Modified balanced truncation preserving ellipsoidal cone-invariance

Christian Grussler and Anders Rantzer

Abstract—We consider model order reduction of stable linear systems which leave ellipsoidal cones invariant. We show how balanced truncation can be modified to preserve cone-invariance. Additionally, this implies a method to perform external positivity preserving model reduction for a large class of systems.

I. INTRODUCTION

Cone-invariance of linear time-invariant systems is a common feature, which is appearing nowadays very frequently in the literature. This is due to an increased interest in systems with compartmental structure as they can be found in bio-medicine, economics, data networks and many more application areas (cf. [6], [8], [15], [20]). For example, consider the linear time-invariant system

\[
G: \begin{cases}
    x(t) = Ax(t) + Bu(t), \\
    y(t) = Cx(t) + Du(t),
\end{cases}
\]

with state vector \( x \in \mathbb{R}^n \), input \( u \in \mathbb{R}^m \) and output \( y \in \mathbb{R}^k \).

Here \( x \) could stand for the temperature in \( n \) rooms within a building, influenced by the temperature of \( m \) radiators \( u \). The temperature in \( k \) sensor locations, e.g. floors, is then represented by \( y \). Consequently, \( x \) and \( y \) are confined to be non-negative, whenever \( u \) is non-negative. In literature, such systems are referred to as being internally positive and as being externally positive if the system is positive from input to output (see Section II).

Naturally, these systems often tend to be of large dimension \( n \) and need to be approximated with the help of model order reduction. Unfortunately, conventional model reduction methods (see e.g. [2], [10], [12], [13], [16]) do not preserve external positivity.

However, working with an approximation that is violating basic physical constraints by allowing for instance negative concentrations of chemical substance always leaves the question of how conclusive results on this basis are. Recently developed methods have tackled this problem by preserving internal positivity (cf. [9], [11], [14], [19], [21]), i.e. the invariance with respect to (w.r.t.) the non-negative orthant.

The main goal of this work is to present a variant of balanced truncation, which guarantees to preserve invariance w.r.t. an ellipsoidal cone (see Section II and III). An immediate consequence of this result is the preservation of external positivity under the assumption of ellipsoidal cone-invariance (see Section IV).

Unlike internal positivity, our definition has the advantage of being computationally tractable and independent of a particular state-space realization (see Sections II and IV). In Section VI we will see that ellipsoidal cone-invariance is often implied by internal positivity. Moreover, numerical experiments indicate that the error-difference between balanced truncation and our method appears to be fairly small (see Section VI).

II. PRELIMINARIES

The following notation for real matrices and vectors \( X = (x_{ij}) \) are used throughout this paper. We say that \( X \in \mathbb{R}^{m \times n}_{\geq 0} \) is non-negative, if all entries are non-negative \((x_{ij} \geq 0 \text{ for all } i, j)\). By \( |X| = (|x_{ij}|) \) we denote the entry-wise absolute value of \( X \) and by \( x_i \) its \( i \)-th column, if not further specified.

If \( X = X^T \), then we write \( X \succeq 0 \), or \( X \geq 0 \) if \( X \) is positive definite, or semi-definite, i.e. the set of eigenvalues of \( X \), \( \sigma(X) \subset [0, \infty] \). We also use these notations to describe the relation between two matrices, e.g. \( A \succeq B \) defines \( A - B \geq 0 \).

A real vector valued function \( u(t) \in \mathbb{R}^m \) is called non-negative if and only if \( u(t) \in \mathbb{R}^{m \times 0}_{\geq 0} \) for all \( t \geq 0 \).

The inertia \((p, z, n)\) of \( X \) is defined by the number of eigenvalues of \( X \) with positive, zero and negative real-parts, respectively counting multiplicities.

Now, let us define cone-invariance.

**Definition 1 (Invariant cone):** Let \( \mathcal{K} \subset \mathbb{R}^n \) be a cone and \( A \in \mathbb{R}^{n \times n} \). \( \mathcal{K} \) is called \( A \)-invariant if and only if \( A \mathcal{K} \subset \mathcal{K} \). \( \mathcal{K} \) is called exponentially \( A \)-invariant if and only if \( \forall t \geq 0 : e^{At} \mathcal{K} \subset \mathcal{K} \).

**Definition 2 (Cone invariance):** \((A, B)\) is called cone-invariant w.r.t. a cone \( \mathcal{K} \) if and only if \( b_i \in \mathcal{K} \), for all \( i \) and \( \mathcal{K} \) is exponentially invariant w.r.t. \( A \).

Similar to the introductory example, cone-invariance says: if the state-vector starts within a cone \( \mathcal{K} \) then it will remain there for all non-negative inputs \( u \). Two important classes of cone-invariant systems are the so-called externally and internally positive systems, which will be discussed in Section IV.

In the following we define ellipsoidal cones, the essential ingredient for our main result. This class has been investigated in [22], [23], which is why we adapt the notations.
Definition 3 (Ellipsoidal cones): Let $Q = Q^T \in \mathbb{R}^{n \times n}$ with inertia $(n-1,0,1)$, then
\[
\mathcal{K}_Q := \{ x : x^T Q x \leq 0 \}
\]
is called an ellipsoidal double-cone. If $p \in \mathbb{R}^n$ is such that
\[
\{ p \}^\perp \cap \mathcal{K}_Q = \{ 0 \}
\]
where $\{ p \}^\perp$ denotes the orthogonal complement of linear span $\{ p \}$ of $p$, then we call $\mathcal{K}_{Q,p} := \{ x : x^T Q x \leq 0, p^T x \geq 0 \}$ an ellipsoidal cone.

It is obvious that $\mathcal{K}_{Q,p}$ and $-\mathcal{K}_{Q,p} := \mathcal{K}_{Q,-p}$ are proper convex cones. In the following, we make the convention that $Q_n := \text{blkdiag}(I_{n-1},-1)$ and $\mathcal{K}_{Q_n, e_n}$ is called the ice-cream cone, where $e_n$ is the $n$-th canonical unit vector.

Lemma 1: Let $\mathcal{K}_Q$ be an ellipsoidal double-cone and $\mathcal{K}_{Q,u_n}$ the corresponding ellipsoidal cone, where $u_n$ is an eigenvector belonging to the negative eigenvalue of $Q$. Their dual sets can be parametrized as $\mathcal{K}^*_Q = \mathcal{K}^{-1}_Q$, $\mathcal{K}^*_{Q,u_n} = \mathcal{K}^{-1}_{Q,u_n}$.

Proof: Let $\mathcal{K}_{Q,u_n}$ be an ellipsoidal cone with $\sigma(Q) = \{ \lambda_1, \ldots, \lambda_n \}$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0 > \lambda_n$. Diagonalizing $Q$ with $U^T Q U = \text{diag}(\lambda_1, \ldots, \lambda_n) = \Delta$ defines a transformation matrix $T = [\Delta^{1/2}] U^T$, which gives $T^T \mathcal{K}_{Q,u_n} = \mathcal{K}_{T^T Q T, T^T u_n} = \mathcal{K}_{Q_{e_n}, e_n}$. Since $\mathcal{K}_{Q_{e_n}, e_n}$ is the $\mathcal{K}_{Q,u_n}$ it follows that $T^T \mathcal{K}_{Q,u_n} = \mathcal{K}_{Q_{e_n}, e_n}$ and therefore $\mathcal{K}^*_{Q,u_n} = T^T \mathcal{K}^*_{Q,u_n} = \mathcal{K}^{-1}_{Q,u_n}$. If $\mathcal{K}^*_{Q,u_n}$ is independent of $u_n$ we conclude the proof.

The boundary of $\mathcal{K}_Q$ is given by
\[
\partial \mathcal{K}_Q = \{ x : x^T Q x = 0, p^T x > 0 \} \cup \{ 0 \}.
\]
From this it follows that $p \in \text{int}(\mathcal{K}^*_Q)$, which by the preceding Lemma is equivalent with $p^T Q^{-1} p < 0$. Thus, for given $Q = Q^T$ with inertia $(n-1,0,1)$ we conclude that $\mathcal{K}_{Q,p}$ is a proper convex cone if and only if $p^T Q^{-1} p < 0$.

Moreover, $p \in \text{int}(\mathcal{K}^*_Q)$ if and only if $\exists \tau > 0 : \forall x \in \mathcal{K}_{Q,p} : x^T Q x + \tau p^T x > 0 \Leftrightarrow Q + \tau p p^T > 0$. Together with the main result in [22], this leads to the following theorem.

Theorem 1: Let $Q = Q^T$ with inertia $(n-1,0,1)$. Then $\mathcal{K}_{Q,p} := \{ x : x^T Q x \leq 0, p^T x \geq 0 \}$ is exponentially invariant w.r.t. $A$ if and only if
\[
\exists \tau, \tau \in \mathbb{R} : A^T Q + QA + 2\gamma Q \leq 0, \quad Q + \tau p p^T \succ 0.
\]
Notice, if $\sigma(A + \gamma I) \cap \mathbb{R} = \{ 0 \}$, then $Q$ exists if and only if the inertia of $A + \gamma I$ and $-Q$ are equal (see e.g. [5]). Equivalently, $A$ has a single dominant real eigenvalue $\lambda_{\text{max}} \in \sigma(A)$ and $\sigma(A + \gamma I) \cap \mathbb{R} \geq 0 = \{ \lambda_{\text{max}} + \gamma \}$.

III. CENTRAL THEORY

In the following we consider asymptotically stable systems as in (1), where $(A,B)$ is invariant w.r.t. to an ellipsoidal double-cone. We assume the reader to be familiar with the concept of standard balanced truncation (BT) (see e.g. [2], [7], [16]).

In general, balanced truncation does not preserve the invariance with respect to an ellipsoidal cone – unless the system is reduced to order $r = 1$. To this end, we will modify the concept of balanced truncation to what we call cone-balanced truncation. For notational simplicity we start by deriving the main results for the case of a controllable system. Nonetheless, the reader should check that the results are still true in the uncontrollable case.

Let us start with the first of two following modifications of balancing a system.

Proposition 1: Given $(A,B)$ and $\gamma > 0$, let $Q = Q^T$ with inertia $(n-1,0,1)$ and $P > 0$ fulfill
\begin{enumerate}
  \item $A^T Q + QA + 2\gamma Q \leq 0$,
  \item $b_j Q b_j < 0$ for all $j$,
  \item $A^T P + P A = -B B^T$.
\end{enumerate}

Then there exists $T \in \mathbb{R}^{n \times n}$ such that
\[
T^{-1} P T^{-T} = \text{blkdiag}(\sigma_1, \sigma_2 I_{k_2}, \ldots, \sigma_s I_{k_s}),
\]
\[
T^T Q T = \text{blkdiag}(-\sigma_1, \sigma_2 I_{k_2}, \ldots, \sigma_s I_{k_s})
\]
where $\sigma_1 > \cdots > \sigma_s > 0$, $k_2 + \cdots + k_s = n-1$ and
\[
\sigma_1 \geq \sqrt{\frac{\sum_{i=1}^n \sigma_i^2}{\sqrt{n}}}
\]

Proof: Assume that $P$ and $Q$ are as in the claim and fulfill i. - iii. We decompose $P$ into $P = U \Sigma_p U^T$ and define $L := U \Sigma_p^2$. By another singular value decomposition of $L^T Q L$ into $L^T Q L = \Sigma^2 V^T V$ we define $T := L \Sigma^{-1} \Sigma^2$. Then we can verify that $\tilde{P} := T^{-1} P T^{-T}$ and $\tilde{Q} := T^T Q T$ fulfill
\[
\tilde{P} = \Sigma^2 V^T L^T L^T V \Sigma^2 = \Sigma,
\]
\[
[\tilde{Q}] = [\Sigma^{-1/2} V^T L^T Q L \Sigma^{-1/2}] = \Sigma,
\]
with $\Sigma = \text{blkdiag}(\sigma_1 I_{k_1}, \ldots, \sigma_s I_{k_s})$, $\sigma_1 > \cdots > \sigma_s > 0$ and $k_1 + \cdots + k_s = n$. By Sylvester’s law of inertia it follows that the inertia of $T^T Q T$ remains invariant, which is why $\tilde{P}$ and $\tilde{Q}$ are equal up to a sign-change on one of the diagonal entries.

We will now show that $\text{tr}(\tilde{Q}) < 0$ implies that the sign-change occurs at $\sigma_1$ and $k_1 = 1$. To this end, assume without loss of generality (w.l.o.g.) that $P = I$ and $|Q| = \Sigma^2$, i.e.
\begin{align*}
A^T Q + QA + 2\gamma Q &\leq 0, \\
b_j^T Q b_j &< 0, \text{ for all } j \\
A + A^T &= -B B^T.
\end{align*}

By substitution of $A = -B B^T - A^T$ in (3) we get
\[
-(B B^T + A) Q - Q(B B^T + A^T) - 2\gamma Q \leq -4\gamma Q.
\]

Taking the trace $\text{tr}(\cdot)$ over (6) and using
\begin{itemize}
  \item $\sum_{j=0}^d b_j^T Q b_j = \text{tr}(B B^T Q) = \text{tr}(Q B B^T) < 0$
  \item $\text{tr}(A^T Q + QA) = \text{tr}(A^T Q + QA)$
\end{itemize}
gives the following inequality
\[
-4\gamma \text{tr}(Q) \geq -2 \text{tr}(B B^T Q) > 0 \Leftrightarrow \text{tr}(Q) \leq \frac{1}{2\gamma} \left( \text{tr}(B B^T Q) \right) < 0.
\]

Therefore, by the inertia of $Q$ and the assumption that $\sigma_1 > \cdots > \sigma_s > 0$ we conclude that the largest magnitude in $Q$ is negative.

If $(A,B,C,D)$ is a system with $(A,B)$, $P$, $Q$, $\gamma$ and $T$ as in Proposition 1, then truncating any of the last $n-1$ states
of \((\bar{A}, \bar{B}, \bar{C}, \bar{D}) := (T^{-1}A, T^{-1}B, CT, D)\) preserves controllability as well as ellipsoidal cone-invariance. But, since \(T^TQ T\) is indefinite we cannot apply the error-bound known from balanced truncation. Instead, we perform another balancing of \((\bar{A}, \bar{B}, \bar{C}, \bar{D})\) which will provide us with such.

**Proposition 2:** Let \((\bar{A}, \bar{B}, \bar{C}, \bar{D})\) be such that \((\bar{A}, \bar{B})\) is invariant w.r.t. \(\mathcal{X}_Q\) and \(\bar{A}\bar{P} + \bar{P}\bar{A} = -BB^T\) for diagonal \(\bar{P} > 0\) with \(\bar{P} = |\bar{Q}|\). Then \(\exists \Delta > 0 : A^T\Delta + \Delta A \preceq -C^T\bar{C}\) with \(\Delta\) being diagonal.

**Proof:** If \((\bar{A}, \bar{B}, \bar{C}, \bar{D})\) is as in the assumptions, then by Lemma 1 and Theorem 1 we conclude that
\[
\bar{A}^T\bar{Q} + \bar{Q}\bar{A} + 2\gamma\bar{Q} \preceq 0,
\]
and multiplying (8) by \(2\gamma\epsilon\) gives
\[
2\gamma\epsilon\bar{A}\bar{P} + 2\gamma\epsilon\bar{P}\bar{A} + 2\gamma\epsilon BB^T = 0,
\]
where \(2\gamma\epsilon \sigma_1 - \sigma_1 > 0\). Adding up (10) and (11) results in
\[
\bar{A}\Delta^{-1} + \Delta^{-1}\bar{A}^T + 2\gamma(\bar{Q}^{-1} + \epsilon BB^T) \preceq 0
\]
with \(\Delta := (2\gamma\epsilon\bar{P} + \bar{Q}^{-1})^{-1} > 0\). Finally, a proper scaling of \(\Delta\) gives a diagonal solution to
\[
\bar{A}^T\Delta + \Delta A \preceq -\bar{C}^T\bar{C}.
\]

Now, if \(\bar{P}\) and \(\Delta\) are as in Proposition 2 we can define a second balancing transformation \(\bar{T} := \text{blkdiag}(1, \bar{P}_{22}, \ldots, \bar{P}_{nn})\) such that \((\bar{A}, \bar{B}, \bar{C}, \bar{D}) := (T^{-1}AT, T^{-1}B, CT, D)\) fulfills \(\bar{A}\bar{P} + \bar{P}\bar{A} = -BB^T\) and \(\bar{A}^T\bar{Q} + \bar{Q}\bar{A} \preceq -\bar{C}^T\bar{C}\), where \(\bar{P}\) and \(\bar{Q}\) are diagonal and equal except for the first diagonal entry.

**Definition 4 (Cone-balanced):** A \((\bar{A}, \bar{B}, \bar{C}, \bar{D})\) is called cone-balanced if \(\exists \gamma > 0, \bar{P}, \bar{Q} > 0\) and \(\bar{K} = \bar{T}^T\) with inertia \((n - 1, 0, 1)\) such that
\[
\bar{A}^T\bar{K} + \bar{K}\bar{A} + 2\gamma\bar{K} \preceq 0,
\]
\[
\bar{A}^T\bar{Q} + \bar{Q}\bar{A} \preceq -\bar{C}^T\bar{C},
\]
\[
\bar{A}^T\bar{P} + \bar{P}\bar{A} = -BB^T,
\]
where \(\bar{P}\) and \(\bar{Q}\) are diagonal with \(k_{11} < 0\) and \(\bar{P}_{22} = \bar{q}_{22} \geq \cdots \geq \bar{P}_{nn} = \bar{q}_{nn}\).

Again, truncating a cone-balanced system preserves ellipsoidal cone-invariance and it is well known (see e.g. [7]) that the error-bound result from standard balanced truncation carries over to the diagonal elements of \(\bar{P}\).

**Theorem 2:** Suppose \((\bar{A}, \bar{B}, \bar{C}, \bar{D})\) is the cone-balanced realization of a stable, minimal cone-invariant system with transfer function \(\bar{G}(s)\) and controllability Gramian \(\bar{P} = \text{blkdiag}(\Sigma_1, \Sigma_2)\),
\[
\bar{\Sigma}_1 = \text{blkdiag}(\bar{\sigma}_1, \bar{l}_k, \ldots, \bar{\sigma}_l l_k),
\]
\[
\bar{\Sigma}_2 = \text{blkdiag}(\bar{\sigma}_{r+1} l_{r+1}, \ldots, \bar{\sigma}_p l_p),
\]
where \(\bar{\sigma}_1 > \cdots > \bar{\sigma}_r > \bar{\sigma}_{r+1} > \cdots > \bar{\sigma}_p > 0\).

Truncating the states corresponding to \(\Sigma_2\) results in an approximation \((\bar{A}_r, \bar{B}_r, \bar{C}_r, \bar{D}_r)\) of order \(1 + \sum_{i=2}^r k_i\) with transfer function \(G_r(s)\), which is cone-balanced, controllable and stable. Moreover, it holds for the \(H_\infty\)-error
\[
\|\tilde{G}(s) - G_r(s)\|_\infty \leq 2 \sum_{i=r+1}^p \bar{\sigma}_i.
\]

It is known (cf. [11]) that the \(\bar{\sigma}_i\) in (13) are always larger than the Hankel singular values. Nevertheless, we will see in Section VI that we can get fairly close to them. The whole algorithm for cone-balanced truncation (CBT) is summarized in Algorithm 1.

**Algorithm 1 Cone balanced truncation (CBT)**

1. Let \((A, B, C, D)\) be a minimal system.
2. If \((A, B)\) fulfills Proposition 1.
3. Find \(T \in \mathbb{R}^{n \times n}\) such that \((\bar{A}, \bar{B}, \bar{C}, \bar{D}) := (T^{-1}AT, T^{-1}B, CT, D)\) has diagonal controllability Gramian \(\bar{P}\) and \((\bar{A}, \bar{B})\) is invariant w.r.t. \(\mathcal{X}_Q\) with \(\bar{P} = |\bar{Q}|\).
4. Minimize \(\sum_{i=1}^m \delta_i\) subject to
\[
A^T\Delta + \Delta A \preceq -C^T\bar{C},
\]
where \(\Delta := (2\gamma\epsilon\bar{P} + \bar{Q}^{-1})^{-1} > 0\).
5. Find a cone-balanced realization \((\bar{A}, \bar{B}, \bar{C}, \bar{D})\) with generalized singular values \(\bar{\sigma}_i := \sqrt{\bar{P}_{ii}\bar{Q}_{ii}}, i > 1\).
6. Choose a reduced order according to (13) and truncate \((\bar{A}, \bar{B}, \bar{C}, \bar{D})\).
7. END

**IV. POSITIVE SYSTEMS**

In the following we formally define externally and internally positive systems and compare them with ellipsoidal cone-invariant systems. After that, it will be evident why our result naturally extends to the class of externally positive systems.

**Definition 5 (External Positivity):** A linear system (1) is called externally positive if and only if its output corresponding to a zero initial state is non-negative for every non-negative input.

**Theorem 3:** A linear system \((A, B, C, D)\) is externally positive if and only if \(\forall t \geq 0 : Ce^{At}B \in \mathbb{R}^{k_x \times m} D \in \mathbb{R}^{k_y \times m}\). [8] It is readily seen that every single-input-single-output (SISO) externally positive system \((A, B, C)\) is invariant with respect to its so-called reachable and observable cone, \(R(A,B) := \{e^{At}B : t \geq 0\}\) and \(O(A,C) := \{x \forall t \geq 0 : Ce^{At}x \geq 0\}\), see e.g. [17].

**Definition 6 (Internal Positivity):** A linear system (1) is called internally positive if and only if its state and output are non-negative for every non-negative input and every non-negative initial state.

Internal positivity of (1) requires that the non-negative or-thant \(\mathbb{R}^n_{\geq 0}\) is exponentially invariant w.r.t. \(A\). In [3] it is
shown, that this is the case if and only if $A$ is Metzler, i.e. $\exists \alpha \geq 0 : A + \alpha I \in \mathbb{R}^{n \times n}_{\geq 0}$.

**Theorem 4:** A continuous linear system $(A, B, C, D)$ is internally positive if and only if $A$ is Metzler, i.e. $\exists \alpha \geq 0 : A + \alpha I \in \mathbb{R}^{n \times n}_{\geq 0}$. Theorem 4: A continuous linear system $(A, B, C, D)$ is internally positive if and only if $A$ is Metzler, i.e. $\exists \alpha \geq 0 : A + \alpha I \in \mathbb{R}^{n \times n}_{\geq 0}$. Thus our method also preserves internally positive realizability. Unfortunately, shown, that this is the case if and only if $A$ is Metzler, i.e. $\exists \alpha \geq 0 : A + \alpha I \in \mathbb{R}^{n \times n}_{\geq 0}$.

**Theorem 4:** A continuous linear system $(A, B, C, D)$ is internally positive if and only if $A$ is Metzler, i.e. $\exists \alpha \geq 0 : A + \alpha I \in \mathbb{R}^{n \times n}_{\geq 0}$. Theorem 4: A continuous linear system $(A, B, C, D)$ is internally positive if and only if $A$ is Metzler, i.e. $\exists \alpha \geq 0 : A + \alpha I \in \mathbb{R}^{n \times n}_{\geq 0}$. Thus our method also preserves internally positive realizability. Unfortunately, shown, that this is the case if and only if $A$ is Metzler, i.e. $\exists \alpha \geq 0 : A + \alpha I \in \mathbb{R}^{n \times n}_{\geq 0}$.

Moreover, we assume that we reduce our system by one order discrete-time linear systems – for ellipsoidal cones see e.g. [41].

positive and cone-balanced w.r.t.

balanced truncation preserves external positivity.

is beneficial to look at externally positive systems, which is known to be computationally difficult (cf. [1], [17]). We restrict ourselves to internally positive systems is a convenient way to deal with this problem. But, as indicated in Theorem 4, internal positivity depends on very specific state-space realizations. Finding such a realization is known to be computationally difficult (cf. [1], [17]). We believe, if one is only interested in external positivity, it is beneficial to look at externally positive systems, which are ellipsoidal cone-invariant. In this case, verification of external positivity can be performed with the help of convex optimization.

**Theorem 5:** Given $(A, B, C, D)$ with $D \in \mathbb{R}^{l \times m}_{\geq 0}$, assume there exists $Q = Q^T$ with inertia $(n, 1, 0, 1)$ and $\gamma, \tau \in \mathbb{R}$ such that

\begin{align*}
1) & \ A^T Q + QA + 2\gamma Q \preceq 0, \\
2) & \ b_j^T Q b_j < 0 \text{ for all } j \\
3) & \ Q + \tau c_i^T c_i > 0 \text{ for all } i \\
4) & \ CB \in \mathbb{R}^{l \times m}_{\geq 0}
\end{align*}

where $c_i$ is the $i$-th row of $C$. Then $(A, B, C, D)$ is externally positive.

**Proof:** The result follows directly from Lemma 1 and Theorem 1, which imply that $(A, B, C, D)$ is invariant w.r.t. $\mathcal{K}_{Q, c_i^T}$, for all $i$.

**Corollary 1:** Assume $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ is a stable and externally positive system fulfilling Theorem 5. Then cone-balanced truncation preserves external positivity.

**Proof:** Assume w.l.o.g. that $(\bar{A}, \bar{B}, \bar{C}, \bar{D})$ is externally positive and cone-balanced w.r.t. $\mathcal{K}_{\bar{Q}, \bar{c}_i} = \mathcal{K}_{Q, e_i}$ for all $i$. Moreover, we assume that we reduce our system by one order with CBT to $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$. Clearly, 1) – 3) in Theorem 5 are preserved after cone-balanced truncation. To see 4), observe that $b_j \in \mathcal{K}_{Q, e_i} \Rightarrow (b_1, j, \ldots, b_{n-1}, 0)^T \in \mathcal{K}_{\bar{Q}, \bar{c}_i} \Rightarrow \bar{c}_j b_j \geq 0$.

We will refer to this method as positive cone-balanced truncation (PCBT).

V. DISCUSSION

In the previous two sections we have formally derived a solution to the following problems

I. Ellipsoidal cone preserving model reduction.

II. Externally positivity preserving model reduction under the constraint of ellipsoidal cone-invariance.

Now, we want to get some further insights into these problems and into the numerical computations involved.

First notice, it is straightforward to extend all results to discrete-time linear systems – for ellipsoidal cones see e.g. [23]. Furthermore, Lemma 1, Proposition 1 and the duality between observability and controllability imply that if a non-minimal system is ellipsoidal cone-invariant then this is also true for its minimal realization.

A major draw-back of our method is the need to solve linear matrix inequalities (LMIs) in order to preserve external positivity – LMI-solvers are usually computational demanding (see e.g. [18]). Given the NP-hardness of the verification problem, this is a small price to pay. Moreover, standard balanced truncation usually requires pre-reduction methods such as [13] to be able to handle large-scale systems. Hence, it is valid to assume a reduced system whose LMIs are sufficiently fast solvable.

Further observe, if one only wants to verify/preserve cone-invariance, it is often enough to consider Lyapunov-equations. To see this, assume $(A, B)$ is controllable with $\sigma(A + \gamma I) \cap i\mathbb{R} = \emptyset$ and no $b_j$ is in the span of the eigenvectors belonging to the non-dominant eigenvalues. If $\mathcal{K}_Q$ is exponentially $A$-invariant then so is $e^{-A t}K_Q$, $t \geq 0$, which by our assumptions implies that $\exists \gamma \geq 0 : b_j \in e^{-A t}K_Q$, for all $j$.

More explicitly, if $B \in \mathbb{R}^n$ one could solve

$$
A^T Q + QA + 2\gamma Q = -R \preceq 0,
$$

$$
AP + PA^T + 2\gamma P = -BB^T.
$$

Since $\frac{\partial (BB^T)}{\partial x_{ij}} = \frac{\partial (BB^T)}{\partial x_{ji}} = \frac{\partial (PR)}{\partial x_{ij}} = p_{ij}$ it follows, if w.l.o.g. $P$ is diagonal with $p_{11} < 0$, that for any $K > 0$ with $k_{11} > 0$ and $k_{ii} > 0$, $i > 1$ the resulting $Q$ fulfills $B^TQB < 0$.

**Corollary 2:** Assume $(A, B, C, D)$ is a minimal symmetric SISO-system i.e. $A = A^T$ and $B = C^T$ and assume that $A$ has single dominant eigenvalue. Then there exists $Q = Q^T$ fulfilling Theorem 5.

**Proof:** By the previous discussion it follows that there exists $Q = Q^T$ and $u_n$, such that $\mathcal{K}_{Q, u_n}$ is exponentially invariant w.r.t. $A$ and $B \in \mathcal{K}_{Q, u_n}$. Hence, $C^T \in \mathcal{K}_{Q, u_n}$ and therefore $\forall t \geq 0 : C^T e^{A t}C^T B = \|e^{A t}B\|^2 > 0$. W.l.o.g. we can assume that $\mathcal{K}_{Q, u_n} = \mathcal{K}_{Q, e_n}$ and $B^T Q e_n B = CQ e_n C^T < 0$. That implies that $C^T \in \text{int}(\mathcal{K}_{Q, e_n})$ which is why $\mathcal{K}_{Q, e_n} = \mathcal{K}_{Q, C^T}$.

It is straightforward to show that all symmetric SISO-systems have an internally positive realization of the same dimension. A practical procedure to deal with large-scale externally positive systems could be the following:

1) Reduce the system with help of a Krylov-subspace method (cf. [2], [13]) to an order where Lyapunov-equations can be solved efficiently.

2) Apply CBT to reduce the system to an order where the LMIs in Theorem 5 can be solved efficiently.

3) Use PCBT to verify external positivity and to reduce the system even further.

In our derivations we deal with externally positive systems $(A, B, C)$ where $CB \in \mathbb{R}^{l \times m}_{\geq 0}$. To treat with zero entries in $CB$ one could pre-approximate the original system with $(A, Be^{A\varepsilon}C)$ for $\varepsilon > 0$. Then $\forall \varepsilon > 0 : Ce^{A\varepsilon}B > 0$ by the assumption of a single dominant real pole and the error between those systems can be made arbitrarily small by the choice of $\varepsilon$.

Finally, notice that if an externally positive system has a strictly dominant real pole of multiplicity 1 then the system possesses a positive realization [1]. Thus our method also preserves internally positive realizability. Unfortunately,
internal positivity is not sufficient to ensure the requirements of Theorem 5.

VI. EXAMPLES & COMPARISON

By considering some numerical examples, we discuss the quality of (positive) cone-balanced truncation. The results are compared to symmetric balanced truncation (SBT) in [11] and standard balanced truncation (BT). Moreover, by the comparisons in [11] and [21] it follows, that even a reduced model of order 1 often outperforms the methods in [9], [14], [19], [21].

Our comparison will always start from a minimal realization, which can be considered a pre-reduction. In order to make the solutions unique, we will add to minimize \( \text{tr}(Q + \tau C^TC) \) in case of PCBT, which turned out to give good results. For CBT we use the same \( \tau \)-shift as determined by PCBT and \( Q \) is given by \( A^TQ + QA + 2\gamma Q = -C^TC \).

A. Heat Equation

We begin with one of the examples given in [11], the 2-dimensional heat equation on a square

\[
\dot{T} = \Delta T = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2}
\]

with control of the Dirichlet boundary conditions of the four edges. Discretisation on a uniform grid leads to the following linear internally positive system:

\[
\dot{T} = AT + Bu \quad \text{with } u \in \mathbb{R}^4 \text{ and } T \in \mathbb{R}^{N^2}
\]

where \( A \) is the \( N^2 \times N^2 \)-Poisson-matrix and \( B := [b_{ij}] \in \mathbb{R}^{N^2 \times 4} \), where \( b_{ij} = 0 \) except for the following cases:

- \( b_{1i} = 1 \), for \( i = 1, 2, \ldots, N \)
- \( b_{2i} = 1 \), for \( i = N, 2N, \ldots, N^2 \)
- \( b_{3i} = 1 \), for \( i = N(N - 1) + 1, N(N - 1) + 2, \ldots, N^2 \)
- \( b_{4i} = 1 \), for \( i = 1, N + 1, \ldots, N(N - 1) + 1 \)

One may think of this example in the same way as in the one given in the introduction. In our first experiment the output is equal to the global average temperature, i.e.

\[
y = \frac{1}{N^2} CT, \quad C := 1^T_{N^2}
\]

where \( 1^T_{N^2} = (1 \ldots 1) \in \mathbb{R}^{1 \times N^2} \).

In this case it was shown that SBT performs very well, because the minimal balanced realization is a symmetric system. Then by Corollary 2 it must be possible to apply PCBT. Repeating this experiment for (P)CBT with \( N = 10 \) gives the \( H_\infty \)-error as shown in Fig. 1. We observe, (P)CBT performs closely to (S)BT, the error-difference is due to numerical issues and a different sorting of the singular values. Moreover, PCBT does not suffer from disregarding the advantage of symmetry, as exploited by SBT. In fact, (P)CBT preserves it and an internally positive realization can be found here as well. Also the error-bounds for (P)CBT lie within a good range as indicated in Fig. 2, where the singular values of \( P(\text{CBT}) \) result from Theorem 2.

Now, we modify this example by using the second and the fourth input only. Furthermore, we split the unit-square into 5 equally spaced vertical stripes and let \( y \) represent the average temperature in each of these zones, i.e.

\[
C = \text{blkdiag}\left(1^T_{N^2}, 1^T_{N^2}, 1^T_{N^2}, 1^T_{N^2}, 1^T_{N^2}\right)
\]

In this case, the minimal balanced system is no longer symmetric. Therefore, SBT will arrive with an approximation of order 1 and the same error as BT. Again, the normalized errors are shown in Fig. 3.

B. Balanced truncation destroying positivity

It is readily verified that

\[
G(s) = \frac{(s + 1)^{10}}{(s + 1) \prod_{k=2}^{9}(s + 2 - e^{\pm \sqrt{k} \pi}) \prod_{k=4}^{9}(s + 2 - \frac{1}{k})}
\]
defines an externally positive systems, which has an ellipsoidal cone-invariant realization. BT does not preserve these properties for the reduced models of order 2 and 4. A comparison of the normalized errors is presented in Fig. 4. Although, both methods perform well, observe the comparably large error of PCBT for order 2. Interestingly, other well established model reduction methods, such as [10] and [13] also destroy positivity for a better error performance.

VII. CONCLUSION

We have not only presented a model reduction method which guarantees to preserve ellipsoidal cone-invariance but also defined a class of systems, which gives a broad intersection of some well studied cone-invariant systems. In fact, it seems that ellipsoidal cone-invariance is often implied by internal positivity. By that a numerical test for external/internal positivity has been established as well as a method for external positivity preserving model order reduction.

REFERENCES