Passivity-Preserving Model Reduction Via A Computationally Efficient Project-And-Balance Scheme*

N. Wong
Department of Electrical and Electronic Engineering
The University of Hong Kong
Pokfulam Road, Hong Kong
nwong@eee.hku.hk

V. Balakrishnan
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN USA
ragu@ecn.purdue.edu

C.-K. Koh
School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN USA
chengkok@ecn.purdue.edu

ABSTRACT

This paper presents an efficient two-stage project-and-balance scheme for passivity-preserving model order reduction. Orthogonal dominant eigenspace projection is implemented by integrating the Smith method and Krylov subspace iteration. It is followed by stochastic balanced truncation wherein a novel method, based on the complete separation of stable and unstable invariant subspaces of a Hamiltonian matrix, is used for solving two dual algebraic Riccati equations at the cost of essentially one. A fast-converging quadruple-shift bulge-chasing SR algorithm is also introduced for this purpose. Numerical examples confirm the quality of the reduced-order models over those from conventional schemes.

Categories and Subject Descriptors

I.6.5 [**Simulation and Modeling**]: Model Development—*modeling methodologies*; J.6 [**Computer-Aided Engineering**]: —*computer-aided design (CAD)*

General Terms

Algorithms

Keywords

Model Reduction, Dominant Eigenspace, Projection, Stochastic Balanced Truncation, Riccati Equation, SR Algorithm

*This work was supported in part by the Hong Kong Research Grants Council, the University Research Committee of The University of Hong Kong, and the National Science Foundation of the United States of America under Grant No. ECS-0200320.

DAC 2004, June 7–11, 2004, San Diego, California, USA.

1. INTRODUCTION

Model reduction is an integral part in modern system design (e.g., [1–15]). Initial state space modeling of a physical system often involves thousands or millions of state variables, rendering the direct simulation and analysis computationally prohibitive. A common strategy is to reduce the order of the original model, or when the original model is big, divide and reduce smaller subsystems which are then connected back for global simulation. Subsequently, it is desirable that the reduced-order models have small approximation error over the frequency and/or time-domain(s). Also, important properties such as stability and passivity[1] must be preserved during the reduction in order for the smaller models to be meaningful (e.g., a non-passive reduced-order model can generate erroneous responses when connected to other passive systems or nonlinear circuits [1, 2]).

In the context of VLSI synthesis, interconnect modeling is becoming more and more critical as circuits are designed with ever-increasing speed and complexity. Using methods such as sparse tableau, nodal formulation and modified nodal analysis (MNA), etc. [16], the RLC interconnect model extracted from a netlist often results in an order not amenable to simulation or analysis. Techniques to carry out model reduction include transfer function moment matching (e.g., asymptotic waveform evaluation (AWE) [3]), Krylov subspace projections (e.g., Padé Approximation via Lanczos (PVL), matrix PVL (MPVL) [4]), and the passivity-preserving congruence transform (e.g., PRIMA [1]). These schemes, usually implemented by Krylov methods, are computationally effective, but suffer from the lack of optimality. Another class of model reduction techniques stems from control theory. Examples of which are optimal Hankel-norm approximation [5], standard balanced truncation (BT) [6–9], and the passivity-preserving stochastic balanced truncation (SBT) [10–13]. A merit of these control-theoretic approaches is the availability of error bounds for the approximation errors [5, 11] and the superior global accuracy. These schemes are, however, expensive to deploy due to the need of solving large-size matrix equations and factorizations. In view of this, a two-stage reduction is proposed to provide tradeoff between computational cost and model quality (e.g., [12, 14, 15]). Such approach starts with (dominant) gramian eigenspace projection to reduce the original

[1]Roughly speaking, a passive system is one that cannot generate energy.

models to some moderately sized ones (say, order less than a hundred), followed by BT or SBT to generate further-reduced models (say, order less than a few tens).

The contribution of this paper is a computationally efficient two-stage project-and-balance model reduction algorithm with stability and passivity-preserving features. The initial projection basis is formed by the dominant eigenspaces of the controllability and observability gramians [7, 14, 15]. Utilizing the Smith method and Krylov subspace iteration in solving Lyapunov equations, we show that orthogonal bases for the eigenspaces are readily obtained. The second stage of the reduction is done by SBT via the solution of a pair of dual continuous-time algebraic Riccati equations (CAREs). A novel observation arising from the Hamiltonian invariant subspaces reveals that two CAREs can be solved simultaneously at the cost of essentially one. The idea relies on completely separating the stable and unstable invariant subspaces, and a fast-converging quadruple-shift SR algorithm (analogous to the QR algorithm [17]) to achieve this is described. Numerical examples show that the proposed scheme has a competitive computational cost and produces low-order models with excellent accuracies.

Section 2 provides the state-space formulation and background on gramian eigenspaces and SBT. Section 3 shows how the bases of gramian eigenspaces are obtained from the solution of Lyapunov equations via the Smith method. The intermediate model thus formed is then stochastically balanced and truncated by simultaneously solving two dual CAREs as in Section 4. Numerical examples in Section 5 demonstrate the effectiveness of the proposed scheme over conventional ones. Section 6 draws the conclusion.

2. PROBLEM SETTINGS

A target application of the proposed algorithm is in the model reduction of large-scale RLC circuits commonly encountered in VLSI interconnect and pin package simulations. Consider a state space model of

$$\dot{x} = Ax + Bu \tag{1a}$$
$$y = Cx + Du \tag{1b}$$

where $A \in \Re^{n\times n}$, $B \in \Re^{n\times m}$, $C \in \Re^{m\times n}$, $D \in \Re^{m\times m}$, B, C are of low ranks, i.e., $m \ll n$, and u, y are power-conjugate[2]. Using $M > 0$ ($M \geq 0$) to denote a positive definite (positive semidefinite) matrix M, we assume $D + D^T > 0$. For RLC state space models in MNA format, we also have $A + A^T \leq 0$, $B = C^T$, and $D = 0$. In [9], it is shown that such a system can be transformed into an equivalent system with $D + D^T > 0$. Moreover, a system in *descriptor* format [2] with a singular E matrix preceeding $\dot{x}$ can be transformed into the standard form in (1) [8], so the settings of (1) are assumed without loss of generality.

The controllability gramian, W_c, and observability gramian, W_o, are solved through the following Lyapunov equations

$$AW_c + W_cA^T + BB^T = 0 \tag{2a}$$
$$A^TW_o + W_oA + C^TC = 0 \tag{2b}$$

The spans (ranges) of W_c and W_o denote the reachable and observable sets of the states, respectively. For many physical systems including RLC circuits, W_c and W_o are of low ranks. On the other hand, the positive real lemma [2] states that a system in (1) is passive if and only if there exists a $P \in \Re^{n\times n}$, $P = P^T \geq 0$, satisfying the linear matrix inequality

$$\begin{bmatrix} A^TP + PA & PB - C^T \\ B^TP - C & -(D + D^T) \end{bmatrix} \leq 0. \tag{3}$$

Using Schur complement, (3) is equivalent to

$$A^TP + PA + (PB - C^T)(D + D^T)^{-1}(B^TP - C) \leq 0 \tag{4}$$

The solution of (4) being zero is a CARE. Taking the matrix root $UU^T = (D + D^T)^{-1}$ and defining $\hat{B} = BU$, $\hat{C} = U^TC$, and $\hat{A} = A - \hat{B}\hat{C}$, the CARE is expressible as

$$F(P) = \hat{A}^TP + P\hat{A} + P\hat{B}\hat{B}^TP + \hat{C}^T\hat{C} = 0 \tag{5}$$

The solution of (5), if it exists, is not unique. Among them there is a unique *stabilizing solution*, P_∞, in the sense that $(\hat{A} + \hat{B}\hat{B}^TP_\infty)$ is stable, i.e., all eigenvalues have negative real parts. In SBT, the stabilizing solutions, P_{min} and P_{max}^{-1}, to the two dual CAREs are solved

$$\hat{A}^TP_{min} + P_{min}\hat{A} + P_{min}\hat{B}\hat{B}^TP_{min} + \hat{C}^T\hat{C} = 0 \tag{6a}$$
$$\hat{A}P_{max}^{-1} + P_{max}^{-1}\hat{A}^T + P_{max}^{-1}\hat{C}^T\hat{C}P_{max}^{-1} + \hat{B}\hat{B}^T = 0 \tag{6b}$$

Let $P_{max}^{-1} = XX^T$, $P_{min} = YY^T$ be any root decompositions, now do the singular value decomposition (SVD)

$$X^TY = U\Sigma V^T \tag{7}$$

where $\Sigma \geq 0$ is an "economy size" k-by-k ($k \leq n$) diagonal matrix with singular values in descending order. Suppose the singular values of Σ are

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \gg \sigma_{r+1} \geq \cdots \geq \sigma_k \tag{8}$$

Define I_m to be the identity matrix of dimension m, $0_{m\times n}$ an $m \times n$ zero matrix, and

$$T_L = \begin{bmatrix} I_r & 0_{r\times(k-r)} \end{bmatrix} \Sigma^{-\frac{1}{2}}V^TY^T \tag{9a}$$
$$T_R = XU\Sigma^{-\frac{1}{2}} \begin{bmatrix} I_r \\ 0_{(k-r)\times r} \end{bmatrix} \tag{9b}$$

The system (T_LAT_R, T_LB, CT_R, D) represents the stochastically balanced reduced-order model whose states are aligned in descending involvement in the energy transfer process [9]. The best bound to date for the frequency domain approximation error can be found in [11]. SBT is preferred over BT because it guarantees passivity, further to stability, of the reduced-order model (e.g., [9, 11, 12]).

3. EIGENSPACE PROJECTION

The first stage of the reduction is to select an appropriate subspace onto which the original high-order system is projected. It is therefore well justified to use the spans (approximate spans) of W_c and W_o as they capture all (most) state activities. This idea appeared as *dominant subspaces projection* in [7] and later as *dominant gramian eigenspaces method* in [15]. To effectively extract the span of, say, W_c, we note that the Smith method, which transforms a continuous-time Lyapunov equation into a discrete one, provides a viable alternative to solving (2a). Specifically, the following two equations solve the same W_c:

$$AW_c + W_cA^T + BB^T = 0 \tag{10a}$$
$$A_zW_cA_z^T - W_c + B_zB_z^T = 0 \tag{10b}$$

[2]For every component of u that is a node voltage (branch current), the corresponding component of y is a branch current (node voltage) so that u^Ty represents the instantaneous power delivered to the system.

where $A_z = (pI + A)(pI - A)^{-1}$, $B_z = -\sqrt{2p}(pI - A)^{-1}B$, $p > 0$ is a shift parameter, and the subscript z means discrete time. Subsequently, $W_c = \sum_{i=0}^{\infty} A_z^i B_z B_z^T (A_z^T)^i$. We want to minimize the norm of A_z so the power terms decay quickly and the infinite summation can well be represented by finite terms. A good choice of p to achieve this is $p = \sqrt{|\lambda_{max}(A)\lambda_{min}(A)|}$ [9], where $\lambda_{max}(\circ)$ and $\lambda_{min}(\circ)$ respectively denote the maximum and minimum eigenvalues, which are found by simple power iterations in practice [8,17]. An important observation is that W_c is naturally cast as a matrix factorization, namely, when the growth of the summation decays to machine precision after τ terms,

$$W_c \approx \sum_{i=0}^{\tau-1} A_z^i B_z B_z^T (A_z^T)^i = \mathcal{K}_\tau(A_z, B_z)\mathcal{K}_\tau(A_z, B_z)^T \quad (11)$$

where the factor $\mathcal{K}_\tau(A, B) = \left[\begin{array}{cccc} B & AB & \cdots & A^{\tau-1}B \end{array}\right]$ is called the τth-order Krylov matrix. Application of the Smith method in standard BT of VLSI models can be found in [8, 9]. Arnoldi and Lanczos algorithms [17] are efficient algorithms to obtain the Krylov matrix. Here we present only the Arnoldi algorithm due to space limitation. The following codes assume a rank-one B_z, but block versions of Arnoldi and Lanczos algorithms are readily available to handle arbitrary ranks (e.g., [1, 2]).

Smith_Arnoldi: Input (A_z, B_z, max_itr, tolerance)
$j = 1$;
$q_1 = B_z/\|B_z\|_2$; $\beta = 1$; $Q_1 = q_1$; $R_1 = \|B_z\|_2$; $H_1 = [\ ]$;
while $j \le$ *max_itr* {
  for $i = 1 : j$
    $h_{ij} = q_i^T A_z q_j$;
  end
  $r_{j+1} = A_z q_j - \Sigma_{i=1}^{j} h_{ij} q_i$;
  $H_j = \left[\begin{array}{c|c} H_{j-1} & \begin{array}{c} h_{1j} \\ \vdots \\ h_{j-1,j} \end{array} \\ \hline \begin{array}{cccc} 0 & \cdots & 0 & \beta \end{array} & h_{jj} \end{array}\right]$;
  if $j > 1$
    $R_j = \left[\begin{array}{c|c} \begin{array}{c} R_{j-1} \\ \hline 0 \ \cdots \ 0 \end{array} & H_j \begin{bmatrix} R_{j-1}(:, j-1) \\ 0 \end{bmatrix} \end{array}\right]$;
    $\omega_j = Q_j R_j(:, j)$;
    if ($\|\omega_j\|_2 <$ *tolerance) break while loop;*
  end if
  $\beta = \|r_{j+1}\|_2$;
  if ($\beta <$ *tolerance) break while loop;*
  $q_{j+1} = r_{j+1}/\beta$;
  $Q_{j+1} = \left[\begin{array}{cc} Q_j & q_{j+1} \end{array}\right]$;
  $j = j + 1$;
end while
$\tau =$ *number of columns in* R_j;
Return Q_τ.

In short, the Arnoldi algorithm iteratively computes the τ orthogonal columns of $Q_\tau \in \Re^{n\times\tau}$ and an upper triangular matrix $R_\tau \in \Re^{\tau\times\tau}$ such that

- $Q_\tau^T Q_\tau = I_\tau$;
- $H_\tau = Q_\tau^T A_z Q_\tau$ is a τ-by-τ upper Hessenberg matrix;
- $\mathcal{K}_\tau(A_z, B_z) = \left[\begin{array}{cccc} B_z & A_z B_z & \cdots & A_z^{\tau-1} B_z \end{array}\right] = Q_\tau R_\tau$ is a QR factorization.

Obviously, Q_τ spans the column range of W_c. A counterpart, Q_υ, corresponding to the column range of W_o is obtained similarly. A Gram-Schmidt (GS) orthogonalization of Q_υ against Q_τ (columns in Q_τ are already orthogonal) produces an orthogonal $Q_k = GS(\left[\begin{array}{cc} Q_\tau & Q_\upsilon \end{array}\right]) \in \Re^{n\times k}$, $k \le \tau + \upsilon$, which serves as a projection basis to generate an intermediate model of order k.

Referring to (3), RLC models obtained from MNA have the properties $A + A^T \le 0$, $B = C^T$, and $D = 0$ [13]. Passivity of the circuit is then borne out by the fact that $P = I$ is a solution satisfying (3). Performing a congruence transformation of compatible dimensions, we have

$$\left[\begin{array}{cc} Q_k^T & 0 \\ 0 & I \end{array}\right] \left[\begin{array}{cc} A + A^T & B - C^T \\ B^T - C & 0 \end{array}\right] \left[\begin{array}{cc} Q_k & 0 \\ 0 & I \end{array}\right] \le 0 \quad (12)$$

It is easily seen that the system $(Q_k^T A Q_k, Q_k^T B, C Q_k, 0)$ inherits passivity from its parent. Careful inspection reveals that the Smith method is mathematically equivalent to the vector alternate direction implicit (ADI) implementation [14, 15] with a single ADI parameter. The Smith method, however, avoids the tridiagonalization of the A matrix which can be numerically unstable. Also, compared to ADI, SVDs of the low-rank square roots of W_c and W_o are unnecessary as Q_τ and Q_υ are already orthogonal.

4. SOLVING DUAL CARES

The intermediate model from projection, say, of order $k < 100$, is then subject to SBT to achieve further reduction with guaranteed passivity (provided the original model is passive). Standard ways of solving a CARE focus on identifying the stable invariant subspace of the corresponding *Hamiltonian* matrix, (e.g., [18, 19]). While this is sufficient for the stabilizing solution, information regarding the unstable invariant subspace is not utilized. We show that with a few extra low-cost steps, the stable and unstable invariant subspaces can be completely separated. This enables simultaneous solution of the pair of dual CAREs, (6a) and (6b), thereby significantly reducing the computational cost. Consider the Hamiltonian matrices, H and H', corresponding to (6a) and (6b) respectively:

$$H = \left[\begin{array}{cc} \hat{A} & \hat{B}\hat{B}^T \\ -\hat{C}^T\hat{C} & -\hat{A}^T \end{array}\right], H' = \left[\begin{array}{cc} \hat{A}^T & \hat{C}^T\hat{C} \\ -\hat{B}\hat{B}^T & -\hat{A} \end{array}\right] \quad (13)$$

It can be seen that if λ is the eigenvalue of a Hamiltonian matrix, then so is $-\lambda$. Since H and H' are real, eigenvalues apart from the real and imaginary axes occur even in quadruple $(\lambda, -\lambda, \overline{\lambda}, -\overline{\lambda})$. Suppose (A, B) is stabilizable and (A, C) is detectable, and H has no eigenvalues on the imaginary axis, then the stable and unstable invariant subspaces can be decoupled, i.e.,

$$H \left[\begin{array}{cc} X_{11} & X_{12} \\ X_{21} & X_{22} \end{array}\right] = \left[\begin{array}{cc} X_{11} & X_{12} \\ X_{21} & X_{22} \end{array}\right] \left[\begin{array}{cc} \Lambda_s & 0 \\ 0 & \Lambda_u \end{array}\right] \quad (14)$$

where Λ_s contains the stable eigenvalues and Λ_u the unstable ones. A well-known fact is that X_{11} is invertible and $P_{min} = P_{min}^T = X_{21}X_{11}^{-1}$. By the relationship $H' = \left[\begin{array}{cc} 0 & I \\ I & 0 \end{array}\right](-H)\left[\begin{array}{cc} 0 & I \\ I & 0 \end{array}\right]$, we get $P_{max}^{-1} = P_{max}^{-T} = X_{12}X_{22}^{-1}$. In other words, all information about P_{min} and P_{max}^{-1} are contained in (14), provided the invariant subspaces are completely separated. Defining $J = \left[\begin{array}{cc} 0 & I \\ -I & 0 \end{array}\right]$, a matrix S is called *symplectic* if $S^T J S = J$. Similarity transform of a Hamiltonian matrix by symplectic matrices preserves its

Hamiltonian structure. Here we present an effective implementation of the SR algorithm [20] to achieve subspace separation. It is assumed that H has already been transformed into J-tridiagonal form (see remarks):

$$H = \left[\begin{array}{cccc|cccc} a_1 & & & & c_1 & b_1 & & \\ & a_2 & & & b_1 & c_2 & \ddots & \\ & & \ddots & & & \ddots & \ddots & b_{k-1} \\ & & & a_k & & & b_{k-1} & c_k \\ \hline q_1 & & & & -a_1 & & & \\ & q_2 & & & & -a_2 & & \\ & & \ddots & & & & \ddots & \\ & & & q_k & & & & -a_k \end{array}\right] \quad (15)$$

SR algorithm converts H into a block J-upper-triangular form to reveal the eigenvalues. The three types of symplectic transforms being used in the SR algorithm are [20]:

• *Algorithm J – Givens Rotations*

$$J(i,c,s) = \left[\begin{array}{cc} \tilde{C} & \tilde{S} \\ -\tilde{S} & \tilde{C} \end{array}\right] \quad (16)$$

Here $\tilde{C}, \tilde{S} \in \Re^{k\times k}$ are the diagonal matrices $\tilde{C} = I_k + (c-1)e_i e_i^T$ and $\tilde{S} = s e_i e_i^T$ (e_i is the ith unit vector). The choice of c and s is standard [17]. Algorithm J zeroes single entries in the lower half of the columns of the Hamiltonian matrix. Given i, $1 \le i \le k$, and $x \in \Re^{2k}$, we have $J(i,c,s)x = y$ where $y_{k+i} = 0$ (subscript indexes the $k+i$ entry).

• *Algorithm H – Householder Transform*

$$H(i,l,w) = \left[\begin{array}{cc} \Psi & 0 \\ 0 & \Psi \end{array}\right] \quad (17)$$

where $\Psi = diag(I_{i-1}, P, I_{k-l-i+1})$ and $P = I_l - 2ww^T/w^Tw$. Again, the choice of $w \in \Re^l$, $2 \le l \le k-i+1$, is standard [17,20]. Algorithm H is used to zero column vectors of length l on the upper half of the Hamiltonian matrix. Given i, $1 \le i \le k$, and $x \in \Re^{2k}$, we have $H(i,l,w)x = y$ where $y_{i+1} = y_{i+2} = \cdots = y_{i+l-1} = 0$.

• *Algorithm G – Gaussian Elimination*

$$G(i,v) = \left[\begin{array}{cc} \Theta & \Phi \\ 0 & \Theta^{-1} \end{array}\right], G(i,v)^{-1} = \left[\begin{array}{cc} \Theta^{-1} & -\Phi \\ 0 & \Theta \end{array}\right] \quad (18)$$

where $\Theta = I_k + ((1+v^2)^{-1/4} - 1)(e_{i-1}e_{i-1}^T + e_i e_i^T)$ and $\Phi = (v(1+v^2)^{-1/4})(e_{i-1}e_i^T + e_i e_{i-1}^T)$. Algorithm G zeroes single entries in the upper half of the columns of the Hamiltonian matrix when $y_{k+i} = 0$ (Algorithm J does not work). Given i, $2 \le i \le k$, $x \in \Re^{2k}$, we have $G(i,v)x = y$ where $y_i = 0$.

The first two (stable) transformations use orthogonal symplectic matrices, while the last one is a nonorthogonal symplectic matrix with a condition number $cond_2(G(i,v)) = (1+v^2)^{1/2} + |v|$.

• *Implicit Quadruple-Shift SR Algorithm*

As in modern QR algorithm implementations [17], an SR counterpart utilizes *Implicit S* bulge-chasing such that all computations are in the real domain. Single and double-shift strategies are investigated in the technical report version of [19], in which the shifts are chosen from the real and imaginary axes only. Our implementation waives this constraint, and complies better with the quadruple occurence of eigenvalues away from the axes. A proven heuristic to speed up convergence is to choose the four shifts as eigenvalues of the 4×4 subblock

$$N_j = \left[\begin{array}{cc|cc} a_j & 0 & c_j & b_j \\ 0 & a_{j+1} & b_j & c_{j+1} \\ \hline q_j & 0 & -a_j & 0 \\ 0 & q_{j+1} & 0 & -a_{j+1} \end{array}\right] \quad (19)$$

Where $j = k$ in the first iteration, and gradually decreases when the J-tridiagonal matrix deflates [17, 20]. Defining $\alpha_j = a_j^2 + c_j q_j$ and $\beta_j = a_{j+1}^2 + c_{j+1}q_{j+1}$, the characteristic polynomial of (19) is:

$$s^4 - (\alpha_j + \beta_j)s^2 + \alpha_j\beta_j - q_j q_{j+1} b_j^2 = 0 \quad (20)$$

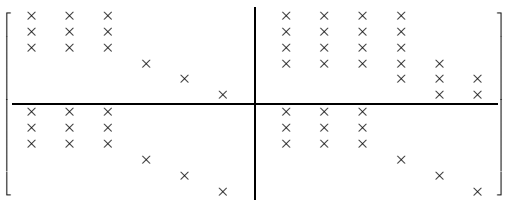
Analogous to the Implicit Q theorem in QR algorithm, the first column of the following matrix product is required for Implicit S similarity transform:

$$\begin{aligned} p(\lambda) &= (H-\lambda)(H+\lambda)(H-\bar{\lambda})(H+\bar{\lambda}) \\ &= H^4 - 2Re(\lambda^2)H^2 + |\lambda|^4 I \\ &= H^4 - (\alpha_j+\beta_j)H^2 + (\alpha_j\beta_j - q_jq_{j+1}b_j^2)I \quad (21) \end{aligned}$$

where the shifts are roots of (20). Putting $j = 1$ in α_j and β_j, and using a Matlab-type representation, the first columns of H^2 and H^4 are

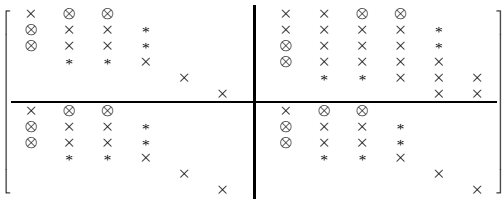
$$H^2(:,1) = \left[\begin{array}{c} \alpha_1 \\ b_1 q_1 \\ 0 \\ \vdots \\ 0 \end{array}\right], H^4(:,1) = \left[\begin{array}{c} \alpha_1^2 + b_1^2 q_1 q_2 \\ b_1 q_1 (\alpha_1 + \beta_1) \\ b_1 q_1 b_2 q_2 \\ 0 \\ \vdots \\ 0 \end{array}\right] \quad (22)$$

Setting $H_1 := H$ and using Algorithm H to find an $H(1,3,w)$ such that $H(1,3,w)p(\lambda)e_1$ is a multiple of e_1, the bulge-chasing begins by forming $H_2 := H(1,3,w)H_1H(1,3,w)^T$ and $\Pi := H(1,3,w)^T$. An example $H_2 \in \Re^{12\times 12}$ looks like:



To restore the J-tridiagonal structure, we refer to the following matrix. Circles represents zeroing of entries and asterisks stand for newly generated entries. The $(9,1)$ entry is zeroed using $H_3 := J(3,c,s)H_2J(3,c,s)^T$ and the update $\Pi := \Pi J(3,c,s)^T$ (entry at $(7,3)$ is automatically zeroed due to the Hamiltonian structure-preserving symplectic transform). Similarly, Algorithm J is used to zero $(8,1)$. Then $(3,1)$ is zeroed by Algorithm H with $H(2,2,w)$, followed by Algorithm G for $(2,1)$. Next, on the right half, $(9,7)$ and $(8,7)$ are zeroed by two times of Algorithm J, and on the upper right partition, $(4,7)$ and $(3,7)$ are zeroed with Algorithm H with $H(2,3,w)$. Consequently, the bulge is pushed to the lower right and the process is continued until it is

completely driven out:



As iteration proceeds, some of the b_js become negligibly small and the problem size deflates as in the QR algorithm. Ultimately, the SR algorithm reduces the J-tridiagonal matrix in into decoupled 2×2 and 4×4 subblocks. Stability and passivity of the intermediate model (Section 3) implies the absence of purely imaginary eigenvalues. Using the procedures in [20], the 2×2 (4×4) subblock can then be transformed into an upper (block) triangular form with the upper left (block) entry containing the eigenvalue(s) with negative real part(s). We introduce here additional symplectic transforms for each type of subblock that serve to completely separate the stable and unstable invariant subspaces.

• 2×2 *Subblock*

Let N_j be an ordered subblock taken from the j, $k+j$ plane of the $2k \times 2k$ matix:

$$N_j = \begin{bmatrix} -\lambda_j & x_j \\ 0 & \lambda_j \end{bmatrix} \tag{23}$$

where $\lambda_j > 0$ and x_j is non-zero (otherwise no processing is required). Define the 2×2 symplectic matrix

$$T_j = \begin{bmatrix} 1/2\lambda_j & x_j \\ 0 & 2\lambda_j \end{bmatrix} \tag{24}$$

It is easy to verify that $T_j^{-1} N_j T_j$ gives the diagonal matrix $diag(-\lambda_j, \lambda_j)$. Lifting T_j into the j, $k+j$ plane and updating Π completes the subspace separation in this subblock.

• 4×4 *Subblock*

Let N_j be an ordered subblock taken from the j, $j+1$, $k+j$, $k+j+1$ plane of the $2k \times 2k$ matix:

$$N_j = \begin{bmatrix} \Delta_j & \Omega_j \\ 0 & -\Delta_j^T \end{bmatrix} \tag{25}$$

where Δ_j, $\Omega_j (=\Omega_j^T)$ are 2×2 matrices. Assume Δ_j contains the stable eigenvalues $-\lambda_j$, $-\overline{\lambda}_j$ whose real parts are negative. The key to separating the subspaces is to realize that the column range of $U_j = (N_j + \lambda_j I)(N_j + \overline{\lambda}_j I)$ spans the unstable invariant subspace. Simple manipulation shows

$$span\,(U_j) = span\left(\begin{bmatrix} \Delta_j \Omega_j - \Omega_j \Delta_j^T + 2\,\mathrm{Re}(\lambda)\Omega_j \\ -4\,\mathrm{Re}(\lambda)\Delta_j^T \end{bmatrix} \right) \tag{26}$$

On the right hand side of (26), denoting the upper part of the partition by Z_1 and the lower part by Z_2, we define

$$F_j = \begin{bmatrix} Z_2^{-T} & Z_1 \\ 0 & Z_2 \end{bmatrix} \tag{27}$$

It is easy to see that F_j is well defined (Z_2 invertible) and symplectic. Moreover, $F_j^{-1} N_j F_j$ gives $diag(\Delta_j, -\Delta_j^T)$. Lifting F_j into the j, $j+1$, $k+j$, $k+j+1$ plane and updating Π completes the subspace separation in this subblock.

Eventually, $H\Pi = \Pi diag(\Lambda_s, -\Lambda_s^T)$, and solutions to the dual CAREs can be extracted from Π as in (14).

Remarks:

1. Techniques and results in this section are independent of the projection in Section 3. In fact, an alternative to perform the first-stage reduction is by an implicitly restarted Lanczos algorithm [19]. In that case, H is readily in J-tridiagonal format, but the projection may not be as good as the eigenspace approach in capturing the state transitions.

2. The JHESS algorithm in [20] is used to transform a Hamiltonian matrix into J-tridiagonal (or Hamiltonian J-Hessenberg) form. Existence of this transformation is strongly dependent on the first column of the similarity transform matrix [19]. The set of these breakdown-free vectors is dense in $\Re^{2k}$. Should breakdown (or near-breakdown) occurs due to high condition numbers in Algorithm G, a different projection basis Q_k in Section 3 is chosen by varying the order and/or number of columns in Q_τ and Q_υ. If the implicitly restarted Lanczos algorithm is used, then it is a simple matter of invoking an implicit start.

3. Convergence of the quadruple-shift SR algorithm is excellent (usually within 10 iterations) under mild conditions. In the few cases where Algorithm G produces a very large condition number (only during early iterates), an exceptional shift is performed and the process is continued [20].

4. The most expensive step in the proposed scheme is the matrix inversion ($O(n^3)$ work) in A_z computation in the Smith method. However, this step is done only once and all other steps are of $O(n^2)$. In the SBT of the intermediate model, the transformation to J-tridiagonal form requires $O(k^3)$ work (not required in implicitly restarted Lanczos), while the SR algorithm is of $O(k^2)$. In practice, $k \ll n$ and the cost of the second stage is insignificant.

5. NUMERICAL EXAMPLES

The first example is a passive RLC system of order 300. It is reduced to models of order 11 using full BT, full SBT, PRIMA [1], low-rank square root method [7], and the proposed scheme (intermediate model order = 20). As shown in Fig. 1, the control-theoretic approaches have better performance. Interestingly, the proposed scheme, resembling the full SBT, has even lower error. This may arise from the numerical conditioning issues in the large-size CAREs being solved in the full-order approach. In terms of computational cost, Table 1 shows that the proposed scheme offers an effective means to achieve tradeoff between the cost and quality of the reduced-order models. The second example in Fig. 2 is a system of order 300 with 80 dominant poles close to the imaginary axis. The intermediate model order is 82 and the reduced order is 22 in the proposed scheme, while other schemes render 31. Again, similar observations are obtained. This confirms the versatility of the project-and-balance implementation.

6. CONCLUSION

This paper has presented a computationally efficient two-stage project-and-balance scheme for passivity-preserving model reduction. It is shown that bases for eigenspaces projection are readily obtained by integrating Smith method and Krylov subspace iteration. The intermediate model is stochastically balanced and truncated. A novel approach for solving two dual CAREs simultaneously, namely, by separating the stable and unstable invariant subspaces of a Hamiltonian matrix, has been introduced. A fast-converging

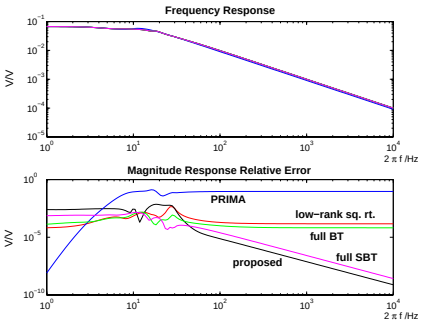


Figure 1: Example RLC system of order 300 (reduced to 11).

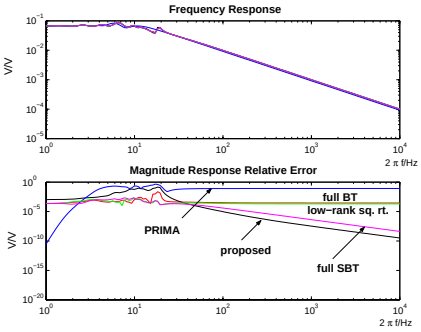


Figure 2: Example system of order 300 with 80 dominant poles (reduced to 22 and 31).

quadruple-shift bulge-chasing SR algorithm has also been described. Numerical examples have verified the effectiveness of the proposed scheme over conventional approaches.

7. REFERENCES

[1] A. Odabasioglu, M. Celik, and L. T. Pileggi, "PRIMA: Passive reduced-order interconnect macromodeling algorithm," *IEEE Trans. Computer-Aided Design*, vol. 17, no. 8, pp. 645–654, Aug. 1998.

[2] Z. Bai, P. M. Dewilde, and R. W. Freund, "Reduced-order modeling," *Numerical Analysis Manuscript No. 02-4-13, Bell Laboratories*, Mar. 2002.

[3] L. T. Pillage and R. A. Rohrer, "Asymptotic waveform evaluation for timing analysis," *IEEE Trans. Computer-Aided Design*, vol. 9, no. 4, pp. 352–366, Apr. 1990.

[4] P. Feldmann and R. W. Freund, "Reduced-order modeling of large linear subcircuits via a block Lanczos algorithm," in *Proc. ACM/IEEE Design, Automation Conf.*, June 1995, pp. 474–479.

[5] K. Glover, "All optimal Hankel-norm approximation of linear multivariable systems and their L^∞-error bounds," *Int. J. Control*, vol. 39, no. 6, pp. 1115–1193, June 1984.

[6] B. Moore, "Principal component analysis in linear systems: Controllability, observability, and model reduction," *IEEE Trans. Automat. Contr.*, vol. 26, no. 1, pp. 17–32, Feb. 1981.

[7] T. Penzl, "Algorithms for model reduction of large dynamical systems," Sonderforschungsbereich 393 Numerische Simulation auf massiv parallelen Rechnern, TU Chemnitz, 09107 Chemnitz, FRG, Tech. Rep. SFB393/99-40, 1999.

[8] Q. Su, V. Balakrishnan, and C.-K. Koh, "Efficient approximate balanced truncation of general large-scale RLC systems via Krylov methods," in *Proc. ASPDAC/Int. Conf. VLSI Design*, Jan. 2002, pp. 311–316.

[9] Q. Su, "Algorithms for model reduction of large scale RLC systems," Ph.D. dissertation, School of ECE, Purdue University, Aug. 2002.

[10] M. Green, "Balanced stochastic realizations," *Linear Algebra Appl.*, vol. 98, pp. 211–247, 1988.

[11] X. Chen and J. T. Wen, "Positive realness preserving model reduction with H_∞ norm error bounds," *IEEE Trans. Circuits Syst. I*, vol. 42, no. 1, pp. 23–29, Jan. 1995.

[12] J. R. Phillips, L. Daniel, and L. M. Silveira, "Guaranteed passive balancing transformations for model order reduction," *IEEE Trans. Computer-Aided Design*, vol. 22, no. 8, pp. 1027–1041, Aug. 2003.

[13] Q. Su, V. Balakrishnan, and C.-K. Koh, "A factorization-based framework for passivity-preserving model reduction of RLC systems," in *Proc. IEEE Design, Automation Conf.*, June 2002, pp. 40–45.

[14] J. Li and J. White, "Efficient model reduction of interconnect via approximate system gramians," in *Proc. IEEE Int. Conf. Computer-Aided Design*, Nov. 1999, pp. 380–383.

[15] J. Li, "Model reduction of large linear systems via low rank system gramians," Ph.D. dissertation, Department of Mathematics, Massachusetts Institute of Technology, Sept. 2000.

[16] J. Vlach and K. Singhal, *Computer Methods for Circuit Analysis and Design*. Kluwer Academic Publishers, July 1993.

[17] G. Golub and C. V. Loan, *Matrix Computations*, 2nd ed. Baltimore: JohnsHopkins Univ. Press, 1989.

[18] G. S. Ammar, P. Benner, and V. Mehrmann, "A multishift algorithm for the numerical solution of algebraic Riccati equations," *Electr. Trans. Num. Anal.*, vol. 1, pp. 33–48, Sept. 1993.

[19] P. Benner and H. Faßbender, "An implicitly restarted symplectic Lanczos method for the Hamiltonian eigenvalue problem," *Linear Algebra and its Applications*, vol. 263, pp. 75–111, 1997.

[20] A. Bunse-Gerstner and V. Mehrmann, "A symplectic QR like algorithm for the solution of the real algebraic Riccati equation," *IEEE Trans. Automat. Contr.*, vol. 31, no. 12, pp. 1104–1113, Dec. 1986.

Table 1: CPU time for various schemes with that of PRIMA normalized to 1.

	PRIMA	LRSR	Proposed	BT	SBT
Eg. 1	1	3.62	8.00	108.75	417.00
Eg. 2	1	8.22	8.31	85.54	387.72