Numerical computation and new output bounds for time-limited balanced truncation of discrete-time systems

Igor Pontes Duff[a], Patrick Kürschner[b]

[a]Max Planck Institute for Dynamics of Complex Technical Systems, Magdeburg, Germany
[b]Group Science, Engineering and Technology, KU Leuven Kulak, Kortrijk, Belgium and Department of Electrical Engineering ESAT/STADIUS, KU Leuven, Leuven, Belgium

Abstract

In this paper, balancing based model order reduction (MOR) for large-scale linear discrete-time time-invariant systems in prescribed finite time intervals is studied. The first main topic is the development of error bounds regarding the approximated output vector within the time limits. The influence of different components in the established bounds will be highlighted. After that, the second part of the article proposes strategies that enable an efficient numerical execution of time-limited balanced truncation for large-scale systems. Numerical experiments illustrate the performance of the proposed techniques.

Keywords: Model reduction, discrete-time systems, large-scale systems, Stein equations, linear systems.
2010 MSC: 15A24; 93A15; 93B99; 93C05; 93C55;

1. Introduction

In this paper, we consider multi-input multi-output (MIMO) linear time-invariant (LTI) discrete-time dynamical systems. These systems are governed by a set of difference equations of the form

$$\mathcal{S}: \begin{cases} x(k+1) &= Ax(k) + Bu(k), \text{ for } k \in \mathbb{N} = \{0, 1, 2, \dots\} \\ y(k) &= Cx(k), \quad x(0) = x_0, \end{cases} \tag{1}$$

where $x(k) \in \mathbb{R}^n$ is the state-variable, $u(k) \in \mathbb{R}^m$ is the input, $y(k) \in \mathbb{R}^p$ is the output for every discrete-time $k \in \mathbb{N}$. Here $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and the leading dimension n is the order of the system. We denote $\mathcal{S} = (A, B, C)$ for the given realization (1). Additionally, we use $\mathbb{N}^*$ to denote the set of positive integers, i.e., $\{1, 2, 3, \dots\}$. We assume that $x_0 = 0$ and the reader is referred to [1, 2, 3, 4] which treat the case of nonzero initial condition for continuous-time systems. Some results of those papers can be extended to discrete-time systems.

In this case, we can represent the output as

$$y(k) = \sum_{j=0}^{k} h(k-j)u(j) = (h * u)(k), \tag{2}$$

where h is the impulse response of the system, given by

$$h(0) = 0, \; h(k) = CA^{k-1}B, \text{ for } k = 1, 2, \dots. \tag{3}$$

We say that $\mathcal{S}$ is (asymptotically) stable if and only if A has its eigenvalues inside the unitary disc, in which case we call the matrix A stable. Otherwise, we say that A is unstable. For stable systems, the infinite reachability and observability Gramians P_∞ and Q_∞ are defined as

$$P_\infty = \sum_{k=1}^{\infty} A^{k-1}B\left(A^{k-1}B\right)^T, \tag{4a}$$

$$Q_\infty = \sum_{k=1}^{\infty} \left(CA^{k-1}\right)^T CA^{k-1}, \tag{4b}$$

and they are the unique solutions of the following Stein equations (discrete-time Lyapunov equations)

$$AP_\infty A^T - P_\infty + BB^T = 0, \tag{5a}$$

$$A^T Q_\infty A - Q_\infty + C^T C = 0. \tag{5b}$$

A LTI discrete-time system (1) is said to be minimal in infinite horizon if $P_\infty Q_\infty$ is nonsingular.

Mathematical models of systems (1) are considered to be large-scale, whenever its order is very large, perhaps $n > 10^5$ or more. This leads to difficulties for tasks involving simulation, optimization, or control of this system, motivating the use of a reduced order model (ROM) of the form

$$\hat{\mathcal{S}} : \begin{cases} \hat{x}(k+1) &= \hat{A}\hat{x}(k) + \hat{B}u(k), \text{ for } k \in \mathbb{N} \\ \hat{y}(k) &= \hat{C}\hat{x}(k), \end{cases} \tag{6}$$

where $\hat{x}(k) \in \mathbb{R}^r$, for $k \in \mathbb{N}$, $\hat{A} \in \mathbb{R}^{r\times r}$, $\hat{B} \in \mathbb{R}^{r\times m}$ and $\hat{C} \in \mathbb{R}^{p\times r}$. The goal is to construct a ROM $\hat{\mathcal{S}}$ such that $r \ll n$ and still $\hat{y} \approx y$, i.e., $\|\hat{y} - y\|$ should be small for some prescribed norm for a large class of inputs u. Projection based model reduction consists in constructing matrices $W, V \in \mathbb{R}^{n\times r}$ with $W^T V = I_r$, such that

$$\hat{A} = W^T A V, \quad \hat{B} = W^T B \quad \text{and} \quad \hat{C} = CV. \tag{7}$$

In order to measure the quality of reduced order models, system norms are defined. Given a stable system $\mathcal{S}$ as in (1) whose impulse response h is given by (3), its h_∞ and h_2 norms are defined as

$$\begin{aligned} \|\mathcal{S}\|_{h_\infty} &= \sup_{w\in[0,2\pi]} \|C(e^{i\omega}I - A)^{-1}B\|_2, \text{ and} \\ \|\mathcal{S}\|_{h_2} &= \left(\sum_{j=0}^{\infty} \|h(j)\|_F^2\right)^{1/2} = \operatorname{tr}\Big(CP_\infty C^T)\Big)^{1/2} = \operatorname{tr}\Big(B^T Q_\infty B)\Big)^{1/2}. \end{aligned} \tag{8}$$

Balanced truncation (BT) is a model order reduction technique introduced in [5], allowing to construct such a reduced order model $\hat{\mathcal{S}}$ by projection. It relies on the concept of simultaneous diagonalization of the reachability and observability Gramians. In other words, the goal is to find a state-space transformation $T \in \mathbb{R}^{n\times n}$ nonsingular, such that

$$TP_\infty T^T = T^{-T} Q_\infty T^{-1} = \Sigma_\infty = \begin{bmatrix} \Sigma_{1,\infty} & 0 \\ 0 & \Sigma_{2,\infty} \end{bmatrix},$$

where $\Sigma_{1,\infty} = \operatorname{diag}(\sigma_{1,\infty}, \ldots, \sigma_{r,\infty})$, $\Sigma_{2,\infty} = \operatorname{diag}(\sigma_{r+1,\infty}, \ldots, \sigma_{n,\infty})$, and $\sigma_{1,\infty} \geq \cdots \geq \sigma_{n,\infty} \geq 0$ are the so-called Hankel singular values. Let

$$TAT^{-1} := A_{\mathcal{B}} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad TB := B_{\mathcal{B}} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad CT^{-1} := C_{\mathcal{B}} = \begin{bmatrix} C_1 & C_2 \end{bmatrix}.$$

The equivalent realization $(A_{\mathcal{B}}, B_{\mathcal{B}}, C_{\mathcal{B}})$ is referred to as a balanced realization. Then, the projection matrices V and W are taken as the first r columns of T and T^{-T}, respectively, and the reduced order model is given by (7).

The reduced order model $\hat{\mathcal{S}}$ obtained by balancing satisfies an a priori error bound in the h_∞ norm which is given by (cf. [6, Theorem 7.10])

$$\|\mathcal{S} - \hat{\mathcal{S}}\|_{h_\infty} \leq 2\left(\sum_{k=r+1}^{n} \sigma_k\right) = 2\operatorname{tr}(\Sigma_{2,\infty}), \tag{9}$$

i.e., the h_∞ norm of the error system is bounded by twice the sum of the neglected Hankel singular values. From now on, we use the following notation $\boldsymbol{\sigma_r} := 2\operatorname{tr}(\Sigma_{2,\infty})$ for the sum of the neglected Hankel singular values. This error bound is also valid in the continuous-time context due to [7, 8].

An error bound a posteriori with respect to the h_2 norm is also available in [9]. It is expressed by

$$\|\mathcal{S} - \hat{\mathcal{S}}\|_{h_2} = \operatorname{tr}\Big(C_2\Sigma_{2,\infty}C_2^T + 2A_{12}\Sigma_{2,\infty}A_{:2}^T Z_\infty\Big) + \operatorname{tr}\Big(C_1(\hat{P}_\infty - \Sigma_{1,\infty})C_1^T\Big) \tag{10a}$$

$$= \operatorname{tr}\Big(B_2^T\Sigma_{2,\infty}B_2 + 2A_{21}^T\Sigma_{2,\infty}A_{2:}Y\Big) + \operatorname{tr}\Big(B_1^T(\hat{Q}_\infty - \Sigma_{1,\infty})B_1\Big) \tag{10b}$$

where $\hat{P}_\infty$ and $\hat{Q}_\infty$ are, respectively, the reachability and observability Gramians of the ROM, which satisfy

$$\begin{aligned} A_{11}\hat{P}_\infty A_{11}^T - \hat{P}_\infty + B_1 B_1^T &= 0, \\ A_{11}^T\hat{Q}_\infty A_{11} - \hat{Q}_\infty + C_1^T C_1 &= 0, \end{aligned}$$

the matrices $Y_\infty, Z_\infty \in \mathbb{R}^{n\times r}$ are the solutions of the Sylvester equations

$$AY_\infty A_{11}^T - Y_\infty + BB_1^T = 0, \tag{11a}$$

$$A^T Z_\infty A_{11} - Z_\infty + C^T C_1 = 0, \tag{11b}$$

and $A_{:2}^T = \begin{bmatrix} A_{12}^T & A_{22}^T \end{bmatrix}$. It is worth noticing that an H_2 error bound for continuous-time systems is also available in [6, Lemma 7.13]. Similar research for stochastic systems can be found in, e.g., [10, 11, 12].

Balanced truncation for continuous- and discrete-time LTI systems was extended by the restriction to given time intervals in [13]. In this context, one aims at a ROM that is an accurate approximation until a finite time horizon $\tau > 0$, but allows the ROM to be inaccurate outside of this time interval. The time-limited (TL) Gramians are defined as

$$P_\tau = \sum_{k=1}^{\tau} A^{k-1} B \left(A^{k-1} B\right)^T, \tag{12a}$$

$$Q_\tau = \sum_{k=1}^{\tau} \left(CA^{k-1}\right)^T CA^{k-1}, \tag{12b}$$

and satisfy the following Stein equations

$$AP_\tau A^T - P_\tau + BB^T = FF^T, \tag{13a}$$

$$A^T Q_\tau A^T - Q_\tau + C^T C = G^T G, \tag{13b}$$

where $F := A^\tau B$ and $G := CA^\tau$. Even if the pairs (A, B) and (A^T, C^T) are reachable, the TL Gramians (12) might be only positive semidefinite. This might happen whenever $\tau < n/m$ or $\tau < n/p$. In this case, one can remove the states that are unreachable and unobservable for the the given time interval, which are given by the kernels of P_τ and Q_τ. As a consequence, the resulting system is reachable and observable for the given time interval and the Gramians in (12) are positive definite matrices. Henceforth, we will assume that the TL Gramians in (12) are positive definite matrices.

The time-limited balanced truncation (TLBT) is performed by balancing P_τ and Q_τ, *i.e.*, finding the state transformation T such that $TP_\tau T^T = T^{-T} Q_\tau T^{-1} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ and neglecting the states associated to small time-limited Hankel singular values. The reader should notice that the Gramians P_τ and Q_τ also exist when the matrix A is unstable because the equations (13a) and (13b) have a unique solution provided $\forall \lambda \in \Lambda(A) \backslash \{0\}$ it holds $1/\lambda \notin \Lambda(A)$. As a consequence, TLBT is also applicable to unstable systems. On the other hand, for stable systems, TLBT is not guaranteed to preserve stability. Still, experimental evidence [14, 15] indicates that this does not deteriorate the approximation quality inside the targeted time interval which will be also confirmed by the experiments in this paper. Moreover, the upcoming error bounds will, to some extent, indicate that the occasionally generated unstable reduced order models still provide accurate output approximations. Some stability preserving variants of time-/ and frequency-limited BT have been proposed in, e.g., [16, 17, 15, 19] leading to so-called modified BT variants. However, enforcing stability via such modified TLBT variants appears to deteriorate the good approximation quality of TLBT within the time interval and, at the same time, is computational more expensive [20, 21, 14] for large-scale systems. Hence, in the study at hand, we will not consider such stability preserving variants. Additionally, readers are referred to [22, 23, 24, 25] for $\mathcal{H}_2$ time-/ and frequency-limited model reduction of continuous-time systems.

In this paper, time-limited balanced truncation for large-scale linear discrete-time time-invariant systems is studied. The main contribution is twofold. In the first part, we develop error bounds regarding the approximated output vector within the time limits. Those error bounds are an extension of those given in [9] to the time-limited case. However, they also hold in the case the original system or the reduced order model are unstable. Additionally, their asymptotic behavior with respect to the time limits is analyzed and some sufficient conditions to preserve stability are provided. The second part of the article proposes computational strategies that enable an efficient numerical execution of time-limited balanced truncation for large-scale systems, a topic which has so far not been considered in the literature. These strategies rely on solvers for the time-limited Stein equations using low-rank factors. Different solvers are proposed, and their performances are compared.

It is worth noticing that BT and TLBT can still be applied to the original system, even if the Gramians are not of full rank. Indeed, let the Gramians be given as factorizations $P_\tau = S^T S$ and $Q_\tau = R^T R$, with S and R full column rank matrices. Now assume the following partitioned SVD of SR^T as

$$SR^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix} \begin{bmatrix} V_1^T & V_2^T \end{bmatrix}.$$

Then, we obtain the reduced order model by (time-limited) balancing using Petrov-Galerkin projections as in (7) with $V = S^T U_1 \Sigma_1^{-1/2}$ and $W = R^T V_1 \Sigma_1^{-1/2}$. This approach is known as square root balancing. It has the advantage of avoiding the computation of the balancing transformation T, which can be an expensive and ill-conditioned

problem (see [6]). In Sections 4 and 5, the above procedure, combined with the Stein equations solutions' low-rank factors, is used to compute reduced order models.

The rest of the paper is organized as follows. In Section 2, the time-limited h_2 inner product and norm are defined and characterized using Gramians. Also, a first error bound is provided based on the discrete-time convolution expression. In Section 3, a specially tailored error bound for time-limited balanced truncation is developed. Additionally, a sufficient condition for stability preservation is provided, and the asymptotic behavior of the error bound is studied. In Section 4, different solvers based on low-rank factors are proposed to compute the TL Gramians approximately. Finally, Section 5 carries out some numerical experiments for large-scale systems and Section 6 concludes the paper.

2. Preliminary results

2.1. TL h_2 inner product and norm

From now on, we consider the finite horizon τ to be fixed and given. In what follows, we recall the definition of the TL h_2 norm and inner-product.

Definition 2.1. (time-limited h_2 norm and inner-product) *Let $\mathcal{S} = (A, B, C)$ and $\hat{\mathcal{S}} = (\hat{A}, \hat{B}, \hat{C})$ be two LTI discrete-time dynamical systems as in* (1)*. Then, the h_2 TL inner-product between $\mathcal{S}$ and $\hat{\mathcal{S}}$ is given by*

$$\langle \mathcal{S}, \hat{\mathcal{S}} \rangle_{h_2,\tau} = \sum_{j=0}^{\tau} \operatorname{tr}\Big(h(j)\hat{h}(j)^T\Big), \tag{14}$$

where $h(0) = 0, h(k) = CA^{k-1}B$ and $\hat{h}(0) = 0, \hat{h}(k) = \hat{C}\hat{A}^{k-1}\hat{B}$ for $k \in \mathbb{N}^*$ are the impulse response of $\mathcal{S}$ and $\hat{\mathcal{S}}$ respectively. Moreover, the h_2 TL norm of $\mathcal{S}$ is given by[1]

$$\|\mathcal{S}\|_{h_2,\tau} = \left(\sum_{j=0}^{\tau} \operatorname{tr}\Big(h(j)h(j)^T\Big)\right)^{1/2} = \left(\sum_{j=0}^{\tau} \|h(j)\|_F^2\right)^{1/2} = \langle \mathcal{S}, \mathcal{S} \rangle_{h_2,\tau}^{\frac{1}{2}}. \tag{15}$$

The reader should notice that if τ goes to infinite, equations (14) and (15) become the classical definition of the inner-product and norm for an infinite time horizon for stable systems. However, the TL norm and inner-product are also well defined for unstable systems. Additionally, they can be characterized by matrix equations, as it follows.

Proposition 2.1. (TL inner-product and norm characterization) *Let $\mathcal{S} = (A, B, C)$ and $\hat{\mathcal{S}} = (\hat{A}, \hat{B}, \hat{C})$ be two discrete-time LTI systems as in* (1)*. Then the h_2 TL inner-product can be computed as*

$$\langle \mathcal{S}, \hat{\mathcal{S}} \rangle_{h_2,\tau} = \operatorname{tr}\Big(CY\hat{C}^T\Big) = \operatorname{tr}\Big(B^T Z \hat{B}\Big), \tag{16}$$

where

$$Y = \sum_{j=1}^{\tau} A^{j-1}B\hat{B}^T(\hat{A}^T)^{j-1} \quad \text{and} \quad Z = \sum_{j=1}^{\tau} (A^T)^{j-1}C^T\hat{C}\hat{A}^{j-1}.$$

Additionally, if $\alpha\beta \neq 1$, for all $\alpha \in \Lambda(A)$ and $\beta \in \Lambda(\hat{A})$, the matrices Y and Z are the unique solution of the following Stein-like matrix equations

$$AY\hat{A}^T - Y + B\hat{B}^T - F\hat{F}^T = 0, \tag{17a}$$

$$A^T Z\hat{A} - Z + C^T\hat{C} - G^T\hat{G} = 0, \tag{17b}$$

where $F = A^\tau B, \hat{F} = \hat{A}^\tau\hat{B}$, $G = CA^\tau$ and $\hat{G} = \hat{C}\hat{A}^\tau$.

Proof. Notice $\langle \mathcal{S}, \hat{\mathcal{S}} \rangle_{h_2,\tau} = \operatorname{tr}\Big(C\left(\sum_{j=1}^{\tau} A^{j-1}B\hat{B}^T(\hat{A}^T)^{j-1}\right)\hat{C}^T\Big) = \operatorname{tr}\Big(CYC^T\Big)$. Then, as an application of the telescopic sum on $AY\hat{A}^T - Y$, one obtains that Y satisfies equation (17a). Moreover, equation (17a) has a unique solution if and only if $\alpha\beta \neq 1$, for all $\alpha \in \Lambda(A)$ and $\beta \in \Lambda(\hat{A})$ (see [26, Theorem 18.2]). The equivalent result for the matrix Z follows similarly. □

[1] Given a matrix $H \in \mathbb{R}^{p\times m}$, its Frobenius norm is defined as $\|H\|_F^2 = \operatorname{tr}\Big(HH^T\Big)$.

Proposition 2.1 states that, if the equations (17) have unique solutions, then the solutions can be used to compute the TL inner-product via formula (16). As a consequence, the TL h_2 norm of a system can be computed via

$$\|\mathcal{S}\|_{h_2,\tau}^2 = \mathrm{tr}\Big(CP_\tau C^T\Big) = \mathrm{tr}\Big(B^T Q_\tau B\Big), \tag{18}$$

where P_τ and Q_τ are the solutions of (13a) and (13b).

Assumption 2.1. *From now on, we assume that $\alpha\beta \neq 1$, for all $\alpha \in \Lambda(A)$ and $\beta \in \Lambda(\hat{A})$, so that the equations* (17) *always have unique solutions.*

2.2. First characterization of error bound

Let us assume the discrete-time system $\mathcal{S} = (A, B, C)$ is the full order model and $\hat{\mathcal{S}} = (\hat{A}, \hat{B}, \hat{C})$ is the reduced order model. The output of the original system $\mathcal{S}$ and the reduced system $\hat{\mathcal{S}}$ can be expressed as

$$y(k) = \sum_{j=0}^{k} h(k-j)u(j), \text{ and } \hat{y}(k) = \sum_{j=0}^{k} \hat{h}(k-j)u(j),$$

where $h(0) = 0$, $h(k) = CA^{k-1}B$, for $k \in \mathbb{N}^*$, is the impulse response of $\mathcal{S}$, and $\hat{h}(0) = 0$, $\hat{h}(k) = \hat{C}\hat{A}^{k-1}\hat{B}$, for $k \in \mathbb{N}^*$, is the impulse response of $\hat{\mathcal{S}}$. Hence, the error between y and $\hat{y}$ can be bounded as

$$\begin{aligned}
\|y(k) - \hat{y}(k)\|_2 &= \left\| \sum_{j=0}^{k} h(k-j)u(j) - \sum_{j=0}^{k} \hat{h}(k-j)u(j) \right\|_2 \\
&\leq \sum_{j=0}^{k} \left\| \Big(h(k-j) - \hat{h}(k-j)\Big) u(j) \right\|_2 \\
&\leq \sum_{j=0}^{k} \left\| h(k-j) - \hat{h}(k-j) \right\|_2 \|u(j)\|_2 \\
&\leq \sum_{j=0}^{k} \left\| h(k-j) - \hat{h}(k-j) \right\|_F \|u(j)\|_2, \\
&\leq \left(\sum_{j=0}^{k} \|h(k) - \hat{h}(k)\|_F^2 \right)^{\frac{1}{2}} \left(\sum_{j=0}^{k} \|u(j)\|_2^2 \right)^{\frac{1}{2}},
\end{aligned}$$

where we have applied the Cauchy-Schwarz inequality in the last step. By recalling that

$$\|\mathcal{S}\|_{h_2,\tau} = \left(\sum_{j=0}^{\tau} \|h(j)\|_F^2 \right)^{1/2}, \text{ and } \langle \mathcal{S}, \hat{\mathcal{S}} \rangle_{h_2,\tau} = \sum_{j=0}^{\tau} \mathrm{tr}\Big(h(j)\hat{h}(j)^T\Big),$$

one can easily see that

$$\max_{j=0,1,\ldots,\tau} \|y(j) - \hat{y}(j)\|_2 \leq \left\| \mathcal{S} - \hat{\mathcal{S}} \right\|_{h_2,\tau} \left(\sum_{j=0}^{\tau} \|u(j)\|_2^2 \right)^{\frac{1}{2}}.$$

Now, let us first use the inner-product expression. Hence,

$$\left\| \mathcal{S} - \hat{\mathcal{S}} \right\|_{h_2,\tau}^2 = \|\mathcal{S}\|_{h_2,\tau}^2 + \left\| \hat{\mathcal{S}} \right\|_{h_2,\tau}^2 - 2\langle \mathcal{S}, \hat{\mathcal{S}} \rangle_{h_2,\tau}.$$

Now, we recall that

$$\begin{aligned}
\|\mathcal{S}\|_{h_2,\tau}^2 &= \mathrm{tr}\Big(CP_\tau C^T\Big) = \mathrm{tr}\Big(B^T Q_\tau B\Big), \\
\left\| \hat{\mathcal{S}} \right\|_{h_2,\tau}^2 &= \mathrm{tr}\Big(C_1 \hat{P}_\tau C_1^T\Big) = \mathrm{tr}\Big(B_1^T \hat{Q}_\tau B_1\Big), \text{ and} \\
\langle \mathcal{S}, \hat{\mathcal{S}} \rangle_{h_2,\tau} &= \mathrm{tr}\Big(CYC_1^T\Big) = \mathrm{tr}\Big(B^T Z B_1\Big).
\end{aligned}$$

As a consequence, the following error bound holds.

Proposition 2.2. *The following error bound holds for time-limited balanced truncation of discrete-time systems*

$$\max_{j=0,1,\ldots,\tau} \|y(j) - \hat{y}(j)\|_2 \leq \mathcal{I} \left(\sum_{j=0}^{\tau} \|u(j)\|_2^2 \right)^{\frac{1}{2}},$$

where

$$\begin{aligned} \mathcal{I}^2 &= \operatorname{tr}\left(C P_\tau C^T + C_1 \hat{P}_\tau C_1^T - 2CYC_1^T \right) \\ &= \operatorname{tr}\left(B^T Q_\tau B + B_1^T \hat{Q}_\tau B_1 - 2B^T Z B_1 \right) \end{aligned}$$

where P_τ and Q_τ are the TL Gramians of the full order system $\mathcal{S}$, $\hat{P}_\tau$ and $\hat{Q}_\tau$ are the TL Gramians of the reduced order system $\hat{\mathcal{S}}$, and Y, Z are the solutions of the matrix equations (17a) *and* (17b)*.*

Proposition 2.2 provides an error bound for the time-limited h_2 norm of the error system $\mathcal{S} - \hat{\mathcal{S}}$. It can be computed in practice by solving two TL Stein equations (as in (13)) for the model $\mathcal{S}$ and the model $\hat{\mathcal{S}}$, and one Stein-like equation (as in (17)). It is worth noting that this bound is valid for every reduced order model $\hat{\mathcal{S}}$. Moreover, it holds even in the case the original model or the reduced order model are unstable. In the next section, we develop an expression of this error bound tailored to a reduced order model arising from TL balanced truncation.

3. Output error bound to time-limited balanced truncation

3.1. Error bound to TL balanced truncation

Let suppose that $\mathcal{S} = (A, B, C)$ is an n-th order balanced systems associated with the time-limited Gramians $P_\tau = Q_\tau = \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$. Let's consider the following partition

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \quad C = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \text{ and } \Sigma = \begin{bmatrix} \Sigma_1 & \\ & \Sigma_2 \end{bmatrix}. \tag{19}$$

As a consequence, we must have

$$A\Sigma A^T - \Sigma + BB^T - F_\tau F_\tau^T = 0, \tag{20a}$$

$$A^T \Sigma A - \Sigma + C^T C - G_\tau^T G_\tau = 0, \tag{20b}$$

where $F_\tau = A^\tau B = \begin{bmatrix} F_1 \\ F_2 \end{bmatrix}$, and $G_\tau = CA^\tau = \begin{bmatrix} G_1 & G_2 \end{bmatrix}$. The reduced order model obtained by time-limited balanced truncation is $\hat{\mathcal{S}} = (\hat{A}, \hat{B}, \hat{C})$, where $\hat{A} = A_{11} \in \mathbb{R}^{r \times r}$, $\hat{B} = B_1 \in \mathbb{R}^{r \times m}$ and $\hat{C} = C_1 \in \mathbb{R}^{p \times r}$.

Hence, the time-limited h_2 norm of the error system is

$$\begin{aligned} \|\mathcal{S}_e\|_{h_{2,\tau}}^2 &= \operatorname{tr}\left(B^T \Sigma B - 2B^T Z B_1 + B_1 \hat{Q}_\tau B_1 \right) \\ &= \operatorname{tr}\left(B^T \Sigma B - 2B_1^T Z_1 B_1 - 2B_2^T Z_2 B_1 + B_1 \hat{Q}_\tau B_1 \right). \end{aligned} \tag{21}$$

By developing the (2,1) term of (20a), we obtain

$$A_{11}\Sigma_1 A_{21}^T + A_{12}\Sigma_2 A_{22}^T + B_1 B_2^T - F_1 F_2^T = 0,$$

and consequently

$$\operatorname{tr}\left(-2B_2^T Z_2 B_1 \right) = \operatorname{tr}\left(-2B_1 B_2^T Z_2 \right) = \operatorname{tr}\left(2A_{11}\Sigma_1 A_{21}^T Z_2 + 2A_{12}\Sigma_2 A_{22}^T Z_2 - 2F_1 F_2^T Z_2 \right).$$

Substituting this into (21) yields

$$\|\mathcal{S}_e\|_{h_{2,\tau}}^2 = \operatorname{tr}\left(B^T \Sigma B - 2B_1^T Z_1 B_1 + 2A_{11}\Sigma_1 A_{21}^T Z_2 + 2A_{12}\Sigma_2 A_{22}^T Z_2 - 2F_1 F_2^T Z_2 + B_1^T \hat{Q} B_1 \right).$$

For developing the term $\operatorname{tr}\left(2A_{11}\Sigma_1 A_{21}^T Z_2 \right)$, consider the $(1, 1)$ entry of (17b):

$$A_{11}^T Z_1 A_{11} + A_{21}^T Z_2 A_{11} - Z_1 + C_1^T C_1 - G_1^T \hat{G} = 0$$

leading to

$$\begin{aligned}\operatorname{tr}\left(2A_{11}\Sigma_1 A_{21}^T Z_2\right) &= \operatorname{tr}\left(2\Sigma_1 A_{21}^T Z_2 A_{11}\right)\\ &= \operatorname{tr}\left(2\Sigma_1 Z_1 - 2\Sigma_1 A_{11}^T Z_1 A_{11} - 2\Sigma_1 C_1^T C_1 + 2\Sigma_1 G_1^T \hat{G}\right).\end{aligned}$$

Hence,

$$\begin{aligned}\|\mathcal{S}_e\|_{h_{2,\tau}}^2 = \operatorname{tr}&\left(B^T\Sigma B - 2B_1^T Z_1 B_1 + 2\Sigma_1 Z_1 + 2\Sigma_1 G_1^T\hat{G} - 2\Sigma_1 A_{11}^T Z_1 A_{11}\right)\\ &+ \operatorname{tr}\left(-2\Sigma_1 C_1^T C_1 + 2A_{12}\Sigma_2 A_{22}^T Z_2 - 2F_1 F_2^T Z_2 + B_1^T\hat{Q}_\tau B_1\right).\end{aligned}$$

From now on, the steps get particularly different from derivations for TLBT for continuous-time systems, because, for discrete-time systems, the reduced order model is not balanced. Recalling that

$$\operatorname{tr}\left(B^T\Sigma B\right) = \operatorname{tr}\left(C\Sigma C^T\right), \text{ and } \operatorname{tr}\left(B_1^T\hat{Q}_\tau B_1\right) = \operatorname{tr}\left(C_1\hat{P}_\tau C_1^T\right),$$

gives

$$\begin{aligned}\|\mathcal{S}_e\|_{h_{2,\tau}}^2 = \operatorname{tr}&\left(2A_{12}\Sigma_2 A_{22}^T Z_2 + C_2\Sigma_2 C_2^T - C_1\Sigma_1 C_1^T + C_1\hat{P}_\tau C_1^T\right)\\ &+ \operatorname{tr}\left(-2B_1^T Z_1 B_1 - 2A_{11}\Sigma_1 A_{11}^T Z_1 + 2\Sigma_1 Z_1\right)\\ &+ \operatorname{tr}\left(2\Sigma_1 G_1^T\hat{G} - 2F_1 F_2^T Z_2\right).\end{aligned}$$

Since

$$A_{11}\Sigma_1 A_{11}^T + A_{12}\Sigma_2 A_{12}^T - \Sigma_1 B_1 B_1^T - F_1 F_1^T = 0$$

it holds

$$\operatorname{tr}\left(-2B_1^T Z_1 B_1 - 2A_{11}\Sigma_1 A_{11}^T Z_1 + 2\Sigma_1 Z_1\right) = \operatorname{tr}\left(2A_{12}\Sigma_2 A_{12}^T Z_1 - 2F_1 F_1^T Z_1\right).$$

Summarizing all of these steps together, we have the following theorem.

Theorem 3.1. *Let* $S = \left(\begin{bmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{bmatrix}, \begin{bmatrix}B_1\\ B_2\end{bmatrix}, \begin{bmatrix}C_1 & C_2\end{bmatrix}\right)$ *be a balanced system and* $\hat{\mathcal{S}} = (A_{11}, B_1, C_1)$ *be the r-th order reduced model obtained by time-limited balanced truncation. The time-limited* h_2 *norm of the error system is given by*

$$\begin{aligned}\|\mathcal{S}_e\|_{h_{2,\tau}}^2 &= \operatorname{tr}\left(C_2\Sigma_2 C_2^T + 2A_{12}\Sigma_2 A_{:2}^T Z\right) + \operatorname{tr}\left(C_1(\hat{P}_\tau - \Sigma_1)C_1^T\right)\\ &\quad +2\operatorname{tr}\left(\Sigma_1 G_1^T\hat{G} - F_1 F^T Z\right),\\ &= \operatorname{tr}\left(B_2^T\Sigma_2 B_2 + 2A_{21}^T\Sigma_2 A_{2:}Y\right) + \operatorname{tr}\left(B_1^T(\hat{Q}_\tau - \Sigma_1)B_1\right)\\ &\quad +2\operatorname{tr}\left(\Sigma_1 F_1^T\hat{F} - G_1 G^T Y\right),\end{aligned} \tag{22}$$

where $A_{:2} = \begin{bmatrix}A_{12}\\ A_{22}\end{bmatrix}$, $A_{2:} = \begin{bmatrix}A_{21} & A_{22}\end{bmatrix}$, $F = A^\tau B = \begin{bmatrix}F_1\\ F_2\end{bmatrix}$, $G = CA^\tau = \begin{bmatrix}G_1 & G_2\end{bmatrix}$, $\hat{G} = CA_{11}^\tau$, *and* Z *and* Y *are the solutions of* (17b) *and, respectively,* (17a) *with* $\hat{A} = A_{11}$.

Theorem 3.1 gives an analytic expression for the error bound provided in Proposition 2.2 in the case the reduced order model is obtained by TLBT. This characterization highlights how the error bound depends on the singular values Σ and the time-limited terms $G, \hat{G}, F, \hat{F}$. It should be emphasized that it holds even if the original and reduced order models are unstable, provided the solvability conditions for the involved matrix equations hold. This expression depends on the partitioned matrix of the balanced full order model, the partitioning of the time-limited Hankel singular value matrices, and the matrices Y and Z appearing in Proposition 2.1 for the computation of the inner product. Readers should notice that the TL error bound differs for the infinite horizon error bound in (10) by the residual time-limited term

$$R_\tau := 2\operatorname{tr}\left(\Sigma_1 G_1^T\hat{G} - F_1 F^T Z\right). \tag{23}$$

As one would expect, we will see that if $\tau \to \infty$ yields $R_\tau \to 0$, and the expression given in (22) will tend to the error bound expression for the infinite time horizon. In what follows, we will study the impact of the TL terms $G, F, \hat{G}$, and $\hat{F}$ in the error bound.

3.2. Time-limited residue impact in error bound

3.2.1. Stability preservation

For the infinite time horizon case, balanced truncation for discrete-time systems always produces a stable reduced order model which is not automatically the case for the time-limited variant. In what follows, we provide a sufficient condition for the reduced order model obtained by TLBT to be stable. We keep the notation used in the last section.

Proposition 3.1. *(***Stability preservation***)* *Suppose that*

$$Q = A_{12}\Sigma_2 A_{12}^T + B_1 B_1^T - F_1 F_1^T \geq 0,$$

and the pair (A_{11}, Q) *is reachable. Then the reduced order model is stable.*

Proof. From the Stein equation (20a) it follows

$$A_{11}\Sigma_1 A_{11}^T + A_{12}\Sigma_2 A_{12}^T - \Sigma_1 + B_1 B_1^T - F_1 F_1^T = 0. \tag{24}$$

Let $v \in \mathbb{C}^r$ and $\mu \in \mathbb{C}$ be an eigenpair of A_{11}^T, *i.e.*, $A_{11}^T v = \mu v$. Then we multiply (24) by v^* (on the left), and v (on the right) to obtain

$$(1 - |\mu|^2) v^* \Sigma_1 v = v^* \underbrace{\left(A_{12}\Sigma_2 A_{12}^T + B_1 B_1^T - F_1 F_1^T\right)}_{=Q} v \geq 0.$$

Since $\Sigma_1 > 0$ this immediately implies $|\mu| \leq 1$.

Now let us assume, by contradiction, that $|\mu| = 1$. In this case, we have $v^* Q = 0$. Moreover, if we multiply (24) by A_{11} (on the left) and A_{11}^T (on the right), we obtain

$$A_{11}^2 \Sigma_1 (A_{11}^T)^2 - A_{11}\Sigma_1 A_{11}^T + A_{11} Q A_{11}^T = 0.$$

Hence, if we multiply the later equation by v^* and v, we obtain

$$0 = (|\mu|^2 - |\mu|^4) v^* \Sigma_1 v = v^* A_{11} Q A_{11}^T v.$$

As a consequence, we have $v^* A_{11} Q = 0$. By induction, we conclude that $v^* A_{11}^{k-1} Q = 0$ for $k > 0$, which implies that the pair (A, Q) is not reachable which contradicts the reachability hypothesis. Then, we must have $|\mu| < 1$ and, hence, that the matrix A_{11} is stable. □

Proposition 3.1 gives a sufficient condition for the ROM produced by TLBT to be stable. It is worth mentioning that this condition relies on the matrices (19) of the balanced realization.

3.2.2. Asymptotic behavior of A^p

Given matrices A and A_{11}, there exist constants $c, \hat{c}, \lambda, \hat{\lambda} > 0$, such that

$$\|A^p\|_2 \leq c \cdot \lambda^p \quad \text{and} \quad \|A_{11}^p\|_2 \leq \hat{c} \cdot \hat{\lambda}^p \tag{25}$$

for all $p \in \mathbb{N}$ and for any matrix norm $\|\cdot\|$. Moreover, if $\Lambda(A)$ and $\Lambda(A_{11})$ lies inside the open unit disc, *i.e.*, A and A_{11} are stable matrices, then λ and $\hat{\lambda}$ can be chosen such that $\lambda < 1$ and $\hat{\lambda} < 1$. If A, A_{11} are assumed to be stable matrices, we know that $A^k \to 0$ and $A_{11}^k \to 0$ whenever $k \to \infty$. Equation (25) describes the asymptotic behavior of the norm $\|\cdot\|$ of those matrix powers, *i.e.*, how fast those sequences of matrices go to zero.

There are different ways to compute c, $\hat{c}$, $\hat{\lambda}$ and λ. For example, in the case where $\|\cdot\|$ is the p induced norm and A is diagonalizable, *i.e.*, $A = XDX^{-1}$ with X nonsingular and D diagonal, we can choose $\lambda = \rho(A)$ to be the spectral radius of A, and $c = \kappa(X) = \|X\|_p \|X^{-1}\|_p$ is the condition number of X in the norm $\|\cdot\|_p$. We refer to [27] for other asymptotic bounds of the form (25). Additionally, the recent paper [28] provides a new improvement on the bounds of matrix functions, which includes matrix powers. The main result of [28] states that

$$\|f(A)\|_2 \leq (1 + \sqrt{2}) \sup_{z \in \Omega} |f(z)|,$$

where $\Omega = \{z \in \mathbb{C}, z = v^H A v, \quad \text{for all} \quad v \in \mathbb{C}^n, \|v\| = 1\}$ is the numerical range of the matrix $A \in \mathbb{C}^{n \times n}$ (also called field of values). Hence, for $f(z) = z^\tau$, the numerical radius $\lambda = r(A) := \max_{z \in \Omega} |z|$, and $c = 1 + \sqrt{2}$, we can use the bound

$$\|A^\tau\| \leq (1 + \sqrt{2}) \cdot \lambda^\tau \tag{26}$$

because $r(A^\tau) \leq r(A)^\tau$. Since in our case, $\tau < \infty$, the above bounds will always be finite even if spectrum or numerical range do not lie inside the unit disc.

From now one, we assume that such c, $\hat{c}$, $\hat{\lambda}$, and λ as in (25) are available.

3.2.3. Asymptotic impact of TL residue

Let us now discuss the impact of R_τ from equation (23) in the error bound of Theorem 3.1. The terms of R_τ can be bounded as

$$2\,\mathrm{tr}\Big(\Sigma_1 G_1^T \hat{G}\Big) \le \|\Sigma_1\|_F \|G_1\|_F \|\hat{G}\|_F,$$
$$2\,\mathrm{tr}\Big(F_1 F^T Z\Big) \le \|Z\|_F \|F_1\|_F \|F\|_F.$$

We recall that, if $V^T = \begin{bmatrix} I_r & 0_{r\times(n-r)} \end{bmatrix}$, then $F = A^\tau B$, $F_1 = V^T A^\tau B$, $G_1 = CA^\tau V$ and $\hat{G} = C_1 A_{11}^\tau$. Additionally , the norms of $\|F_1\|_F$, $\|F\|_F$ are bounded by $c\lambda^\tau\|B\|_F$, $\|G_1\|_F$ is bounded by $c\lambda^\tau\|C\|_F$, and $\|\hat{G}\|_F$ is bound by $\hat{c}\hat{\lambda}^\tau\|C_1\|_F$, where and $\lambda, \hat{\lambda}, c, \hat{c} > 0$ are suitable constants. Moreover, if we assume that $\Lambda(A)$ and $\Lambda(A_{11})$ lie inside the open unit disc, the norms decay fast whenever the value of τ increases and the term $R_\tau \to 0$ if τ goes to infinity. As a consequence, the error bound formulas provided in Theorem 3.1 coincides with those for the infinite time horizon (see equation (10)) in the limit $\tau \to \infty$.

Remark 3.1. *For the infinite time horizon case, if the original and the reduced order model are stable, the error bound in* (10) *can be bounded by*

$$\|\mathcal{S} - \hat{\mathcal{S}}\|_{h_2} \le \mathrm{tr}\Big(C_2 \Sigma_{2,\infty} C_2^T + 2A_{12}\Sigma_{2,\infty} A_{;2}^T Z_\infty\Big), \tag{27}$$

because the term $\mathrm{tr}\Big(C_1(\hat{P}_\infty - \Sigma_{1,\infty})C_1^T\Big) \le 0$. *Indeed, the matrix* $E_\infty = P_\infty - \Sigma_{1,\infty}$ *is negative definite, since it satisfies the following Stein equation*

$$A_{11}E_\infty A_{11}^T - E_\infty - A_{12}\Sigma_{2,\infty}A_{12}^T = 0,$$

and $-A_{12}\Sigma_{2,\infty}A_{12}^T$ *is a negative semi-definite matrix. As a consequence, we directly observe in equation* (27) *that the decay of the singular values will lead to a decay in the error for the infinite time horizon case. We believe that this expression is new and that it was not presented in [9].*

3.2.4. Error bound depending on Σ_2 and asymptotic parameters

Now we wish to explicitly describe the dependency of the expression (22) on the neglected singular values Σ_2 and on the time-limited terms F, $\hat{F}$, G, and $\hat{G}$. From now on, we will assume that A and A_{11} are stable, *i.e.*, that their eigenvalues lie inside the open unit disc. Additionally, we assume that $|\lambda| < 1$ and $\hat{\lambda} < 1$. We will discuss the case where A or A_{11} are unstable in Remark 3.2.

Let us first write $E = \hat{P}_\tau - \Sigma_1$. As a consequence, E satisfies the following Stein equation

$$A_{11}EA_{11}^T - E - A_{12}\Sigma_2 A_{12}^T + F_1F_1^T - \hat{F}\hat{F}^T = 0.$$

Consider the composition $E = E_{\Sigma_2} + E_{TL}$, where

$$A_{11}E_{\Sigma_2}A_{11}^T - E_{\Sigma_2} - A_{12}\Sigma_2 A_{12}^T = 0,$$
$$A_{11}E_{TL}A_{11}^T - E_{TL} + F_1F_1^T - \hat{F}\hat{F}^T = 0.$$

Since A_{11} is stable and $-A_{12}\Sigma_2 A_{12}^T$ is a symmetric negative semi-definite matrix, E_{Σ_2} is also symmetric negative semi-definite. As a consequence, we can rewrite the term $\mathrm{tr}\Big(C_1(\hat{P}_\tau - \Sigma_1)C_1^T\Big)$ as

$$\mathrm{tr}\Big(C_1(\hat{P}_\tau - \Sigma_1)C_1^T\Big) = \mathrm{tr}\Big(C_1(E_{\Sigma_2} + E_{TL})C_1^T\Big) \le \mathrm{tr}\Big(C_1(E_{TL})C_1^T\Big). \tag{28}$$

Since $A_{11}E_{TL}A_{11}^T + F_1F_1^T - \hat{F}\hat{F}^T = E_{TL}$ and A_{11} is stable, E_{TL} can be written as the following infinite series

$$E_{TL} = \sum_{j=1}^{\infty} A_{11}^{j-1}\mathcal{F}_{TL}(A_{11}^T)^{j-1}, \text{ with } \mathcal{F}_{TL} = F_1F_1^T - \hat{F}\hat{F}^T. \tag{29}$$

Consequently,

$$\|E_{TL}\|_2 \le \sum_{j=1}^{\infty} \|A_{11}^{j-1}\|_2 \|\mathcal{F}_{TL}\|_2 \|(A_{11}^T)^{j-1}\|_2$$
$$\le \|\mathcal{F}_{TL}\|_2 \sum_{j=1}^{\infty} \hat{c}^2 \cdot (\hat{\lambda}^{j-1})^2 = \|\mathcal{F}_{TL}\|_2 \frac{\hat{c}^2}{1-\hat{\lambda}^2}.$$

Using similar steps one can show that

$$\|Z\|_2 \leq \frac{c \cdot \hat{c}}{1 - \lambda\hat{\lambda}} \|\mathcal{M}\|_2, \text{ with } \mathcal{M} = C^T C_1 - G^T \hat{G}. \tag{30}$$

Finally, we can bound

$$\begin{aligned}
\left|\operatorname{tr}\left(C_1(\hat{P}_\tau - \Sigma_1)C_1^T\right)\right| &\leq p\|C_1\|_2^2\|E_{TL}\|_2 \leq \frac{p \cdot \hat{c}^2}{1 - \hat{\lambda}^2}\|C_1\|_2^2\|\mathcal{F}_{TL}\|_2, \\
\left|\operatorname{tr}\left(2A_{12}\Sigma_2 A_{:2}^T Z\right)\right| &\leq 2r\|A_{12}\|_2\|\Sigma_2\|_2\|A_{:2}\|_2\|Z\|_2 \\
&\leq \sigma_{r+1}\frac{2r \cdot c \cdot \hat{c}}{1 - \lambda\hat{\lambda}}\|A_{12}\|_2\|A_{:2}\|_2\|\mathcal{M}\|_2 \\
|\operatorname{tr}\left(C_2\Sigma_2 C_2^T\right)| &\leq p\|C_2\|_2^2\|\Sigma_2\|_2 = p\|C_2\|_2^2\sigma_{r+1}, \\
|2\operatorname{tr}\left(\Sigma_1 G_1^T \hat{G}\right)| &\leq 2p \cdot \sigma_1\|G_1\|_2\|\hat{G}\|_2, \\
|2\operatorname{tr}\left(F_1 F^T Z\right)| &\leq \frac{2m \cdot c \cdot \hat{c}}{1 - \lambda\hat{\lambda}}\|F_1\|_2\|F\|\|\mathcal{M}\|_2,
\end{aligned}$$

Additionally, using (25), we have

$$\begin{aligned}
\|\mathcal{F}_{TL}\|_2 &\leq \|F_1\|_2^2 + \|\hat{F}\|_2^2 \leq c^2\lambda^{2\tau}\|B\|_2 + \hat{c}^2\hat{\lambda}^{2\tau}\|B_1\|_2, \\
\|\mathcal{M}\|_2 &\leq \|C\|_2\|C_1\|_2 + \|G\|_2\|\hat{G}\|_2 \leq \|C\|_2\|C_1\|_2(1 + c\hat{c}\lambda^\tau\hat{\lambda}^\tau).
\end{aligned}$$

The following theorem assembles all these results.

Theorem 3.2. *Let* $S = \left(\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \begin{bmatrix} C_1 & C_2 \end{bmatrix}\right)$ *be a balanced system and* $\hat{\mathcal{S}} = (A_{11}, B_1, C_1)$ *be the order-r reduced model obtained by TLBT,* $B_1 \in \mathbb{R}^{r\times m}$*, and* $C_1 \in \mathbb{R}^{p\times r}$*. Let* $c, \hat{c}, \lambda, \hat{\lambda} > 0$ *be constants such that* (25) *holds. Then the following bound holds.*

$$\|S_e\|_{\mathcal{H}_{2,\tau}}^2 \leq J(\tau) \cdot \sigma_{r+1} + J_{TL}(\tau), \tag{31}$$

where

$$\begin{aligned}
J(\tau) &= p\|C_2\|_2^2 + \frac{2rc\hat{c}(1 + c\hat{c}\lambda^\tau\hat{\lambda}^\tau)}{1 - \lambda\hat{\lambda}}\|A_{12}\|_2\|A_{:2}\|_2\|C\|_2\|C_1\|_2, \\
J_{TL}(\tau) &= \frac{p \cdot \hat{c}^2}{1 - \hat{\lambda}^2}\|C_1\|_2^2(c^2\lambda^{2\tau}\|B\|_2 + \hat{c}^2\hat{\lambda}^{2\tau}\|B_1\|_2) + 2p \cdot \sigma_1 c\hat{c}\lambda^\tau\hat{\lambda}^\tau\|C\|_2\|C_1\|_2 \\
&\quad + \frac{2m \cdot c \cdot \hat{c}}{1 - \lambda\hat{\lambda}}c^2\lambda^{2\tau}\|B\|_2^2\|C\|_2\|C_1\|_2(1 + c\hat{c}\lambda^\tau\hat{\lambda}^\tau).
\end{aligned}$$

Theorem 3.2 splits the bounds from (22) into $J(\tau)\sigma_{r+1}$ and $J_{TL}(\tau)$. The term $J(\tau)\sigma_{r+1}$ depends linearly on σ_{r+1}, *i.e.*, the largest neglected Hankel singular value. The term $J_{TL}(\tau)$ represents the time-limited terms. If τ goes to infinite we have

$$J_{TL}(\tau) \to 0 \text{ and } \; J(\tau) \to J_\infty = p\|C_2\|_2^2 + \frac{2rc\hat{c}}{1 - \lambda\hat{\lambda}}\|A_{12}\|_2\|A_{:2}\|_2\|C\|_2\|C_1\|_2.$$

Remark 3.2. *In the case* $\lambda \geq 1$*, or* $\hat{\lambda} \geq 1$ *, the equations* (28), (29) *and* (30) *do not hold anymore and, consequently, Theorem3.2 is not valid in this form. However, we can still bound the terms*

$$\begin{aligned}
\left|\operatorname{tr}\left(C_1(\hat{P}_\tau - \Sigma_1)C_1^T\right)\right| &\leq p\|C_1\|_2^2\|\hat{P}_\tau - \Sigma_1\|_2, \\
\left|\operatorname{tr}\left(2A_{12}\Sigma_2 A_{:2}^T Z\right)\right| &\leq 2r\sigma_{r+1}\|A_{12}\|_2\|A_{:2}\|_2\|Z\|_2, \\
|2\operatorname{tr}\left(F_1 F^T Z\right)| &\leq 2m\|Z\|_2\|F_1\|_2\|F\|_2.
\end{aligned}$$

Hence, the equivalent to Theorem 3.2 has explicit dependencies on Z, $\hat{P}_\tau$ *and* Σ_1*.*

Table 1: Summary of the error bounds for small-scale example

Eq. (10) for BT	Prop. 2.2 for BT	Prop. 2.2 for TLBT	σ_r TLBT	σ_r BT
9.18e-04	4.37e-04	1.92e-07	7.04e-07	3.1e-03

Table 2: Constants for asymptotic behavior

	c	λ	$\hat{c}$	$\hat{\lambda}$	Thm 3.2 bound
Eig. Value Decomp.	12.26	0.97	2.95	0.97	276.91
Field of values	2.41	1.06	2.41	0.99	9.93

Remark 3.3. *For generalized state-space systems*

$$\begin{aligned} Mx(k+1) &= Ax(k) + Bu(k), \text{ for } k \in \mathbb{N} = \{0, 1, 2, \dots\} \\ y(k) &= Cx(k), \quad x(0) = x_0, \end{aligned} \tag{32}$$

with a nonsingular matrix $M \in \mathbb{R}^{n\times n}$, the results established so far hold as well with minor modifications that we give next without derivations as those follow the same reasoning as in the continuous-time situation [14, 15]. In particular, the time-limited Gramians are P_τ, $M^T Q_\tau M$ and are now obtained from the solutions of the generalized Stein equations

$$AP_\tau A^T - MP_\tau M^T + BB^T = F_M F_M^T, \quad F_M := M(AM^{-1})^\tau B \tag{33a}$$

$$A^T Q_\tau A^T - M^T Q_\tau M + C^T C = G_M^T G_M, \quad G_M := C(M^{-1}A)^\tau \tag{33b}$$

Obviously, the infinite Gramians of (32) *are given by omitting the terms F_M, G_M in* (33)*. Since in balanced coordinates M is transformed to the identity and $M_{11} = I_r$, the matrix equations for Gramians $\hat{P}_\tau$ and $\hat{Q}_\tau$ of the reduced system remain unchanged. The Sylvester equations* (17) *transform to*

$$AY\hat{A}^T - MY + B\hat{B}^T - F_M\hat{F}^T = 0, \tag{34a}$$

$$A^T Z\hat{A} - M^T Z + C^T\hat{C} - G_M^T\hat{G} = 0, . \tag{34b}$$

Consequently, by using the adapted Gramians and matrix equations, the error bounds still hold.

3.3. Small-scale example

We illustrate the obtained results by applying BT and TLBT to a small-scale system and compute the infinite time horizon (equation (10)) and time-limited bounds (Proposition 2.2 or Theorem 3.1). For this, we consider a random stable single-input single-output (SISO) system of order $n = 10$, generated by the MATLAB command rss and convert it to a discrete-time system using a zero-order hold procedure (command c2d) with discretization step dt = 1sec. We considered a time horizon of $\tau = 20$. Then the infinite horizon and time-limited Gramians and error bounds are computed using the MATLAB direct solver (command dlyap). Finally, two reduced models of order $r = 6$ are computed using BT and TLBT, and, in this case, the two models are stable.

We compare the time-domain response of the corresponding two reduced models. For this, we use $u(1) = 1, u(k) = 0$, for $k > 1$ as the control input. The results of the absolute errors are depicted in Figure 1, including the bounds from Equation (10) and Proposition 2.2 for BT and TLBT, and twice the sum of the neglected (time-limited) Hankel singular values $\boldsymbol{\sigma}_r$ for (TL)BT. By inspecting the time-domain error between the original response and the two reduced order models, we observe that the TLBT generally produces better results compared to BT in the given time-limited interval. Additionally, the errors satisfy the bounds from Equation (10) and Proposition 2.2 (see Table 1 for the numerical values).

Now, we compute also the bounds from Theorem 3.2 for TLBT. For this goal, we first need an estimation of the constants $c, \hat{c}, \lambda, \hat{\lambda}$. We considered two sets of constants, one obtained using the eigenvalue decomposition and the other one using the field of values and the inequality (26). Those values and the error bounds are displayed in the Table 2. Notice that for the values related to the eigenvalue decomposition, we have that $\lambda < 1$ and $\hat{\lambda} < 1$. As a consequence, Theorem 3.2 holds, and we use it to compute the displayed error bounds. However, for the field of values, we have that $\lambda > 1$, and so Theorem 3.2 does not hold anymore. To circumvent the issue, we use the ideas in Remark 3.2 to compute the bound. By inspecting Tables 1 and 2, we conclude that the bounds depending on the asymptotic parameters are less sharp. Indeed, they were developed in order to study the asymptotic behavior of the error with respect to τ, and their value is rather theoretical than practical nature.

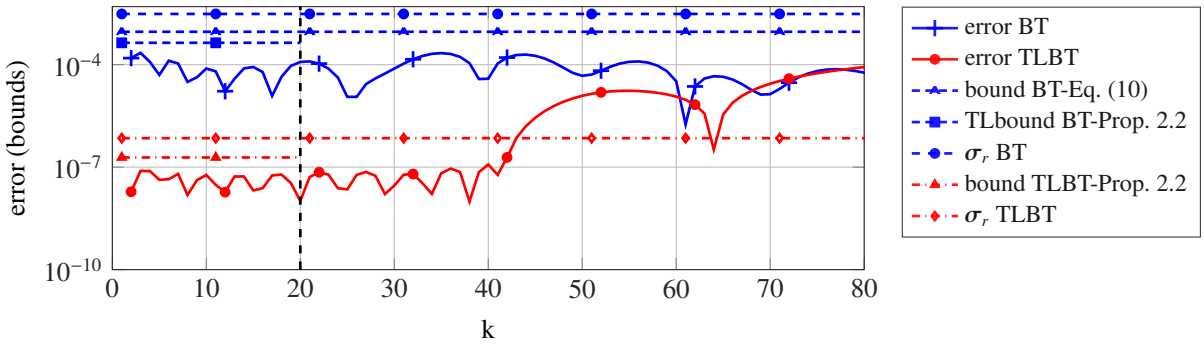

Figure 1: Output errors $|y(k)-\hat{y}(k)|$, error bounds, and sum of neglected HSVs σ_r for (TL)BT reduction of small-scale example to order $n = 10$, reduced order $r = 6$ with time limit $\tau = 20$.

Algorithm 1: Smith-Arnoldi method for Stein equations

Input : $A,\ B$ as in (4), tolerances $0 < \varepsilon \ll 1$.

Output: $Q_k Q_k^T \approx P_\infty$ with $Q_k \in \mathbb{R}^{n\times\ell}$, $Y_k \in \mathbb{R}^{\ell\times\ell}$, $\ell \leq mk \ll n$.

$B = q_1\beta$ s.t. $q_1^T q_1 = I_m$, $Q_1 = q_1$, $r_{1:m,1:m} = \beta E_1$

for $k = 1, 2, \ldots$ **do**

  $g = Aq_k$, $h_{k+1,1:k} = Q_k^T g$, $g_+ = g - Q_k h_{k+1,1:k}$.

  $q_{k+1} = g_+ h_{k+1,k}$ s.t. $q_{k+1}^T q_{k+1} = I_m$.

  $Q_{k+1} = [Q_k,\ q_{k+1}]$.

  $r_{1:k+1,1:k+1} = H_{1:k+1,1:k} r_{1:k+1,1:k}$ (next block column of R_k)

$Q_k = Q_k R_k$.

4. Computational Aspects

4.1. Numerical Computation of the Gramians

As for BT for continuous-time systems, the solution of the large-scale discrete-time Lyapunov equations (5), (13) is the computationally most demanding step. We will restrict the following discussion to the reachability Gramians, since from there, the results for the observability are easily given by replacing $A,\ B$ by $A^T,\ C^T$. Especially in the large-scale situation, directly computing and storing the Gramians is infeasible because, in general, they are large, dense matrices. The common practice when $m, p \ll n$ is to compute approximations of low-rank, e.g. $P_\infty \approx QYQ^T$ with $Q \in \mathbb{R}^{n\times k}$, $Y = Y^T \in \mathbb{R}^{k\times k}$, $k \ll n$ which is motivated by the typically rapid singular value decay of the Gramians, see e.g., [29, 30, 31]. BT is then carried out with the low-rank solution factors of the (infinite or time-limited) Gramians instead of exact Cholesky factors.

There exist different algorithms for computing the low-rank solution factors $Q,\ Y$ by using techniques from large-scale, numerical linear algebra. For the Stein equations, the expressions (4) directly motivate the Smith method [32, 33] for computing low-rank factors:

$$\begin{aligned} P_k &= AP_{k-1}A^T + BB^T, \quad k \geq 1,\ X_0 = 0 \\ &= \sum_{j=0}^{k-1} A^j BB^T (A^T)^j = Z_{k-2}Z_{k-2}^T + A^{k-1}BB^T(A^T)^{k-1} = Z_k Z_k^T \approx P_\infty, \\ Z_k &:= [B, AB, \ldots, A^{k-1}B]. \end{aligned} \tag{35}$$

Underlying the Smith iteration (35) is the (block) Krylov subspace of order k:

$$\operatorname{range}(Z_k) = \mathcal{K}_k(A, B) = \operatorname{range}([B, AB, \ldots, A^\tau B]).$$

Hence, we can also find approximate solutions of (4) via a block Arnoldi process [34, 35]. Let $Q_k = [q_1, \ldots, q_k] \in \mathbb{R}^{n\times km}$, $q_i \in \mathbb{R}^{n\times m}$ span a orthonormal basis of $\mathcal{K}_k(A, B)$ with $B = q_1\beta$, $\beta \in \mathbb{R}^{m\times m}$. Suppose the Arnoldi relation $AQ_k = Q_k H_k + q_{k+1}h_{k+1,k}E_k^T$ holds, where $H_k = Q_k^T A Q_k = [h_{ij}]$ is block upper Hessenberg, and $E_k = e_k \otimes I_m$. Then we have $Z_k = Q_k R_k$ for a block upper triangular matrix $R_k = [r_1, \ldots, r_k] \in \mathbb{R}^{mk\times mk}$ and $r_i = H_k(r_{i-1})$, $2 \leq i \leq k$, $r_1 = E_1\beta$. Algorithm 1 illustrates this procedure. Alternatively, we can impose a Galerkin orthogonality condition

on the Lyapunov residual associated to a low-rank approximation of the form $Q_k Y_k Q_k^T$. This enforces that Y_k has to be the solution of a projected version of (4), i.e.,

$$H_k Y_k H_k^T - Y_k + (Q_k^T B)(B^T Q_k) = 0, \quad H_k := Q_k^T A Q_k, \tag{36}$$

which can be solved by standard dense methods. If the quality of the approximation $Q_k Y_k Q_k^T$ is not sufficient, Q_k is orthogonally expanded by continuing the Arnoldi process. The convergence rate of the Smith iteration depends on the spectral radius of A and can be very slow if $\rho(A) \approx 1$. To overcome this issues, so called squared Smith methods were discussed in [33, 29, 36] with limited success.

The occurrence of the matrix functions in time- and frequency-limited BT, or more precisely the action of $f(A)$ to B, adds an additional computational difficulty. At a first glance, the required monomials $f(z) = z^\tau$ in time-limited discrete-time BT appear to be a comparatively simple situation, especially if τ is very small relatively to n and the required τ matrix vector products with A are affordable. In that case we can directly use the iteration (35) or Algorithm 1 for the time-limited Gramians (12). Running (1) for an additional step allows to read off F from the last block column of Q_k. Alternatively, we can use the Galerkin projection framework mentioned above, i.e., we build Galerkin approximations $F \approx H_k^\tau (Q_k^T B)$ and $P_\tau \approx Q_k Y_k Q_k^T$, where Y_k solves

$$H_k Y_k H_k^T - Y_k + (Q_k^T B)(B^T Q_k) - H_k^\tau (Q_k^T B)(B^T Q_k)(H_k^\tau)^T = 0. \tag{37}$$

These approximations are exact if range $(Q_k) = \mathcal{K}_{\tau+1}(A, B)$ because then range $(A^\tau B) \in$ range (Q_k) and range $(Z_{\tau+1}) \in$ range (Q_k) with $Z_{\tau+1}$ from (35).

Unfortunately, for large values $\tau \approx n$ and if $\rho(A) \approx 1$, this basic Galerkin approach or Algorithm 1 become impractical as they would both require prohibitively large subspace dimensions. Note that getting the powers of A via approaches like binary powering [37, Chapter 4.1] are not feasibly for large A, since successively squaring A would destroy its sparsity and, hence, the required matrix-matrix multiplications would become too costly.

For achieving accurate low-rank approximations for the Gramians with smaller subspace dimensions, rational Krylov subspaces

$$\text{range}\,(Q_k) = \mathcal{RK}_k(A, B, \boldsymbol{\xi}) = \text{range}\left([B, (A + \xi_2 I)^{-1} B, \ldots, \prod_{j=2}^{k} (A + \xi_j I)^{-1} B] \right) \tag{38}$$

have been proven to be a viable choice [38, 39, 40], provided adequate shift parameters $\xi_i \in \mathbb{C}$ are available. The majority of literature regarding rational Krylov subspace methods for solving large matrix equations is focused on the continuous-time case. Although the discrete-time case can be dealt with similarly, to the authors knowledge not much is known about the shift parameter selection in this case.

A low-rank ADI iteration for Stein equations (4) was proposed in [41, 21]. Note that the low-rank ADI iteration is related to both the Smith method as well as to rational Krylov subspace methods. Rational Krylov subspace as well as ADI methods for (4) can be used directly to the time-limited equations (13) if F and G or approximations thereof are given which, however, is a crucial point because they have to be computed first.

In the present work, we follow an approach similar to the one proposed in [20, 21, 14] for the continuous-time setting. We employ the rational Krylov subspace method illustrated in 2 that iteratively computes approximations of both $F = A^\tau B$ and P_τ.

We shall next describe some important aspects this method. Obviously, by omitting all parts related to $F = A^\tau B$, Algorithm 2 is applicable to the infinite Gramians (4) as well.

Solution of the projected problems. Two approaches for dealing with (37) in line 5 are discussed. Following the algorithmic strategy proposed in [20, 21, 14], at first an approximation of $F = A^\tau B$ is computed by a projection principle: $F \approx F_k := Q_k \hat{F}_k$, $\hat{F}_k := H_k^\tau (Q_k^T B)$. Since H_k is of size $mk \ll n$, the powers of H_k can be efficiently computed by binary powering which requires $\lfloor \log_2 \tau \rfloor$ matrix-matrix multiplications. This can be less costly compared to the computation of the more complicated matrix functions (matrix exponentials and logarithms) that occur in other time- or frequency-limited BT methods. Since the goal is to approximate the associated term of the inhomogeneity of (12), we use a relative norm wise change $\mathfrak{F} := \|\hat{F}_k \hat{F}_k^T - \hat{F}_{k-1} \hat{F}_{k-1}^T\| / \|\hat{F}_{k-1}\|^2$ to assess the accuracy of the current approximation F_k. Once $\mathfrak{F} \le \varepsilon_f \le \varepsilon \ll 1$, where ε is a given threshold, the computed approximation is accepted and we start solving the Galerkin system (37) for Y_k. This can be done by, e.g., direct (Bartels-Stewart type) methods [42] or the Smith iteration 35. The rational Krylov method is continued until the scaled residual norm $\mathfrak{R}$ with respect to the low-rank solution $Q_k Y_k Q_k^T$ falls below ε. During these steps, the quality of the approximation F_k of F can be further refined by computing a new $\hat{F}_k$ before solving the compressed Stein equation (37).

Algorithm 2: Rational Krylov subspace method for time-limited DALEs (13)

Input : $A,\ B,\ \tau$ as in (13), tolerances $0 < \varepsilon \ll 1$.

Output: $Q_k Y_k Q_k^T \approx P_\tau$ with $Q_k \in \mathbb{R}^{n\times\ell}$, $Y_k \in \mathbb{R}^{\ell\times\ell}$, $\ell \le mk \ll n$.

1 $B = q_1\beta$ s.t. $q_1^T q_1 = I_m$, $Q_1 = q_1$.
2 **for** $k = 1, 2, \ldots$ **do**
3 $\quad H_k = Q_k^T A Q_k$, $B_k = Q_k^T B$.
4 $\quad \hat{F}_k = H_k^\tau B_k$, $F_k = Q_k \hat{F}_k$.
5 $\quad$ Compute Gramian defined by $H_k,\ B_k,\ \hat{F}_k$ (e.g., solve (37)).
6 $\quad$ Set $\mathfrak{R}_k := \|A(Q_k Y_k Q_k^T)A^T - (Q_k Y_k Q_k^T) + BB^T - F_k F_k^T\|$ with $F_k \approx A^\tau B$.
7 $\quad$ **if** $\mathfrak{R}_k < \varepsilon \|BB^T - F_k F_k^T\|$ **then**
8 $\quad\quad$ Return $P_{\tau,k} = Q_k Y_k Q_k^T$ (truncate if necessary).
9 $\quad$ Select next shift s_{k+1}.
10 $\quad$ Solve $(A - s_{k+1} I)g = q_k$ for g.
11 $\quad g_+ = g - Q_k(Q_k^T g)$, $q_{k+1} = g_+\beta_k$ s.t. $q_{k+1}^T q_{k+1} = I_m$.
12 $\quad Q_{k+1} = [Q_k,\ q_{k+1}]$.

Alternatively, the separate computation of $\hat{F}_k$ can be avoided since (37) can be entirely dealt with by τ steps of the Smith iteration 35. Depending on the sizes of τ and H_k, this can be less costly compared to the first approach, where (37) is solved by a direct method. As mentioned earlier, $\hat{F}_k = H_k^\tau B_k$ can still be obtained as by-product of the Smith iteration. Note that for the infinite situation $\tau = \infty$, using the Smith iteration for (36) requires that the restriction H_k is stable, which is theoretically ensured if the numerical range of A lies in the unit disc.

The computational effort of both of these two strategies can be further reduced by solving the projected matrix equation (37) only in each μth iteration step (e.g., $\mu = 5$) of Algorithm 2.

Computing the residual of the Stein equations. We wish to assess the accuracy of the low-rank approximation $Q_k Y_k Q_k^T$ by means of the norm $\mathfrak{R}$ of the Lyapunov residual matrix. However, directly computing the residual norm $\mathfrak{R}$ is impractical since the Lyapunov residual matrix is a large, dense matrix. The following Lemma reveals an efficient way to compute the residual norm.

Lemma 4.1. *The residual matrix at step k of Algorithm 2 is given by*

$$\begin{aligned} R_k &:= A(Q_k Y_k Q_k^T) - Q_k Y_k Q_k^T + BB^T - Q_k \hat{F}_k \hat{F}_k \hat{Q}_k^T \\ &= [g_k, w_k] \begin{bmatrix} \psi_k^T Y_k \psi_k & I_m \\ I_m & 0_m \end{bmatrix} [g_k, w_k]^T, \quad g_k := s_{k+1} q_{k+1} - (I - Q_k Q_k^T) A v_{k+1}, \\ w_k &:= Q_k H_k Y_k \psi_k, \quad \psi_k = \Psi_k^{-T} E_k \psi_{k+1,k}, \end{aligned}$$

where $\Psi_k = [\psi_{ij}] \in \mathbb{R}^{km\times km}$ is the matrix of orthonormalization coefficients $\psi_{ij} \in \mathbb{R}^{m\times m}$ accumulated from line 11. Hence, $\|R_k\| = \|S_k \begin{bmatrix} \psi_k^T Y_k \psi_k & I_m \\ I_m & 0_m \end{bmatrix} S_k^T\|$ and $S_k \in \mathbb{R}^{2m\times 2m}$ is the triangular factor of a thin QR-factorization of $[g_k, w_k]$. The result is also valid for the infinite Gramians by omitting the term $Q_k \hat{F}_k \hat{F}_k \hat{Q}_k^T$.

Proof. The result can be easily established by combining the corresponding results for rational Arnoldi methods for continuous-time equations [38] with those related to standard Arnoldi methods for discrete-time equations [35, 43]. □

Shift parameter selection. Having suitable shift parameters $\xi_2, \ldots, \xi_k$ available is crucial for a rapid convergence of the rational Arnoldi method. Two selection strategies are employed here. At first, alternating shifts $\xi_j = (-1)^j$, $2 \le j \le k$ are used as in [44, Section 3.3] which formally corresponds to the extended Krylov subspace setting ($\xi_{2j-1} = \infty$, $\xi_{2j} = 0$) from [45], and applying a Cayley transformation to map $\overline{\mathbb{C}_-}$ into the closed unit disc. Because only two different coefficient matrices $A \pm I$ occur with this choice in the linear systems in line 10, precomputing and reusing sparse LU-factorizations $L_\pm U_\pm = A \pm I$ in each step afterwards can substantially reduce the computation times for solving the linear systems.

The second shift selection approach is more general and modifies the strategy proposed in [38] by selecting shifts adaptively from the boundary of the unit disc. Suppose the unit circle is discretized into $h_s \in \mathbb{N}$ points, $\Xi := \{\exp \frac{j2\Pi i}{h_s},\ 1 \le i \le h_s\}$, set $m = 1$ for simplicity, and let $\theta_i \in \Lambda(H_k)$. The next shift s_{k+1} is then obtained by

maximizing the rational function associated to the current space $\mathcal{RK}_k$, i.e.,

$$s_{k+1} = \underset{\xi \in \Xi}{\operatorname{argmax}} |r_k(\xi)|, \quad r_k(\xi) = \prod_{j=1}^{k} \frac{\xi - \theta_j}{\xi - \xi_j}, \tag{39}$$

For $m > 1$, there are mk Ritz values θ_i and each previous shift s_j is taken m times in the denominator of r_k in (39) as in [38]. Note that the returned shifts are complex numbers. By requiring that each complex shift is followed by its complex conjugate, the amount of complex arithmetic operations can be reduced following the machinery in, e.g., [46], which will ensure the construction of real, low-rank solution factors Q_k, Y_k.

Generalized systems. Generalized Stein equations (33) corresponding to generalized state-space systems (32) are handled as in the continuous-time setting by implicitly using the algorithms on an equivalent standard state-space system defined by, e.g., $A_M := L_M^{-1} A U_M^{-1}$, $B_M := L_M^{-1} B$, $C := C U_M^{-1}$ with a precomputed sparse factorization $M = L_M U_M$. Afterwards, the obtained low-rank solution factors Z have to be transformed back via $U_M^{-1} Q_k$.

4.2. Computing the error bounds

We will briefly discuss the practical usage of the proposed error bounds. For computing the error bounds in (10) and Proposition 2.2 after a reduction of large-scale systems, the low-rank Gramian approximation are used in the associated places, e.g., in $\operatorname{tr}\left(B^T Q_\tau B\right) \approx \operatorname{tr}\left(B^T Z_{Q_\tau} Z_{Q_\tau}^T B\right)$. The computation of the terms involving the Gramians of the reduced order model requires solving Stein equations of dimension r which can be done direct, dense methods. The terms involving the mixed Gramians require solving Sylvester equations (11), (17), where one of the coefficient is large and sparse but the other one is small and dense. For this particular situation, specialized solvers are available in, e.g. [47], that require r sparse linear systems to be solved. We emphasize that, since only approximate Gramians are used, the expression involving the reachability Gramians might not be identical to the expression with the observability Gramians. For the error bound in Proposition 2.2 for TLBT, this effect might be more pronounced because only approximations of F, G are available in practice which enter (17). Therefore, we use the average of both expressions in the following. Another frequent observation is that the traces with positive and negative signs are very close to each other, e.g., $\operatorname{tr}\left(CP_\tau C^T + C_1 \hat{P}_\tau C_1^T\right) \approx \operatorname{tr}\left(2CYC_1^T\right)$, which can lead to numerical cancellation or even negative values for the complete trace. This seems to be especially an issue if the reduced order model is already very accurate. Hence, we take absolute values $|\operatorname{tr}\left(CP_\tau C^T + C_1 \hat{P}_\tau C_1^T - 2CYC_1^T\right)|$ to circumvent these effects.

It is clear that the bound in Theorem 3.1 is not accessible for large-scale systems because the neglected quantities such as B_2, A_{12}, C_2 are not available in a practical implementation of TLBT.

Some of these unknown quantities are also present in the bound in Theorem 3.2. Additionally, the constants $c, \hat{c}, \lambda, \hat{\lambda}$ are required. Here, the approach used for bounding the powers of A, A_{11} matters. When λ, $\hat{\lambda}$ represent the largest magnitude eigenvalues they can be easily computed for A_{11} and estimated for A by, e.g., an Arnoldi process. The constants $c, \hat{c}$ are then the condition numbers of the eigenvector matrices, which is a difficult to get quantity for large matrices unless the matrices are normal, i.e. $c = 1$. Applying the Crouzeix-Palencia result [28], however, simply uses $c = \hat{c} = 1 + \sqrt{2}$ and the largest value of z^τ on the numerical range of A. It holds $\sup |z^\tau| \leq \alpha^\tau$, where $\alpha := \sup |z|$ is the numerical radius of A which can also be efficiently estimated by approaches utilizing an Arnoldi process see, e.g., [48].

5. Numerical Experiments

In this section, we test the model order reduction methods and the algorithms for computing low-rank factors of the Gramians. All experiments are carried out with implementations in MATLAB® 2016a on a Intel®Core™2 i7-7500U CPU @ 2.7GHz with 16 GB RAM.

5.1. Used test cases

As test cases we use some discrete-time systems from [49] as well as artificially generated and freely scalable test systems, summarized in Table 3 which also gives additional information such as the spectral radius $\rho = \max_{z \in \Lambda(A,M)} |z|$. Consider a positive definite, diagonally dominant matrix $S = L + U + D$, where L, U and D are its strictly upper, lower, and, respectively, diagonal part. Here, S is the matrix associated to a centered finite difference discretization of the Laplace operator on the unit disc. The Jacobi (*Jac*) iteration $v_{k+1} = D^{-1}(L + U)v_k + D^{-1}b$ for the linear system $Sv = b$ represents a basic generalized discrete-time system with coefficients $A = L + U$ and $M = D$ satisfying $\Lambda(D^{-1}(L + U)) = \Lambda(L + U, D) \subset \mathbb{D}$, see, e.g. [50, Chapter 11.2]. Likewise, the Gauss-Seidel

Table 3: Overview of examples

Example	n	$m,\ p$	details	ρ
skl	24389	4, 6	discrete-time system "sparse-skewlap3d-mod-1" from [49][2]	0.91372
Jac	31064	5, 5	Jacobi iteration for S :=`delsq(numgrid('D',200))`	0.99985
GS	31064	5, 5	Gauss-Seidel iteration for S	0.9997

(*GS*) iteration is given by $A = L$, $M = U + D$ with $\Lambda(L, U + D) \subset \mathbb{D}$ but does not require diagonal dominance of S. The input and output maps $B,\ C$ for the *Jac*, *GS* examples are chosen randomly from a uniform distribution on $[0, 1]$.

5.2. Approximation of Gramians and matrix powers

We start testing the approximation of the infinite and time-limited Gramians as well as $F = A^\tau B$ by the methods described in Section 4: the Smith method from Algorithm 1 and the rational Krylov subspace method (Algorithm 2) using two types of shifts: $\xi_j = (-1)^j$ (RKSM(± 1)) and the adaptive selection on the unit circle (RKSM($\mathbb{D}$)). We also compare with the LR-ADI iteration for discrete-time Lyapunov equations [41, 21]. For the time-limited equations (12) this is done via a hybrid approach, where the approximation F obtained from RKSM(± 1) is used to set up the inhomogeneity. The time-limited Gramians are considered with two different time limits to gain insight on how τ influences the computations. The desired accuracy for all cases is

$$\Re := \|AP_kA^T - P_k + BB^T - F_kF_k^T\|/\|BB^T - F_kF_k^T\| \le \tau_P := 10^{-8}$$

and $\|BB^T - F_kF_k^T\|/\|F_kF_k^T\| \le \tau_f = 10^{-8}$ is used for the approximation of F computed in Algorithm 2. The exception is the Smith method for the time-limited Gramians, which is carried out for exactly τ steps, hence providing exact results (up to round-off). After termination, the computed Gramian approximations are truncated by means of an eigenvalue decomposition and keeping only those eigenpairs with $\sum(\lambda_i(P)) > 10^{-12}\lambda_{\max}(P)$. The results are summarized in Table 4, where the approximation of F obtained by the Smith method is used for the final residual norms $\Re$ regarding the time-limited Gramians. Apparently, for small final times τ the Smith method can be competitive in terms of the computation time especially for the *skl* and jac examples. It requires, e.g., the least amount of time for $\tau = 50$ and the *skl* example among all tested methods. Due to the comparatively small spectral radius of A in the *skl* example, the Smith method also delivers competitive times for the infinite Gramians, but fails to deliver the required accuracy for the other two test systems. For all examples, the LR-ADI iteration appears to require the smallest computation times for the infinite Gramians. Considering the dimensions of the built up subspaces, however, indicates that the Smith method generates substantially larger spaces compared to the other approaches. Also, the obtained ranks after truncation seem to be somewhat higher than for the other methods. For approximating the time-limited Gramians, the RKSM approaches seem to be a viable choice with respect to both computational time and the subspace sizes, especially for larger values of τ. The used shift generation strategy has a noticeable influence: while for the *skl* example using the shifts ± 1 leads to less consumed time than the shifts from the unit circle ($\mathbb{D}$), it is the other way around for the *jac* example, and for the *GS* example both shift approaches lead to similar results. The obtained subspace dimensions generated with RKSM($\mathbb{D}$) are in almost all cases smaller compared to RKSM(± 1). The smaller computation times of RKSM(± 1) for the *skl* and *GS* examples are a result of the reuse of LU-factorizations of $A \pm M$ for the linear systems as explained before. For the *jac* example, these savings in solving the linear systems were nullified by the substantially higher subspace dimensions which resulted in much higher times for solving the projected problems.

However, for most examples the substantial discrepancy between subspace dimension and rank after truncation indicates that further enhancements by selecting better shifts are possible. We plan to pursue this topic in future research. For the time-limited Gramians, the hybrid approach of RKSM and LR-ADI appears to yield similar results than the pure RKSM approach.

To conclude this first experimental phase, for small final times τ (and/or a small spectral radius of A), the Smith method can be a viable choice for generating the low-rank factors of the (time-limited) Gramians. For larger τ (and/or spectral radii close to one), the rational Krylov approach appears to be superior, even with the basic shift selection strategies employed here. If $\tau = \infty$, the LR-ADI iteration is often the fastest method.

5.3. Model reduction results and error bounds

Now we carry out infinite and time-limited balanced truncation employing low-rank Gramian approximations generated from the experiments before. It is noteworthy that, apart from different computations times, the obtained

[5] Available at Mert Gürbüzbalaban's webpage `http://mert-g.org/software/`

Table 4: Column dimension d of built up low-rank factors before truncation, rank rk after truncation, final residual norm $\mathfrak{R}$, and computation time t_c (in seconds) of the approximation of P, P_τ by different methods.

		P_∞				$P_{\tau=50}$				$P_{\tau=150}$			
Ex.	method	d	rk	$\mathfrak{R}$	t_c	d	rk	$\mathfrak{R}$	t_c	d	rk	$\mathfrak{R}$	t_c
skl	Smith	680	149	9.4e-09	27.4	200	105	1.5e-15	1.7	600	145	1.4e-15	15.0
	RKSM(±1)	140	91	2.1e-10	21.4	240	64	7.3e-13	9.5	240	85	4.9e-13	9.5
	RKSM($\mathbb{D}$)	128	91	3.0e-10	42.3	168	64	7.3e-13	41.4	184	85	4.9e-13	45.6
	ADI	64	64	5.9e-09	15.5	176	69	3.1e-09	36.6	152	75	4.3e-09	32.6
		P_∞				$P_{\tau=100}$				$P_{\tau=200}$			
Jac	Smith	1500	246	7.0e-01	347.3	500	201	3.2e-13	53.2	1000	230	7.4e-13	173.4
	RKSM(±1)	700	156	1.9e-09	381.7	700	111	1.0e-10	172.5	700	121	2.3e-10	179.0
	RKSM($\mathbb{D}$)	630	156	2.1e-09	342.8	360	111	1.8e-10	52.6	380	121	2.3e-10	69.6
	ADI	230	218	8.9e-09	13.0	570	117	5.5e-09	92.9	540	127	7.0e-09	79.5
		P_∞				$P_{\tau=150}$				$P_{\tau=250}$			
GS	Smith	1500	209	6.0e-01	291.4	750	191	3.9e-13	60.6	1250	206	5.7e-13	166.4
	RKSM(±1)	325	105	4.3e-09	188.4	375	92	8.6e-10	24.2	325	97	8.6e-10	18.5
	RKSM($\mathbb{D}$)	410	105	4.3e-09	59.8	280	92	8.6e-10	20.4	280	97	1.1e-09	20.6
	ADI	135	135	3.8e-09	5.0	310	86	9.9e-09	25.1	290	97	3.0e-09	20.6

Table 5: Results and error bounds for BT and TLBT model reduction to fixed orders r.

		BT				TLBT			
Ex.	r	$\mathcal{E}_{\max}$	bound	σ_r	ρ	$\mathcal{E}_{\max}$	bound	σ_r	ρ
skl, $\tau = 50$	20	3.5e-01	9.3e-01	1.1e+01	0.9728	2.2e-01	6.9e-01	6.5e+00	0.9584
	40	1.0e-02	3.3e-02	2.4e-01	0.9717	1.3e-03	3.6e-03	2.2e-02	0.9852
	60	6.2e-05	1.8e-02	2.7e-03	0.9650	4.6e-07	6.9e-04	1.1e-05	1.0008
jac, $\tau = 200$	40	8.1e-01	2.7e+00	5.9e+01	0.99985	1.8e-01	6.5e-01	6.8e+00	1.00019
	60	1.9e-01	1.7e+00	6.7e+00	0.99986	8.9e-03	6.1e-01	1.9e-01	1.00030
	80	2.1e-02	2.0e+00	6.8e-01	0.99985	1.5e-04	6.6e-01	6.0e-03	1.00000
GS, $\tau = 150$	40	1.3e-01	2.2e+00	2.7e+00	0.99971	1.4e-02	3.6e-01	3.2e-01	0.99964
	60	5.5e-03	2.3e+00	1.0e-01	0.99971	2.2e-04	5.0e-01	4.6e-03	0.99988
	80	1.3e-04	1.5e+00	3.8e-03	0.99971	3.9e-06	7.2e-01	5.4e-05	1.00592

reduction results were largely unaffected by the employed method for generating the low-rank factors, provided the accuracy threshold was achieved. Table 5 lists the results obtained by reducing the systems to different order r: the largest output error in the considered time interval $\mathcal{E}_{\max} := \max_{0\le k\le\tau} \|y(k) - \hat{y}(k)\|_2$, the error bounds (10) and Proposition 2.2 for BT and, respectively, TLBT, twice the sum of the neglected (time-limited) Hankel singular values, $2\sum_{i\ge r+1}\sigma_i$, and the spectral radius of A_r to access stability. For selected reduced orders, Figures 2–4 illustrate the output errors $\|y(k)-\hat{y}(k)\|_2$ against time k as well as the error bounds and HSV sums. From the gathered data, it is apparent that TLBT achieves always smaller reduction errors in the targeted time interval $[0, \tau]$ and, moreover, also the error bound from Proposition 2.2 takes smaller values than the counterpart (10) for unrestricted BT. The doubled sum of neglected HSVs is smaller for TLBT. All this is visually visible in Figures 2–4. As it is expected, after passing the time limit τ, the accuracy of the TLBT models worsens to a point $\tilde{k} \ge \tau$ where BT is more accurate. We also observe from Table 5 that approximately half of the reduced order models generated by TLBT are unstable. We see this, e.g., in Figures 3–4, where the output error drastically increases after passing the time limit τ. Comparing the largest output errors $\mathcal{E}_{\max}$ and the sums of neglected HSVs for TLBT in Table 5 suggests that, although a bound of the form (9) is not given for TLBT, the HSV sum could be used for adaptively determining a suitable reduced order r, exactly as it is often done in unlimited BT. To underline this point, we repeat the model reduction experiment but let (TL)BT determine to reduced orders r adaptively such that

$$2\sum_{k=r+1}\sigma_k \le \epsilon_{hsv}, \tag{40}$$

for different given reduction tolerances $0 < \epsilon_{hsv} < 1$. The results are summarized in Table 6 and indicate that this adaptive determination of the reduced order works for TLBT as fine as for unlimited BT. Moreover, TLBT appears to yield smaller reduced order models of similar accuracy compared to BT. This is a similar observation as for continuous-time TLBT [15].

6. Conclusion

In this paper, we studied time-limited balanced truncation for discrete-time systems. The contributions of this work were divided into two parts. The first part was dedicated to developing output bounds for TLBT. To this aim, we defined the TL h_2 norm and its characterization using matrix equations. By means of this norm, we were able to establish an error bound for the output. Afterwards, we have analyzed the asymptotic behavior of these error

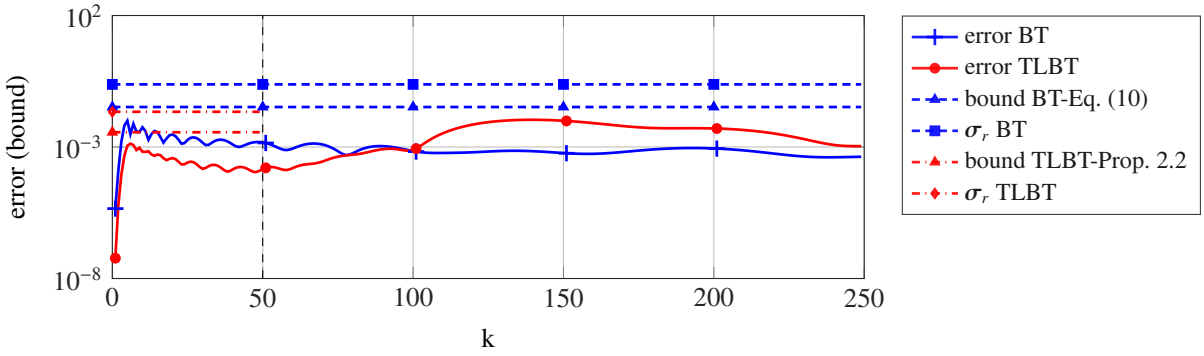

Figure 2: Output errors $\|y(k) - \hat{y}(k)\|_2$, error bounds, and sum of neglected HSVs $\boldsymbol{\sigma}_r$ for (TL)BT reduction of *skl* example to order $r = 40$ with time limit $\tau = 50$.

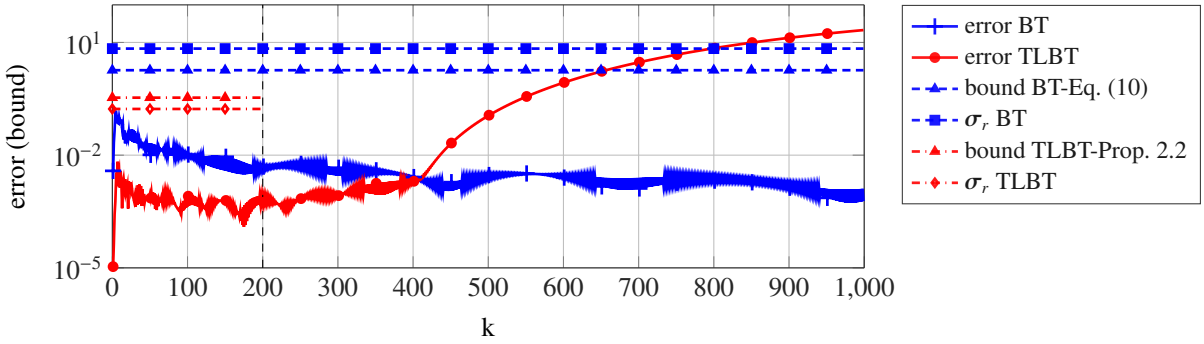

Figure 3: Output errors $\|y(k) - \hat{y}(k)\|_2$, error bounds, and sum of neglected HSVs $\boldsymbol{\sigma}_r$ for (TL)BT reduction of *jac* example to order $r = 60$ with time limit $\tau = 200$.

bounds regarding the time-horizon, highlighting differences to the infinite time horizon as well as the continuous-time situation. The obtained bounds furthermore indicated that the neglected Hankel singular values can be used for an automatic reduced order determination.

The second part of this work was dedicated to computational aspects in large-scale settings. Therein, approximate solutions of the TL Stein equations were obtained by using low-rank factorizations. Rational Krylov subspace methods were proposed for computing the low-rank solution factors. Furthermore, we discussed the residual and error bound computations as well as the selection of shift parameters for the rational Krylov subspace methods. Finally, the algorithms were tested on large-scale examples, and the results were compared with other methods. The time-limited BT approach typically led to more accurate ROMs in the restricted time interval of interest compared to infinite BT, which was also revealed by the smaller values of the corresponding output error bounds. As in the continuous-time case, TLBT occasionally returned unstable ROMS, which might be circumvented in investigations along the lines of, e.g., [16]. The proposed low-rank methods for the arising Stein equations returned satisfactory results for the application in the MOR context with respect to both computing time and accuracy. However, further research is required to bring them to the same level of efficiency as their continuous-time counterparts [38, 21]. Especially the shift parameter selection for discrete-time problems should be improved in future research endeavors.

Acknowledgements

Thank goes to Stefano Massei (TU Eindhoven) and Stefan Guettel (U Manchester) for helpful hints regarding the shift parameter selection for discrete-time problems, and to Michiel Hochstenbach (TU Eindhoven) for providing a MATLAB routine for estimating the numerical radius of a matrix. This work was done while Patrick Kürschner was affiliated with the MPI Magdeburg. The authors also thank the anonymous referee whose comments help improving the paper.

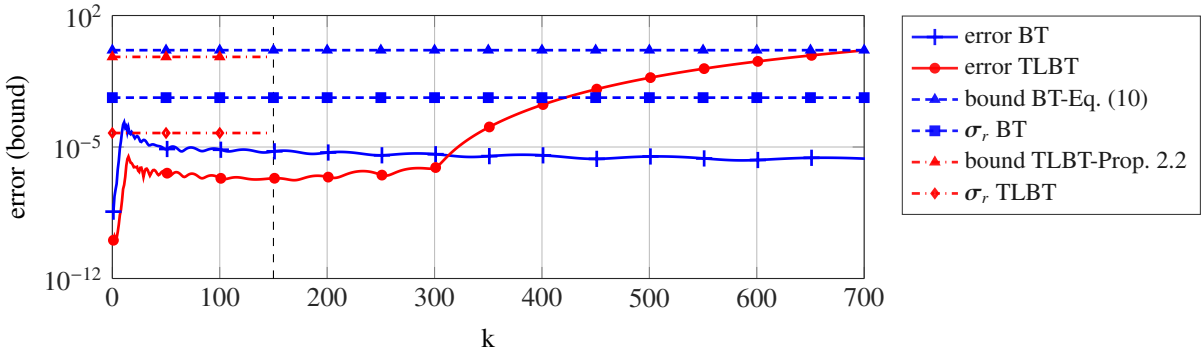

Figure 4: Output errors $\|y(k) - \hat{y}(k)\|_2$, error bounds, and sum of neglected HSVs σ_r for (TL)BT reduction of *GS* example to order $r = 80$ with time limit $\tau = 150$.

Table 6: Results and error bounds for BT and TLBT model reduction, where the reduced orders are adaptively determined via (40) for different ϵ_{hsv}.

Ex.		BT					TLBT				
	ϵ_{hsv}	r	$\mathcal{E}_{\max}$	bound	σ_r	ρ	r	$\mathcal{E}_{\max}$	bound	σ_r	ρ
skl, $\tau = 50$	1.0e-01	45	3.0e-03	2.0e-02	7.9e-02	0.96659	36	5.0e-03	8.8e-03	8.0e-02	0.96958
	1.0e-02	55	3.3e-04	1.8e-02	8.8e-03	0.96885	43	3.4e-04	3.6e-03	7.5e-03	0.97627
	1.0e-03	65	4.9e-05	1.8e-02	8.3e-04	0.96501	49	5.7e-05	1.8e-03	9.0e-04	0.97347
jac, $\tau = 150$	1.0e-01	97	2.2e-03	1.8e+00	9.5e-02	0.99985	64	2.1e-03	2.8e-01	9.2e-02	1.00513
	1.0e-02	115	1.6e-04	1.7e+00	9.3e-03	0.99985	78	2.2e-04	5.9e-01	8.9e-03	1.00049
	1.0e-03	133	2.7e-05	1.9e+00	9.4e-04	0.99985	90	2.6e-05	3.1e-01	9.4e-04	1.00059
GS, $\tau = 150$	1.0e-01	61	5.5e-03	2.6e+00	9.9e-02	0.99971	46	7.1e-03	6.3e-01	9.8e-02	1.00073
	1.0e-02	75	2.1e-04	2.0e+00	9.3e-03	0.99971	58	3.3e-04	7.0e-01	8.4e-03	0.99987
	1.0e-03	88	2.5e-05	1.1e+00	9.7e-04	0.99971	69	2.4e-05	6.8e-01	8.1e-04	0.99981

References

[1] M. Heinkenschloss, T. Reis, A. C. Antoulas, Balanced truncation model reduction for systems with inhomogeneous initial conditions, Automatica J. IFAC 47 (3) (2011) 559–564. doi:10.1016/j.automatica.2010.12.002.

[2] U. Baur, P. Benner, L. Feng, Model order reduction for linear and nonlinear systems: a system-theoretic perspective, Archives of Computational Methods in Engineering 21 (4) (2014) 331–358.

[3] C. Beattie, S. Gugercin, V. Mehrmann, Model reduction for systems with inhomogeneous initial conditions, Systems & Control Letters 99 (2017) 99–106.

[4] C. Schröder, M. Voigt, Balanced Truncation Model Reduction with A Priori Error Bounds for LTI Systems with Nonzero Initial Value, arXiv e-print:2006.02495.

[5] B. C. Moore, Principal component analysis in linear systems: controllability, observability, and model reduction, IEEE Trans. Autom. Control AC–26 (1) (1981) 17–32. doi:10.1109/TAC.1981.1102568.

[6] A. C. Antoulas, Approximation of large-scale dynamical systems, Vol. 6, SIAM, 2005.

[7] K. Glover, All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error norms, Internat. J. Control 39 (6) (1984) 1115–1193. doi:10.1080/00207178408933239.

[8] D. F. Enns, Model reduction with balanced realizations: An error bound and a frequency weighted generalization, in: Proc. 23rd IEEE Conf. Decision Contr., Vol. 23, 1984, pp. 127–132. doi:10.1109/CDC.1984.272286.

[9] Y. Chahlaoui, A posteriori error bounds for discrete balanced truncation, Linear Algebra and Its Applications 436 (8) (2012) 2744–2763.

[10] P. Benner, M. Redmann, Model reduction for stochastic systems, Stochastics Partial Differential Equations: Analysis and Computations 3 (3) (2015) 291–338. doi:10.1007/s40072-015-0050-1.

[11] P. Benner, M. Redmann, An $\mathcal{H}_2$-type error bound for balancing- related model order reduction of linear systems with Lévy noise, Systems & Control Letters 105 (2017) 1–5. doi:10.1016/j.sysconle.2017.04.004.

[12] M. Redmann, M. A. Freitag, Balanced model order reduction for linear random dynamical systems driven by Lévy noise, Journal of Computational Dynamics 5 (1& 2) (2018) 33. doi:10.3934/jcd.2018002.
URL http://aimsciences.org//article/id/183c1785-bf47-4e2f-a4aa-52825f824da5

[13] W. Gawronski, J.-N. Juang, Model reduction in limited time and frequency intervals, International Journal of Systems Science 21 (2) (1990) 349–376.

[14] P. Kürschner, Balanced truncation model order reduction in limited time intervals for large systems, Advances in Computational Mathematics 44 (6) (2018) 1821–1844. doi:10.1007/s10444-018-9608-6.

[15] M. Redmann, P. Kürschner, An output error bound for time-limited balanced truncation, Syst. Control Lett. 121 (2018) 1–6.

[16] S. Gugercin, A. C. Antoulas, A survey of model reduction by balanced truncation and some new results, Internat. J. Control 77 (8) (2004) 748–766. doi:10.1080/00207170410001713448.

[17] K. S. Haider, A. Ghafoor, M. Imran, F. M. Malik, Model reduction of large scale descriptor systems using time limited gramians, Asian Journal of Control 19 (3) (2017) 1217–1227. doi:10.1002/asjc.1444.

[18] M. Redmann, P. Kürschner, An output error bound for time-limited balanced truncation, Syst. Control Lett. 121 (2018) 1–6.

[19] M. Imran, A. Ghafoor, U. Zulfiqar, V. Sreeram, Model reduction of discrete time systems using time limited gramians, in: 2018 Australian New Zealand Control Conference (ANZCC), 2018, pp. 22–26. doi:10.1109/ANZCC.2018.8606584.

[20] P. Benner, P. Kürschner, J. Saak, Frequency-limited balanced truncation with low-rank approximations, SIAM J. Sci. Comput. 38 (1) (2016) A471–A499. doi:10.1137/15M1030911.

[21] P. Kürschner, Efficient low-rank solution of large-scale matrix equations, Dissertation, Otto-von-Guericke-Universität, Magdeburg, Germany, Shaker Verlag, ISBN 978-3-8440-4385-3 (2016).
URL http://hdl.handle.net/11858/00-001M-0000-0029-CE18-2
[22] P. Goyal, M. Redmann, Towards time-limited $\mathcal{H}_2$-optimal model order reduction, Applied Mathematics and Computation 355 (2019) 184 – 197. doi:10.1016/j.amc.2019.02.065.
[23] K. Sinani, S. Gugercin, $\mathcal{H}_2(t_f)$ optimality conditions for a finite-time horizon, Automatica 110 (2019) 108604. doi:j.automatica.2019.108604.
[24] P. Vuillemin, C. Poussot-Vassal, D. Alazard, Poles residues descent algorithm for optimal frequency-limited $\mathcal{H}_2$ model approximation, in: Proc. European Control Conf., 2014, pp. 1080–1085.
[25] D. Petersson, J. Löfberg, Model reduction using a frequency-limited $\mathcal{H}_2$-cost, Syst. Control Lett. 67 (2014) 32–39.
[26] H. Dym, Linear algebra in action, Vol. 78, American Mathematical Soc., 2013.
[27] N. J. Higham, P. A. Knight, Matrix powers in finite precision arithmetic, SIAM journal on matrix analysis and applications 16 (2) (1995) 343–358.
[28] M. Crouzeix, C. Palencia, The numerical range is a $(1 + \sqrt{2})$-spectral set, SIAM Journal on Matrix Analysis and Applications 38 (2) (2017) 649–655.
[29] P. Benner, G. E. Khoury, M. Sadkane, On the Squared Smith Method for Large-Scale Stein Equations, Numer. Lin. Alg. Appl. 21 (5) (2014) 645–665.
[30] M. Sadkane, A low-rank Krylov squared Smith method for large-scale discrete-time Lyapunov equations, Linear Algebra and its Applications 436 (8) (2012) 2807–282. doi:10.1016/j.laa.2011.07.021.
[31] B. Beckermann, A. Townsend, On the Singular Values of Matrices with Displacement Structure, SIAM J. Matrix Anal. Appl. 38 (4) (2017) 1227–1248. doi:10.1137/16M1096426.
[32] R. A. Smith, Matrix equation $XA + BX = C$, SIAM J. Appl. Math. 16 (1) (1968) 198–201.
[33] T. Penzl, Eigenvalue decay bounds for solutions of Lyapunov equations: the symmetric case, Syst. Control Lett. 40 (2000) 139–144. doi:10.1016/S0167-6911(00)00010-4.
[34] Y. Saad, Numerical solution of large Lyapunov equation, in: M. A. Kaashoek, J. H. van Schuppen, A. C. M. Ran (Eds.), Signal Processing, Scattering, Operator Theory and Numerical Methods, Birkhäuser, 1990, pp. 503–511.
[35] I. M. Jaimoukha, E. M. Kasenally, Krylov subspace methods for solving large Lyapunov equations, SIAM J. Numer. Anal. 31 (1) (1994) 227–251. doi:10.1137/0731012.
[36] T. Li, P. C. Y. Weng, E. K. Chu, W. W. Lin, Large-scale Stein and Lyapunov equations, Smith method, and applications, Numer. AlgorithmsDOI: 10.1007/s11075-012-9650-2.
[37] N. J. Higham, Functions of Matrices: Theory and Computation, Applied Mathematics, SIAM Publications, Philadelphia, PA, 2008. doi:10.1137/1.9780898717778.
[38] V. Druskin, V. Simoncini, Adaptive rational Krylov subspaces for large-scale dynamical systems, Syst. Control Lett. 60 (8) (2011) 546–560. doi:10.1016/j.sysconle.2011.04.013.
[39] V. Druskin, L. Knizhnerman, V. Simoncini, Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation, SIAM J. Numer. Anal. 49 (5) (2011) 1875–1898. doi:10.1137/100813257.
[40] B. Beckermann, An Error Analysis for Rational Galerkin Projection Applied to the Sylvester Equation, SIAM J. Numer. Anal. 49 (6) (2011) 2430–2450. doi:10.1137/110824590.
[41] P. Benner, P. Kürschner, Computing real low-rank solutions of Sylvester equations by the factored ADI method, Comput. Math. Appl. 67 (9) (2014) 1656–1672. doi:10.1016/j.camwa.2014.03.004.
[42] A. Y. Barraud, A numerical algorithm to solve $A^T XA - X = Q$, IEEE Trans. Autom. Control 22 (1977) 883–885.
[43] A. Bouhamidi, M. Heyouni, K. Jbilou, Block Arnoldi-based methods for large scale discrete-time algebraic Riccati equations, J. Comput. Appl. Math. 236 (6) (2011) 1531–1542.
[44] D. Kressner, P. Kürschner, S. Massei, Low-rank updates and divide-and-conquer methods for quadratic matrix equations, Numer. Alg. 84 (2020) 717741. doi:10.1007/s11075-019-00776-w.
[45] V. Simoncini, A new iterative method for solving large-scale Lyapunov matrix equations, SIAM J. Sci. Comput. 29 (3) (2007) 1268–1288. doi:10.1137/06066120X.
[46] A. Ruhe, The Rational Krylov algorithm for nonsymmetric eigenvalue problems. III: Complex shifts for real matrices, BIT Numerical Mathematics 34 (1) (1994) 165–176. doi:10.1007/BF01935024.
[47] P. Benner, M. Köhler, J. Saak, Sparse-dense Sylvester equations in H_2-model order reduction, Preprint MPIMD/11-11, Max Planck Institute Magdeburg (Dec. 2011).
[48] M. E. Hochstenbach, Fields of values and inclusion regions for matrix pencils, Electron. Trans. Numer. Anal. 38 (2011) 98–112.
[49] N. Guglielmi, M. Gürbüzbalaban, M. L. Overton, Fast approximation of the H_∞ norm via optimization over spectral value sets, SIAM J. Matrix Anal. Appl. 34 (2) (2013) 709–737.
[50] G. H. Golub, C. F. Van Loan, Matrix Computations, 4th Edition, Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, 2013.