low rank Smith method for large sparse Lyapunov equations*, SIAM J. Sci. Comp. (to appear).

[34] E. Wachspress, *The ADI model problem*, Monograph, NA Digest vol. 96, issue 36 (1996).

[35] J. Li, F. Wang, J. White, *An efficient Lyapunov equation-based approach for generating reduced-order models of interconnect*, Proc. 36th IEEE/ACM Design Automation Conference, New Orleans, LA, (1999).

[36] M. Kamon, F. Wang, and J. White, *Generating nearly optimal compact models from Krylov-subspace based reduced-order models*, IEEE Trans. on Circuits and Systems-II, **CAS 47**: 239-248 (2000).

[37] P. Van Dooren, *The Lanczos algorithm and Padé approximations*, Short Course, Benelux Meeting on Systems and Control, (1995).

[38] P. Van Dooren, *Gramian based model reduction of large-scale dynamical systems*, in Numerical Analysis 2000, pages 231-247 (2000).

8 Appendix: the proof of theorem 3.2

The $\mathcal{H}_2$ norm of the error system is the square root of the following expression:

$$\| \mathbf{\Sigma}_e \|^2_{\mathcal{H}_2} = \text{trace} \left[\begin{pmatrix} C_1 & C_2 & C_1 \end{pmatrix} \begin{pmatrix} X_{11} & X_{12} & -Y_1 \\ X_{12}^T & X_{22} & -Y_2 \\ -Z_1 & -Z_2 & \hat{X} \end{pmatrix} \begin{pmatrix} B_1 \\ B_2 \\ B_1 \end{pmatrix} \right] \Rightarrow$$

$$\begin{aligned} \| \mathbf{\Sigma}_e \|^2_{\mathcal{H}_2} \quad = \quad & \text{trace}[\, C_1 X_{11} B_1 + C_1 X_{12} B_2 - C_1 Y_1 B_1 \,] \\ & + \text{trace}[\, C_2 X_{21} B_1 + C_2 X_{22} B_2 \underbrace{- C_2 Y_2 B_1}_{\diamond} \,] \\ & + \text{trace}[\, -C_1 Z_1 B_1 \underbrace{- C_1 Z_2 B_2}_{\diamond} + C_1 \hat{X} B_1 \,] \end{aligned}$$

Using the off-diagonal entries of (16)

From the off-diagonal entries of the Sylvester equation (16) follow the relationships:

$$\begin{aligned} A_{11}X_{12} + A_{12}X_{22} + X_{11}A_{12} + X_{12}A_{22} + B_1C_2 &= 0 \;\Rightarrow \\ -B_1C_2Y_2 &= [A_{11}X_{12} + A_{12}X_{22} + X_{11}A_{12} + X_{12}A_{22}]Y_2 \\ A_{21}X_{11} + A_{22}X_{21} + X_{21}A_{11} + X_{22}A_{21} + B_2C_1 &= 0 \;\Rightarrow \\ -Z_2B_2C_1 &= Z_2\,[A_{21}X_{11} + A_{22}X_{21} + X_{21}A_{11} + X_{22}A_{21}] \end{aligned}$$

Substituting in the original expression for the $\mathcal{H}_2$-norm of $\mathbf{\Sigma}_e$, we get:

$$\begin{aligned} \| \mathbf{\Sigma}_e \|_{\mathcal{H}_2}^2 \;=\; & \operatorname{trace}[\,C_1X_{11}B_1 + C_1X_{12}B_2 - C_1Y_1B_1\,] \\ & + \operatorname{trace}[\,C_2X_{21}B_1 + C_2X_{22}B_2 + A_{11}X_{12}Y_2 + A_{12}X_{22}Y_2 + \underbrace{X_{11}A_{12}Y_2}_{\triangleleft} + X_{12}A_{22}Y_2\,] \\ & + \operatorname{trace}[\,-C_1Z_1B_1 + \underbrace{Z_2A_{21}X_{11}}_{\triangleleft} + Z_2A_{22}X_{21} + Z_2X_{21}A_{11} + Z_2X_{22}A_{21} + C_1\hat{X}B_1\,] \end{aligned}$$

Using the (1,1) entries of (17) and (18)

We now substitute the two $\triangleleft$ with the expressions below which come from the (1,1) entries of (17) and (18):

$$\begin{aligned} A_{11}Y_1 + A_{12}Y_2 + Y_1A_{11} + B_1C_1 = 0 \quad &\Rightarrow\; X_{11}A_{12}Y_2 = -X_{11}(A_{11}Y_1 + Y_1A_{11} + B_1C_1) \\ A_{11}Z_1 + Z_1A_{11} + Z_2A_{21} + B_1C_1 = 0 \quad &\Rightarrow\; Z_2A_{21}X_{11} = -(A_{11}Z_1 + Z_1A_{11} + B_1C_1)X_{11} \end{aligned}$$

to obtain

$$\begin{aligned} \| \mathbf{\Sigma}_e \|_{\mathcal{H}_2}^2 = \operatorname{trace}[\,& \underbrace{C_1X_{11}B_1}_{T} + \underbrace{C_1X_{12}B_2}_{\circ} - C_1Y_1B_1\,] \\ + \operatorname{trace}[\,& \underbrace{C_2X_{21}B_1}_{\bullet} + C_2X_{22}B_2 + \underbrace{A_{11}X_{12}Y_2}_{\circ} + A_{12}X_{22}Y_2 - X_{11}A_{11}Y_1 - X_{11}Y_1A_{11} \underbrace{-X_{11}B_1C_1}_{T} + \underbrace{X_{12}A_{22}Y_2}_{\circ}\,] \\ + \operatorname{trace}[\,& -C_1Z_1B_1 - A_{11}Z_1X_{11} - Z_1A_{11}X_{11} - B_1C_1X_{11} + \underbrace{Z_2A_{22}X_{21}}_{\bullet} + \underbrace{Z_2X_{21}A_{11}}_{\bullet} + Z_2X_{22}A_{21} + C_1\hat{X}B_1\,] \end{aligned}$$

Using the (2,1) entry of (17) and the (1,2) entry of (18)

Next, all terms containing X_{12} and X_{21} are collected; making use of the (2,1) entry of (17) and of the (1,2) entry of (18), we get:

$$\begin{array}{lllll} \circ & C_1X_{12}B_2 + A_{11}X_{12}Y_2 + X_{12}A_{22}Y_2 &= (B_2C_1 + Y_2A_{11} + A_{22}Y_2)X_{12} &= -A_{21}Y_1X_{12} \\ \bullet & C_2X_{21}B_1 + A_{22}X_{21}Z_2 + X_{21}A_{11}Z_2 &= X_{21}(B_1C_2 + Z_2A_{22} + A_{11}Z_2) &= -X_{21}Z_1A_{12} \end{array}$$

The error thus simplifies as follows (note that the two $*$ terms cancel):

$$\begin{aligned} \| \mathbf{\Sigma}_e \|_{\mathcal{H}_2}^2 \;=\; & \operatorname{trace}[\,\underbrace{-A_{21}Y_1X_{12} - C_1Y_1B_1}_{\circ}\underbrace{-X_{21}Z_1A_{12}}_{\bullet}\,] \\ & + \operatorname{trace}[\,C_2X_{22}B_2 + A_{12}X_{22}Y_2 \underbrace{-X_{11}A_{11}Y_1 - X_{11}Y_1A_{11}}_{\circ}\,] \\ & + \operatorname{trace}[\,\underbrace{-C_1Z_1B_1 - A_{11}Z_1X_{11} - Z_1A_{11}X_{11}}_{\bullet} - B_1C_1X_{11} + Z_2X_{22}A_{21} + C_1\hat{X}B_1\,] \end{aligned}$$

Using the diagonal entries of (16)

We now collect the $\circ$ and $\bullet$ terms, and use the diagonal entries of (16):

$$\begin{array}{lll} -B_1C_1Y_1 - X_{12}A_{21}Y_1 - A_{11}X_{11}Y_1 - X_{11}A_{11}Y_1 &= -(B_1C_1 + X_{12}A_{21} + A_{11}X_{11} + X_{11}A_{11})Y_1 \\ & = A_{12}X_{21}Y_1 \;\circ \\ -Z_1B_1C_1 - Z_1A_{12}X_{21} - Z_1X_{11}A_{11} - Z_1A_{11}X_{11} &= -Z_1(B_1C_1 + A_{12}X_{21} + X_{11}A_{11} + A_{11}X_{11}) \\ & = Z_1X_{12}A_{21} \;\bullet \end{array}$$

Collecting these expressions we finally obtain (20). ■