TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 04/16 pp151-168
Volume 15, Number 2, April 2010

Recent Advance in Non-Krylov Subspace Model Order Reduction of Interconnect Circuits*

Sheldon X.-D. Tan**, Boyuan Yan, Hai Wang

Department of Electrical Engineering, University of California, Riverside, CA 92521, USA

Abstract: Model order reduction of interconnect circuits is an important technique to reduce the circuit complexity and improve the efficiency of post-layout verification process in the nanometer VLSI design. Existing works using the Krylov subspace method are very efficient, but the resulting models are less compact and lack global accuracy. Also, existing methods cannot handle interconnect circuits with large input and output ports. Recent advances in reduction techniques using non-Krylov subspace techniques such as truncated balanced realization (TBR) hold some promise to solve these problems. In this paper, we first review the classic TBR-based reduction methods and then present the recent developments in fast TBR-based reduction and techniques such as PMTBR, SBPOR, and ETBR methods. These newly proposed methods try to avoid the expensive computing steps in traditional TBR methods at some cost to accuracy to boost efficiency and scalability, which is critical to reduce large interconnect parasitics modeled as RLCK circuits. The ETBR method can also reduce circuits with massive ports by considering the input signals. We show the pros and cons of each method and compare them on a set of large interconnect circuits, and finally point to some new research directions for this area.

Key words: model order reduction; balanced realization; interconnect

Introduction

Model order reduction (MOR) is an efficient technique to reduce circuit complexity while preserving the input and output behavior. In the reduction process, we expect that the algorithms have the following properties. First, the approximation error should be small, and ideally there exists a global error bound. Second, system properties, like stability and passivity should be preserved. Third, the procedure is numerically stable and computationally efficient. There are many ways to classify MOR algorithms. For instance, from a formulation point of view, MOR techniques can be classified into first-order based methods (using modified nodal analysis, MNA) and second-order based methods (using nodal analysis, NA). In terms of projection subspace, these approaches are divided into two broad categories, namely, moment matching based methods (Krylov subspace methods) and balanced truncation based methods. In the former case, the system is projected onto a subspace to match dominant moments of transfer functions, while in the latter case the system is projected onto a subspace that is both easily controllable and easily observable. A more detailed survey of model order reduction techniques can be found[1,2].

Moment-matching or Krylov subspace based approaches have received much attention and development in applications for VLSI design in the past[3-9]. The AWE method[3] first introduces the explicit

Received: 2009-11-25; revised: 2010-01-28
* Supported in part by National Science Foundation (NSF) (Nos. CCF-0448534 and OISE-0929699) and in part by the National Natural Science Foundation of China (No. 60828008)
** To whom correspondence should be addressed.
E-mail: stan@ee.ucr.edu; Tel: 1-951-8272425

moment-matching technique for fast interconnect modeling (mainly delay calculation). But AWE suffers from numerical instability owing to explicit moment-matching. To mitigate this problem, Krylov subspace based methods were proposed[4,5], where implicit moment-matching is performed in a projection framework. Furthermore, to ensure the stability of the simulation process, PRIMA[6] was developed based on the Arnoldi process. PRIMA exploits the positive semi-definiteness of matrices in MNA formulation so that passivity can be easily preserved via congruency transformation[10]. More recently, SPRIM[8] further exploits the block structure of RLC formulation such that, in addition to passivity, structure information inherent to RLC circuits can be also preserved. At the same time, second-order moment-matching based approaches have been successfully developed, such as ENOR[7] to SAPOR[9].

While suitable for reduction of large-scale circuits, these techniques do not necessarily generate models as compact as desired[11]. Therefore, another approach, truncated balanced realization (TBR), or balanced truncation, which was developed in the control community[12-16], has been studied intensively for interconnect modeling[17-26]. In theory, these methods can produce overall good approximates with wide frequency band accuracy. Existing first-order balanced truncation methods in the control community include Lyapunov balancing[12] and stochastic balancing (include Riccati balancing)[14]. Stochastic balancing is more computationally demanding than Lyapunov balancing but it is highly appreciated by preserving the positive realness (passivity)[15] without posing any constraints on the internal structure of the state-space.

Recent developments of balanced truncation for general dynamic systems include approximate balancing by iterative solution of Lyapunov equation[17,18,27], Sylvester equation[28], Riccati equation[20,22], the AISAID method[29], and its derivative method, which are based on the power method, for fast gramian product approximation[21], and descriptor system balancing[30-32]. Specifically, the recent book[2] by Antoulas provides excellent coverage of classic TBR methods such as Lyapunov balancing, stochastic balancing, bounded real balancing, positive balancing, and frequency weighted balanced reduction (Chapter 7), as well as connections between the Krylov subspace methods and TBR methods (Chapter 12). It also discusses the iterative solutions of Lyapunov equations for approximating balanced truncation. Research work in Ref. [27] presents an approximate balancing method (the low-rank Smith method) by iterative solutions of Lyapunov equations. A recent paper in Ref. [33] offers a comparison of TBR and Krylov subspace based methods such as PMTBR, Riccati balanced truncation, PRIMA, the special zero method and the optimal H_2 method. The paper in Ref. [34] introduces various gramian definitions in the frequency domain and discusses the resulting balancing schemes. The paper in Ref. [35] presents an excellent review of TBR methods and a new error bound for positive-real balanced truncation. Research work in Ref. [28] studies the Sylvester equations, the cross gramian, and the approximate balancing. A recent talk by Benner offers a good tutorial on the balanced truncation for circuit simulation[36].

Standard balanced truncation methods are known to be too expensive for direct application to large integrated circuit problems, owing to the cubic cost of solving two Lyapunov equations. In addition, it takes considerable knowledge of control theory and numerical procedures to implement balanced truncation in a stable way[37,38]. Especially for nonstandard systems, additive decompositions and special treatments are causally needed[19,30,39]. To remedy this problem, several gramian approximation methods have been proposed[23,25,26,40], where the approximated dominant subspace of a gramian can be obtained in a very efficient way. However, no rigorous error bounds exist for gramian approximation methods. The single gramian approximation (SGA) technique (also called Poor Man's TBR or PMTBR)[23] was first proposed to reduce the system by projecting onto the approximated dominant subspace of the controllability gramian. This method works well for RC circuits, which can be naturally formulated in a first-order form with matrices both symmetric and positive-definite. In this case, the controllability and observability gramians are equivalent and can be simultaneously diagonalized via a congruency transformation. As a result, both accuracy and passivity can be preserved simultaneously. However, for general RLCK circuits, which models the on-chip global interconnects with fast signals, the first-order formulation could be either symmetric or positive-definite, but not both. Therefore, preserving high accuracy and passivity cannot be achieved at the same

time.

There are several methods proposed to mitigate this problem. One of them, SBPOR[24], is based on the second-order formulation, which is both symmetric and positive-definite for RLCK interconnect circuits. In SBPOR, second-order gramians are defined based on a symmetric first-order realization. As a result, both second-order gramians, which are also the leading blocks of the gramians of first-order realization, become the same and can be simultaneously diagonalized by a congruence transformation. As a result, it achieves passivity without sacrificing accuracy (it still approximates both controllability and observability gramians). Further, a fast SBPOR method, called SOGA, was proposed[26]. It computes the approximate gramians of one second-order formation from SBPOR to make the algorithm more computationally efficient.

Other gramian approximation methods also covered in this survey include double gramian approximation (DGA), cross-gramian approximation (CGA) and the recently proposed response gramian approximation (RGA), called ETBR[25], for power grid network analysis. CGA[23] combines both controllability and observability into a single cross-gramian. ETBR is used to reduce and simulate circuits with a large number of independent sources, which cannot be reduced by the conventional multi-port model reduction methods. ETBR considers both systems as well as the input sources for reduction by defining an approximated response gramian.

Another quite different approach to circuit complexity reduction is by means of local node elimination and realization[41-45]. The major advantage of these methods over projection-based methods is that the reduction can be done in a local manner and no overall solution of the entire circuit is required and reduced models can be easily realized using RLCM elements. This idea was first explored by selective node elimination for RC circuits[41,42], where time-constant analysis is used to select nodes for elimination. Node reduction for magnetic coupling interconnect (RLCM) circuits has recently become an active research area. Generalized Y-Δ transformation[44], RLCK circuit crunching[43], and branch merging[45] have been developed based on NA, where inductance becomes susceptance in the admittance matrix. Since mutual inductance is coupled via branch currents, to perform nodal reduction, an equivalent 6-susceptance NA model is introduced in Ref. [44] to reduce two coupling current variables and template matching via geometrical programming is used to realize the model order reduced admittances, but its accuracy depends heavily on the selection of templates and only 1-port realization has been reported. Meanwhile, RLCK circuit crunching and branch merging methods are first-order approximation based on the nodal time-constant analysis. A more general node-elimination algorithm based on the general s-domain hierarchical graph-based symbolic reduction technique[11,46,47] was proposed, which deals with circuits with large number of ports more efficiently than the projection based reduction techniques. But in this paper, we still focus on projection based reduction techniques.

We remark that model order reduction by the truncated balanced realization method and other non-Krylov subspace methods has a large body of literature from many different fields. This survey paper, however, will not be a comprehensive treatment on this topic. Instead, we try to focus on recent advances in interconnect modeling and reduction, especially TBR based reduction techniques. Relevant publications not cited in this paper do not diminish their contributions to this field.

This paper is organized as follows: Section 1 introduces some basic concepts such as dynamic systems, model order reduction and passivity. Section 2 introduces classic balanced truncation-based reduction methods developed for general dynamic systems. Section 3 presents recently proposed fast TBR-based reduction methods using gramian approximation methods developed for large-scale VLSI circuits. Numerical results are given in Section 4 and Section 5 to summarize the paper.

1 Model Order Reduction in a Nutshell

1.1 Dynamic system models

The behavior of linear time-invariant (LTI) systems in many engineering problems can be described by state-space equations in descriptor form $(\boldsymbol{E}, \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D})$,

$$\begin{aligned} \boldsymbol{E}\dot{\boldsymbol{x}}(t) &= \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t), \\ \boldsymbol{y}(t) &= \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) \end{aligned} \tag{1}$$

where $\boldsymbol{E}$, $\boldsymbol{A}\in\mathbf{R}^{n\times n}$, $\boldsymbol{B}\in\mathbf{R}^{n\times p}$, $\boldsymbol{C}\in\mathbf{R}^{p\times n}$, $\boldsymbol{D}\in\mathbf{R}^{p\times p}$, $\boldsymbol{x}(t)\in\mathbf{R}^{n}$,

$\boldsymbol{u}(t)$, $\boldsymbol{y}(t)\in\mathbf{R}^p$. When $\boldsymbol{E}$ equals identity matrix $\boldsymbol{I}$, the state-space equations become the standard form $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$,

$$\dot{\boldsymbol{x}}(t)=\boldsymbol{A}\boldsymbol{x}(t)+\boldsymbol{B}\boldsymbol{u}(t),$$
$$\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{x}(t)+\boldsymbol{D}\boldsymbol{u}(t) \quad (2)$$

In fact, many LTI systems can also be described by a set of second-order differential equations $(\boldsymbol{M},\boldsymbol{D},\boldsymbol{K},\boldsymbol{B},\boldsymbol{C})$,

$$\boldsymbol{M}\ddot{\boldsymbol{a}}(t)+\boldsymbol{D}\dot{\boldsymbol{a}}(t)+\boldsymbol{K}\boldsymbol{a}(t)=\boldsymbol{B}\boldsymbol{u}(t),$$
$$\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{a}(t) \quad (3)$$

where $\boldsymbol{M},\boldsymbol{D},\boldsymbol{K}\in\mathbf{R}^{n\times n}$, $\boldsymbol{B}\in\mathbf{R}^{n\times p}$, $\boldsymbol{C}\in\mathbf{R}^{p\times n}$, $\boldsymbol{u}(t)$, $\boldsymbol{y}(t)\in\mathbf{R}^p$, $\boldsymbol{a}(t)\in\mathbf{R}^n$. The behavior of the systems can be completely characterized by the state-space equations. However, in some cases, we are only interested in the input-output behavior. In such cases, transfer functions are needed. The transfer function associated with the first-order system (1) in the Laplace domain is given by

$$\boldsymbol{H}(s)=\boldsymbol{C}(s\boldsymbol{E}-\boldsymbol{A})^{-1}\boldsymbol{B} \quad (4)$$

which becomes

$$\boldsymbol{H}(s)=\boldsymbol{C}(s\boldsymbol{I}-\boldsymbol{A})^{-1}\boldsymbol{B} \quad (5)$$

for a standard system (2). For the second-order model (3), the transfer function is given by

$$\boldsymbol{H}(s)=\boldsymbol{C}(\boldsymbol{M}s^2+\boldsymbol{D}s+\boldsymbol{K})^{-1}\boldsymbol{B} \quad (6)$$

1.2 Model order reduction

The complexity of the system can be characterized by the size, n, of the model. In electrical engineering, civil engineering, or aeronautics, the size, n, is often very, even prohibitively, large that many analysis and design problems cannot be solved within a reasonable computing time. Model order reduction is the technique to solve this problem by constructing a reduced model $\boldsymbol{H}_{\mathrm{r}}(s)$ of size $r\ll n$ like

$$\boldsymbol{E}_{\mathrm{r}}\dot{\boldsymbol{x}}(t)=\boldsymbol{A}_{\mathrm{r}}\boldsymbol{x}(t)+\boldsymbol{B}_{\mathrm{r}}\boldsymbol{u}(t),$$
$$\boldsymbol{y}(t)=\boldsymbol{C}_{\mathrm{r}}\boldsymbol{x}(t)+\boldsymbol{D}\boldsymbol{u}(t) \quad (7)$$

where $\boldsymbol{E}_{\mathrm{r}},\boldsymbol{A}_{\mathrm{r}}\in\mathbf{R}^{r\times r}$, $\boldsymbol{B}_{\mathrm{r}}\in\mathbf{R}^{r\times p}$, $\boldsymbol{C}_{\mathrm{r}}\in\mathbf{R}^{p\times r}$, $\boldsymbol{D}\in\mathbf{R}^{p\times p}$, $\boldsymbol{x}(t)\in\mathbf{R}^r$, $\boldsymbol{u}(t)$, $\boldsymbol{y}(t)\in\mathbf{R}^p$, for the first-order model (1), or

$$\boldsymbol{M}_{\mathrm{r}}\dot{\boldsymbol{a}}(t)+\boldsymbol{D}_{\mathrm{r}}\boldsymbol{a}(t)+\boldsymbol{K}_{\mathrm{r}}\int\boldsymbol{a}(t)=\boldsymbol{B}_{\mathrm{r}}\boldsymbol{u}(t),$$
$$\boldsymbol{y}(t)=\boldsymbol{C}_{\mathrm{r}}\boldsymbol{a}(t) \quad (8)$$

where $\boldsymbol{M}_{\mathrm{r}},\boldsymbol{D}_{\mathrm{r}},\boldsymbol{K}_{\mathrm{r}}\in\mathbf{R}^{r\times r}$, $\boldsymbol{B}_{\mathrm{r}}\in\mathbf{R}^{r\times p}$, $\boldsymbol{C}_{\mathrm{r}}\in\mathbf{R}^{p\times r}$, $\boldsymbol{u}(t),\boldsymbol{y}(t)\in\mathbf{R}^p$, $\boldsymbol{a}(t)\in\mathbf{R}^r$, for the second-order model (3).

Such a low order system will have approximately the same output $\boldsymbol{y}$ as the original system to the input $\boldsymbol{u}$ of interest. The transfer functions are often used as a metric for approximation. If $\|\boldsymbol{H}(s)-\boldsymbol{H}_{\mathrm{r}}(s)\|<\varepsilon$ in some appropriate norm, for some given allowable error ε and frequency range of interest s, the reduced model is accepted as accurate.

In addition, it is important to preserve system properties like stability and passivity in model order reduction. Given a passive system, we hope the reduced system is also passive. Otherwise, the reduced system may cause nonphysical behavior when it is simulated with other subsystems even if it is stable.

Currently, most model order reduction methods are projection based. Given two projection matrices $\boldsymbol{W},\boldsymbol{V}\in\mathbf{R}^{n\times r}$, for the first-order model (7), we have

$$\boldsymbol{E}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{V},\ \boldsymbol{A}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V},$$
$$\boldsymbol{B}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{B},\quad \boldsymbol{C}_{\mathrm{r}}=\boldsymbol{C}\boldsymbol{V} \quad (9)$$

and for the second-order model (8), we have

$$\boldsymbol{M}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{V},\ \boldsymbol{D}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{V},\ \boldsymbol{K}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{V},$$
$$\boldsymbol{B}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{B},\ \boldsymbol{C}_{\mathrm{r}}=\boldsymbol{C}\boldsymbol{V} \quad (10)$$

where $\boldsymbol{W}$ is the left projector and $\boldsymbol{V}$ is the right projector.

Typically, $\boldsymbol{W}$ and $\boldsymbol{V}$ span useful subspaces. Different choices of $\boldsymbol{W}$ and $\boldsymbol{V}$ will result in different model reduction approaches (Krylov subspace based methods, balanced truncation methods, etc). If $\boldsymbol{W}\neq\boldsymbol{V}$, the projection is an oblique (Petrov-Galerkin) projection. If $\boldsymbol{W}=\boldsymbol{V}$, the projection is an orthogonal (Galerkin) projection. Usually, oblique projection is better in terms of accuracy as both subspaces are used (e.g., PVL[4] and TBR[12]). However, orthogonal projection is widely used in practice because it can be used to preserve important properties like passivity for systems with special state-space formulation (e.g., PRIMA[6] and PMTBR[23]).

1.3 Passivity

Passivity is an important property of many physical systems. Passive systems cannot produce energy internally. When modeling passive systems, non-passive reduced models may generate unbounded responses in transient simulation. For linear dynamic systems, passivity requires the transfer functions to be positive-real when the input and output signals are port voltages and currents. For scattering-parameter (s-parameter) systems passivity requires bounded-real for s-parameter matrices.

1.3.1 Necessary and sufficient condition

Condition 1 The system is passive if and only if its transfer function $\boldsymbol{H}(s)$ is positive real[48], which means

- $\boldsymbol{H}(s)$ is analytic for $\mathrm{Re}\,(s)>0$,
- $\overline{\boldsymbol{H}(s)}=\boldsymbol{H}(\overline{s})$ for $s\in\mathbf{C}$,
- $\boldsymbol{H}(s)+\boldsymbol{H}(s)^{\mathrm{H}}\geqslant 0$ for $\mathrm{Re}\,(s)>0$,

where $\overline{\boldsymbol{H}}$ denotes complex conjugate, $\boldsymbol{H}^{\mathrm{H}}$ denotes Hermitian (complex conjugate and transpose), and $\geqslant 0$ denotes positive semi-definiteness in a matrix context.

1.3.2 Sufficient condition

Condition 2 Given a dynamic system model, if the system matrices are positive semi-definite and the input matrix and output matrix equal, then the state-space model is in a passive form Refs. [6,7,10].

- For first-order model (1), $\boldsymbol{A}$ and $\boldsymbol{E}$ are positive semi-definite and $\boldsymbol{B}=\boldsymbol{C}^{\mathrm{T}}$.
- For second-order model (3), $\boldsymbol{M}$, $\boldsymbol{D}$, and $\boldsymbol{K}$ are positive semi-definite and $\boldsymbol{B}=\boldsymbol{C}^{\mathrm{T}}$.

In such a passive form, the transfer function will be positive-real, which means the system is passive. This sufficient condition is important because RLCK circuits can be formulated into such a passive form. Since the passive form can be inherited by the reduced model via an orthogonal projection, passivity can be easily preserved[6-10,23,26].

2 Review of Classic Balanced Truncation Methods for Reduction

In this section, we review the classical balanced truncation methods developed in the control community for general dynamic systems.

2.1 Lyapunov balancing

Lyapunov balancing was introduced to the system and control society by Ref. [12]. Given a stable minimal linear time invariant (LTI) system in standard state-space form $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$ in Eq. (2), the controllability gramian $\boldsymbol{X}$ and observability gramian $\boldsymbol{Y}$ are as follows:

$$\boldsymbol{X}=\int_0^{\infty}\mathrm{e}^{\boldsymbol{A}\tau}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\mathrm{e}^{\boldsymbol{A}^{\mathrm{T}}\tau}\mathrm{d}\tau,$$

$$\boldsymbol{Y}=\int_0^{\infty}\mathrm{e}^{\boldsymbol{A}^{\mathrm{T}}\tau}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\mathrm{e}^{\boldsymbol{A}\tau}\mathrm{d}\tau \tag{11}$$

It is easy to verify that they are the unique symmetric positive definite solutions to the Lyapunov equations,

$$\boldsymbol{A}\boldsymbol{X}+\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}=0,$$

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y}+\boldsymbol{Y}\boldsymbol{A}+\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}=0 \tag{12}$$

The controllability and observability gramians $\boldsymbol{X}$ and $\boldsymbol{Y}$ are related to the energy demanded to control and observe the system.

2.1.1 Controllability

Given any state $\boldsymbol{x}_0$ at $t=0$, if the system is controllable, there is a signal $\boldsymbol{u}(t)$ with the smallest energy (measured by L_2 norm),

$$\|\boldsymbol{u}(t)\|_2=\sqrt{\int_{-\infty}^{0}\boldsymbol{u}^{\mathrm{T}}(t)\boldsymbol{u}(t)\mathrm{d}t} \tag{13}$$

which could drive the system from zero initial condition at $t=-\infty$ to $\boldsymbol{x}_0$. Assuming $\boldsymbol{x}(-\infty)=0$, the zero-state response is

$$\boldsymbol{x}(t)=\int_{-\infty}^{t}\mathrm{e}^{\boldsymbol{A}(t-\tau)}\boldsymbol{u}(\tau)\mathrm{d}\tau \tag{14}$$

The controllability gramian $\boldsymbol{X}$ is connected to the solution of the minimum L_2 norm problem,

$$\min_{\boldsymbol{u}\in L_2[-\infty,0]}\|\boldsymbol{u}(t)\|_2^2,$$

$$\text{subject to } \boldsymbol{x}(0)=\int_{-\infty}^{0}\mathrm{e}^{-\boldsymbol{A}\tau}\boldsymbol{u}(\tau)\mathrm{d}\tau=\boldsymbol{x}_0 \tag{15}$$

The solution to this problem is

$$\boldsymbol{u}(t)=\boldsymbol{B}^{\mathrm{T}}\mathrm{e}^{-\boldsymbol{A}^{\mathrm{T}}t}\left(\int_{-\infty}^{0}\mathrm{e}^{-\boldsymbol{A}\tau}\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}\mathrm{e}^{-\boldsymbol{A}^{\mathrm{T}}\tau}\mathrm{d}\tau\right)^{-1}\boldsymbol{x}_0=$$

$$\boldsymbol{B}^{\mathrm{T}}\mathrm{e}^{-\boldsymbol{A}^{\mathrm{T}}t}\boldsymbol{X}^{-1}\boldsymbol{x}_0 \tag{16}$$

So the minimal energy needed to reach $\boldsymbol{x}_0$ is

$$\|\boldsymbol{u}(t)\|_2^2=\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{X}^{-1}\boldsymbol{x}_0 \tag{17}$$

Now the optimization problem becomes a quadratic form, which means the size of the eigenvalues of $\boldsymbol{X}$ describes how much input energy is needed to control the associated state eigenvector. In other words, if $\boldsymbol{x}_0$ is picked as one of the eigenvectors of $\boldsymbol{X}$, the energy needed in the input will be exactly the inverse of the corresponding eigenvalue. As a result, the largest eigenvalue will correspond to the state which is easiest to control.

2.1.2 Observability

Observability shares the similar definition of controllability. Given any state $\boldsymbol{x}_0$ at $t=0$, we want to observe how much energy (measured by L_2 norm) there will be from the output signal if the system is released from $\boldsymbol{x}_0$ with zero input for $t\geqslant 0$. The observability gramian $\boldsymbol{Y}$ is related to the solution of the maximum L_2 norm problem

$$\max_{\boldsymbol{y}\in L_2[0,+\infty]}\|\boldsymbol{y}(t)\|_2^2,$$

$$\text{subject to } \boldsymbol{x}(0)=\boldsymbol{x}_0 \tag{18}$$

The zero-input response is

$$\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{x}(t)=\boldsymbol{C}\boldsymbol{x}(0)\mathrm{e}^{\boldsymbol{A}t} \tag{19}$$

The L_2 norm of the output signal when the system is released from $\boldsymbol{x}_0$ is

$$\|\boldsymbol{y}(t)\|_2^2=\boldsymbol{x}_0^{\mathrm{T}}\left(\int_0^{\infty}\mathrm{e}^{\boldsymbol{A}^{\mathrm{T}}t}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}\mathrm{e}^{\boldsymbol{A}t}\mathrm{d}t\right)\boldsymbol{x}_0=\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{x}_0 \tag{20}$$

which means the size of the eigenvalues of $\boldsymbol{Y}$ describes how much output energy is produced when the associated state eigenvector is in free evolution. In other words, if $\boldsymbol{x}_0$ is picked as one of the eigenvectors of $\boldsymbol{Y}$, the energy observed in the output will be exactly the corresponding eigenvalue. As a result, the largest eigenvalue will correspond to the state which is easiest to observe.

2.1.3 Balanced truncation

Given a dynamic system, the state-space representation is not unique. Any nonsingular linear transformation $\boldsymbol{x}=T\tilde{\boldsymbol{x}}$ can be applied to the system $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$ to obtain a new state-space representation $(\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}},\tilde{\boldsymbol{C}},\boldsymbol{D})$,

$$\begin{aligned}\dot{\tilde{\boldsymbol{x}}}(t)&=\tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}}(t)+\tilde{\boldsymbol{B}}\boldsymbol{u}(t),\\ \boldsymbol{y}(t)&=\tilde{\boldsymbol{C}}\tilde{\boldsymbol{x}}(t)+\boldsymbol{D}\boldsymbol{u}(t)\end{aligned} \tag{21}$$

where

$$\tilde{\boldsymbol{A}}=\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T},\quad \tilde{\boldsymbol{B}}=\boldsymbol{T}^{-1}\boldsymbol{B},\quad \tilde{\boldsymbol{C}}=\boldsymbol{C}\boldsymbol{T} \tag{22}$$

Such a transformation is known as a similarity transformation, which does not change the input-output behavior of the system. It is easy to see that both representations $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$ and $(\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}},\tilde{\boldsymbol{C}},\boldsymbol{D})$ have the same transfer function $\boldsymbol{H}(s)$.

A balanced realization is a special state-space representation, where the controllability and observability gramians are diagonal and equal. The balancing transformation can be computed by calculating the eigenmodes of the gramian product $\boldsymbol{XY}$,

$$\boldsymbol{X}\boldsymbol{Y}=\boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}^{-1} \tag{23}$$

It can be seen that the eigenvetors of $\boldsymbol{XY}$ are the basis vectors that describe the balancing transformation as follows. From Eqs. (11) and (22), we obtain the following expressions for the gramians of the transformed system

$$\begin{aligned}\tilde{\boldsymbol{X}}&=\boldsymbol{T}^{-1}\boldsymbol{X}\boldsymbol{T}^{-\mathrm{T}},\\ \tilde{\boldsymbol{Y}}&=\boldsymbol{T}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{T}\end{aligned} \tag{24}$$

For a balanced system, we require $\tilde{\boldsymbol{X}}=\tilde{\boldsymbol{Y}}=\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is a diagonal matrix. From Eq. (24), we can write

$$\begin{aligned}\boldsymbol{T}^{-1}\boldsymbol{X}&=\boldsymbol{\Sigma}\boldsymbol{T}^{\mathrm{T}},\\ \boldsymbol{Y}\boldsymbol{T}&=\boldsymbol{T}^{-\mathrm{T}}\boldsymbol{\Sigma}\end{aligned} \tag{25}$$

or

$$\boldsymbol{T}^{-1}\boldsymbol{X}\boldsymbol{Y}\boldsymbol{T}=\boldsymbol{\Sigma}^2 \tag{26}$$

which means the transformation $\boldsymbol{T}$, which balances the system, contains the eigenvectors of the gramian product $\boldsymbol{XY}$ as its columns.

From the gramian expression (23) and (26), it can be seen that the the eigenvalues λ_i contained in the diagonal matrix $\boldsymbol{\Lambda}$ are positive real numbers, and $\sigma_i=\sqrt{\lambda_i}$ are known as the Hankel singular values of the system. The eigenvectors of $\boldsymbol{XY}$ correspond to states through which the input is transmitted to the output. The magnitudes of the Hankel singular values describe the relative importance of these states and are independent of the particular realization of the system. States corresponding to the small Hankel singular values are difficult to control and difficult to observe. Such states are less involved in the energy transfer from inputs to outputs.

Therefore, a general idea of balanced truncation is to transform the system into a balanced form $(\tilde{\boldsymbol{A}},\tilde{\boldsymbol{B}},\tilde{\boldsymbol{C}},\boldsymbol{D})$, where the states which are difficult to control are also difficult to observe. Then, the parts of the dynamics that correspond to those week states will be discarded. We may partition $\boldsymbol{\Sigma}$ into

$$\boldsymbol{\Sigma}=\begin{bmatrix}\boldsymbol{\Sigma}_1 & 0\\ 0 & \boldsymbol{\Sigma}_2\end{bmatrix} \tag{27}$$

and conformingly partition the transformed matrices as

$$\begin{aligned}\tilde{\boldsymbol{A}}&=\begin{bmatrix}\tilde{\boldsymbol{A}}_{11} & \tilde{\boldsymbol{A}}_{12}\\ \tilde{\boldsymbol{A}}_{21} & \tilde{\boldsymbol{A}}_{22}\end{bmatrix},\\ \tilde{\boldsymbol{B}}&=\begin{bmatrix}\tilde{\boldsymbol{B}}_1\\ \tilde{\boldsymbol{B}}_2\end{bmatrix},\\ \tilde{\boldsymbol{C}}&=[\tilde{\boldsymbol{C}}_1\quad \tilde{\boldsymbol{C}}_2]\end{aligned} \tag{28}$$

The reduced model of order r $(\boldsymbol{A}_{\mathrm{r}},\boldsymbol{B}_{\mathrm{r}},\boldsymbol{C}_{\mathrm{r}},\boldsymbol{D})$ is obtained by taking the $r\times r$, $r\times p$, $q\times r$ leading blocks of $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{B}}$, $\tilde{\boldsymbol{C}}$, respectively.

$$\boldsymbol{A}_{\mathrm{r}}=\tilde{\boldsymbol{A}}_{11},\quad \boldsymbol{B}_{\mathrm{r}}=\tilde{\boldsymbol{B}}_1,\quad \boldsymbol{C}_{\mathrm{r}}=\tilde{\boldsymbol{C}}_1 \tag{29}$$

This truncation leads to a balanced reduced-order system $(\boldsymbol{A}_{\mathrm{r}},\boldsymbol{B}_{\mathrm{r}},\boldsymbol{C}_{\mathrm{r}},\boldsymbol{D})$.

The TBR method has two important properties. One is the stability preservation, the other is the existence of the error bound throughout the frequency domain. Specifically, the error in the transfer function of the order r approximation is bounded by Ref. [13],

$$\|\boldsymbol{H}(s)-\boldsymbol{H}_{\mathrm{r}}(s)\|\leqslant 2\sum_{i=r+1}^{n}\sigma_i \tag{30}$$

To summarize, the standard balanced truncation algorithm flow chart[37] is shown in Fig. 1. An approach with improved numerical properties may be found in Ref. [38].

Algorithm: Standard Balanced Truncation Method.
Input: $\boldsymbol{H}:(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$
Output: $\boldsymbol{H}_{\mathrm{r}}:(\boldsymbol{A}_{\mathrm{r}},\boldsymbol{B}_{\mathrm{r}},\boldsymbol{C}_{\mathrm{r}},\boldsymbol{D})$
(1) Compute $\boldsymbol{X}>0$ and $\boldsymbol{Y}>0$
(2) Cholesky factorization $\boldsymbol{X}=\boldsymbol{L}_x\boldsymbol{L}_x^{\mathrm{T}}$ and $\boldsymbol{Y}=\boldsymbol{L}_y\boldsymbol{L}_y^{\mathrm{T}}$
(3) Compute SVD $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}=\boldsymbol{L}_y^{\mathrm{T}}\boldsymbol{L}_x$
(4) Compute $\boldsymbol{T}=\boldsymbol{L}_x\boldsymbol{V}\boldsymbol{\Sigma}^{-1/2}$ and $\boldsymbol{T}^{-1}=\boldsymbol{\Sigma}^{-1/2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{L}_y^{\mathrm{T}}$
(5) Compute the balanced realizations $\tilde{\boldsymbol{A}}=\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}$, $\tilde{\boldsymbol{B}}=\boldsymbol{T}^{-1}\boldsymbol{B}$, $\tilde{\boldsymbol{C}}=\boldsymbol{C}\boldsymbol{T}$.
(6) Truncate to form the reduced system $(\boldsymbol{A}_{\mathrm{r}},\boldsymbol{B}_{\mathrm{r}},\boldsymbol{C}_{\mathrm{r}},\boldsymbol{D})$

Fig. 1 Balanced truncation algorithm

2.1.4 Balancing in descriptor form

Given a state-space model in descriptor form $(\boldsymbol{E},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$ in Eq. (1), if the matrix $\boldsymbol{E}$ is singular, the system may not be proper. In this case, there are infinite eigenvalues and the transfer function can be represented as a sum of proper transfer function and a matrix of polynomials

$$\boldsymbol{G}(s)=\boldsymbol{G}_{\mathrm{p}}(s)+\sum_{i>0}\boldsymbol{G}_i s^i \tag{31}$$

where $\boldsymbol{G}_{\mathrm{p}}(s)$ is a matrix of proper rational functions of s. The proper and polynomial parts of the transfer function can be separated by projecting the system onto deflating subspaces of the pair $(\boldsymbol{E},\boldsymbol{A})$ corresponding to finite and infinite eigenvalues, respectively[30]. The polynomial terms should be exactly preserved by the reduced system and the proper rational term $\boldsymbol{G}_{\mathrm{p}}(s)$, where $\boldsymbol{E}$ is nonsingular, can be reduced by classical balanced truncation.

Given the system in descriptor form with nonsingular $\boldsymbol{E}$, controllability and observability gramians[21] can be computed by solving generalized Lyapunov equations,

$$\begin{aligned}&\boldsymbol{E}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}+\boldsymbol{A}\boldsymbol{X}\boldsymbol{E}^{\mathrm{T}}+\boldsymbol{B}\boldsymbol{B}^{\mathrm{T}}=0,\\&\boldsymbol{E}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{A}+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{E}+\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C}=0\end{aligned} \tag{32}$$

Similarly, we want to find a matrix $\boldsymbol{T}$ to perform a similarity transformation $(\boldsymbol{T}^{-1}\boldsymbol{E}\boldsymbol{T},\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T},\boldsymbol{T}^{-1}\boldsymbol{B},\boldsymbol{C}\boldsymbol{T})$ to diagonalize the product $\boldsymbol{X}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{E}$. After the similarity transformation, the system is balanced. We may conformingly partition the transformed matrices as

$$\boldsymbol{T}^{-1}\boldsymbol{E}\boldsymbol{T}=\begin{bmatrix}\tilde{\boldsymbol{E}}_{11}&\tilde{\boldsymbol{E}}_{12}\\\tilde{\boldsymbol{E}}_{21}&\tilde{\boldsymbol{E}}_{22}\end{bmatrix},$$

$$\boldsymbol{T}^{-1}\boldsymbol{A}\boldsymbol{T}=\begin{bmatrix}\tilde{\boldsymbol{A}}_{11}&\tilde{\boldsymbol{A}}_{12}\\\tilde{\boldsymbol{A}}_{21}&\tilde{\boldsymbol{A}}_{22}\end{bmatrix},$$

$$\boldsymbol{T}^{-1}\boldsymbol{B}=\begin{bmatrix}\tilde{\boldsymbol{B}}_1\\\tilde{\boldsymbol{B}}_2\end{bmatrix},$$

$$\boldsymbol{C}\boldsymbol{T}=[\tilde{\boldsymbol{C}}_1\quad\tilde{\boldsymbol{C}}_2] \tag{33}$$

and $(\tilde{\boldsymbol{E}}_{11},\tilde{\boldsymbol{A}}_{11},\tilde{\boldsymbol{B}}_1,\tilde{\boldsymbol{C}}_1,\boldsymbol{D})$ is the reduced order system. In fact, this reduction is mathematically equivalent to performing balanced truncation on the system $(\boldsymbol{E}^{-1}\boldsymbol{A},\boldsymbol{E}^{-1}\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$. However, the computation steps are numerically better conditioned via generalized Lyapunov equations.

2.1.5 Balancing from a projection point of view

Note that, the interpretation of balanced truncation is to project the system onto a subspace both easily controllable and easily observable, which is just the dominant eigenspace of the matrix $\boldsymbol{X}\boldsymbol{Y}$ corresponding to the r largest eigenvalues. If we partition $\boldsymbol{T}$ and $\boldsymbol{T}^{-1}$ as

$$\boldsymbol{T}^{-1}=\begin{bmatrix}\boldsymbol{W}^{\mathrm{T}}\\\tilde{\boldsymbol{W}}^{\mathrm{T}}\end{bmatrix},\quad\boldsymbol{T}=[\boldsymbol{V}\quad\tilde{\boldsymbol{V}}] \tag{34}$$

and substitute Eq. (34) into Eq. (33), then we have

$$\begin{aligned}&\tilde{\boldsymbol{E}}_{11}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{V},\ \tilde{\boldsymbol{A}}_{11}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V},\\&\tilde{\boldsymbol{B}}_1=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{B},\quad\tilde{\boldsymbol{C}}_1=\boldsymbol{C}\boldsymbol{V}\end{aligned} \tag{35}$$

which unifies the balancing and truncating operations into one projection step. Since the left projector $\boldsymbol{W}$ and right projector $\boldsymbol{V}$ are not equal generally, i.e., $\boldsymbol{W}\neq\boldsymbol{V}$, the projection is an oblique projection.

2.1.6 Balancing a symmetric system

Given a state-space model in descriptor form $(\boldsymbol{E},\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$ in Eq. (1), the state-space model is symmetric if

$$\boldsymbol{E}=\boldsymbol{A}^{\mathrm{T}},\ \boldsymbol{E}=\boldsymbol{E}^{\mathrm{T}},\ \boldsymbol{B}=\boldsymbol{C}^{\mathrm{T}} \tag{36}$$

In this symmetric case, both Lyapunov Eq. (32) are the same and both gramians are equivalent $\boldsymbol{Y}=\boldsymbol{X}$. Since the gramian $\boldsymbol{X}$ is symmetric, it is orthogonally diagonalizable, i.e., there exists $\boldsymbol{T}^{-1}=\boldsymbol{T}^{\mathrm{T}}$ such that $\boldsymbol{T}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{T}=\boldsymbol{\Sigma}$. Then, we have

$$\boldsymbol{T}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}\boldsymbol{T}=\boldsymbol{T}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{T}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{T}=\boldsymbol{\Sigma}^2 \tag{37}$$

which means, in this symmetric case, the eigenspace of gramian product $\boldsymbol{X}\boldsymbol{X}$ is exactly the eigenspace of each gramian $\boldsymbol{X}$. In this case, we only need to project onto the dominant eigenspace of one gramian. Since either gramian is symmetric, the left projector $\boldsymbol{W}$ and right projector $\boldsymbol{V}$ are equal, $\boldsymbol{W}=\boldsymbol{V}$, and the projection (35) becomes an orthogonal projection,

$$\begin{aligned}&\tilde{\boldsymbol{E}}_{11}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{E}\boldsymbol{V},\ \tilde{\boldsymbol{A}}_{11}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{V},\\&\tilde{\boldsymbol{B}}_1=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{B},\quad\tilde{\boldsymbol{C}}_1=\boldsymbol{C}\boldsymbol{V}\end{aligned} \tag{38}$$

2.2 Riccati balancing

Lyapunov balancing preserves the stability of the system, but passivity might not be preserved. To keep the passivity properties of a system, Riccati balancing[14] is needed. If a system $(\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D})$ is positive real (passive), it will satisfy the positive real (PR) equations[49],

$$\begin{aligned}\boldsymbol{AP}+\boldsymbol{PA}^{\mathrm{T}}&=-\boldsymbol{B}_l\boldsymbol{B}_l^{\mathrm{T}},\\ \boldsymbol{PC}^{\mathrm{T}}-\boldsymbol{B}&=-\boldsymbol{B}_l\boldsymbol{D}_l^{\mathrm{T}},\\ -\boldsymbol{D}-\boldsymbol{D}^{\mathrm{T}}&=-\boldsymbol{D}_l\boldsymbol{D}_l^{\mathrm{T}}\end{aligned}\tag{39}$$

where $\boldsymbol{P}=\boldsymbol{P}^{\mathrm{T}}>0$. A dual pair of positive real equations are as follows:

$$\begin{aligned}\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Q}+\boldsymbol{QA}&=-\boldsymbol{C}_{\mathrm{r}}^{\mathrm{T}}\boldsymbol{C}_{\mathrm{r}},\\ \boldsymbol{QB}-\boldsymbol{C}^{\mathrm{T}}&=-\boldsymbol{C}_{\mathrm{r}}^{\mathrm{T}}\boldsymbol{D}_{\mathrm{r}},\\ -\boldsymbol{D}-\boldsymbol{D}^{\mathrm{T}}&=-\boldsymbol{D}_{\mathrm{r}}^{\mathrm{T}}\boldsymbol{D}_{\mathrm{r}}\end{aligned}\tag{40}$$

where $\boldsymbol{Q}=\boldsymbol{Q}^{\mathrm{T}}>0$. The above equations can be rewritten as a dual pair of Riccati equations, and then solved for $\boldsymbol{P}$ and $\boldsymbol{Q}$,

$$\boldsymbol{AP}+\boldsymbol{PA}^{\mathrm{T}}+(\boldsymbol{PC}^{\mathrm{T}}-\boldsymbol{B})\cdot(\boldsymbol{D}+\boldsymbol{D}^{\mathrm{T}})^{-1}(\boldsymbol{CP}-\boldsymbol{B}^{\mathrm{T}})=0\tag{41}$$

$$\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Q}+\boldsymbol{QA}+(\boldsymbol{QB}-\boldsymbol{C}^{\mathrm{T}})\cdot(\boldsymbol{D}+\boldsymbol{D}^{\mathrm{T}})^{-1}(\boldsymbol{B}^{\mathrm{T}}\boldsymbol{Q}-\boldsymbol{C})=0\tag{42}$$

Riccati balancing can now be achieved by substituting $(\boldsymbol{P},\boldsymbol{Q})$ with $(\boldsymbol{X},\boldsymbol{Y})$ in the balanced truncation algorithm. Since the reduced system also satisfies the (PR) equations, passivity is preserved. Riccati balancing has been applied to interconnect reduction in the positive-real TBR (PR-TBR) method[19].

Similar to Lyapunov balancing, Riccati balancing also has physical interpretations in terms of energy. Let $\boldsymbol{s}(\boldsymbol{u}(t),\boldsymbol{y}(t))$ be the supply function, which describes the rate at which power is supplied to the system and typically is defined such that $\boldsymbol{s}(\boldsymbol{u}(t),\boldsymbol{y}(t))>0$ implies a positive amount of energy input, while $\boldsymbol{s}(\boldsymbol{u}(t),\boldsymbol{y}(t))<0$ means energy is extracted from the system back to the environment. When the system inputs and outputs are currents or voltages, i.e., when the system transfer function represents impedance or admittance matrices, we may use the supply function $\boldsymbol{s}(\boldsymbol{u}(t),\boldsymbol{y}(t))=\boldsymbol{u}(t)^{\mathrm{T}}\boldsymbol{y}(t)$.

The input energy gramian $\boldsymbol{P}$ is associated with the following optimization problem:

$$\inf\left(\int_{-\infty}^{0}\boldsymbol{s}(\boldsymbol{u}(t),\boldsymbol{y}(t))\mathrm{d}t\right)=\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{P}^{-1}\boldsymbol{x}_0\tag{43}$$

which minimizes the amount of energy that must be injected into the system, in order to control the system to state $\boldsymbol{x}_0$ at time 0. In this setting, the sizes of the eigenvalues of $\boldsymbol{R}$ describe how much energy is needed to control the associated state eigenvector. Small eigenvalues of $\boldsymbol{P}$ implies that a large amount of energy is needed to reach the associated mode.

Similarly, the output energy gramian $\boldsymbol{Q}$ is associated with the following optimization problem

$$\sup\left(-\int_0^{\infty}\boldsymbol{s}(\boldsymbol{u}(t),\boldsymbol{y}(t))\mathrm{d}t\right)=\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{x}_0\tag{44}$$

which maximizes the amount of energy which can be extracted from the system in free evolution from $\boldsymbol{x}_0$ at time 0. Also, the sizes of the eigenvalues of $\boldsymbol{Q}$ describe how much energy can be extracted from the system in free evolution. Small eigenvalues of $\boldsymbol{Q}$ implies that a small amount of energy can be extracted from the associated model.

For the positive-real case, the error bound is given by Ref. [50].

$$\|\boldsymbol{H}(s)-\boldsymbol{H}_{\mathrm{r}}(s)\|\leqslant\lambda_{\max}(\boldsymbol{D}+\boldsymbol{D}^{\mathrm{T}})\cdot\sum_{i=r+1}^{n}\frac{2\sigma_k}{(1-\sigma_k)^2}\left(1+\sum_{j=1}^{k}\frac{2\sigma_j}{1-\sigma_j}\right)^2\tag{45}$$

2.3 Second-order balancing

Consider a second-order LTI stable system $(\boldsymbol{M},\boldsymbol{D},\boldsymbol{K},\boldsymbol{B},\boldsymbol{C})$ in Eq. (3) with $\boldsymbol{M}$ assumed to be nonsingular, the general idea of reducing the second-order system is to transform the second-order system into the equivalent first-order system, from which the balancing matrices are obtained. The second-order gramians[16] were defined based on the first-order realization in a standard state-space form $(\boldsymbol{\mathcal{A}},\boldsymbol{\mathcal{B}},\boldsymbol{\mathcal{C}})$ in Eq. (2) with $2n$-dimensional state $\boldsymbol{x}^{\mathrm{T}}=[\boldsymbol{q}^{\mathrm{T}}\ \dot{\boldsymbol{q}}^{\mathrm{T}}]$, where

$$\begin{aligned}\boldsymbol{\mathcal{A}}&=\begin{bmatrix}0&\boldsymbol{I}\\ -\boldsymbol{M}^{-1}\boldsymbol{K}&-\boldsymbol{M}^{-1}\boldsymbol{D}\end{bmatrix},\\ \boldsymbol{\mathcal{B}}&=\begin{bmatrix}0\\ \boldsymbol{M}^{-1}\boldsymbol{B}\end{bmatrix},\\ \boldsymbol{\mathcal{C}}&=[\boldsymbol{P}\quad\boldsymbol{Q}]\end{aligned}\tag{46}$$

The first-order realization has the same input-output behavior as the second-order system. Although a first-order MOR approach, like classic balanced truncation[12], can be applied to reduce Eq. (46), the reduced model is no longer a second-order. To perform the reduction directly on the second-order Eq. (3), one needs to define gramians for second-order systems. Similar to the first order gramian definition (15), the second order controllability gramian definition is based

on the following optimization problem[16],

$$\min_{\dot{\boldsymbol{a}}(0)\in\mathbf{R}^n,\boldsymbol{u}\in L_2[-\infty,0]}\left(\int_{-\infty}^{0}\boldsymbol{u}^{\mathrm{T}}(t)\boldsymbol{u}(t)\mathrm{d}t\right),$$
$$\text{subject to } \boldsymbol{M}\ddot{\boldsymbol{a}}(t)+\boldsymbol{D}\dot{\boldsymbol{a}}(t)+\boldsymbol{K}\boldsymbol{a}(t)=\boldsymbol{B}\boldsymbol{u}(t),$$
$$\boldsymbol{a}(0)=\boldsymbol{a}_0 \tag{47}$$

which minimizes the necessary energy to reach the given $\boldsymbol{a}_0$ over all past inputs $\boldsymbol{u}\in L_2[-\infty,0]$ and initial $\dot{\boldsymbol{a}}(0)\in\mathbf{R}^n$. First, we minimize the energy over all past inputs $\boldsymbol{u}\in L_2[-\infty,0]$, the solution of which has been available based on the optimization problem related to the first-order gramian (15),

$$\min_{\dot{\boldsymbol{a}}(0)\in\mathbf{R}^n}\left(\min_{\boldsymbol{u}\in L_2[-\infty,0]}\left(\int_{-\infty}^{0}\boldsymbol{u}^{\mathrm{T}}(t)\boldsymbol{u}(t)\mathrm{d}t\right)\right)=$$
$$\min_{\dot{\boldsymbol{a}}(0)\in\mathbf{R}^n}(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\mathcal{X}}^{-1}\boldsymbol{x}_0) \tag{48}$$

If we compatibly partition the controllability gramian of the first-order realization (46) defined in Eq. (11) $\boldsymbol{\mathcal{X}}$ and its inverse $\boldsymbol{\mathcal{X}}^{-1}$ as

$$\boldsymbol{\mathcal{X}}=\begin{bmatrix}\boldsymbol{R}_1 & \boldsymbol{R}_2\\ \boldsymbol{R}_2^{\mathrm{T}} & \boldsymbol{R}_3\end{bmatrix},\quad \boldsymbol{\mathcal{X}}^{-1}=\begin{bmatrix}\tilde{\boldsymbol{R}}_1 & \tilde{\boldsymbol{R}}_2\\ \tilde{\boldsymbol{R}}_2^{\mathrm{T}} & \tilde{\boldsymbol{R}}_3\end{bmatrix} \tag{49}$$

then we minimize the energy over initial $\dot{\boldsymbol{a}}(0)\in\mathbf{R}^n$,

$$\min_{\dot{\boldsymbol{a}}(0)\in\mathbf{R}^n}(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\mathcal{X}}^{-1}\boldsymbol{x}_0)=$$
$$\min_{\dot{\boldsymbol{a}}_0\in\mathbf{R}^n}\left([\boldsymbol{a}_0^{\mathrm{T}}\quad \dot{\boldsymbol{a}}_0^{\mathrm{T}}]\begin{bmatrix}\tilde{\boldsymbol{R}}_1 & \tilde{\boldsymbol{R}}_2\\ \tilde{\boldsymbol{R}}_2^{\mathrm{T}} & \tilde{\boldsymbol{R}}_3\end{bmatrix}\begin{bmatrix}\boldsymbol{a}_0\\ \dot{\boldsymbol{a}}_0\end{bmatrix}\right) \tag{50}$$

By annihilating the gradient, we can obtain the minimum energy $\boldsymbol{a}_0^{\mathrm{T}}(\tilde{\boldsymbol{R}}_1-\tilde{\boldsymbol{R}}_2\tilde{\boldsymbol{R}}_3^{-1}\tilde{\boldsymbol{R}}_2^{\mathrm{T}})\boldsymbol{a}_0$. Since $\tilde{\boldsymbol{R}}_1-\tilde{\boldsymbol{R}}_2\tilde{\boldsymbol{R}}_3^{-1}\tilde{\boldsymbol{R}}_2^{\mathrm{T}}$ is the Schur complement of $\tilde{\boldsymbol{R}}_3$, we have $\tilde{\boldsymbol{R}}_1-\tilde{\boldsymbol{R}}_2\tilde{\boldsymbol{R}}_3^{-1}\tilde{\boldsymbol{R}}_2^{\mathrm{T}}=\boldsymbol{R}_1^{-1}$ and

$$\min_{\dot{\boldsymbol{a}}(0)\in\mathbf{R}^n}(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\mathcal{X}}^{-1}\boldsymbol{x}_0)=\boldsymbol{a}_0^{\mathrm{T}}\boldsymbol{R}_1^{-1}\boldsymbol{a}_0 \tag{51}$$

So the optimum for the problem (47) is $\boldsymbol{a}_0^{\mathrm{T}}\boldsymbol{R}_1^{-1}\boldsymbol{a}_0$ and thus the controllability gramian of the second-order system is $\boldsymbol{X}_2=\boldsymbol{R}_1$. Similarly, the second-order observability gramian definition is based on the following optimization problem:

$$\max_{\dot{\boldsymbol{a}}(0)\in\mathbf{R}^n,\boldsymbol{y}\in L_2[0,\infty]}\left(\int_0^{\infty}\boldsymbol{y}^{\mathrm{T}}(t)\boldsymbol{y}(t)\mathrm{d}t\right),$$
$$\text{subject to } \boldsymbol{M}\ddot{\boldsymbol{a}}(t)+\boldsymbol{D}\dot{\boldsymbol{a}}(t)+\boldsymbol{K}\boldsymbol{a}(t)=\boldsymbol{B}\boldsymbol{u}(t),$$
$$\boldsymbol{a}(0)=\boldsymbol{a}_0 \tag{52}$$

If we compatibly partition the observability gramian of the first-order realization (46) as

$$\boldsymbol{\mathcal{Y}}=\begin{bmatrix}\boldsymbol{N}_1 & \boldsymbol{N}_2\\ \boldsymbol{N}_2^{\mathrm{T}} & \boldsymbol{N}_3\end{bmatrix} \tag{53}$$

then the observability gramian of the second-order system is $\boldsymbol{Y}_2=\boldsymbol{N}_1$. The eigenvalues of the gramian product $\boldsymbol{X}_2\boldsymbol{Y}_2$ are invariant under a similarity transformation. Let $\boldsymbol{W}$ and $\boldsymbol{V}$ be the dominant left and right eigenvectors of the gramian product $\boldsymbol{X}_2\boldsymbol{Y}_2$. A reduced second-order model can be obtained as $(\boldsymbol{M}_{\mathrm{r}},\boldsymbol{D}_{\mathrm{r}},\boldsymbol{K}_{\mathrm{r}},\boldsymbol{B}_{\mathrm{r}},\boldsymbol{C}_{\mathrm{r}})$ in which

$$\boldsymbol{M}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{V},\ \boldsymbol{D}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{V},\ \boldsymbol{K}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{V},$$
$$\boldsymbol{B}_{\mathrm{r}}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{B},\ \boldsymbol{C}_{\mathrm{r}}=\boldsymbol{C}\boldsymbol{V} \tag{54}$$

However, in order to preserve the symmetry and stability of the original system, an orthogonal projection is performed in Ref. [16] as follows:

$$\boldsymbol{M}_{\mathrm{r}}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{V},\ \boldsymbol{D}_{\mathrm{r}}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{V},\ \boldsymbol{K}_{\mathrm{r}}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{V},$$
$$\boldsymbol{B}_{\mathrm{r}}=\boldsymbol{V}^{\mathrm{T}}\boldsymbol{B},\ \boldsymbol{C}_{\mathrm{r}}=\boldsymbol{C}\boldsymbol{V} \tag{55}$$

where the equations are left multiplied by $\boldsymbol{V}$ instead of $\boldsymbol{W}$. Unfortunately, since $\boldsymbol{W}\neq\boldsymbol{V}$ for a non-symmetric system (46), the resulting gramian product $\boldsymbol{X}_2\boldsymbol{Y}_2$ will not be balanced and accuracy is sacrificed. In fact, this issue has been resolved[26], which is to be presented in the following section, where second-order systems are in a symmetric form,

$$\boldsymbol{M}=\boldsymbol{M}^{\mathrm{T}},\ \boldsymbol{D}=\boldsymbol{D}^{\mathrm{T}},\ \boldsymbol{K}=\boldsymbol{K}^{\mathrm{T}},\ \boldsymbol{B}=\boldsymbol{C}^{\mathrm{T}} \tag{56}$$

3 Fast TBR-Based Reduction Methods by Gramian Approximation

Although balanced truncation methods have been demonstrated to produce overall good approximations, they are, in general, too expensive for direct application to large interconnect circuit problems. In this section, we present efficient approaches to mitigating the high computing costs of standard TBR methods by gramian approximation. We show that the fast TBR method, and its various derivative methods, can deliver both high accuracy and reduction efficiency required for interconnect reductions for high performance integrated circuits.

3.1 Circuit formulation using susceptance matrices

RLCK circuits are a special class of dynamic systems. Corresponding circuit formulations are dynamic system models with special internal structures. In this section, we present circuit formulation that use the susceptance matrix for RLCK circuits, which stamp the $\boldsymbol{L}^{-1}$ (susceptance) matrix instead of the inductance matrix $\boldsymbol{L}$. This is because susceptance decays much faster than inductance, and thus, the L^{-1} matrix is easier to sparsify without causing stability issues[51].

3.1.1 First-order passive circuit formulation $(\mathcal{C},\mathcal{G},\mathcal{B})$

Recently, the branch vector potential $\boldsymbol{A}_l=\int \boldsymbol{E}_l\boldsymbol{v}\mathrm{d}t$ (and $\partial\boldsymbol{A}_l/\partial t=\boldsymbol{E}_l\boldsymbol{v}$) is introduced as a new state variable to obtain a first-order admittance that contains $\boldsymbol{L}^{-1}$ elements[52]. $\boldsymbol{E}_l$ here is the incident matrix for inductance in the MNA formulation. If we define $\boldsymbol{x}^{\mathrm{T}}=[\boldsymbol{v}^{\mathrm{T}},\boldsymbol{A}_l^{\mathrm{T}}]$, the first-order formulation, called vector-potential based nodal analysis (VNA)[52], is as follows:

$$\begin{aligned}\mathcal{C}\dot{\boldsymbol{x}}(t)&=-\mathcal{G}\boldsymbol{x}(t)+\mathcal{B}\boldsymbol{i}(t),\\ \boldsymbol{y}(t)&=\mathcal{B}^{\mathrm{T}}\boldsymbol{x}(t)\end{aligned}\tag{57}$$

where

$$\mathcal{C}=\begin{bmatrix}\boldsymbol{C}&0\\0&\boldsymbol{L}^{-1}\end{bmatrix},\quad \mathcal{G}=\begin{bmatrix}\boldsymbol{G}&\boldsymbol{E}_l\boldsymbol{L}^{-1}\\-\boldsymbol{L}^{-1}\boldsymbol{E}_l^{\mathrm{T}}&0\end{bmatrix},\quad \mathcal{B}=\begin{bmatrix}\boldsymbol{B}\\0\end{bmatrix}\tag{58}$$

Like MNA formulation, VNA matrices have the following properties:

$$\mathcal{C}=\mathcal{C}^{\mathrm{T}}\geqslant 0,\quad \mathcal{G}+\mathcal{G}^{\mathrm{T}}\geqslant 0\tag{59}$$

Hence the formulation is in a passive form described by the sufficient conditions of passivity in Condition 2. However, such a formulation is not in a symmetric form in Eq. (36) because $\mathcal{G}$ is not symmetric.

3.1.2 First-order symmetric circuit formulation $(\mathcal{C}_{\mathrm{s}},\mathcal{G}_{\mathrm{s}},\mathcal{B})$

It is easy to see that the formulation (58) can be rewritten into a symmetric formulation $(\mathcal{C}_{\mathrm{s}},\mathcal{G}_{\mathrm{s}},\mathcal{B})$,

$$\begin{aligned}\mathcal{C}_{\mathrm{s}}\dot{\boldsymbol{x}}(t)&=-\mathcal{G}_{\mathrm{s}}\boldsymbol{x}(t)+\mathcal{B}\boldsymbol{i}(t),\\ \boldsymbol{y}(t)&=\mathcal{B}^{\mathrm{T}}\boldsymbol{x}(t)\end{aligned}\tag{60}$$

where

$$\mathcal{C}_{\mathrm{s}}=\begin{bmatrix}\boldsymbol{C}&0\\0&-\boldsymbol{L}^{-1}\end{bmatrix},\quad \mathcal{G}_{\mathrm{s}}=\begin{bmatrix}\boldsymbol{G}&\boldsymbol{E}_l\boldsymbol{L}^{-1}\\\boldsymbol{L}^{-1}\boldsymbol{E}_l^{\mathrm{T}}&0\end{bmatrix},\quad \mathcal{B}=\begin{bmatrix}\boldsymbol{B}\\0\end{bmatrix}\tag{61}$$

Since both $\mathcal{G}_{\mathrm{s}}$ and $\mathcal{C}_{\mathrm{s}}$ are symmetric, this formulation falls into the class of systems in descriptor form Eq. (1) with additional symmetric conditions in Eq. (36). However, since $\mathcal{G}_{\mathrm{s}}$ and $\mathcal{C}_{\mathrm{s}}$ are no longer positive semi-definite, the sufficient conditions of passivity in Condition 2 are violated.

3.1.3 Second-order circuit formulation $(\boldsymbol{C},\boldsymbol{G},\boldsymbol{\Gamma},\boldsymbol{B})$

ENOR[7] stamps the nodal susceptance $\boldsymbol{\Gamma}=\boldsymbol{E}_l\boldsymbol{L}^{-1}\boldsymbol{E}_l^{\mathrm{T}}$ ($\boldsymbol{E}_l$ is the incident matrix for inductance) and results in a second-order NA (nodal analysis) form, which can be passively reduced[7,9]. The second-order formulation is as follows:

$$\begin{aligned}\boldsymbol{C}\dot{\boldsymbol{v}}(t)+\boldsymbol{G}\boldsymbol{v}(t)+\boldsymbol{\Gamma}\int\boldsymbol{v}(t)\mathrm{d}t&=\boldsymbol{B}\boldsymbol{i}(t),\\ \boldsymbol{y}(t)&=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{v}(t)\end{aligned}\tag{62}$$

where $\boldsymbol{i}(t)$, $\boldsymbol{y}(t)\in\mathbf{R}^p$ are input currents and output voltages; $\boldsymbol{v}(t)\in\mathbf{R}^n$ are nodal voltages; $\boldsymbol{G},\boldsymbol{C},\boldsymbol{\Gamma}\in\mathbf{R}^{n\times n}$ are matrices of conductance, capacitance and susceptance respectively; $\boldsymbol{B}\in\mathbf{R}^{n\times p}$ is the input matrix and its transpose $\boldsymbol{B}^{\mathrm{T}}\in\mathbf{R}^{p\times n}$ is the output matrix. An important property in second order formulation is that the system matrices are both symmetric and positive semi-definite,

$$\boldsymbol{C}=\boldsymbol{C}^{\mathrm{T}}\geqslant 0,\quad \boldsymbol{G}=\boldsymbol{G}^{\mathrm{T}}\geqslant 0,\quad \boldsymbol{\Gamma}=\boldsymbol{\Gamma}^{\mathrm{T}}\geqslant 0\tag{63}$$

which means the formulation fulfills both the sufficient conditions of passivity in Condition 2 and the symmetric conditions (56) for second-order systems.

It is easy to verify the formulations (57), (60), and (62), have the same transfer function,

$$\boldsymbol{H}(s)=\boldsymbol{B}^{\mathrm{T}}(\boldsymbol{C}s+\boldsymbol{G}+\boldsymbol{\Gamma}/s)^{-1}\boldsymbol{B}\tag{64}$$

Hence they are equivalent in terms of input-output behavior and either Eq. (57) or Eq. (60) can be viewed as a first-order realization of Eq. (62).

There is always a tradeoff in the first-order circuit formulation, either symmetric (implying accuracy) or positive semi-definite (implying passivity). Both can be obtained simultaneously only when the circuits are RC/RL circuits, where the formulations (62), (57), and (60) equal

$$\begin{aligned}\boldsymbol{C}\dot{\boldsymbol{v}}(t)&=-\boldsymbol{G}\boldsymbol{v}(t)+\boldsymbol{B}\boldsymbol{i}(t),\\ \boldsymbol{y}(t)&=\boldsymbol{B}^{\mathrm{T}}\boldsymbol{v}(t)\end{aligned}\tag{65}$$

with $\boldsymbol{C}=\boldsymbol{C}^{\mathrm{T}}\geqslant 0$ and $\boldsymbol{G}=\boldsymbol{G}^{\mathrm{T}}\geqslant 0$.

3.2 Single gramian approximation (SGA)

The first-order passive circuit formulation $(\mathcal{C},\mathcal{G},\mathcal{B})$ (58) falls into the class of systems in descriptor form in Eq. (1). In frequency domain, the controllability gramian $\mathcal{X}$ can also be computed from the expression,

$$\mathcal{X}=\frac{1}{2\pi}\int_{-\infty}^{+\infty}(\mathrm{j}\omega\mathcal{C}+\mathcal{G})^{-1}\mathcal{B}\mathcal{B}^{\mathrm{T}}(\mathrm{j}\omega\mathcal{C}+\mathcal{G})^{-\mathrm{H}}\mathrm{d}w\tag{66}$$

where superscript H denotes Hermitian transpose. A

gramian approximation process was proposed in Ref. [23]. Let ω_k be the k-th sampling point. If we define

$$z_k = (\mathrm{j}\omega_k \boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{G}})^{-1}\boldsymbol{\mathcal{B}} \tag{67}$$

then $\boldsymbol{\mathcal{X}}$ can be approximated as

$$\hat{\boldsymbol{\mathcal{X}}} = \sum w_k z_k z_k^{\mathrm{H}} = \boldsymbol{\mathcal{Z}}\boldsymbol{W}^2\boldsymbol{\mathcal{Z}}^{\mathrm{H}} \tag{68}$$

where $\boldsymbol{\mathcal{Z}} = [z_1, z_2, \ldots, z_q]$ and $\boldsymbol{W}$ is a diagonal matrix with diagonal entries $w_{kk} = \sqrt{w_k}$ and w_k comes from a specific numerical quadrature method. If we perform a singular value decomposition $\boldsymbol{\mathcal{Z}}\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}$, then we have

$$\hat{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{Z}}\boldsymbol{W}^2\boldsymbol{\mathcal{Z}}^{\mathrm{H}} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V})(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V})^{\mathrm{T}} =$$
$$\boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathrm{T}} = [\boldsymbol{U}_1 \quad \boldsymbol{U}_2]\begin{bmatrix}\boldsymbol{\Sigma}_1^2 & 0\\ 0 & \boldsymbol{S}_2^2\end{bmatrix}\begin{bmatrix}\boldsymbol{U}_1^{\mathrm{T}}\\ \boldsymbol{U}_2^{\mathrm{T}}\end{bmatrix} \tag{69}$$

where $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I}$. $\boldsymbol{U}$ converges to the eigenspace of $\boldsymbol{\mathcal{X}}$ and the dominant eigenvectors $\boldsymbol{U}_1$ can be used as the projection matrix. Note that since we are only interested in the dominant subspace in $\boldsymbol{\mathcal{Z}}$, the actual values of weights w_k do not matter as they are just a constant multiplied to each z_k. As a result, we will drop them in the rest of the paper. The reduced model $(\boldsymbol{\mathcal{C}}_{\mathrm{r}}, \boldsymbol{\mathcal{G}}_{\mathrm{r}}, \boldsymbol{\mathcal{B}}_{\mathrm{r}})$ can be obtained via an orthogonal projection

$$\boldsymbol{\mathcal{C}}_{\mathrm{r}} = \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{\mathcal{C}}\boldsymbol{U}_1, \quad \boldsymbol{\mathcal{G}}_{\mathrm{r}} = \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{\mathcal{G}}\boldsymbol{U}_1, \quad \boldsymbol{\mathcal{B}}_{\mathrm{r}} = \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{\mathcal{B}} \tag{70}$$

Since the orthogonal projection preserves the definiteness of matrices such that

$$\boldsymbol{\mathcal{C}}_{\mathrm{r}} = \boldsymbol{\mathcal{C}}_{\mathrm{r}}^{\mathrm{T}} \geqslant 0, \quad \boldsymbol{\mathcal{G}}_{\mathrm{r}} + \boldsymbol{\mathcal{G}}_{\mathrm{r}}^{\mathrm{T}} \geqslant 0 \tag{71}$$

the reduced-order system $(\boldsymbol{\mathcal{C}}_{\mathrm{r}}, \boldsymbol{\mathcal{G}}_{\mathrm{r}}, \boldsymbol{\mathcal{B}}_{\mathrm{r}})$ also fulfills the sufficient condition of passivity in Condition 2, which means passivity is preserved.

Given q sampling points and p inputs, the cost of SVD on matrix $\boldsymbol{\mathcal{Z}}_{n\times pq}$ is $O(n(pq)^2)$. In addition, it takes q matrix factorizations and pq matrix solutions. The total cost is $O(n(pq)^2 + qn^{\beta} + pqn^{\alpha})$ (typically, $1.1 \leqslant \beta \leqslant 1.5$ and $1 \leqslant \alpha \leqslant 1.2$ for circuits)[23], which is dominated by $O(qn^{\beta})$.

However, since the system is projected onto the dominant invariant subspace of controllability gramian only instead of onto the gramian product, the system is not balanced and accuracy will be sacrificed.

3.3 Double gramian approximation (DGA)

In fact, if the passivity constraint is removed, controllability and observability gramians can be approximated simultaneously. In order to achieve this, we start from the first-order symmetric circuit formulation $(\boldsymbol{\mathcal{C}}_{\mathrm{s}}, \boldsymbol{\mathcal{G}}_{\mathrm{s}}, \boldsymbol{\mathcal{B}})$ in Eq. (61), which falls into the class of systems in descriptor form, Eq. (1), with additional symmetric conditions like Eq. (36).

In this symmetric case, both gramians are equivalent, $\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}$. Based on the discussion of balancing a symmetric system in Section 2, we only need to find the dominant invariant subspace of an approximated gramian $\hat{\boldsymbol{\mathcal{X}}}$. As a result, the same gramian approximation process from Eqs. (66) to (70) can be performed.

Since both controllability and observability gramians are approximated simultaneously, accuracy will be improved. However, passivity is not guaranteed as the reduced matrices $\boldsymbol{\mathcal{C}}_{\mathrm{r}}$ and $\boldsymbol{\mathcal{G}}_{\mathrm{r}}$ are no longer positive semi-definite.

3.4 Second-order gramian approximation (SOGA)

In order to balance accuracy and passivity, a second-order balanced truncation method was proposed[26]. Given the second-order circuit formulation $(\boldsymbol{C}, \boldsymbol{G}, \boldsymbol{\Gamma}, \boldsymbol{B})$ in Eq. (62), the symmetric formulation $(\boldsymbol{\mathcal{C}}_{\mathrm{s}}, \boldsymbol{\mathcal{G}}_{\mathrm{s}}, \boldsymbol{\mathcal{B}})$ in Eq. (61) is employed as a first-order realization in descriptor form to define second-order gramians. Controllability and observability gramians in descriptor form can be computed from a pair of generalized Lyapunov Eq. (32). However, in this symmetric case, both gramians become the same and only one equation is to be solved,

$$\boldsymbol{\mathcal{C}}_{\mathrm{s}}\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{G}}_{\mathrm{s}} + \boldsymbol{\mathcal{G}}_{\mathrm{s}}\boldsymbol{\mathcal{X}}\boldsymbol{\mathcal{C}}_{\mathrm{s}}^{\mathrm{T}} - \boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^{\mathrm{T}} = 0 \tag{72}$$

If we compatibly partition the gramians as

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{Y}} = \begin{bmatrix}\boldsymbol{R}_1 & \boldsymbol{R}_2\\ \boldsymbol{R}_2^{\mathrm{T}} & \boldsymbol{R}_3\end{bmatrix} \tag{73}$$

then the second-order gramians, which measure the contribution of the node voltages v with respect to the transfer function[26], are also equal

$$\boldsymbol{X}_2 = \boldsymbol{Y}_2 = \boldsymbol{R}_1 \tag{74}$$

Since both gramians are equal, balanced truncation can be performed based on one gramian only. The reduced model $(\boldsymbol{C}_{\mathrm{r}}, \boldsymbol{G}_{\mathrm{r}}, \boldsymbol{\Gamma}_{\mathrm{r}}, \boldsymbol{B}_{\mathrm{r}})$ can be obtained by projecting onto the dominant invariant subspace of the gramian $\boldsymbol{R}_1$,

$$\boldsymbol{C}_{\mathrm{r}} = \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{C}\boldsymbol{U}_1, \boldsymbol{G}_{\mathrm{r}} = \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{G}\boldsymbol{U}_1,$$
$$\boldsymbol{\Gamma}_{\mathrm{r}} = \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{U}_1, \quad \boldsymbol{B}_{\mathrm{r}} = \boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{B} \tag{75}$$

This kind of projection is known as orthogonal projection, which preserves symmetry and definiteness of matrices such that $\boldsymbol{C}_{\mathrm{r}} = \boldsymbol{C}_{\mathrm{r}}^{\mathrm{T}} \geqslant 0$, $\boldsymbol{G}_{\mathrm{r}} = \boldsymbol{G}_{\mathrm{r}}^{\mathrm{T}} \geqslant 0$, $\boldsymbol{\Gamma}_{\mathrm{r}} =$

$\boldsymbol{\Gamma}_{\mathrm{r}}^{\mathrm{T}} \geqslant 0$, implying the reduced-order system also satisfies the sufficient passive conditions in Condition 2, and thus passivity is preserved[7].

Now, we approximate the invariant subspace of the second-order gramian. Given the first-order realization (60), in frequency domain, the controllability gramian $\boldsymbol{\mathcal{X}}$ can also be computed from the expression,

$$\boldsymbol{\mathcal{X}} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} (\mathrm{j}\omega\boldsymbol{\mathcal{C}}_{\mathrm{s}} + \boldsymbol{\mathcal{G}}_{\mathrm{s}})^{-1}\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^{\mathrm{T}}(\mathrm{j}\omega\boldsymbol{\mathcal{C}}_{\mathrm{s}} + \boldsymbol{\mathcal{G}}_{\mathrm{s}})^{-\mathrm{H}}\mathrm{d}w \tag{76}$$

Let ω_k be k-th sampling point. If we define

$$\boldsymbol{z}_k = (\mathrm{j}\omega_k\boldsymbol{\mathcal{C}}_{\mathrm{s}} + \boldsymbol{\mathcal{G}}_{\mathrm{s}})^{-1}\boldsymbol{\mathcal{B}} \tag{77}$$

then $\boldsymbol{\mathcal{X}}$ can be approximated as $\hat{\boldsymbol{\mathcal{X}}} = \sum \boldsymbol{z}_k \boldsymbol{z}_k^{\mathrm{H}} = \boldsymbol{\mathcal{Z}}\boldsymbol{\mathcal{Z}}^{\mathrm{H}}$ where $\boldsymbol{\mathcal{Z}} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_q]$. If we compatibly partition $\boldsymbol{\mathcal{Z}}^{\mathrm{H}} = [\boldsymbol{\mathcal{Z}}_1^{\mathrm{H}} \quad \boldsymbol{\mathcal{Z}}_2^{\mathrm{H}}]$, then we have

$$\hat{\boldsymbol{\mathcal{X}}} = \begin{bmatrix} \hat{\boldsymbol{R}}_1 & \hat{\boldsymbol{R}}_2 \\ \hat{\boldsymbol{R}}_2^{\mathrm{T}} & \hat{\boldsymbol{R}}_3 \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_1\boldsymbol{Z}_1^{\mathrm{H}} & \boldsymbol{Z}_1\boldsymbol{Z}_2^{\mathrm{H}} \\ \boldsymbol{Z}_2\boldsymbol{Z}_1^{\mathrm{H}} & \boldsymbol{Z}_2\boldsymbol{Z}_2^{\mathrm{H}} \end{bmatrix} \tag{78}$$

So the approximated second-order gramian $\hat{\boldsymbol{R}}_1$ equals $\boldsymbol{Z}_1\boldsymbol{Z}_1^{\mathrm{H}}$ and can be diagonalized as follows:

$$\hat{\boldsymbol{R}}_1 = \boldsymbol{Z}_1\boldsymbol{Z}_1^{\mathrm{H}} = (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V})(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V})^{\mathrm{T}} = \boldsymbol{U}\boldsymbol{\Sigma}^2\boldsymbol{U}^{\mathrm{T}} = [\boldsymbol{U}_1 \quad \boldsymbol{U}_2]\begin{bmatrix} \boldsymbol{\Sigma}_1^2 & 0 \\ 0 & \boldsymbol{\Sigma}_2^2 \end{bmatrix}\begin{bmatrix} \boldsymbol{U}_1^{\mathrm{T}} \\ \boldsymbol{U}_2^{\mathrm{T}} \end{bmatrix} \tag{79}$$

which means $\boldsymbol{U}_1$ is the dominant invariant subspace of $\hat{\boldsymbol{R}}_1$ and a reduced model $(\boldsymbol{C}_{\mathrm{r}}, \boldsymbol{G}_{\mathrm{r}}, \boldsymbol{\Gamma}_{\mathrm{r}}, \boldsymbol{B}_{\mathrm{r}})$ can be obtained by projecting onto $\boldsymbol{U}_1$ like Eq. (75).

3.5 Cross-gramian approximation (CGA)

The cross-gramian $\boldsymbol{X}_{\mathrm{CG}}$ is introduced, which contains both controllability and observability information in a single matrix. Given a system in descriptor form (1), $\boldsymbol{X}_{\mathrm{CG}}$ can be calculated from the Sylvester equation,

$$\boldsymbol{A}\boldsymbol{X}_{\mathrm{CG}}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{X}_{\mathrm{CG}}\boldsymbol{A} = -\boldsymbol{B}\boldsymbol{C} \tag{80}$$

The cross-gramian reduction method is based on projecting onto the eigenspaces related to the dominant eigenvalues of $\boldsymbol{X}_{\mathrm{CG}}$.

For symmetric systems, both of the two Lyapunov equations used in balanced truncation will be the same as the Sylvester equation in the cross-gramian method. Also, both controllability and observability gramians are identical to the cross-gramian and $\boldsymbol{X}\boldsymbol{Y} = \boldsymbol{X}_{\mathrm{CG}}^2$. These properties make the cross-gramian method equivalent to the balanced truncation method for symmetric models.

Given the circuit formulation (58), in the frequency domain, $\boldsymbol{\mathcal{X}}_{\mathrm{CG}}$ is expressed as

$$\boldsymbol{\mathcal{X}}_{\mathrm{CG}} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} (\mathrm{j}\omega\boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{G}})^{-1}\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{B}}^{\mathrm{T}}(\mathrm{j}\omega\boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{G}})^{-1}\mathrm{d}w \tag{81}$$

Let ω_k be the k-th sampling point. If we define

$$\boldsymbol{z}_{\mathrm{c}_k} = (\mathrm{j}\omega_k\boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{G}})^{-1}\boldsymbol{\mathcal{B}}, \quad \boldsymbol{z}_{\mathrm{o}_k} = (\mathrm{j}\omega_k\boldsymbol{\mathcal{C}} + \boldsymbol{\mathcal{G}}^{\mathrm{T}})^{-1}\boldsymbol{\mathcal{B}} \tag{82}$$

then $\hat{\boldsymbol{\mathcal{X}}}_{\mathrm{CG}}$ can be computed as

$$\hat{\boldsymbol{\mathcal{X}}}_{\mathrm{CG}} = \sum \boldsymbol{z}_{\mathrm{c}_k}\boldsymbol{z}_{\mathrm{o}_k}^{\mathrm{H}} = \boldsymbol{\mathcal{Z}}_{\mathrm{c}}\boldsymbol{\mathcal{Z}}_{\mathrm{o}}^{\mathrm{H}} \tag{83}$$

where $\boldsymbol{\mathcal{Z}}_{\mathrm{c}}$ and $\boldsymbol{\mathcal{Z}}_{\mathrm{o}}$ are matrices whose columns are $\boldsymbol{z}_{\mathrm{c}_k}$ and $\boldsymbol{z}_{\mathrm{o}_k}$, respectively. In order to find the eigenvectors of $\boldsymbol{\mathcal{X}}_{\mathrm{CG}}$, we need to do the eigen-decomposition on $\boldsymbol{\mathcal{Z}}_{\mathrm{c}}\boldsymbol{\mathcal{Z}}_{\mathrm{o}}^{\mathrm{H}}$ and a reduced order model can be obtained by a projection onto the left and right dominant eigenspaces $\boldsymbol{W}_1$ and $\boldsymbol{V}_1$.

Since $\boldsymbol{\mathcal{Z}}_{\mathrm{c}}\boldsymbol{\mathcal{Z}}_{\mathrm{o}}^{\mathrm{H}}$ is not symmetric, the projection is done by an oblique projection $\boldsymbol{W}_1 \neq \boldsymbol{V}_1$ and the passivity of the reduced model is not guaranteed. In addition, since an eigenanalysis of a full matrix is still needed, the computational cost can not be significantly reduced.

The comparison of four gramian methods that we have reviewed are summarized in Table 1. In terms of the gramian used, DGA takes into consideration both controllability and observability gramians. SOGA considers second-order controllability and observability gramians, which are the leading blocks of the corresponding first-order gramians. CGA considers controllability and observability in a single cross-gramian. SGA includes the controllability gramian only. Passivity is not preserved in either DGA or CGA. In the first case, the circuit matrix is symmetric but not positive semi-definite. In the second case, the projection is not orthogonal.

Table 1 Time complexity comparison ($1.1 \leqslant \beta \leqslant 1.5$ for sparse matrices. q is the number of sampling points)

Methods	Gramians used	Passivity	CPU cost
SGA	contr.	yes	$O(qn^{\beta})$
DGA	contr. observ.	no	$O(qn^{\beta})$
CGA	cross	no	$O(n^3)$
SOGA	contr. observ (2nd)	yes	$O(qn^{\beta})$

3.6 Response gramian approximation (RGA)

Model order reduction of circuits with many ports remains a difficult problem. Fundamentally, with more ports, we need to have more states in order to

observe all the port behaviors. However, when input information is available as *a priori*, this issue can be resolved by including the input signals as part of the system and converting a MIMO system into a SIMO system.

As a result, the reduction essentially becomes a simulation process as we need the input signals, and the reduced models cannot work for different input signals. But such a reduction based approach is still a viable simulation approach for a linear system. Early approaches using such a strategy in the Krylov subspace framework, called (improved) extended Krylov subspace methods (EKS/IEKS) have been proposed[53,54].

In the ETBR (extended TBR) method proposed in Ref. [25], a response gramian in the frequency domain is defined as

$$\boldsymbol{\mathcal{X}}_{\mathrm{r}}=\frac{1}{2\pi}\int_{-\infty}^{+\infty}(\mathrm{j}\omega\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{G}})^{-1}\boldsymbol{\mathcal{B}}\boldsymbol{u}(\mathrm{j}\omega)\cdot \boldsymbol{u}^{\mathrm{T}}(\mathrm{j}\omega)\boldsymbol{\mathcal{B}}^{\mathrm{T}}(\mathrm{j}\omega\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{G}})^{-\mathrm{H}}\mathrm{d}\omega \tag{84}$$

Let ω_k be the k-th sampling point over the frequency range. If we further define

$$\boldsymbol{z}_k=(\mathrm{j}\omega_k\boldsymbol{\mathcal{C}}+\boldsymbol{\mathcal{G}})^{-1}\boldsymbol{\mathcal{B}}\boldsymbol{u}(\mathrm{j}\omega_k) \tag{85}$$

then $\boldsymbol{\mathcal{X}}_{\mathrm{r}}$ can be approximately computed as

$$\hat{\boldsymbol{\mathcal{X}}}_{\mathrm{r}}=\sum \boldsymbol{z}_k\boldsymbol{z}_k^{\mathrm{H}}=\boldsymbol{\mathcal{Z}}_{\mathrm{r}}\boldsymbol{\mathcal{Z}}_{\mathrm{r}}^{\mathrm{H}} \tag{86}$$

where $\boldsymbol{\mathcal{Z}}_{\mathrm{r}}$ is a matrix whose columns are $\boldsymbol{z}_k$. The projection matrix can be obtained by performing a singular value decomposition on $\boldsymbol{\mathcal{Z}}_{\mathrm{r}}$. After this, we can reduce the original matrices into smaller ones, and then perform the transient analysis on the reduced circuit matrices. Note that we need the frequency spectra of input signal $\boldsymbol{u}$ in Eq. (85). This can be obtained by fast Fourier transformation on the input signals in time domain.

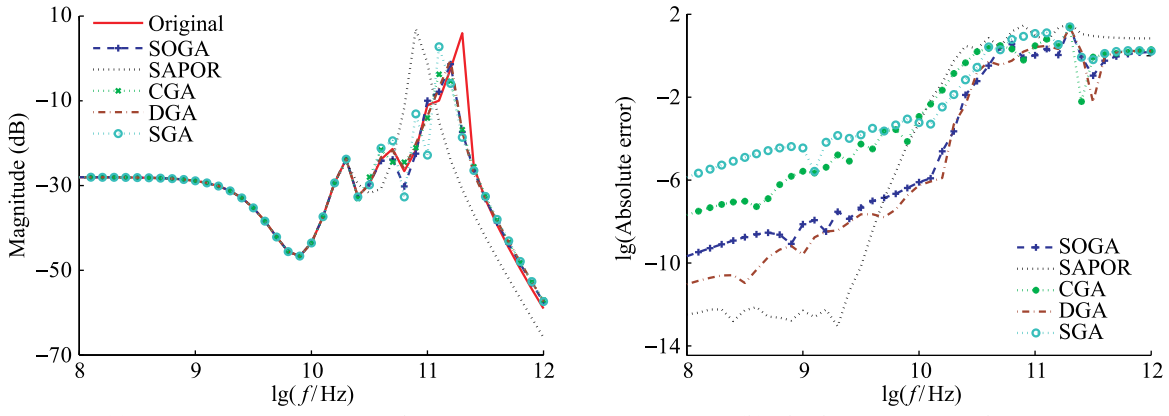
4 Numerical Results

In this section, we present some numerical results to compare the gramian approximation techniques presented in this paper. All the proposed methods have been implemented in Matlab 7.0. Besides the TBR based methods, the second-order Krylov subspace based reduction method, SAPOR[9], is also implemented and compared.

We remark that we do not want to seek comprehensive numerical results and complete comparison due to the nature of this paper. Instead, we want to present some numerical data so that observations and insight can be made to inspire future research and investigation of this important topic.

4.1 Performance of gramian based modeling methods

The original model is an RLC mesh with original order 640, where the values of R,L,C are randomly generated at the order of $10^{-1}\,\Omega$, 10^{-12} H, 10^{-12} F, respectively. The reduced order is set to be 20 for all methods. As shown in Fig. 2, gramian approximation based methods (SGA, DGA, CGA, SOGA) offers better approximation than the Krylov subspace based second-order method SAPOR[9] over a wide frequency band. However, SAPOR is more accurate at low frequencies close to the expansion point (0.1 GHz). In this example, we observe that DGA is the most accurate method, as it takes into consideration both controllability and observability gramians. SOGA is

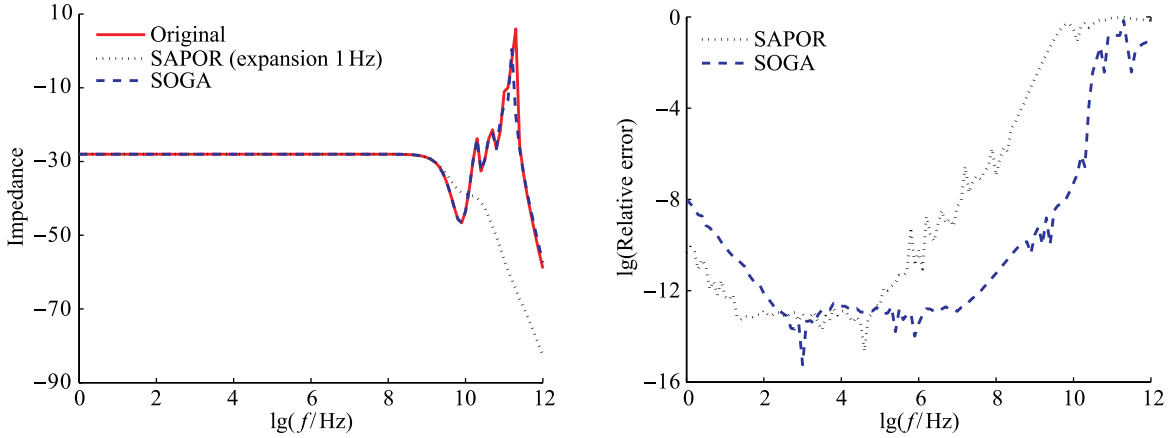


(a) Frequency response comparison (b) Absolute error comparison

Fig. 2 Comparison of different gramian approximation methods with SAPOR

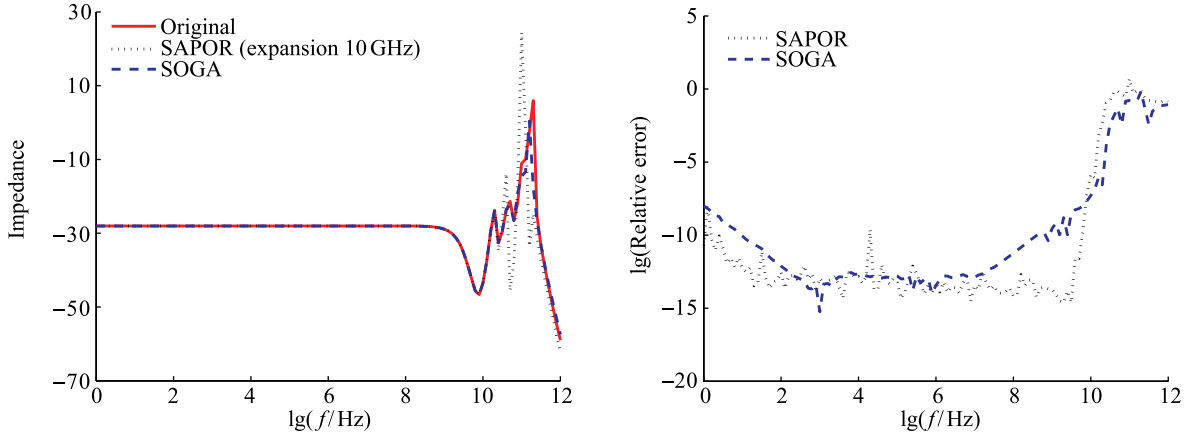
ranked in second place, and it considers second-order controllability and observability gramians, which are the leading blocks of the corresponding first-order gramians. CGA is ranked in third place, which considers controllability and observability in a single cross-gramian. SGA is in last place, which includes controllability gramian only.

Since DC performance is more important in practice, the expansion point of the Krylov subspace (moment matching) based methods is often chosen at low frequencies. In fact, the performance of these methods are dependent on the choices of expansion points, and choices at higher frequencies may result in better global performance. However, if a compact model is desired, gramian approximation methods often achieve better global accuracy than moment-matching based methods. Now we show this point as follows: First, the expansion point is chosen at 1 Hz. As shown in Fig. 3, SAPOR is more accurate than SOGA at low frequencies around the expansion point but it cannot match the dynamics at high frequencies at all. Next, we shift the expansion point to 10 GHz. As shown in Fig. 4, the performance of SAPOR is much better than in the first case. However, it cannot match the high frequency dynamics well either. After that, we further shift the expansion point to higher frequencies (100 GHz). In this case, while the high frequency performance is much better, the low frequency accuracy deteriorates. Noticeable error can be observed at low frequencies as shown in Fig. 5.



(a) Frequency response comparison (b) Relative error comparison

Fig. 3 Accuracy comparison between SOGA and SAPOR expanded at 1 Hz



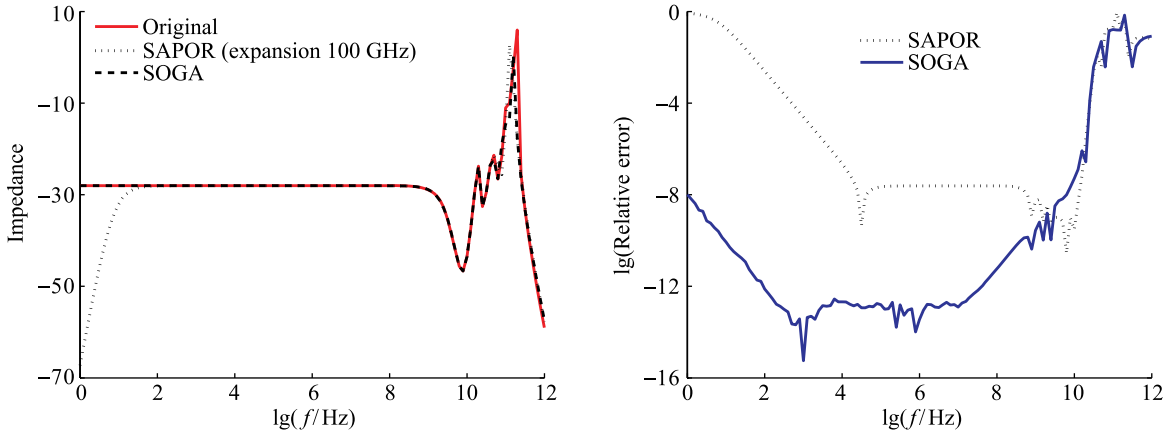
(a) Frequency response comparison (b) Relative error comparison

Fig. 4 Accuracy comparison between SOGA and SAPOR expanded at 10 GHz

4.2 Performance of gramian based simulation methods

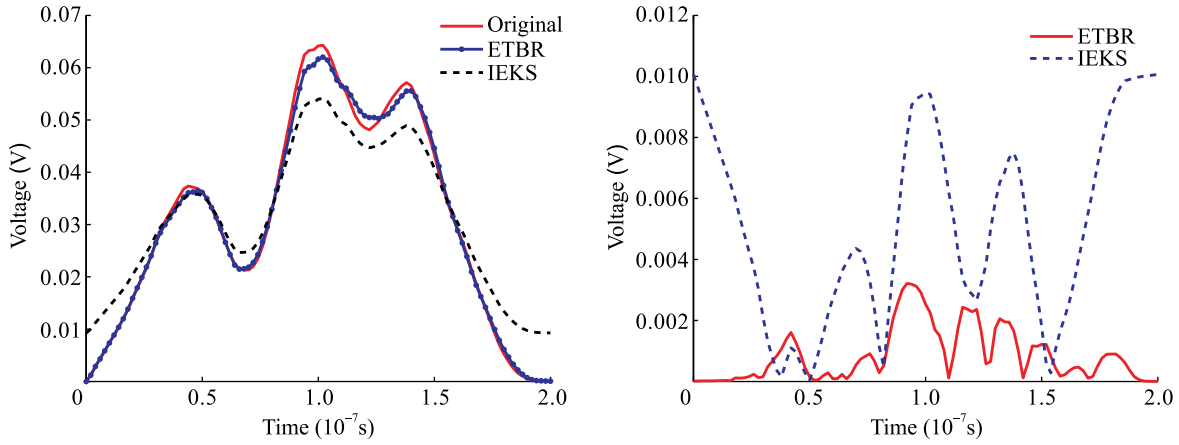
We then compare ETBR[25] with the Krylov subspace based simulation method EKS[53,54]. The test circuit has 10 000 nodes and 100 sources. The reduction order q is set to 6 for EKS and the number of frequency samples used for ETBR is also set to 6. Figure 6 shows

the simulation results of ETBR and EKS at the 200th node. The simulation errors compared with SPICE results are also shown in Fig. 6. We can see that ETBR is more accurate than EKS over the entire simulation time.



(a) Frequency response comparison (b) Relative error comparison

Fig. 5 Accuracy comparison between SOGA and SAPOR expanded at 100 GHz



(a) Transient simulation comparison (b) Absolute error comparison

Fig. 6 Accuracy comparison between ETBR and EKS

Table 2 shows the CPU times of both ETBR (including the cost of FFT) and EKS on the given set of circuits using the same reduction order $q = 6$. We find that EKS is a bit faster for small circuits. But for Ckt6 and larger circuits, the CPU times are almost the same for both methods. For the largest circuit Ckt8, EKS is unable to even finish owing to memory constraint, while ETBR runs through all the circuits. This clearly shows that ETBR is more memory efficient than EKS using a non-LU decomposition solver.

Table 2 CPU times (in seconds) comparison of ETBR and EKS ($q = 6$)

Test	#Nodes	#Sources	ETBR (s)	EKS (s)
Ckt1	1000	100	0.23	0.08
Ckt2	10 000	100	1.28	0.89
Ckt3	10 000	1000	1.80	1.40
Ckt4	100 000	1000	20.4	18.8
Ckt5	100 000	4000	28.6	25.3
Ckt6	500 000	5000	152.0	151.0
Ckt7	500 000	20 000	162.0	160.0
Ckt8	1 000 000	50 000	562.0	out of memory

5 Summary and Discussion

In this paper, we have presented the recent development in non-Krylov subspace based reduction techniques. We first reviewed classic truncated balanced realization methods as a backdrop. Then we presented several recently proposed reduction methods using fast truncated balanced realization methods. Among them are gramian approximation techniques such as single gramian approximation (SGA), double

gramian approximation (DGA), second-order gramian approximation (SOGA), cross-gramian approximation (CGA), and response gramian approximation (RGA) for model order reduction of RLCK circuits. We presented the reduction methods based on both first-order formulation and second-order formulations. In addition, we furher provided some numerical comparison results for the presented methods. We showed that fast TBR methods typically produce a better approximation over a wide frequency range than the corresponding Krylov subspace based methods. We show that each fast TBR method has their pros and cons in terms of accuracy, efficiency, numerical stability, structure-preserving and passivity-preserving properties.

Still many challenging problems remain unsolved such as reduction for interconnect circuits with massive ports, wide frequency band, and high-fidelity accuracy reduction for RF/analog circuits. More research efforts are still required to develop new efficient and accurate reduction techniques to meet the increasing demands of analyzing massive interconnect parasites in nanometer VLSI systems.

References

[1] Tan S X-D, He L. Advanced Model Order Reduction Techniques in VLSI Design. Cambridge University Press, 2007.

[2] Antoulas A C. Approximation of Large-Scale Dynamical Systems. The Society for Industrial and Applied Mathematics (SIAM), 2005.

[3] Pillage L T, Rohrer R A. Asymptotic waveform evaluation for timing analysis. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 1990: 352-366.

[4] Feldmann P, Freund R W. Efficient linear circuit analysis by Pade approximation via the Lanczos process. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 1995, **14**: 639-649.

[5] Silveira M, Kamon M, Elfadel I, et al. A coordinate transformed Arnoldi algorithm for generating uaranteed stable reduced-order models of RLC circuits. In: Proc. Int. Conf. on Computer Aided Design (ICCAD), 1996.

[6] Odabasioglu A, Celik M, Pileggi L. PRIMA: Passive reduced-order interconnect macromodeling algorithm. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 1998, **17**(8): 645-654.

[7] Sheehan B N. ENOR: Model order reduction of RLC circuits using nodal equations for efficient factorization. In: Proc. Design Automation Conf. (DAC), 1999.

[8] Freund R W. SPRIM: Structure-preserving reduced-order interconnect macromodeling. In: Proc. Int. Conf. on Computer Aided Design (ICCAD), 2004.

[9] Su Y, Wang J, Zeng X, et al. SAPOR: Second-order Arnoldi method for passive order reduction of RCS circuits. In: Proc. Int. Conf. on Computer Aided Design (ICCAD), 2004.

[10] Kerns K J, Yang A T. Stable and efficient reduction of large, multiport RC network by pole analysis via congruence transformations. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 1998, **16**: 734-744.

[11] Qi Z, Yu H, Liu P, et al. Wideband passive multi-port model order reduction and realization of RLCM circuits. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2006, **25**: 1496-1509.

[12] Moore B. Principal component analysis in linear systems: Controllability, and observability, and model reduction. *IEEE Trans. Automat. Contr.*, 1981, **26**(1): 17-32.

[13] Glover K. All optimal Hankel-norm approximations of linear multi-variable systems and their L_∞ error bounds. *Int. J. Control*, 1984, **36**: 1115-1193.

[14] Desai U, Pal D. A transformation approach to stochastic model reduction. *IEEE Trans. Automat. Contr.*, 1984, **29**: 1097-1100.

[15] Harshavardhana P, Jonckheere E, Silverman L. Stochastic balancing and approximation-stability and minimality. *IEEE Trans. Automat. Contr.*, 1984, **29**: 744-746.

[16] Meyer D G, Srinivasan S. Balancing and model reduction for second-order form linear systems. *IEEE Trans. Automat. Contr.*, 1996, **AC-41**: 1632-1644.

[17] Li J R, Wang F, White J. An efficient Lyapunov equation-based approach for generating reduced-order models of interconnect. In: Proc. Design Automation Conf. (DAC), 1999.

[18] Li J R. Model reduction of large linear systems via low rank system gramians [Dissertation]. Cambridge, MA, USA: MIT, 2002.

[19] Phillips J R, Daniel L, Silveira L M. Guaranteed passive balancing transformation for model order reduction. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2003, **22**(8): 1027-1041.

[20] Wang N, Balakrishnan V. Fast balanced stochastic truncation via a quadratic extension of the alternating direction implicit iteration. In: Proc. Int. Conf. on Computer Aided Design (ICCAD), 2005.

[21] Vasilyev D, White J. A more reliable reduction algorithm for behavioral model extraction. In: Proc. Int. Conf. on

Computer Aided Design (ICCAD), 2005.

[22] Wang N, Balakrishnan V, Koh C-K. Passivity preserving model reduction via a computationally efficient projection-and-balance scheme. In: Proc. Design Automation Conf. (DAC), 2004.

[23] Phillips J R, Silveira L M. Poor man's TBR: A simple model reduction scheme. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2005, **24**(1): 43-55.

[24] Yan B, Tan S X-D, Liu P, et al. SBPOR: Second-order balanced truncation for passive model order reduction of RLC circuits. In: Proc. Design Automation Conf. (DAC), 2007.

[25] Li D, Tan S X-D, McGaughy B. ETBR: Extended truncated balanced realization method for on-chip power grid network analysis. In: Proc. European Design and Test Conf. (DATE), 2008.

[26] Yan B, Tan S X-D, McGaughy B. Second-order balanced truncation for passive-model order reduction of RLCK circuits. *IEEE Trans*. *on Circuits and Systems II*: *Express Briefs*, 2008, **55**(9): 942-946.

[27] Gugercin S, Sorensen D C, Antoulas A C. A modified low-rank Smith method for large-scale Lyapunov equations. *Numerical Algorithms*, 2003, **32**: 27-55.

[28] Sorensen D C, Antoulas A C. The Sylvester equation and approximate balanced reduction. *Linear Algebra and Its Application*, 2002, **351-352**: 671-700.

[29] Zhou Y. Numerical methods for large scale matrix equations with applications in LTI system model reduction [Dissertation]. Rice University, 2002.

[30] Stykel T. Grammian-based model order reduction for descriptor systems. *Math. Control Signals Systems*, 2004, **16**: 297-319.

[31] Reis T, Stykel T. Positive real and bounded real balancing for model reduction of descriptor systems. http://www.math.tuberlin.de/stykel/, 2008.

[32] Reis T, Stykel T. Passivity-preserving balanced truncation for electrical circuits. http://www.math.tu-berlin.de/stykel/, 2008.

[33] Ionutiu R, Lefteriu S, Antoulas A C. Comparison of model reduction methods with applications to circuit simulation. In: Scientific Computing in Electrical Engineering. Berlin: Springer, 2007: 3-24.

[34] Sorensen D C, Antoulas A C. On model reduction of structured systems. In: Dimension Reduction of Large-Scale Systems. 2005: 125-138.

[35] Gugercin S, Antoulas A C. A survey of model reduction by balanced truncation and some new results. *Int. J. Control*, 2004, **77**: 748-766.

[36] Banner P. Advances in balaning-related model reduction for circuit simulation. http://www-user.tuchemnitz.de/benner/talks/Benner SCEE2008.pdf. 2008.

[37] Laub A J, Heath M T, Paige C C, et al. Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms. *IEEE Trans. Automat. Contr.*, 1987, **32**: 115-122.

[38] Safonov M G, Chiang R Y. A Schur method for balanced truncation model reduction. *IEEE Trans. Automat. Contr.*, 1989, **34**: 729-733.

[39] Kagstrom B, Dooren P V. A generalized state-space approach for the additive decomposition of a transfer matrix. *Journal of Linear Algebra and Applications*, 1992, **1**: 165-181.

[40] Willcox K, Peraire J. Balanced model reduction via the proper orthogonal decomposition. *AIAA Journal*, 2002, **40**(11): 2323-2330.

[41] Elias P, van der Meijs N. Including higher-order moments of RC interconnections in layout-to-circuit extraction. In: Proc. European Design and Test Conf. (DATE), 1996.

[42] Sheehan B N. TICER: Realizable reduction of extracted RC circuits. In: Proc. Int. Conf. on Computer Aided Design (ICCAD), 1999.

[43] Amin C S, Chowdhury M H, Ismail Y I. Realizable RLCK circuit crunching. In: Proc. Design Automation Conf. (DAC), 2003.

[44] Qin Z, Cheng C. Realizable parasitic reduction using generalized Y - transformation. In: Proc. Design Automation Conf. (DAC), 2003.

[45] Sheehan B N. Branch merge reduction of RLCM networks. In: Proc. Int. Conf. on Computer Aided Design (ICCAD). 2003.

[46] Tan S X-D. A general s-domain hierarchical network reduction algorithm. In: Proc. Int. Conf. on Computer Aided Design (ICCAD), 2003.

[47] Tan S X-D, Qi Z, Li H. Hierarchical modeling and simulation of large analog circuits. In: Proc. European Design and Test Conf. (DATE), 2004.

[48] Brune O. Synthesis of a finite two-terminal network whose driving point impedance is a prescribed function of frequency. *Journal of Math. and Phys.*, 1931, **10**: 191-236.

[49] Green M. Balanced stochastic realization. *Linear Algebra and Its Application*, 1988, **98**: 211-247

[50] Chen X, Wen J T. Positive realness preserving model reduction with H_1 norm error bounds. *IEEE Trans. on*

Circuits and Systems I: *Fundamental Theory and Applications*, 1995, **42**: 23-29.

[51] He Z, Celik M, Pillegi L. SPIE: Sparse partial inductance extraction. In: Proc. Design Automation Conf. (DAC), 1997.

[52] Yu H, Shi Y, He L, et al. A fast block structure preserving model order reduction in inverse inductance circuits. In: Proc. Int. Conf. on Computer Aided Design (ICCAD), 2006.

[53] Wang J M, Nguyen T V. Extended Krylov subspace method for reduced order analysis of linear circuit with multiple sources. In: Proc. Design Automation Conf. (DAC), 2000.

[54] Lee Y, Cao Y, Chen T, et al. HiPRIME: Hierarchical and passivity preserved interconnect macromodeling engine for RLKC power delivery. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 2005, **24**(6): 797-806.