A posteriori error bounds for discrete balanced truncation[☆]

Y.Chahlaoui

Centre for Interdisciplinary Computational and Dynamical Analysis (CICADA), School of Mathematics, The University of Manchester, UK.

Abstract

Balanced truncation of discrete linear time-invariant systems is an automatic method once an error tolerance is specified and it yields an a priori error bound, which is why it is widely used in engineering for simulation and control. We derive a discrete version of Antoulas's $\mathcal{H}_2$-norm error formula and show how to adapt it to some special cases. We present an a posteriori computable upper bound for the $\mathcal{H}_2$-norm of the error system defined as the system whose transfer function corresponds to the difference between the transfer function of the original system and the transfer function of the reduced system. We also present a generalization of the $\mathcal{H}_2$-norm error formula to any projection of dynamics method. The main advantage of our results is that we use the information already available in the model reduction algorithm in order to compute the $\mathcal{H}_2$-norm instead of computing a new Gramian of the corresponding error system, which is computationally expensive. The a posteriori bound gives insight into the quality of the reduced system and it can be used to solve many problems accompanying the order reduction operation. Moreover, it is often more accurate in floating point arithmetic.

Keywords: model reduction, balanced truncation, a posteriori error bound, Gramians, Stein equations, $\mathcal{H}_2$-norm
2010 MSC: 15A24, 65P99, 93B40, 93C55, 93D99

1. Introduction

Modeling real world physical processes gives rise to mathematical systems of increasing complexity. Good mathematical models have to reproduce the original process as precisely as possible but the computing time and the storage resources needed to simulate the mathematical model are limited. As a consequence, there must be a tradeoff between accuracy and computational

[☆]This work was supported by EPSRC grant EP/E050441/1.
Email address: Younes.Chahlaoui@manchester.ac.uk (Y.Chahlaoui)
URL: http://www.maths.manchester.ac.uk/∼chahlaoui/ (Y.Chahlaoui)

constraints. One often has to deal with systems that have an unacceptably high level of complexity. It is then desirable to approximate such systems by systems of lower complexity. This is the model reduction problem.

Balanced truncation is one of the best known methods for model reduction of linear systems [1–4]. It is characterized by the principle of projection of dynamics. Balanced truncation is widely used in practice for four main reasons. First, it automatically preserves stability if the original system is stable. Second, for a reasonably small system order, say a few hundred, it gives a satisfactory approximation in the majority of cases without having to solve a complicated minimization problem or having to choose a set of essential system parameters first. Third, this approximation can be obtained at relatively reasonable computational cost. Fourth, an a priori upper bound for the error between the original plant and the reduced-order model exists for the $\mathcal{H}_\infty$-norm, the preferred measure of approximation accuracy in engineering. Recently, an $\mathcal{H}_2$-norm error formula was presented by Antoulas [5, p. 218]. Here we will derive a discrete version of this formula. This version has interesting properties that we will use later to deduce some a posteriori error bounds. Here the a posteriori distinction is made because our bounds require computation of the projection matrices. We will show how to adapt all these results to the special case of the square systems. After that we will generalize our results to general case of projection of dynamics methods. These error bounds are computable upper bounds for the $\mathcal{H}_2$-norm of the error system defined as the system whose transfer function corresponds to the difference between the transfer function of the original system and the transfer function of the reduced system. The main advantage of our results is that we use the information already available in the model reduction algorithm in order to compute the $\mathcal{H}_2$-norm instead of computing a new Gramian of the corresponding error system. There is always a computational restriction on solving high-dimensional Lyapunov equations for Gramians [5].

The a posteriori bounds give insight into the quality of the reduced system and can be used to solve many problems associated with the order reduction task, such as the choice of the best reduced order for a given tolerance. The purpose of the model often determines the "acceptable" reduced order in an implicit way, and no explicit criterion can be formulated without an a priori prohibitively expensive analysis and ranking of the dynamics involved. Our results could be implemented into the model reduction algorithm in order to check if the chosen reduced order is the best choice or needs to be modified before stopping the reduction algorithm. Another possible benefit from our results is related to the approximate balanced truncation method [5]. It is a hybrid method obtained from balanced truncation, where we approximate rather than accurately compute the solutions of the Stein equations and use these approximations to build new projection matrices. Our results give a hint on how to choose these projections in order to achieve a better $\mathcal{H}_2$ error norm or any other related problem to the order reduction.

We consider discrete-time systems

$$\mathcal{S} \quad \left\{ \begin{array}{rcl} x_{k+1} & = & Ax_k + Bu_k \\ y_k & = & Cx_k \end{array} \right. \tag{1}$$

with input $u_k \in \mathbb{R}^m$, state $x_k \in \mathbb{R}^N$ and output $y_k \in \mathbb{R}^p$, and $m, p \ll N$. We assume that the matrices A, B, and C are of appropriate dimensions. We will assume also the system $\mathcal{S}$ to be stable (i.e., all eigenvalues of the matrix A are strictly inside the unit circle). The transfer function corresponding to the system $\mathcal{S}$ is

$$H(z) = C(zI - A)^{-1}B.$$

The controllability and observability Gramians related to $\mathcal{S}$ are defined by

$$\mathcal{G}_c = \sum_{k=0}^{\infty} (A^k B)(A^k B)^T, \quad \mathcal{G}_o = \sum_{k=0}^{\infty} (CA^k)^T (CA^k),$$

and they are solutions of the Stein equations

$$A\mathcal{G}_c A^T - \mathcal{G}_c + BB^T = 0, \quad A^T \mathcal{G}_o A - \mathcal{G}_o + C^T C = 0. \tag{2}$$

This paper is organized as follow. In Section 2, we review the balanced truncation method. Section 3 is dedicated to the presentation of the new error formulas and some new a posteriori bounds of the $\mathcal{H}_2$ norm of the error system corresponding to the balanced truncation method. We also discuss some features of these formulas and bounds. We end this section by specializing the bounds to the square case. In Section 4, we generalize the error formula and bound to any projection of dynamics method. We also show how to use a low-rank approximations of the Gramians when they are not available for large scale systems. In Section 5, we present some numerical examples to show the relevance of our results. We finish with some further discussion and concluding remarks in Section 6.

2. Balanced truncation

The method of balanced truncation is well established for model reduction of linear systems. It is a special case of the projection of dynamics methods (also known as transform and truncate methods). The main idea is to rewrite the system $\mathcal{S}$, which we suppose stable, controllable and observable[1] [2, 6], using a similarity transformation T called the balancing transformation. The balanced system has some desirable sensitivity properties with respect to poles, zeros, truncation errors in digital filter implementations, and so on [2, 6]. It is therefore recommended whenever the choice of a realization (A, B, C) is not

[1] This means essentially that the Gramians are full rank.

specified by the user. The transformation T can be obtained from the Cholesky factorizations

$$\mathcal{G}_c = S^T S, \quad \mathcal{G}_o = R^T R,$$

as follows:

$$T^{-1} = \Sigma^{-1/2} V^T R, \quad T = S^T U \Sigma^{-1/2},$$

where $SR^T = U\Sigma V^T$ is the SVD of SR^T. In this coordinate system one has [7]

$$T\mathcal{G}_c T^T = T^{-T}\mathcal{G}_o T^{-1} = \Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_N),$$

where the σ_i are the Hankel singular values of $\mathcal{S}$ [6].

The balancing transformation T ensures that each state is as controllable as it is observable in the new coordinate system. After balancing the system, a reduced model is obtained by truncating the new state $x = (x_1, \ldots, x_N)^T$ to $\hat{x} = (x_1, \ldots, x_n)^T$, where $n \ll N$. The truncated states are the least controllable and observable states, corresponding to the smallest Hankel singular values and having little effect on the input/output behavior. This truncation is equivalent to projecting the system with a rank n projection $\left[\begin{array}{c|c} I_n & 0 \end{array}\right] \in \mathbb{R}^{n\times N}$. The so-called truncation matrices Π_r and Π_l are

$$\Pi_l = R^T V_1 \Sigma_1^{-1/2}, \quad \Pi_r = S^T U_1 \Sigma_1^{-1/2}, \tag{3}$$

where $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$, and U_1 and V_1 are the n first columns of U and V respectively. We can easily see that $\Pi_l^T \mathcal{G}_c \mathcal{G}_o \Pi_r = \Sigma_1^2$.

An a priori error bound in the induced 2-norm can be given for the error between the original and the reduced system [6]

$$\sigma_{n+1} \leq \|\mathcal{S} - \hat{\mathcal{S}}\|_{\mathcal{H}_\infty} \leq 2(\sigma_{n+1} + \cdots + \sigma_N). \tag{4}$$

This result says that the $\mathcal{H}_\infty$-norm of the error system is bounded above by twice the sum of the neglected Hankel singular values.

More recently, a new result was derived by Antoulas [5, p. 218] for the $\mathcal{H}_2$ norm. It is a computable $\mathcal{H}_2$ norm of the error system which yields also a computable upper bound for this norm. A convenient expression for the $\mathcal{H}_2$ norm is

$$\|\mathcal{S}\|_{\mathcal{H}_2}^2 = \mathrm{trace}\big(B^T \mathcal{G}_o B\big) = \mathrm{trace}\big(C\mathcal{G}_c C^T\big). \tag{5}$$

A result that follows immediately is given by the following proposition.

Proposition 1. *Let (A, B, C) be a balanced realization of the system $\mathcal{S}$, and σ_1 its first Hankel singular value. We have*

$$\sigma_1 \max(\alpha\|C\|_2^2, \beta\|B\|_2^2) \leq \|\mathcal{S}\|_{\mathcal{H}_2}^2 \leq \sigma_1 \min(p\|C\|_2^2, m\|B\|_2^2)$$

where $\alpha = \|C_{:1}\|_2^2/\|C\|_2^2$, $\beta = \|B_{1:}\|_2^2/\|B\|_2^2$, $C_{:1}$ the first column of C and $B_{1:}$ the first row of B and $\|.\|_2$ is the spectral norm ($\|A\|_2 = \sqrt{\lambda_{max}(A^*A)}$).

PROOF. Let us prove first that

$$\sigma_1 \alpha \|C\|_2^2 \leq \|\mathcal{S}\|_{\mathcal{H}_2}^2 \leq \sigma_1 p \|C\|_2^2.$$

Using the formula (5) and the fact that $C^T C$ is a positive matrix, we have

$$\|\mathcal{S}\|_{\mathcal{H}_2}^2 = \mathrm{trace}\big(C\mathcal{G}_c C^T\big) = \mathrm{trace}\big(C^T C \mathcal{G}_c\big) \leq \|\mathcal{G}_c\|_2 \mathrm{trace}\big(C^T C\big) = \|\mathcal{G}_c\|_2 \|C\|_F^2.$$

As the system $\mathcal{S}$ is balanced we have $\|\mathcal{G}_c\|_2 = \sigma_1$. Moreover we have [8] $\|C\|_F \leq \sqrt{p}\|C\|_2$. Then we deduce the upper bound

$$\|\mathcal{S}\|_{\mathcal{H}_2}^2 \leq \sigma_1 p \|C\|_2^2.$$

For the lower bound, it is sufficient to remark that $\mathcal{G}_c = \Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$, then

$$\|\mathcal{S}\|_{\mathcal{H}_2}^2 = \mathrm{trace}\big(C\mathcal{G}_c C^T\big) = \sum_{i=1}^{N} \sigma_i \|C_{:i}\|_F^2 \geq \sigma_1 \|C_{:1}\|_F^2$$

where $C_{:i}$ is the ith column of C. It follows that

$$\|\mathcal{S}\|_{\mathcal{H}_2}^2 \geq \sigma_1 \|C_{:1}\|_2^2 = \sigma_1 \alpha \|C\|_2^2.$$

Similarly we show that

$$\sigma_1 \beta \|B\|_2^2 \leq \|\mathcal{S}\|_{\mathcal{H}_2}^2 \leq \sigma_1 m \|B\|_2^2,$$

and the proposition follows easily.

One should remark here that for SISO systems (i.e. $p = m = 1$) the previous proposition will be simplified as $\|B\|_F = \|B\|_2 = \|C\|_2 = \|C\|_F$.

If the original system is of order[2] N and the reduced order is n, the order of the error system will be $N + n$. To compute the $\mathcal{H}_2$ norm of the error system we have to solve another Stein equation for a new Gramian of this error system, and so the cost will be of the order of $(N + n)^3$ added to the cost of the model order reduction method. With Antoulas's formula, one needs only the Gramian of the original system. This Gramian is supposed to be available already from the balanced truncation method. So the cost will be only the cost of the double product of the Gramian by the input matrix (or equivalently the output matrix) and its transpose, and the computation of the trace of that product.

In the following section we give a discrete-time version of this formula, and show how to adapt it to some special cases. The discrete-time version presents some interesting features that we will discuss later.

[2] Also called the McMillan degree.

3. $\mathcal{H}_2$ norm of the error system for balanced truncation

In this section we derive a computable a posteriori upper bound for the $\mathcal{H}_2$ norm of the error system for balanced truncation. For simplicity, let us assume henceforth that the system $\mathcal{S}$ is already in balanced form, and partition the matrices A, B and C as follows:

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad C = \begin{bmatrix} C_1 & C_2 \end{bmatrix},$$

where $\hat{A} \doteq A_{11} \in \mathbb{C}^{n\times n}$, $\hat{B} \doteq B_1 \in \mathbb{C}^{n\times m}$ and $\hat{C} \doteq C_1 \in \mathbb{C}^{p\times n}$. We will use the notation $A_{:2} = \begin{bmatrix} A_{12} \\ A_{22} \end{bmatrix}$. Since the system $\mathcal{S}$ is balanced its controllability and observability Gramians are diagonal and equal:

$$\mathcal{G}_c = \mathcal{G}_o = \mathcal{G} = \begin{bmatrix} \mathcal{G}_1 & 0 \\ 0 & \mathcal{G}_2 \end{bmatrix}, \quad \text{where} \quad \mathcal{G}_1 \in \mathbb{R}^{n\times n}.$$

We have $\mathcal{G}_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ and $\mathcal{G}_2 = \operatorname{diag}(\sigma_{n+1}, \ldots, \sigma_N)$, where σ_i are the Hankel singular values. The unified Gramian $\mathcal{G}$ then solves the Stein equations

$$A\mathcal{G}A^T - \mathcal{G} + BB^T = 0, \quad A^T\mathcal{G}A - \mathcal{G} + C^TC = 0. \tag{6}$$

To obtain the result, we consider the error system $\mathcal{S}_e$, defined as the system which has the transfer function $H_e(z) := H(z) - \hat{H}(z) = C(zI - A)^{-1}B - C_1(zI - A_{11})^{-1}B_1$, where $H(z)$ is the transfer function of $\mathcal{S}$ and $\hat{H}(z)$ is the transfer function of $\hat{\mathcal{S}}$. A realization of the system $\mathcal{S}_e$ is given by

$$\left\{ \begin{bmatrix} A & 0 \\ 0 & A_{11} \end{bmatrix}, \begin{bmatrix} B \\ -B_1 \end{bmatrix}, \begin{bmatrix} C & C_1 \end{bmatrix} \right\}. \tag{7}$$

The bound on the approximation error $\|\mathcal{S} - \hat{\mathcal{S}}\|_{\mathcal{H}_2} = \|\mathcal{S}_e\|_{\mathcal{H}_2}$ is obtained directly by bounding the $\mathcal{H}_2$ norm of $\mathcal{S}_e$. Let us first note that the controllability Gramian $\mathcal{G}_{c_e}$ and the observability Gramian $\mathcal{G}_{o_e}$ of $\mathcal{S}_e$ are given by

$$\mathcal{G}_{c_e} = \begin{bmatrix} \mathcal{G} & -Y \\ -Y^T & \hat{\mathcal{G}}_c \end{bmatrix}, \qquad \mathcal{G}_{o_e} = \begin{bmatrix} \mathcal{G} & Z \\ Z^T & \hat{\mathcal{G}}_o \end{bmatrix},$$

where $\hat{\mathcal{G}}_c$ and $\hat{\mathcal{G}}_o$ are the controllability and observability Gramians of the reduced model $\hat{\mathcal{S}}$, respectively, which solve

$$A_{11}\hat{\mathcal{G}}_cA_{11}^T - \hat{\mathcal{G}}_c + B_1B_1^T = 0, \qquad A_{11}^T\hat{\mathcal{G}}_oA_{11} - \hat{\mathcal{G}}_o + C_1^TC_1 = 0, \tag{8}$$

and where $Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$ and Y are solutions of

$$AYA_{11}^T - Y + BB_1^T = 0, \qquad A^TZA_{11} - Z + C^TC_1 = 0. \tag{9}$$

The $\mathcal{H}_2$ norm of the error system is given by

$$\begin{aligned}\|\mathcal{S}_e\|_{\mathcal{H}_2}^2 &\doteq \operatorname{trace}\left(\begin{bmatrix} B^T & -B_1^T \end{bmatrix}\begin{bmatrix} \mathcal{G} & Z \\ Z^T & \hat{\mathcal{G}}_o \end{bmatrix}\begin{bmatrix} B \\ -B_1 \end{bmatrix}\right)\\ &= \operatorname{trace}\Big(B^T\mathcal{G}B - 2B^TZB_1 + B_1^T\hat{\mathcal{G}}_oB_1\Big)\\ &= \operatorname{trace}\Big(B^T\mathcal{G}B - 2B_1^TZ_1B_1 - 2B_2^TZ_2B_1 + B_1^T\hat{\mathcal{G}}_oB_1\Big). \quad (10)\end{aligned}$$

Now, from (6), we obtain

$$A_{11}\mathcal{G}_1A_{21}^T + A_{12}\mathcal{G}_2A_{22}^T + B_1B_2^T = 0,$$

and consequently

$$\operatorname{trace}\big(-2B_2^TZ_2B_1\big) = \operatorname{trace}\big(-2B_1B_2^TZ_2\big) = \operatorname{trace}\big(2A_{11}\mathcal{G}_1A_{21}^TZ_2 + 2A_{12}\mathcal{G}_2A_{22}^TZ_2\big).$$

Substituting in (10) yields

$$\|\mathcal{S}_e\|_{\mathcal{H}_2}^2 = \operatorname{trace}\Big(B^T\mathcal{G}B - 2B_1^TZ_1B_1 + 2A_{11}\mathcal{G}_1A_{21}^TZ_2 + 2A_{12}\mathcal{G}_2A_{22}^TZ_2 + B_1^T\hat{\mathcal{G}}_oB_1\Big).$$

From (9), we have

$$A_{11}^TZ_1A_{11} + A_{21}^TZ_2A_{11} - Z_1 + C_1^TC_1 = 0,$$

and consequently

$$\begin{aligned}\operatorname{trace}\big(2A_{11}\mathcal{G}_1A_{21}^TZ_2\big) &= \operatorname{trace}\big(2\mathcal{G}_1A_{21}^TZ_2A_{11}\big)\\ &= \operatorname{trace}\big(-2\mathcal{G}_1A_{11}^TZ_1A_{11} + 2\mathcal{G}_1Z_1 - 2\mathcal{G}_1C_1^TC_1\big).\end{aligned}$$

Combining this with the definition of the $\mathcal{H}_2$ norms of $\mathcal{S}$ and $\hat{\mathcal{S}}$,

$$\|\mathcal{S}\|_{\mathcal{H}_2}^2 = \operatorname{trace}\big(B^T\mathcal{G}B\big) = \operatorname{trace}\big(C\mathcal{G}C^T\big),$$

$$\|\hat{\mathcal{S}}\|_{\mathcal{H}_2}^2 = \operatorname{trace}\Big(B_1^T\hat{\mathcal{G}}_oB_1\Big) = \operatorname{trace}\Big(C_1\hat{\mathcal{G}}_cC_1^T\Big),$$

gives

$$\begin{aligned}\|\mathcal{S}_e\|_{\mathcal{H}_2}^2 &= \operatorname{trace}\Big(2A_{12}\mathcal{G}_2A_{22}^TZ_2 + C_2\mathcal{G}_2C_2^T - C_1\mathcal{G}_1C_1^T + C_1\hat{\mathcal{G}}_cC_1^T\Big)\\ &\quad + \operatorname{trace}\big(-2B_1B_1^TZ_1 - 2A_{11}\mathcal{G}_1A_{11}^TZ_1 + 2\mathcal{G}_1Z_1\big).\end{aligned}$$

The $(1,1)$ block of (6) gives

$$A_{11}\mathcal{G}_1A_{11}^T + A_{12}\mathcal{G}_2A_{12}^T - \mathcal{G}_1 + B_1B_1^T = 0,$$

from which it follows that

$$\operatorname{trace}\big(-2B_1B_1^TZ_1 - 2A_{11}\mathcal{G}_1A_{11}^TZ_1 + 2\mathcal{G}_1Z_1\big) = \operatorname{trace}\big(2A_{12}\mathcal{G}_2A_{12}^TZ_1\big).$$

Finally, we obtain

$$\begin{aligned}\|\mathcal{S}_e\|_{\mathcal{H}_2}^2 &= \operatorname{trace}\left(C_2\mathcal{G}_2C_2^T + C_1(\hat{\mathcal{G}}_c - \mathcal{G}_1)C_1^T + 2A_{12}\mathcal{G}_2\begin{bmatrix} A_{12}^T & A_{22}^T \end{bmatrix}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}\right)\\ &= \operatorname{trace}\big(C_2\mathcal{G}_2C_2^T\big) + \operatorname{trace}\Big(C_1(\hat{\mathcal{G}}_c - \mathcal{G}_1)C_1^T\Big) + 2\operatorname{trace}\big(A_{12}\mathcal{G}_2A_{:2}^TZ\big).\end{aligned}$$

Theorem 2. *Let* $\mathcal{S} = \left\{ \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \begin{bmatrix} C_1 & C_2 \end{bmatrix} \right\}$ *be a balanced system and* $\hat{\mathcal{S}} = \{A_{11}, B_1, C_1\}$ *be the n-truncated model. The* $\mathcal{H}_2$ *norm of the error system is given by both*

$$\|\mathcal{S}_e\|_{\mathcal{H}_2}^2 = \mathrm{trace}\big(C_2\mathcal{G}_2C_2^T\big) + \mathrm{trace}\Big(C_1(\hat{\mathcal{G}}_c - \mathcal{G}_1)C_1^T\Big) + 2\mathrm{trace}\big(A_{12}\mathcal{G}_2A_{:2}^TZ\big) \quad (11)$$

and

$$\|\mathcal{S}_e\|_{\mathcal{H}_2}^2 = \mathrm{trace}\big(B_2^T\mathcal{G}_2B_2\big) + \mathrm{trace}\Big(B_1^T(\hat{\mathcal{G}}_o - \mathcal{G}_1)B_1\Big) + 2\mathrm{trace}\big(A_{12}\mathcal{G}_2A_{:2}^TY\big) \quad (12)$$

where $\mathcal{G}_2$ *is the* $(N-n)\times(N-n)$ *trailing principal submatrix of the unified Gramian of* $\mathcal{S}$, $\hat{\mathcal{G}}_c$ *and* $\hat{\mathcal{G}}_o$ *are respectively the controllability and the observability Gramians of* $\hat{\mathcal{S}}$, *and* Z *and* Y *are the solutions of the Stein equations*

$$A^TZA_{11} - Z + C^TC_1 = 0, \quad AYA_{11}^T - Y + BB_1^T = 0.$$

Remark 1. The second formula is obtained if we use the C matrices instead of the B matrices in the definition of the $\mathcal{H}_2$ norm of the error system (10).

From the Cauchy–Schwarz inequality we obtain

$$\big|\mathrm{trace}\big(C_2\mathcal{G}_2C_2^T\big)\big| \le \sigma_{n+1}p\|C_2\|_2^2, \quad \text{where} \quad \sigma_{n+1} = \|\mathcal{G}_2\|_2,$$

$$\Big|\mathrm{trace}\Big(C_1(\hat{\mathcal{G}}_c - \mathcal{G}_1)C_1^T\Big)\Big| \le \|\hat{\mathcal{G}}_c - \mathcal{G}_1\|_2 p\|C_1\|_2^2,$$

$$\big|\mathrm{trace}\big(2A_{12}\mathcal{G}_2A_{:2}^TZ\big)\big| \le 2\sigma_{n+1}\|A_{12}\|_2\|A_{:2}\|_2\|Z\|_2.$$

As Z is the solution of the Stein equation (9), it has the form

$$Z = \sum_{i=0}^{\infty}(A^T)^iC^TC_1(A_{11})^i,$$

and so

$$\|Z\|_2 \le \|C\|_2^2\sum_{i=0}^{\infty}\|A^i\|_2\|(A_{11})^i\|_2.$$

Moreover, the difference $E := \hat{\mathcal{G}}_c - \mathcal{G}_1$ satisfies the Stein equation

$$A_{11}^TEA_{11} - E + A_{21}^T\mathcal{G}_2A_{21} = 0, \quad (13)$$

which yields the formula

$$E = \hat{\mathcal{G}}_c - \mathcal{G}_1 = \sum_{i=0}^{\infty}(A_{11}^T)^iA_{21}^T\mathcal{G}_2A_{21}(A_{11})^i.$$

Finally, we have

$$\|\hat{\mathcal{G}}_c - \mathcal{G}_1\|_2 \le \sigma_{n+1}\sum_{i=0}^{\infty}\|(A_{11})^i\|_2^2\|A_{21}\|_2^2.$$

This analysis yields the following result.

Theorem 3. *Let* $\mathcal{S} = \left\{ \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \begin{bmatrix} C_1 & C_2 \end{bmatrix} \right\}$ *be a balanced system and* $\hat{\mathcal{S}} = \{A_{11}, B_1, C_1\}$ *be the n-truncated model. The* $\mathcal{H}_2$ *norm of the error system satisfies the a posteriori bound*

$$\alpha \|C\|_2^2 \sigma_{n+1} \leq \|\mathcal{S}_e\|_{\mathcal{H}_2}^2 := \|\mathcal{S} - \mathcal{S}_n\|_{\mathcal{H}_2}^2 \leq cp\sigma_{n+1}\|C\|_2^2$$

where

$$c = 1 + 3\|A\|_2^2 \sum_{i=0}^{\infty} \|A^i\|_2 \|(A_{11})^i\|_2, \quad \alpha = \|C_{:1}\|_2^2 / \|C\|_2^2,$$

and $C_{:1}$ *is the first column of* C

Proof. Here the left inequality follows from Proposition 1 and the lemma in [6, p. 185] which gives bounds on the Hankel singular values of the error system (i.e. $\sigma_i(\mathcal{S}_e)$) in function of those of the original system $\sigma_i(\mathcal{S})$. From this lemma we have

$$\sigma_1(\mathcal{S}_e) \geq \sigma_{n+1}(\mathcal{S}).$$

Another bound could be obtained as follows. Reconsider the Stein equations (9) and (13)

$$A^T Z A_{11} - Z + C^T C_1 = 0, \quad A_{11}^T E A_{11} - E + A_{21}^T \mathcal{G}_2 A_{21} = 0,$$

and let $A = UDU^{-1}$, and $A_{11} = U_1 D_1 U_1^{-1}$ be the eigenvalue decompositions of A and A_{11}. The Stein equations can be rewritten as

$$\begin{aligned} DU^{-1}ZU_1D_1 - U^{-1}ZU_1 + U^{-1}C^TC_1U_1 = 0, \\ D_1U_1^{-1}EU_1D_1 - U_1^{-1}EU_1 + U_1^{-1}A_{21}^T\mathcal{G}_2A_{21}U_1 = 0. \end{aligned}$$

From this, it can be easily seen that

$$\|Z\|_2 \leq \frac{\|C\|_2^2 \kappa_2(U)\kappa_2(U_1)}{1 - \rho(A)\rho(A_{11})}, \quad \|E\|_2 \leq \frac{\sigma_{n+1}\|A_{21}\|_2^2 \kappa_2^2(U_1)}{1 - \rho(A_{11})^2}, \tag{14}$$

where $\rho(\cdot)$ denotes the spectral radius and $\kappa_2(M) = \|M^{-1}\|_2\|M\|_2$ is the condition number. We have

$$\rho(A) = \max_i |d_{ii}|, \quad \rho(A_{11}) = \max_i |\hat{d}_{ii}|,$$

where $D = (d_{ij})_{i,j=1}^N$ and $D_1 = (\hat{d}_{ij})_{i,j=1}^n$.

Theorem 4. *Let* $\mathcal{S} = \left\{ \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \begin{bmatrix} C_1 & C_2 \end{bmatrix} \right\}$ *be a balanced system and* $\hat{\mathcal{S}} = \{A_{11}, B_1, C_1\}$ *be the n-truncated model. The* $\mathcal{H}_2$ *norm of the error system satisfies the a posteriori bound*

$$\alpha \|C\|_2^2 \sigma_{n+1} \leq \|\mathcal{S}_e\|_{\mathcal{H}_2}^2 := \|\mathcal{S} - \mathcal{S}_n\|_{\mathcal{H}_2}^2 \leq c_1 p\sigma_{n+1}\|C\|_2^2,$$

where

$$c_1 = 1 + \frac{\|A\|_2^2 \kappa_2(U)\kappa_2(U_1)}{1 - \rho(A_{11})^2} + \frac{2\|A\|_2^2 \kappa_2^2(U_1)}{1 - \rho(A)\rho(A_{11})}.$$

PROOF. Recall from (11) that

$$\begin{aligned}\|\mathcal{S}_e\|_{\mathcal{H}_2}^2 &= \text{trace}\big(C_2\mathcal{G}_2C_2^T\big) + \text{trace}\Big(C_1(\hat{\mathcal{G}}_c - \mathcal{G}_1)C_1^T\Big) + 2\text{trace}\big(A_{12}\mathcal{G}_2A_{:2}^TZ\big)\\ &\leq p\|C_2\|_2^2\|\mathcal{G}_2\|_2 + p\|C_1\|_2^2\|E\|_2 + 2\|A_{12}\|_2\|\mathcal{G}_2\|_2\|A_{:2}\|_2\|Z\|_2.\end{aligned}$$

Using the bounds (14), we have

$$\begin{aligned}\|\mathcal{S}_e\|_{\mathcal{H}_2}^2 &\leq p\|C_2\|_2^2\sigma_{n+1} + p\|C_1\|_2^2\frac{\sigma_{n+1}\|A_{21}\|_2^2\kappa_2(U)\kappa_2(U_1)}{1-\rho(A_{11})^2}\\ &\quad +2\|A_{12}\|_2\|\mathcal{G}_2\|_2\|A_{:2}\|_2\frac{\|C\|_2^2\kappa_2^2(U_1)}{1-\rho(A)\rho(A_{11})}\\ &\leq p\|C\|_2^2\sigma_{n+1} + p\|C\|_2^2\frac{\sigma_{n+1}\|A\|_2^2\kappa_2(U)\kappa_2(U_1)}{1-\rho(A_{11})^2}\\ &\quad +2\|A\|_2^2\sigma_{n+1}\frac{\|C\|_2^2\kappa_2^2(U_1)}{1-\rho(A)\rho(A_{11})},\end{aligned}$$

which gives the upper bound. The lower bound follows from Proposition 1.

3.1. Discussion

First, notice that in Theorems 3 and 4, the term $\|C\|_2^2$ could be replaced by $\|B\|_2^2$ as a result of Theorem 2 and the definition of the $\mathcal{H}_2$-norm. The discussion will focus then on the results with C.

In Theorem 2, the first term $\text{trace}\big(C_2\mathcal{G}_2C_2^T\big)$ is the $\mathcal{H}_2$-norm of the neglected subsystem of the original system; the second term $\text{trace}\Big(C_1(\hat{\mathcal{G}}_c - \mathcal{G}_1)C_1^T\Big)$ 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 $\text{trace}\big(A_{12}\mathcal{G}_2A_{:2}^TZ\big)$ is the inner product of the non-dominant block of the Gramian with the Z (non square matrix) weighted by non-dominant submatrices of A. $\mathcal{G}_2$ is diagonal and its spectral norm is supposed to be negligible compared to the spectral norm of $\mathcal{G}_1$. Then as the first and last terms are proportional to $\mathcal{G}_2$, they will be very small; the mid-term has the major contribution to the value of the norm. As $E = \hat{\mathcal{G}}_c - \mathcal{G}_1$ is the solution of the Stein equation

$$A_{11}^TEA_{11} - E + A_{21}^T\mathcal{G}_2A_{21} = 0,$$

if either the non dominant Gramian $\mathcal{G}_2$ or the off-diagonal block of A are small (zero), then E will be small (zero). As a conclusion, the quality of the reduced model will be a function of the smallness of the off-diagonal blocks of A and the smallness of σ_{n+1}, the largest neglected Hankel singular value. The last dependence is known but the first one is quite unusual. It can be interpreted as follows. The reduced order model will be a good approximation of the original system if and only if firstly there is a gap between the kept Hankel singular values of the original system and the neglected ones and secondly if the truncated states have no major contribution to the dynamics of the other states.

In Theorems 3 and 4, if the matrix A is close to normal we will have

$$\|A\|_2 \approx \rho(A) \approx \rho(A_{11}) < 1, \quad \lim_{i \mapsto \infty} \|A^i\|_2 = 0, \quad \kappa_2(U) \simeq \kappa_2(U_1) \simeq 1.$$

The constant c_1 can be taken as

$$c_1 = 1 + \frac{3\|A\|_2^2}{1 - \rho(A)^2}.$$

Moreover, the two constants c and c_1 should be of the same order in this case. Note that usually the matrix A results from the finite-element method applied to a partial differential equation, which yields in general a matrix that is close to being normal or symmetric.

We end this discussion by discussing the utility of these formulas and bounds, and even more specifically the utility of the discrete case. First, a relationship between the discrete and continuous time $\mathcal{H}_2$ norms can be derived by introducing the relationship between discrete and continuous time Gramians. One obtains

$$\|\mathcal{S}_c\|_{\mathcal{H}_2}^2 = \frac{1}{\sqrt{\Delta t}} \|\mathcal{S}_d\|_{\mathcal{H}_2}^2,$$

where $\mathcal{S}_c$ is a continuous system and $\mathcal{S}_d$ its discretization corresponding to the sampling time Δt. As a result of this formula, the discrete time $\mathcal{H}_2$ norm does not converge to the continuous time $\mathcal{H}_2$ norm when the sampling time approaches zero.

One key utility of the discrete case is that the spectral radii of the matrices A and A_{11} are smaller than 1. This follows from the stability of both systems: the original and the reduced. If A is close to normal, this property will make both coefficients c and c_1 in Theorems 3 and 4 reasonably small. For c, notice that the terms $\|A^i\|_2$ and $\|A_{11}^i\|_2$ will vanish very quickly as A has its spectral radius smaller than 1 and A_{11} is a sub-matrix of A. Both coefficients c and c_1 are only functions of A and A_{11}. Moreover, if A is normal we can bound c as follows:

$$c \leq 1 + 3\|A\|_2^2 \sum_{i=0}^{\infty} \rho(A)^{2i}.$$

This leads to the conclusion that our error bounds are only functions of σ_{n+1}, the matrix A (its 2-norm and spectral radius) and the matrix C. This is simpler in comparison with the continuous case [5] where one has to consider another residual system and computes its $\mathcal{H}_\infty$-norm. Moreover, the quality of the bound will be only a function of the smallness of σ_{n+1} as the term $c\|C\|_2$ is constant and not a function of the reduced order system. So in this case the bound is an a priori bound.

Our formulas in Theorem 2 are (like Antoulas's formula) computable. We use the data already available from balanced truncation and solve a Stein equation for a thin matrix which is much less expensive than evaluating directly the $\mathcal{H}_2$-norm. The direct evaluation of the $\mathcal{H}_2$-norm, as for example by the function `normh2` of MATLAB's Control System Toolbox, means that one has to compute

the error system, find a realization of this error system, then solve a Lyapunov or a Stein equation for one Gramian in order to evaluate the $\mathcal{H}_2$-norm.

3.2. A special case: square system

For square systems ($m = p$) one can define the cross Gramian X of $\mathcal{S}$ as the solution of the Stein equation

$$AXA - X + BC = 0. \tag{15}$$

The $\mathcal{H}_2$ norm of the system $\mathcal{S}$ is given in this case by

$$\|\mathcal{S}\|^2_{\mathcal{H}_2} = \text{trace}(CXB).$$

In this case, the $\mathcal{H}_2$ norm of the error system $\mathcal{S}_e$ (7) is

$$\|\mathcal{S}_e\|^2_{\mathcal{H}_2} = \text{trace}\left(\begin{bmatrix} C & C_1 \end{bmatrix} \begin{bmatrix} X & Y \\ Z & -\hat{X} \end{bmatrix} \begin{bmatrix} B \\ -B_1 \end{bmatrix} \right), \tag{16}$$

where Y and Z are solutions of the Stein equations

$$AYA_{11} - Y + BC_1 = 0, \quad A_{11}ZA - Z - B_1C = 0, \tag{17}$$

and $\hat{X}$ is the cross Gramian of the n reduced system by balanced truncation $\hat{\mathcal{S}}$. $\hat{X}$ is also solution of a Stein equations given by

$$A_{11}\hat{X}A_{11} - \hat{X} + B_1C_1 = 0. \tag{18}$$

Theorem 5. *The $\mathcal{H}_2$ norm of the error system is given by*

$$\begin{aligned} \|\mathcal{S}_e\|^2_{\mathcal{H}_2} &= \text{trace}(C_2X_{22}B_2) + \text{trace}\Big(C_1(\hat{X} - X_{11})B_1\Big) \\ &\quad + \text{trace}\Big(A_{12} \begin{bmatrix} X_{21} & X_{22} \end{bmatrix} AY\Big) - \text{trace}\left(A_{21}ZA \begin{bmatrix} X_{12} \\ X_{22} \end{bmatrix} \right). \end{aligned}$$

Proof. To show this result we need to expand the formula (16) as

$$\begin{aligned} \|\mathcal{S}_e\|^2_{\mathcal{H}_2} &= \text{trace}(C_1X_{11}B_1 + C_2X_{21}B_1 + C_1Z_1B_1 + C_1X_{12}B_2 + C_2X_{22}B_2) \\ &\quad + \text{trace}\Big(C_1Z_2B_2 - C_1Y_1B_1 - C_2Y_2B_1 + C_1\hat{X}B_1\Big). \end{aligned} \tag{19}$$

From the (1,2) and (2,1) blocks of (15) we have respectively

$$B_2C_1Z_2 = (X_{21} - A_{21}X_{11}A_{11} - A_{22}X_{21}A_{11} - A_{21}X_{12}A_{21} - A_{22}X_{22}A_{21})\, Z_2,$$

and

$$B_1C_2Y_2 = (X_{12} - A_{11}X_{11}A_{12} - A_{12}X_{21}A_{12} - A_{11}X_{12}A_{22} - A_{12}X_{22}A_{22})\, Y_2.$$

Then from the second blocks of the equations (17) we have

$$(A_{11}X_{12}A_{22} - X_{12})\, Y_2 = -X_{12}A_{21}Y_1A_{11} - X_{12}B_2C_1,$$

and

$$(X_{21} - A_{22}X_{21}A_{11})\,Z_2 = X_{21}A_{11}Z_1A_{12} - X_{21}B_1C_2.$$

Collecting all this in the formula (19) we get

$$\begin{aligned}\|\mathcal{S}_e\|^2_{\mathcal{H}_2} = &\operatorname{trace}\Big(C_1X_{11}B_1 + C_1Z_1B_1 + C_2X_{22}B_2 - C_1Y_1B_1 + C_1\hat{X}B_1\Big)\\ &-\operatorname{trace}(A_{21}X_{11}A_{11}Z_2 - A_{21}X_{12}A_{21}Z_2 - A_{22}X_{22}A_{21}Z_2 + A_{11}X_{11}A_{12}Y_2)\\ &+\operatorname{trace}(A_{12}X_{21}A_{12}Y_2 + A_{12}X_{22}A_{22}Y_2 - X_{12}A_{21}Y_1A_{11} + X_{21}A_{11}Z_1A_{12}).\end{aligned} \tag{20}$$

From the (1,1) block of (15) we have

$$B_1C_1 = X_{11} - A_{11}X_{11}A_{11} - A_{12}X_{21}A_{11} - A_{11}X_{12}A_{21} - A_{12}X_{22}A_{21}$$

Injecting this in (20) and using the first leading blocks of (17), i.e.,

$$Z_1 - A_{11}Z_1A_{11} - A_{11}Z_2A_{21} = -B_1C_1,$$

and

$$-Y_1 + A_{11}Y_1A_{11} + A_{12}Y_2A_{11} = -B_1C_1,$$

we get finally

$$\begin{aligned}\|\mathcal{S}_e\|^2_{\mathcal{H}_2} &= \operatorname{trace}\Big(-C_1X_{11}B_1 + C_2X_{22}B_2 + C_1\hat{X}B_1 - A_{11}X_{12}A_{21}Z_1 - A_{12}X_{22}A_{21}Z_1\Big)\\ &\quad+\operatorname{trace}(A_{12}X_{21}A_{11}Y_1 + A_{12}X_{22}A_{21}Y_1 - A_{21}X_{12}A_{21}Z_2 - A_{22}X_{22}A_{21}Z_2)\\ &\quad+\operatorname{trace}(A_{12}X_{21}A_{12}Y_2 + A_{12}X_{22}A_{22}Y_2)\\ &= \operatorname{trace}\Big(C_1(\hat{X} - X_{11})B_1 + C_2X_{22}B_2 - A_{21}Z_1A_{11}X_{12} - A_{21}Z_1A_{12}X_{22}\Big)\\ &\quad-\operatorname{trace}(A_{21}Z_2A_{21}X_{12} - A_{21}Z_2A_{22}X_{22} + A_{12}X_{21}A_{11}Y_1 + A_{12}X_{22}A_{21}Y_1)\\ &\quad+\operatorname{trace}(A_{12}X_{21}A_{12}Y_2 + A_{12}X_{22}A_{22}Y_2)\\ &= \operatorname{trace}\left(C_1(\hat{X} - X_{11})B_1 + C_2X_{22}B_2 - A_{21}\begin{bmatrix} Z_1 & Z_2 \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} X_{12} \\ X_{22} \end{bmatrix}\right)\\ &\quad+\operatorname{trace}\left(A_{12}\begin{bmatrix} X_{21} & X_{22} \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}\right),\end{aligned}$$

which proves the result.

In this theorem, the first term 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 difference 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 and A).

Notice that the difference $\hat{X} - X_{11}$ satisfies the Stein equation

$$A_{11}(X_{11} - \hat{X})A_{11} - (X_{11} - \hat{X}) + A_{12}X_{21}A_{11} + A_{11}X_{12}A_{21} + A_{12}X_{22}A_{21} = 0,$$

if the cross Gramian is block diagonal, i.e., $X_{12} = 0$ and $X_{21} = 0$. The first consequence of this assumption is that $X_{11} - \hat{X}$ as solution of the Stein equation

$$A_{11}(X_{11} - \hat{X})A_{11} - (X_{11} - \hat{X}) + A_{12}X_{21}A_{11} = 0,$$

is given by the formula

$$X_{11} - \hat{X} = \sum_{i=0}^{\infty} A_{11}^i A_{12} X_{22} A_{21} A_{11}^i.$$

As for the last term it becomes $A_{12}X_{22}A_{2:}Y - A_{21}ZA_{:2}X_{22}$. We obtain the following corollary.

Corollary 6. *If the cross Gramian is block diagonal, the $\mathcal{H}_2$ norm of the error system is given by*

$$\|\mathcal{S}_e\|_{\mathcal{H}_2}^2 = \text{trace}(C_2 X_{22} B_2) + \text{trace}\Big(C_1(\hat{X} - X_{11})B_1\Big) + \text{trace}(A_{12}X_{22}A_{2:}Y - A_{21}ZA_{:2}X_{22}).$$

Using the same analysis as the previous section (for Theorems 3 and 4) we obtain the following results.

Theorem 7. *The $\mathcal{H}_2$ norm of the error system satisfies the following a posteriori bound*

$$\alpha \|C\|_2 \|B\|_2 \sigma_{n+1} \leq \|\mathcal{S}_e\|_{\mathcal{H}_2}^2 := \|\mathcal{S} - \mathcal{S}_n\|_{\mathcal{H}_2}^2 \leq pc\sigma_{n+1}\|C\|_2\|B\|_2$$

where

$$c = 1 + 3\|A\|_2^2 \sum_{i=0}^{\infty} \|A^i\|_2 \|(A_{11})^i\|_2.$$

Theorem 8. *The $\mathcal{H}_2$ norm of the error system satisfies the following a posteriori bound*

$$\alpha \|C\|_2 \|B\|_2 \sigma_{n+1} \leq \|\mathcal{S}_e\|_{\mathcal{H}_2}^2 := \|\mathcal{S} - \mathcal{S}_n\|_{\mathcal{H}_2}^2 \leq pc_1\sigma_{n+1}\|C\|_2\|B\|_2$$

where

$$c_1 = 1 + \frac{\|A\|_2^2 \kappa_2(U)\kappa_2(U_1)}{1 - \rho(A_{11})^2} + \frac{2\|A\|_2^2 \kappa_2^2(U_1)}{1 - \rho(A)\rho(A_{11})}.$$

Here also our error bounds are only functions of the matrix A (its 2-norm and spectral radius) and the matrices B and C. We will illustrate later all this discussion in the numerical examples.

4. Generalization

The previous results obtained for balanced truncation can be generalized to any other projection of dynamics method. First, let us suppose that a reduced model $\{Y^TAX, Y^TB, CX\}$ is obtained from $\mathcal{S}$ by applying the projection matrices $X, Y \in \mathbb{R}^{N\times n}$ ($Y^TX = I_n$).

In order to exploit the ideas in the previous sections, we need to find a similarity transformation in which X and Y are embedded. We consider the matrices

$$T_l = \left[\begin{array}{c|c} Y & Y_1 \end{array}\right], \quad T_r = \left[\begin{array}{c|c} X & X_1 \end{array}\right]$$

such that

1. $Y^T X_1 = 0$,
2. $Y_1^T X = 0$,
3. $Y_1^T X_1 = I_{N-n}$,
4. $\mathrm{rank}(T_l) = \mathrm{rank}(T_r) = N$,
5. $XY^T + X_1 Y_1^T = I_N$,
6. $Y_1^T \mathcal{G}_c Y = 0, \quad X_1^T \mathcal{G}_o X = 0.$

The five first conditions insure that T_l and T_r are both similarities transformations with $T_l^T T_r = I_N$. Both matrices X_1 and Y_1 can be constructed as the orthogonal complements to X and Y, respectively, verifying the fourth last property above.

If we apply now the similarity transformations to the system $\mathcal{S}$, we obtain an equivalent[3] system $\bar{\mathcal{S}} = \{\bar{A}, \bar{B}, \bar{C}\} = \{T_l^T A T_r, T_l^T B, C T_r\}$. The corresponding Stein equations are also transformed to

$$\bar{A}\bar{\mathcal{G}}_c\bar{A}^T - \bar{\mathcal{G}}_c + \bar{B}\bar{B}^T = 0, \quad \bar{A}^T\bar{\mathcal{G}}_o\bar{A} - \bar{\mathcal{G}}_o + \bar{C}^T\bar{C} = 0,$$

where

$$\bar{\mathcal{G}}_c = \begin{bmatrix} \bar{\mathcal{G}}_{c1} & 0 \\ 0 & \bar{\mathcal{G}}_{c2} \end{bmatrix} = T_l^T \mathcal{G}_c T_l, \quad \bar{\mathcal{G}}_o = \begin{bmatrix} \bar{\mathcal{G}}_{o1} & 0 \\ 0 & \bar{\mathcal{G}}_{o2} \end{bmatrix} = T_r^T \mathcal{G}_o T_r.$$

Note that we have also $\mathcal{G}_c = T_r \bar{\mathcal{G}}_c T_r^T$ and $\mathcal{G}_o = T_l \bar{\mathcal{G}}_o T_l^T$. Now we decompose the matrices

$$\bar{A} = T_l^T A T_r = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix}, \quad \bar{B} = T_l^T B = \begin{bmatrix} \bar{B}_1 \\ \bar{B}_2 \end{bmatrix}, \quad \bar{C} = C T_r = \begin{bmatrix} \bar{C}_1 & \bar{C}_2 \end{bmatrix},$$

where $\bar{A}_{11} = Y^T A X$, $\bar{B}_1 = Y^T B$ and $\bar{C} = CX$. We obtain the following result.

Theorem 9. *Let $\mathcal{S} = \{A, B, C\}$ be a stable system and $\bar{\mathcal{S}} = \{Y^T A X, Y^T B, CX\}$ be any reduced system where $X, Y \in \mathbb{R}^{N \times n}$ ($Y^T X = I_n$). The $\mathcal{H}_2$ norm of the error system is given by*

$$\begin{aligned} \|\mathcal{S}_e\|_{\mathcal{H}_2}^2 &= \mathrm{trace}\big(C(I_N - XY^T)\mathcal{G}_c(I_N - XY^T)^T C^T\big) + \mathrm{trace}\Big(CX(\hat{\mathcal{G}}_c - Y^T\mathcal{G}_c Y)X^T C^T\Big) \\ &\quad + 2\mathrm{trace}\big(Y^T A(I_N - XY^T)\mathcal{G}_c(I_N - XY^T)^T A^T T_l Z\big) \end{aligned}$$

where $\mathcal{G}_c$ and $\hat{\mathcal{G}}_c$ are respectively the controllability Gramians of $\mathcal{S}$ and $\bar{\mathcal{S}}$, Z is the solution of the Stein equation

$$A^T Z Y^T A X - Z + C^T C X = 0,$$

and $T_l = \left[\begin{array}{c|c} Y & Y_1 \end{array}\right]$, where Y_1 is chosen such that

$$Y_1^T X = 0, \quad \mathrm{rank}(T_l) = N, \quad Y_1^T \mathcal{G}_c Y = 0.$$

[3]Here, the equivalence is in the system sense: two systems are equivalents if their transfer functions are equal.

The previous result can be also expressed similarly as a function of the observability Gramian. The proof of this theorem is very similar to the proof of Theorem 2. The only difference is that in Theorem 2 we have a unified diagonal Gramian for the system $\mathcal{S}$, and here we have two different Gramians. But the transformed Gramians are block diagonal due to the conditions $Y_1^T \mathcal{G}_c Y = 0$, $X_1^T \mathcal{G}_o X = 0$.

In the theorem, the first and third terms are functions of $I_N - XY^T$. So their values will be reasonably small if the projector XY^T is selecting the dominant part of the Gramian $\mathcal{G}_c$. The third term is the inner product of the non-dominant block of the Gramian with the Z (non square matrix) weighted by non-dominant submatrix $Y^T A(I_N - XY^T)$. Similarly to the discussions earlier in this paper, we can conclude that the quality of the reduced model will be a function of the smallness of the trailing blocks of A and the smallness of $(I_N - XY^T)\mathcal{G}_c(I_N - XY^T)^T$, the largest neglected approximated Hankel singular value.

For large-scale systems the full Gramians of the original systems typically can not be computed exactly. Instead, low-rank approximations of these Gramians can be obtained via recursive methods, among the most popular of which are Smith methods [9, 10], the alternating direction implicit (ADI) iteration method [11], Krylov subspace ideas [5, 12–15] and recently two approximate balanced truncation algorithms [16]. For example, using the results in the latter reference which proposes two approximated balanced truncation algorithms, we have that the original Gramians $\mathcal{G}_c$ and $\mathcal{G}_o$ can be approximated as follows

$$\mathcal{G}_c = SS^T + E_c E_c^T, \quad \mathcal{G}_o = RR^T + E_o E_o^T,$$

where S and R are low-rank approximations of the Cholesky factors of the Gramians, and E_c and E_o are solutions of some Lyapunov equations [16]. Moreover we have

$$\|E_c\|_2^2 \leq \sigma_{n+1}^2 \frac{\kappa_2(A)^2}{1-\rho(A)^2}, \quad \|E_o\|_2^2 \leq \sigma_{n+1}^2 \frac{\kappa_2(A)^2}{1-\rho(A)^2}.$$

Recall that the projection matrices are

$$Y = RV\Sigma^{-\frac{1}{2}}, \quad X = SU\Sigma^{-\frac{1}{2}}, \quad S^T R = U\Sigma V^T, \quad \text{where} \quad U, \Sigma, V \in \mathbb{R}^{n\times n}.$$

In this case the reduced system is balanced and the reduced Gramians are $\hat{\mathcal{G}}_c = \hat{\mathcal{G}}_o = \Sigma$. Notice here that the original system is not assumed to be balanced. We have the following theorem.

Theorem 10. *Let $\mathcal{S} = \{A, B, C\}$ be a stable system and $\bar{\mathcal{S}} = \{Y^T AX, Y^T B, CX\}$ be an approximate balanced truncated reduced system where $X, Y \in \mathbb{R}^{N\times n}$ are obtained from low-rank approximations of the Gramians $\mathcal{G}_c = SS^T + E_c E_c^T$ by*

$$Y = RV\Sigma^{-\frac{1}{2}}, \quad X = SU\Sigma^{-\frac{1}{2}}, \quad S^T R = U\Sigma V^T, \quad \textit{where} \quad U, \Sigma, V \in \mathbb{R}^{n\times n}.$$

The $\mathcal{H}_2$ norm of the error system is given by

$$\begin{aligned} \|\mathcal{S} - \bar{\mathcal{S}}\|_{\mathcal{H}_2}^2 &= \operatorname{trace}\big(CS(I-UU^T)S^T C^T\big) + \operatorname{trace}\big(CE_c E_c^T C^T\big) \\ &\quad + 2\operatorname{trace}\big(Y^T AS(I-UU^T)S^T A^T T_l Z\big), \end{aligned}$$

where Z is the solution of the Stein equation

$$A^T Z Y^T A X - Z + C^T C X = 0,$$

and $T_l = \left[\begin{array}{c|c} Y & Y_1 \end{array}\right]$, *where* Y_1 *is chosen such that*

$$Y_1^T X = 0, \quad \operatorname{rank}(T_l) = N, \quad Y_1^T \mathcal{G}_c Y = 0.$$

PROOF. It was shown in [16] that the constructed X and Y give the equalities

$$Y^T E_c = 0, \quad X^T E_o = 0, \quad Y^T S S^T Y = X^T R R^T X = \Sigma,$$

$$S S^T = X \Sigma X^T, \quad R R^T = Y \Sigma Y^T.$$

Moreover, we have

$$(I - XY^T)S = S - XY^T S = S - SU\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}V^T R^T S = S - SUU^T.$$

Then the terms in Theorem 9 become

$$\begin{aligned} C(I - XY^T)\mathcal{G}_c(I - XY^T)^T C^T &= C(I - XY^T)(SS^T + E_c E_c^T)(I - XY^T)^T C^T \\ &= CS(I - UU^T)S^T C^T + CE_c E_c^T C^T, \\ CX(\hat{\mathcal{G}}_c - Y^T \mathcal{G}_c Y)X^T C^T &= CX(\Sigma - Y^T(SS^T + E_c E_c^T)Y)X^T C^T \\ &= 0, \end{aligned}$$

$$Y^T A(I - XY^T)\mathcal{G}_c(I - XY^T)^T A^T T_l Z = Y^T AS(I - UU^T)S^T A^T T_l Z.$$

The result follows easily.

5. Numerical examples

In this section we illustrate the relevance of our formulas and bounds. The results reported here are obtained using the Control System Toolbox (version 6.0) of MATLAB. This version uses the SLICOT libraries for the numerical engine, resulting in faster and more accurate computations, especially for the solvers for stable Stein and Lyapunov equations. We use three different dynamical systems: a building model, a CD player model, and an International Space Station model. These benchmarks are described in more details in [17–19]. These models are continuous, so we discretize each system using a sampling time equal to 1 (we use the MATLAB Control System Toolbox function `c2d`), then we balance each system using `balreal`. In Table 1 we give the order of the system (N), the number of inputs (m) and outputs (p), and the spectral radius, 2-norm and condition number of the matrix A of the discrete system.

For each example, we compute the $\mathcal{H}_2$-norm of the error system $\mathcal{S}_e$ using different formulas. We construct both the original and the reduced systems using the MATLAB Control System Toolbox function `ss`, then we construct the error system $\mathcal{S}_e$ and we compute its $\mathcal{H}_2$-norm using `normh2`. The MATLAB procedure constructs the error system $\mathcal{S}_e$ as a new object, then it extracts a

Table 1: Summary of the benchmark models.

	N	m	p	$\rho(A)$	$\|A\|_2$	$\kappa(A)$
building model	48	1	1	0.769	2.157	$1.312.10^3$
CD player model	120	2	2	0.975	0.995	$1.782.10^{23}$
ISS 1R model	270	3	3	0.996	1.451	10.433

new realization for which a new Gramian (either controllability or observability Gramians) is computed. Then it uses this Gramian to compute the $\mathcal{H}_2$-norm of the system $\mathcal{S}_e$ using the formula (5). This value is shown as a reference of the quality for our results. But one should notice that this operation is much more expensive than the other $\mathcal{H}_2$-norm computations considered here. Second, we use the formula of Theorem 2 to compute the same norm. Third, we compute both controllability and observability Gramians for the realization (7), and we use the formula (5) to compute $\|\mathcal{S}_e\|_{\mathcal{H}_2}$. Fourth, we compute the constants c and c_1 in Theorems 3 and 4, respectively, and hence evaluate the error bounds of the theorems. Table 2 gives the minimum and the maximum values of c and c_1 as the reduced order n varies. The values of c and c_1 are stagnant after few iterations at c_{st} and c_{1st}, respectively. The table gives also the value of $1 + \dfrac{3\|A\|_2^2}{1-\rho(A)^2} = c_2$. It seems that the constants c and c_1 are not dependent on the reduced order for the three examples, even for the CD player example for which the matrix A is not normal. This can be explained as follows. First, note that both formulas for c and c_1 are functions of A_{11}, the submatrix of A. As n is taken larger, this submatrix is closer to A. Second, as the system is balanced, the matrix A has been transformed in such a way that the principal square submatrices A_{11} have the most significant part of A for the system at each value of n. We conclude that a good numerical approximation of both constants is given by

$$c \approx c_1 \approx 1 + \frac{3\|A\|_2^2}{1-\rho(A)^2}.$$

Table 2:

	building model		CD player model		ISS 1R model	
	min	max	min	max	min	max
c	3.677334	55.184797	1.016176	7.755511	$0.552425.10^2$	$6.395860.10^2$
c_1	2.016561	21.680204	1.016307	7.796813	$0.526626.10^2$	$6.123425.10^2$
c_{st}	42.687381		7.755511		$6.395860.10^2$	
c_{1st}	21.680204		7.796813		$6.123425.10^2$	
c_2	35.265914		63.600353		$1.017945.10^3$	

Figures 5.1 – 5.3 show the evolution of these values as a function of the

reduced order. One should expect that as the reduced order becomes closer to the original order N, the $\mathcal{H}_2$-norm should decay. In the three examples, all exact $\mathcal{H}_2$-norm formulas start by decaying before stagnating at a certain level. And even if the reduced order is taken larger and larger, the $\mathcal{H}_2$-norms seem to be not changing. This is due to the machine tolerance implemented in different MATLAB functions used. Comparing the instant at which different formulas for the $\mathcal{H}_2$-norm are stabilizing and the Hankel singular values of each model in the Figure 5.4, we can deduce that different methods used consider that up to a certain tolerance the states corresponding to the remaining Hankel singular values are not adding anything to the system. This is relevant in the following sense. When this tolerance is reached, say at $n = n_1$, there will be no numerical difference between a reduced system of order $k \geq n_1$ and any other reduced system of order $k_1 \geq n_1$ ($k \neq k_1$), even if in theory this is not true. This is due to the sensitivity of the Stein equation [20]. The reference (the MATLAB procedure) is more accurate than the other formulas of the $\mathcal{H}_2$-norm because the computed Gramian for this procedure is more accurate and it is less sensitive to round-off errors.

The bounds in Theorems 3 and 4 seem to be following the behavior of the exact $\mathcal{H}_2$-norms and continue to decay. After a certain order, they are better than the reference (the MATLAB procedure). For the ISS example, this does not happen as quickly as for the two other examples, but we can see easily that it does happen later on as the two bounds are decaying and the reference is stagnating.

In Theorem 2, the matrix Z is a non-square matrix solution of a Stein equation. As Z is not symmetric in some of our numerical tests, the trace of the term involving Z shows an imaginary term, nevertheless neglectable. Moreover, the term $\operatorname{trace}\Big(C_1(\hat{\mathcal{G}}_c - \mathcal{G}_1)C_1^T\Big)$ even if very small could have negative sign. This is related to the still open problem of over-approximation and under-approximation of the Gramians [5, 20].

6. Concluding remarks

We have presented computable error formulas and bounds for the response approximation for the most used projection based method in model reduction of linear time-invariant dynamical systems, balanced truncation and the general case of projection of dynamics. The advantage of these results is that we are using the already given results by balanced truncation and we do not need any additional computation. This has the feature that it can be included into the order reduction loop in order to improve the quality of the reduced order model by choosing the optimal reduced order before ending the model reduction algorithm. We also presented the special case of square systems. These systems have the property that only one Gramian needs to be computed to evaluate the $\mathcal{H}_2$-norm. We generalized the result obtained for balanced truncation to any

LEGEND FOR FIGURES 5.1 – 5.3:
——— $\|\mathcal{S}_e\|_{\mathcal{H}_2} = \texttt{normh2}(\mathcal{S}_e)$, $\cdots$ $\|\mathcal{S}_e\|_{\mathcal{H}_2}$ by (11), —●— bound of Theorem 3,
– · – $\|\mathcal{S}_e\|^2_{\mathcal{H}_2} = \mathrm{trace}\big(C_e \mathcal{G}_{c_e} C_e^T\big)$ from (5), – – $\|\mathcal{S}_e\|^2_{\mathcal{H}_2} = \mathrm{trace}\big(B_e^T \mathcal{G}_{o_e} B_e\big)$ from (5).

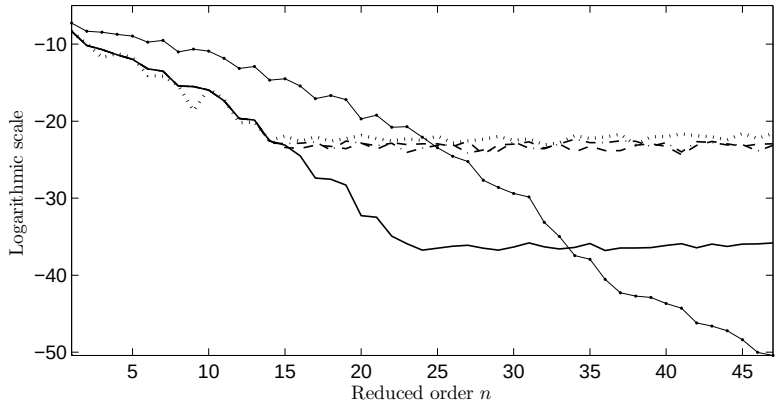

Figure 1: Building model

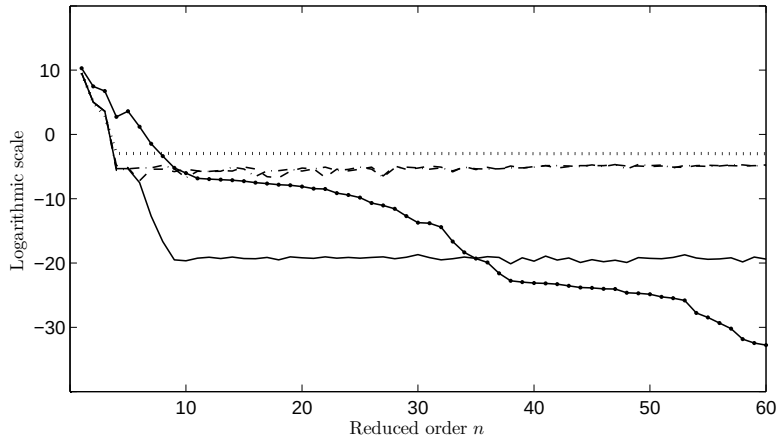

Figure 2: CD player model

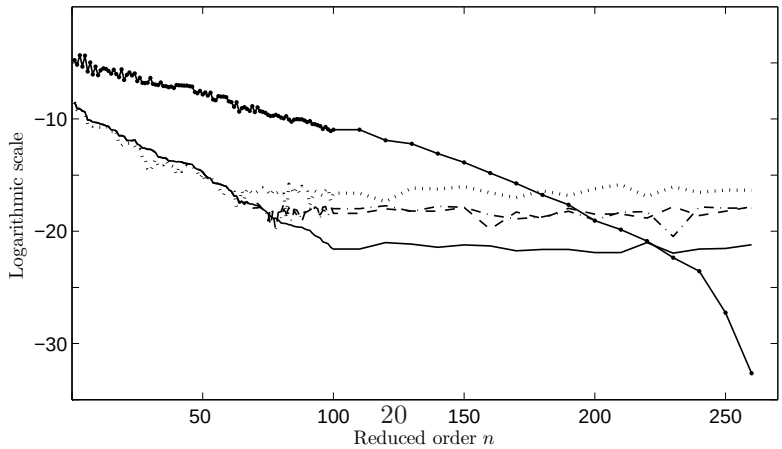

Figure 3: ISS 1R model

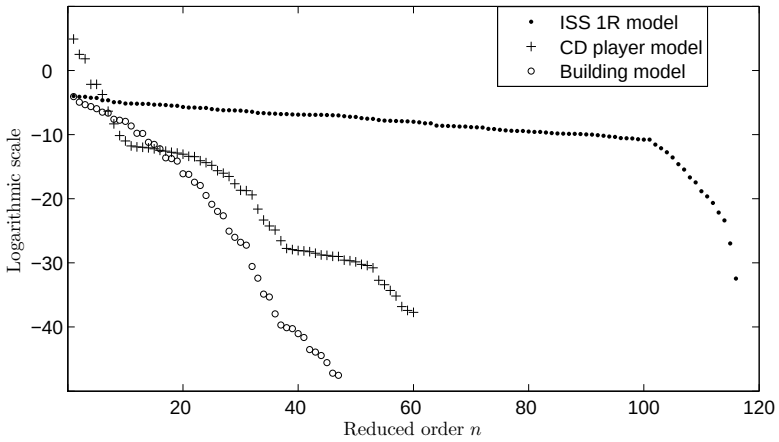

Figure 4: σ_{n+1} for the three models.

other projection of dynamics method. The bounds presented are less sensitive to the round-off errors than the exact formulas.

Despite the obviously desirable features of the easily computable bound for the $\mathcal{H}_2$-norm of the error system, many open questions remain. There are a number of refinements with respect to performance, convergence, and accuracy which require more theoretical and algorithmic analysis. There are two particularly interesting features. The first is how to modify the balanced truncation projection matrices to achieve a better $\mathcal{H}_2$-norm and still be the best for the $\mathcal{H}_\infty$-norm. The second is how to choose the projection matrices from a projection of dynamics method in order to have $\mathcal{H}_2$-norm optimality.

Acknowledgements

This work was initiated and largely influenced by an informal seminar that Prof. D.C. Sorensen gave during his sabbatical months spent at CESAME, UCLouvain in 2003. Prof. Sorensen gave a course on "Numerical linear algebra for systems and control" for the Graduate School in Systems and Control during his stay at CESAME. I hoped that at least a fraction of his insight and intuition have rubbed on me. His precision seemed nearly infinite. Different conversations with him saved me a lot of time to focus on the right directions. I am very thankful to him for that. I also gratefully acknowledge the helpful remarks and suggestions of Nick Higham, Françoise Tisseur which significantly improved the presentation of this paper.

References

[1] D. F. Enns, Model reduction with balanced realizations: An error bound and frequency weighted generalization, Proc. of the IEEE Conference on

Decision and Control (1981) 127–132.

[2] B. C. Moore, Principal component analysis in linear systems: controllability, observability, and model reduction, IEEE Trans. Automat. Control 26 (1981) 17–31.

[3] L. Pernebo, L. M. Silverman, Model reduction via balanced state space representations, IEEE Trans. Automat. Control 27 (2) (1982) 382–387.

[4] M. G. Safonov, R. Y. Chiang, A Schur method for balanced-truncation model reduction, IEEE Trans. Automat. Control 34(7) (1989) 729–733.

[5] A. C. Antoulas, Approximation of Large-Scale Dynamical Systems, SIAM, Philadelphia, USA, 2005.

[6] K. Zhou, J. C. Doyle, K. Glover, Robust and optimal control, Prentice Hall, 1995.

[7] K. Glover, All optimal Hankel norm approximations of linear multivariable systems and their $\mathcal{L}^\infty$-error bounds, Internat. J. Control 39 (1984) 1115–1193.

[8] N. J. Higham, Accuracy and stability of numerical algorithms. 2nd ed., Philadelphia, PA: SIAM., 2002.

[9] T. Penzl, Numerical solution of generalized Lyapunov equations, Advances in Comp. Math 8 (1998) 33–48.

[10] T. Penzl, A cyclic low-rank Smith method for large sparse Lyapunov equations, Siam J. Sci. Comput. 21 (4) (2000) 1404–1418.

[11] E. Wachspress, Iterative solution of the Lyapunov matrix equation, Appl. Math. Lett. 1 (1) (1988) 87–90.

[12] D. Hu, L. Reichel, Krylov-subspace methods for the Sylvester equation, Linear Algebra Appl 172 (1992) 283–313.

[13] I. Jaimoukha, E. Kasenally, Krylov subspace methods for solving large Lyapunov equations, SIAM J. Numer. Anal. 31 (1) (1994) 227–251.

[14] Y. Saad, Numerical solutions of large Lyapunov equations, M.A. Kaashoek, J.H. Van Schuppen, A.C. Ran (Eds.), Signal Processing, Scattering, Operator Theory, and Numerical Methods. Birkhauser, Basel (1990) 503–511.

[15] A. Scottedward Hodel, Numerical methods for the solution of large and very large, sparse Lyapunov equations, Ph.D. thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA (1989).

[16] Y. Chahlaoui, Two efficient SVD/Krylov algorithms for model order reduction of large scale systems, submitted to Electronic Transactions On Numerical Analysis (ETNA).
URL `http://eprints.ma.man.ac.uk/1392`

[17] Y. Chahlaoui, P. Van Dooren, A collection of benchmark examples for model reduction of linear time invariant dynamical systems, SLICOT Working Note 2002-2 (2002) Available from ftp://wgs.esat.kuleuven.ac.be/pub/WGS/REPORTS/SLWN2002–2.ps.Z.

[18] Y. Chahlaoui, P. Van Dooren, Benchmark examples for model reduction of linear time invariant dynamical systems., Benner, Peter (ed.) et al., Dimension reduction of large-scale systems. Springer. Lecture Notes in Computational Science and Engineering. 45 (2005) 379–392.

[19] S. Gugercin, A. C. Antoulas, N. Bedrossian, Approximation of the international space station 1r and 12a flex models, Proceedings of the 40th IEEE Conference on Decision and Control.

[20] D. C. Sorensen, Y. Zhou, Bounds on eigenvalue dacay rates and sensitivity of solutions to lyapunov equations, Tehnical report TR02-07, (2002) http://www.caam.rice.edu/caam/content/techrep.html.