Matrix Equations and Model Reduction

Peter Benner, Tobias Breiten and Lihong Feng

Abstract Model order reduction methods for linear time invariant systems are reviewed in this lecture. The basic ideas of the methods, such as the Padé approximation method, the rational interpolation method, the modal truncation method, the standard balanced truncation method, and the balancing related methods, are presented. The numerical algorithms of implementing the methods are discussed. For the balanced truncation method and the balancing related methods, Lyapunov equations or Riccati equations need to be solved. Algorithms for solving these matrix equations are introduced.

1 Intruduction

Model order reduction (MOR) is a technique of reducing the complexity of large-scale complex systems, so that the input-output relations can be reproduced in acceptable time and with ignorable error. In today's real-life applications, large-scale complex systems can be time-varying, nonlinear, parametric, or stochastic, which propose big challenges for model order reduction. Although model order reduction techniques have been developed for these systems, and proved to be promising in various applications, due to time limitation, the lecture focuses on model order reduction methods for linear time invariant (LTI) systems in the following form (if without pointed out),

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1}$$

Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstr.1, 39106 Magdeburg, Germany. e-mail: benner@mpi-magdeburg.mpg.de, breiten@mpi-magdeburg.mpg.de, feng@mpi-magdeburg.mpg.de.

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$ and $D \in \mathbb{R}^{p\times m}$. Here, $x(t) \in \mathbb{R}^n$ is the state of the system, $u(t) \in \mathbb{R}^m$ is the input, and $y(t) \in \mathbb{R}^p$ is the output. When $m = p = 1$, the system is called single-input and single-output (SISO) system; otherwise if $m, p > 1$, it is called a multiple-input and multiple-output (MIMO) system.

The basic idea of model order reduction is based on projection. Assuming that the trajectory of x in (1) is contained in a low-dimensional subspace $\mathcal{V}$, and $\mathcal{W}^\perp$ is a complementary subspace of $\mathcal{V}$, i.e. $\mathcal{V}\oplus\mathcal{W}^\perp = \mathbb{R}^n$, $\mathcal{V}\cap\mathcal{W}^\perp = \{0\}$. Let $\mathcal{W}$ be the orthogonal complementary subspace of $\mathcal{W}^\perp$. Let the columns of the matrix $V \in \mathbb{R}^{n\times q}$ form the basis of $\mathcal{V}$, and the columns of $W \in \mathbb{R}^{n\times q}$ be the basis of the subspace $\mathcal{W}$, and they satisfy $W^TV = I$, then VW^T is a projector, which projects x onto $\mathcal{V}$, along $\mathcal{W}^\perp$. The reduced-order model is obtained by approximating the state x by its projection $x \approx VW^Tx$,

$$\begin{aligned} \dot{\hat{x}}(t) &= \hat{A}\hat{x}(t) + \hat{B}u(t), \\ \hat{y}(t) &= \hat{C}\hat{x}(t) + \hat{D}u(t), \end{aligned} \tag{2}$$

where $\hat{x}(t) = W^Tx(t) \in \mathbb{R}^q$, $\hat{A} = W^TAV \in \mathbb{R}^{q\times q}$, $\hat{B} = W^TB \in \mathbb{R}^{q\times m}$, $\hat{C} = CV \in \mathbb{R}^{p\times m}$, and $\hat{D} = D \in \mathbb{R}^{p\times m}$. The above process of getting the reduced model is in fact a Petrov-Galerkin projection. That is, replacing x with the approximation $x \approx V\hat{x} =: \tilde{x}$ and then forcing the residual $r = \dot{\tilde{x}} - A\tilde{x} - Bu$ to be zero in a test subspace $\mathcal{W}$, i.e. $W^Tr = 0$, so that the first equation in (2) is derived. The second equation follows directly by replacing x with its approximation $\tilde{x} = V\hat{x}$. When $W = V$, it reduces to a Galerkin projection.

The goals of model order reduction method include

- The output of the large-scale system should be approximated by a reduced model that can be evaluated significantly faster.
- The reduced model should be automatically generated.
- There should be a computable error bound/estimate for the reduced model.
- Physical properties of the original system, such as stability, minimum phase, and/or passivity should be preserved during the MOR process.

The model order reduction methods discussed in this lecture are based on concepts from (numerical) linear algebra and systems and control theory, where matrix decompositions, Krylov subspaces, iterative solvers, matrix equations play important roles. The outline of this summary is as follows. In the next section, the mathematical basics are summarized. In Section 3-6 basic model reduction methods for LTI systems are presented. Numerical algorithms for solving matrix equations are discussed in Section 7. Model reduction related software is introduced in Section 8. Conclusions are given in the end.

2 Mathematical Basics

The singular value decomposition

One essential tool from (numerical) linear algebra for data compression and dimension reduction is the singular value decomposition (SVD) of a matrix. The SVD exists for any matrix as the following theorem shows.

Theorem 1. *Let $A \in \mathbb{R}^{m\times n}$, then there exist orthogonal $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$, such that*

$$A = U\Sigma V^T, \Sigma = \begin{cases} \begin{bmatrix} \Sigma_1 \\ 0 \end{bmatrix}, & m \geq n \\ \begin{bmatrix} \Sigma_1 & 0 \end{bmatrix}, & m \leq n \end{cases} \text{ and } \Sigma_1 = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_{min(m,n)} \end{bmatrix}$$

with $\sigma_1 \geq \ldots \geq \sigma_s > \sigma_{s+1} = \ldots = \sigma_{min(m,n)} = 0$ for $s = rank(A)$.

The singular value decomposition of matrices is the core of the balanced truncation MOR method. It is also used in many other model reduction methods to assist the derivation of the reduced model.

The Laplace transform

Definition 1. The Laplace transform of a time domain function $f \in L_{1,\mathrm{loc}}$ (f is locally integrable, i.e. $\int_K |f(t)|dt < \infty$, $\forall$ compact subset K of $\mathrm{dom}(f)$.) with $\mathrm{dom}(f) = \mathbb{R}_0^+$ is

$$\mathcal{L} : f(t) \mapsto F(s) := \mathcal{L}\{f(t)\}(s) := \int_0^\infty e^{-st} f(t)\, dt, \quad s \in \mathbb{C}.$$

F is a function in the (Laplace or) frequency domain.

For frequency domain evaluations ("frequency response analysis"), one takes $\mathrm{Re}\,(s) = 0$ and $\mathrm{Im}\,(s) \geq 0$. Then $\omega := \mathrm{Im}\,(s)$ takes the role of a frequency (in [rad/s], i.e., $\omega = 2\pi v$ with v measured in [Hz]).

Lemma 1. *Applying Laplace transform to the derivative of $f(t)$ results in $sF(s)$,*

$$\mathcal{L}\{\dot{f}(t)\}(s) = sF(s).$$

For ease of notation, in the following we will use lower-case letters for both a function and its Laplace transform.

Linear systems in frequency domain

Applying Laplace transform ($x(t) \mapsto x(s)$, $\dot{x}(t) \mapsto sx(s)$) to the linear system in (1) with $x(0) = 0$ yields,

$$sx(s) = Ax(s) + Bu(s), \quad y(s) = Cx(s) + Du(s).$$

We get the input-output relation in frequency domain,

$$y(s) = \Big(\underbrace{C(sI - A)^{-1}B + D}_{=:G(s)} \Big) u(s),$$

where $G(s)$ is defined as the transfer function of (1).

In systems and control theory, the error bound of the reduced model is established through the transfer function, i.e.

$$||y - \hat{y}||_2 \leq ||G(s) - \hat{G}(s)||_\infty ||u||_2,$$

where the 2-norm stands for the $\mathcal{L}_2$ ($\mathcal{H}_2$) norm in the frequency domain, or the L_2 norm in the time domain. $||\cdot||_\infty$ is the $\mathcal{H}_\infty$ norm of a matrix-valued function (see the analysis in the subsection "system norms"). $\hat{G}(s) = \hat{C}(sI - \hat{A})^{-1}B + \hat{D}$ is the transfer function of the reduced-order model. The details of deriving the error bound are discussed at the end of this section.

Properties of linear systems

Definition 2. A linear system

$$\dot{x}(t) = Ax(t) + Bu(t), \quad y(t) = Cx(t) + Du(t)$$

is stable if its transfer function $G(s)$ has all its poles in the left half plane and it is asymptotically (or Lyapunov, or exponentially) stable if all poles are in the open left half plane $\mathbb{C}^- := \{z \in \mathbb{C} \,|\, \mathrm{Re}\,(z) < 0\}$.

Lemma 2. *The sufficient condition for asymptotic stability is that A is asymptotically stable (or Hurwitz), i.e., the spectrum of A, denoted by $\Lambda(A)$, satisfies $\Lambda(A) \subset \mathbb{C}^-$.*

Note that by abuse of notation, often "stable system" is used for asymptotically stable systems. For what follows, we need to define the concepts of controllability and observability, see [1].

Definition 3. Given a linear system (A, B, C, D). A state $x_* \in \mathbb{R}^n$, is controllable to the zero state if there exist an input function $u_*(t)$ and a time $t_* < \infty$, such that the solution of the linear dynamical system vanishes at

time t_*, i.e., $\Phi(u_*; x_*; t_*) = 0$. The controllable subspace X^{contr} of the system is the set of all controllable states. The system is (completely) controllable if $X^{\text{contr}} = \mathbb{R}^n$.

Definition 4. ([1]) Given a linear system (A, B, C, D). A state $x_* \in \mathbb{R}^n$ is unobservable if $y(t) = 0$ for all $t \geq 0$, i.e., if x_* is indistinguishable from the zero state for all $t \geq 0$. The *unobservable subspace* X^{unobs} is the set of all unobservable states of the system. The system is (completely) observable if $X^{\text{unobs}} = \{0\}$.

Controllability and observability, characterized by the controllability matrix $K(A, B) = [B, AB, A^2B, \ldots, A^{n-1}B] \in \mathbb{R}^{n \times nm}$, and the observability matrix $\mathcal{O}(A, C) = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \in \mathbb{R}^{np \times n}$, are two important properties of the system, based on which the standard balanced truncation method and the balancing related MOR methods are developed.

Lemma 3. *The LTI system is controllable if and only if $K(A, B)$ has full rank n. Analogously, the LTI system is observable if and only if $\mathcal{O}(A, C)$ has full rank n.*

The controllability and observability of the system can also be examined through the infinite Gramians P and Q of the system. The controllability Gramian matrix P and the observability Gramian matrix Q are defined as [1],

$$\begin{array}{l} P = \int_0^\infty e^{At} B B^T e^{A^T t} dt, \\ Q = \int_0^\infty e^{A^T t} C^T C e^{At} dt. \end{array} \tag{3}$$

Lemma 4. *The LTI system is controllable if and only if P is positive definite. The LTI system is observable if and only if Q is positive definite.*

Please refer to [1] for more discussions on controllability and observability, and other properties of linear systems, such as stabilizability, detectability etc..

Realizations of linear systems

Definition 5. For a linear time-invariant system

$$\Sigma : \begin{cases} \dot{x}(t) = Ax(t) + Bu(t), \\ y(t) = Cx(t) + Du(t), \end{cases}$$

with transfer function $G(s) = C(sI-A)^{-1}B+D$, the quadruple $(A, B, C, D) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n} \times \mathbb{R}^{p\times m}$ is called a realization of Σ.

It can be easily verified that the transfer function is invariant under state-space transformations,

$$\mathcal{T} : \begin{cases} x & \to & Tx, \\ (A, B, C, D) & \to & (TAT^{-1}, TB, CT^{-1}, D). \end{cases}$$

The transfer function is also invariant under addition of uncontrollable or unobservable states as below,

$$\begin{aligned} \tfrac{d}{dt}\begin{bmatrix} x \\ x_1 \end{bmatrix} &= \begin{bmatrix} A & 0 \\ 0 & A_1 \end{bmatrix}\begin{bmatrix} x \\ x_1 \end{bmatrix} + \begin{bmatrix} B \\ B_1 \end{bmatrix} u(t), \quad y(t) = \begin{bmatrix} C\ 0 \end{bmatrix}\begin{bmatrix} x \\ x_1 \end{bmatrix} + Du(t), \\ \tfrac{d}{dt}\begin{bmatrix} x \\ x_2 \end{bmatrix} &= \begin{bmatrix} A & 0 \\ 0 & A_2 \end{bmatrix}\begin{bmatrix} x \\ x_2 \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u(t), \quad y(t) = \begin{bmatrix} C\ C_2 \end{bmatrix}\begin{bmatrix} x \\ x_2 \end{bmatrix} + Du(t), \end{aligned}$$

for arbitrary $A_j \in \mathbb{R}^{n_j\times n_j}$, $j = 1, 2$, $B_1 \in \mathbb{R}^{n_1\times m}$, $C_2 \in \mathbb{R}^{p\times n_2}$ and any $n_1, n_2 \in \mathbb{N}$. Hence, the following four quadruples

$$\begin{aligned} &(A, B, C, D), && \left(\begin{pmatrix} A & 0 \\ 0 & A_1 \end{pmatrix}, \begin{pmatrix} B \\ B_1 \end{pmatrix}, \begin{pmatrix} C\ 0 \end{pmatrix}, D\right), \\ &(TAT^{-1}, TB, CT^{-1}, D), && \left(\begin{pmatrix} A & 0 \\ 0 & A_2 \end{pmatrix}, \begin{pmatrix} B \\ 0 \end{pmatrix}, \begin{pmatrix} C\ C_2 \end{pmatrix}, D\right), \end{aligned}$$

are all realizations of Σ. Therefore, the realizations are not unique.

Definition 6. The McMillan degree of Σ is the unique minimal number $\hat{n} \geq 0$ of states necessary to describe the input-output behavior completely. A minimal realization is a realization $(\hat{A}, \hat{B}, \hat{C}, \hat{D})$ of Σ with order $\hat{n}$.

Theorem 2. *A realization (A, B, C, D) of a linear system is minimal if and only if it is controllable and observable.*

Balanced realizations

Definition 7. A realization (A, B, C, D) of a stable linear system Σ is balanced if its controllability/observability Gramians P, Q satisfy

$$P = Q = \operatorname{diag}\{\sigma_1, \ldots, \sigma_n\} \quad (\text{w.l.o.g. } \sigma_j \geq \sigma_{j+1},\ j = 1, \ldots, n-1).$$

Notice that $\sigma_1, \ldots, \sigma_n \geq 0$ as $P, Q \geq 0$ by definition, and $\sigma_1, \ldots, \sigma_n > 0$ in case of minimality. In general, even for unbalanced systems, the so-called *Hankel singular values* σ_i^{HSV} can be computed by means of the Gramians P

and Q. We have $\sigma_i^{\mathrm{HSV}} = \Lambda_i(PQ)^{\frac{1}{2}}$, i.e., the Hankel singular values are given as the positive square roots of the eigenvalues of the product of the Gramians P and Q. For more information on the precise definition of the Hankel singular values and their relation to the Hankel operator of the system, we refer to, e.g., [1]. The following theorem shows how to obtain a balanced realization. Assume A is Hurwitz, i.e. $\Lambda(A) \subset \mathbb{C}^-$. Then:

Theorem 3. *Given a stable minimal linear system $\Sigma : (A, B, C, D)$, a balanced realization is obtained by the state-space transformation with*

$$T_b := \Sigma^{-\frac{1}{2}} V^T R,$$

where $P = S^T S$, $Q = R^T R$ (e.g., Cholesky decompositions) and $SR^T = U\Sigma V^T$ is the SVD of SR^T.

Theorem 4. *The controllability/observability Gramians P/Q satisfy the Lyapunov equations*

$$AP + PA^T + BB^T = 0, \quad A^T Q + QA + C^T C = 0. \tag{4}$$

In the following, only the case for the controllability Gramian is proved; that for the observability Gramian is analogous.

Proof. From the definition of P in (3),

$$\begin{aligned} AP + PA^T + BB^T &= A \int_0^\infty e^{At} BB^T e^{A^T t} dt + \int_0^\infty e^{At} BB^T e^{A^T t} dt\, A^T + BB^T \\ &= \int_0^\infty \underbrace{Ae^{At} BB^T e^{A^T t} + e^{At} BB^T e^{A^T t} A^T}_{= \frac{d}{dt} e^{At} BB^T e^{A^T t}} dt + BB^T \\ &= \underbrace{\lim_{t\to\infty} e^{At} BB^T e^{A^T t}}_{=0} - \underbrace{e^{A\cdot 0}}_{=I_n} BB^T \underbrace{e^{A^T \cdot 0}}_{=I_n} + BB^T \\ &= 0. \end{aligned}$$

□

Theorem 5. *The Hankel singular values (HSVs) of a stable minimal linear system are system invariants, i.e. they are unaltered by state-space transformations.*

Proof. In balanced coordinates, the HSVs are $\Lambda(PQ)^{\frac{1}{2}}$. Now let

$$(\hat{A}, \hat{B}, \hat{C}, D) = (TAT^{-1}, TB, CT^{-1}, D)$$

be any transformed realization with associated controllability Lyapunov equation

$$0 = \hat{A}\hat{P} + \hat{P}\hat{A}^T + \hat{B}\hat{B}^T = TAT^{-1}\hat{P} + \hat{P}T^{-T}A^T T^T + TBB^T T^T.$$

This is equivalent to

$$0 = A(T^{-1}\hat{P}T^{-T}) + (T^{-1}\hat{P}T^{-T})A^T + BB^T.$$

The uniqueness of the solution of the Lyapunov equation implies that $\hat{P} = TPT^T$ and, analogously, $\hat{Q} = T^{-T}QT^{-1}$. Therefore,

$$\hat{P}\hat{Q} = TPQT^{-1},$$

showing that $\Lambda(\hat{P}\hat{Q}) = \Lambda(PQ) = \{\sigma_1^2, \ldots, \sigma_n^2\}$. □

For non-minimal systems, the Gramians can also be transformed into diagonal matrices with leading $\hat{n} \times \hat{n}$ submatrices equal to $\mathrm{diag}(\sigma_1, \ldots, \sigma_{\hat{n}})$, and

$$\hat{P}\hat{Q} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{\hat{n}}^2, 0, \ldots, 0),$$

see [22, 29].

System norms

Definition 8. The $L_2^n(-\infty, +\infty)$ space is the vector-valued function space $f : \mathbb{R} \mapsto \mathbb{R}^n$, with the norm

$$\|f\|_{L_2^n} = \left(\int_{-\infty}^{\infty} ||f(t)||^2 dt \right)^{1/2}.$$

Here and below, $||\cdot||$ denotes the Euclidean vector or spectral matrix norm.

Definition 9. The frequency domain $\mathcal{L}_2(\jmath\mathbb{R})$ space is the matrix-valued function space $F : \mathbb{C} \mapsto \mathbb{C}^{p\times m}$, with the norm

$$||F||_{\mathcal{L}_2} = \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} ||F(\jmath\omega)||^2 d\omega \right)^{1/2},$$

where $\jmath = \sqrt{-1}$ is the imaginary unit. The maximum modulus theorem [23] will be used in this subsection.

Theorem 6. *Let $f(z) : \mathbb{C}^n \mapsto \mathbb{C}$ be a regular analytic, or holomorphic, function of n complex variables $z = (z_1, \ldots, z_n), n \geq 1$, defined on an (open) domain $\mathbb{D}$ of the complex space $\mathbb{C}^n$, which is not a constant, $f(z) \neq$ const. Let*

$$max_f = \sup\{|f(z)| : z \in D\}.$$

If $f(z)$ is continuous in a finite closed domain $\mathbb{D}$, then max_f can only be attained on the boundary of $\mathbb{D}$.

Consider the transfer function

$$G(s) = C(sI - A)^{-1}B + D$$

and input functions $u \in \mathcal{L}_2(\jmath\mathbb{R})$, with the $\mathcal{L}_2$-norm

$$||u||^2_{\mathcal{L}_2} := \frac{1}{2\pi}\int_{-\infty}^{\infty} u(\jmath\omega)^H u(\jmath\omega)\, d\omega.$$

Assume A is (asymptotically) stable: $\Lambda(A) \subset \mathbb{C}^- := \{z \in \mathbb{C} : \mathrm{Re}\,(z) < 0\}$. Then for all $s \in \mathbb{C}^+ \cup \jmath\mathbb{R}$, following the maximal modulus theorem, $G(s)$ is bounded: $||G(s)|| \leq M < \infty$, so we have

$$\begin{aligned}\int_{-\infty}^{\infty} y(\jmath\omega)^H y(\jmath\omega)\, d\omega &= \int_{-\infty}^{\infty} u(\jmath\omega)^H G(\jmath\omega)^H G(\jmath\omega) u(\jmath\omega)\, d\omega \\ &= \int_{-\infty}^{\infty} ||G(\jmath\omega)u(\jmath\omega)||^2\, d\omega \leq \int_{-\infty}^{\infty} M^2 ||u(\jmath\omega)||^2 d\omega \\ &= M^2 \int_{-\infty}^{\infty} u(\jmath\omega)^H u(\jmath\omega)\, d\omega \;<\; \infty,\end{aligned}$$

So that $y = Gu \in \mathcal{L}_2(\jmath\mathbb{R})$.

Consequently, the $\mathcal{L}_2$-induced operator norm

$$||G||_\infty := \sup_{||u||_2 \neq 0} \frac{||Gu||_{\mathcal{L}_2}}{||u||_{\mathcal{L}_2}} \tag{5}$$

is well defined. It can be further proved that

$$||G||_\infty = \sup_{\omega\in\mathbb{R}} ||G(\jmath\omega)|| = \sup_{\omega\in\mathbb{R}} \sigma_{max}\left(G(\jmath\omega)\right).$$

Definition 10. The Hardy space $\mathcal{H}_\infty$ is the function space of matrix-, scalar-valued functions that are analytic and bounded in $\mathbb{C}^+ := \{z \in \mathbb{C} : \mathrm{Re}\,(z) > 0\}$.

The $\mathcal{H}_\infty$-norm is defined as

$$||F||_\infty := \sup_{\mathrm{Re}\,(s)>0} \sigma_{\max}(F(s)) = \sup_{\omega\in\mathbb{R}} \sigma_{max}\left(F(\jmath\omega)\right).$$

The second equality follows from the maximum modulus theorem.

Definition 11. The Hardy space $\mathcal{H}_2(\mathbb{C}^+)$ is the function space of matrix-, scalar-valued functions that are analytic in $\mathbb{C}^+$ and bounded w.r.t. the $\mathcal{H}_2$-norm defined as

$$\begin{aligned}||F||_2 &:= \tfrac{1}{2\pi}\left(\sup_{\mathrm{re}\sigma>0} \int_{-\infty}^{\infty} ||F(\sigma+\jmath\omega)||_F^2 d\omega\right)^{\frac{1}{2}} \\ &= \tfrac{1}{2\pi}\left(\int_{-\infty}^{\infty} ||F(\jmath\omega)||_F^2\, d\omega\right)^{\frac{1}{2}}.\end{aligned} \tag{6}$$

The last equality in (6) follows Theorem 6.

Following [2], for inputs $u(t)$ with $\int_0^\infty \|u(t)\|_2^2 \mathrm{d}t \leq 1$, the $\mathcal{H}_2$ approximation error gives the following bound

$$\max_{t>0} \|y(t) - \hat{y}(t)\|_\infty \leq \|G - \hat{G}\|_{\mathcal{H}_2}, \tag{7}$$

where G and $\hat{G}$ are original and reduced transfer functions.

Theorem 7. *Practical Computation of the $\mathcal{H}_2$-norm follows*

$$||F||_2^2 = tr(B^T Q B) = tr(CPC^T),$$

where P, Q are the controllability and observability Gramians of the corresponding LTI system.

it the statement for the Paley-Wiener Theorem ok?

Theorem 8. *Paley-Wiener Theorem (Parseval's equation/Plancherel Theorem) The Fourier transform of $f \in L_2^n(-\infty, \infty)$:*

$$F(\xi) = \int_{-\infty}^{\infty} f(t) e^{-\xi t} dt$$

is a Hilbert space isomorphism between $L_2^n(-\infty, \infty)$ and $\mathcal{L}_2(\jmath\mathbb{R})$. Furthermore, the Fourier transform maps $L_2^n(0, \infty)$ onto $\mathcal{H}_2(\mathbb{C}^+)$. In addition it is an isometry, that is, it preserves distances:

$$L_2^n(-\infty, \infty) \cong \mathcal{L}_2(\jmath\mathbb{R}), \quad L_2^n(0, \infty) \cong \mathcal{H}_2(\mathbb{C}^+).$$

Consequently, L_2^n-norm in time domain and $\mathcal{L}_2$-norm, $\mathcal{H}_2$-norm in frequency domain coincide.

Therefore the output error bound (obtained from (5)),

$$||y - \hat{y}||_2 = ||Gu - \hat{G}u||_2 \leq ||G - \hat{G}||_\infty ||u||_2, \tag{8}$$

holds in time and frequency domain due to Paley-Wiener theorem, i.e. the $||\cdot||_2$ in (8) can be the L_2^n-norm in time domain, or the $\mathcal{L}_2$-norm, $\mathcal{H}_2$-norm in frequency domain. Model order reduction aims to compute reduced-order model such that either $||G - \hat{G}||_\infty < tol$ (8) or $||G - \hat{G}||_{\mathcal{H}_2} < tol$ (7), where tol is the acceptable error.

3 Methods based on Padé approximation and rational interpolation

The MOR methods based on Padé approximation [4, 10, 12] and rational interpolation [16, 18] are motivated by observing the series expansion of the transfer function. The LTI system considered is more general,

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cu(t) + Du(t), \end{aligned} \tag{9}$$

where $E \in \mathbb{R}^{n\times n}$ can be singular, and only $\lambda E - A (\forall \lambda \in \mathbb{C})$ is required to be regular. An LTI system with singular E is called descriptor system, and is more complex than the standard state space system in (1).

Methods based on Padé approximation

To consider the transfer function $G(s)$, for simplicity and without loss of generality, we leave $D = 0$. Let $s = s_0 + \sigma$, then within the convergence radius of the series,

$$\begin{aligned} &G(s_0+\sigma) = C[(s_0+\sigma)E - A]^{-1}B \\ &= C[\sigma E + (s_0E - A)]^{-1}B \\ &= C[I + \sigma(s_0E-A)^{-1}E]^{-1}[(s_0E-A)]^{-1}B \\ &= C[I - \sigma(s_0E-A)^{-1}E + \sigma^2[(s_0E-A)^{-1}E]^2 - \ldots]\times \\ &\quad (s_0E-A)^{-1}B \\ &= \sum_{i=0}^{\infty} \underbrace{C[-(s_0E-A)^{-1}E]^i(s_0E-A)^{-1}B}_{:=m_i(s_0)}\ \sigma^i, \end{aligned}$$

where $m_i(s_0)$, $i = 0, 1, 2, \ldots$ are called the moments of the transfer function. Note that the series expansion follows from the Neumann Lemma. If the expansion point is chosen as $s_0 = 0$, then the moments are simply $m_i(0) = -C(A^{-1}E)^iA^{-1}B$. For $s_0 = \infty$ and $E = I$, the moments are also called Markov parameters, $m_i(\infty) = CA^iB$. In fact, for $s_0 < \infty$, the moments $m_i(s_0)$ are nothing but the ith derivative of $G(s)$ at s_0, multiplied with an appropriate scalar $\frac{1}{i!}$.

The projection matrices $V \in \mathbb{R}^{n\times r}$ and $W \in \mathbb{R}^{n\times r}$ are computed from the moments $m_i(s_0)$,

$$\begin{aligned} \text{range}\{V\} &= \text{span}\{\tilde{B}(s_0),\ \tilde{A}(s_0)\tilde{B}(s_0),\ \ldots,\ \tilde{A}^{q-1}(s_0)\tilde{B}(s_0)\}, \\ \text{range}\{W\} &= \text{span}\{C^T,\ \tilde{A}^T(s_0)C^T,\ \ldots,\ (\tilde{A}^T(s_0))^{q-1}C^T\}, \end{aligned} \tag{10}$$

where $\tilde{A}(s_0) = (s_0E - A)^{-1}E$, $\tilde{B}(s_0) = (s_0E - A)^{-1}B$ and $q \ll n$. The following theorem shows that the transfer function of the reduced model computed by the above V and W interpolates the transfer function of the original system up to the $2q$-1th derivative of $G(s)$ at s_0 [10].

Theorem 9. *For a SISO system, if the columns of W and V are bases of the subspace in (10), then the transfer function $\hat{G}(s)$ of the reduced model matches the first $2q$ moments of the transfer function of the original system, i.e.*

$$m_i(s_0) = \hat{m}_i(s_0),\ i = 0, 1, \ldots, 2q-1,$$

where $\hat{m}_i(s_0) = \hat{C}[-(s_0\hat{E} - \hat{A})^{-1}\hat{E}]^i(s_0\hat{E} - \hat{A})^{-1}\hat{B},\ i = 0, 1, \ldots, 2q-1$ are the ith order moments of $\hat{G}$ and $\hat{E} = W^TEV$.

It is shown in [10], that the transfer function $\hat{G}(s)$ is a Padé approximant [3] of $G(s)$ for the SISO system.

For a MIMO system, it is hard to calculate the exact number of moments matched, especially for descriptor systems (when E is singular). In [13], it is shown that $\hat{G}(s)$ matches at least the first $\lfloor r/m \rfloor + \lfloor r/p \rfloor$ moments of $G(s)$, and it is a matrix Padé approximant of $G(s)$. Here r is the order of the reduced model, or equivalently, the number of the columns in V or W.

Methods based on rational interpolation

Instead of using a single expansion point, multiple expansion points can be used to have multiple series expansions of $G(s)$ around expansion points s_i, $i = 1, \ldots k$. The matrices V, W can be computed by the combined Krylov subspaces for each s_i, e.g.

$$\begin{aligned} \text{range}(V) &= \bigcup_{i=1}^{k} \mathcal{K}_{q_i}(\tilde{A}_B(s_i), \tilde{B}(s_i)), \\ \text{range}(W) &= \bigcup_{i=1}^{k} \mathcal{K}_{q_i}(\tilde{A}_C(s_i), \tilde{C}^T(s_i)), \end{aligned} \tag{11}$$

where $\tilde{A}_B(s_i) = (s_i E - A)^{-1} E$, $\tilde{A}_C(s_i) = (s_i E - A)^{-T} E^T$, $\tilde{C}^T(s_i) = (s_i E - A)^{-T} C^T$, and $\mathcal{K}_{q_i}(M, R)$ is the block Krylov subspace $\mathcal{K}_{q_i}(M, R) = \{R, MR, \ldots, M^{q_i - 1} R\}$ generated by a square matrix $M \in \mathbb{R}^{n \times n}$, and a rectangular matrix $R \in \mathbb{R}^{n \times n_R}$.

The resulting reduced model matches the first $2q_i$ moments $m_0(s_i), \ldots, m_{2q_i - 1}(s_i)$ at each s_i, $i = 1, \ldots, k$ for both SISO and MIMO systems [16]. In other words, the transfer function $\hat{G}(s)$ interpolates $G(s)$ at s_j, $j = 1, \ldots, k$, till the $2q_i$-1th order derivative. Notice that the starting matrix (vector) for W in (11) is $\tilde{C}(s_i) = (s_i E - A)^{-T} C^T$, rather than C^T in (10) used by the Padé approximation method. And $\tilde{A}_C(s_i)$ is not the transpose of $\tilde{A}_B(s_i)$, which is also different from the use of $\tilde{A}(s_0)$ and $\tilde{A}^T(s_0)$ in (10).

When the system is single-input and single-output, B and C are vectors and the matrices V, W in (10) can be simultaneously computed by the Lanczos algorithm [10], such that $W^T V = I$, i.e. the columns of W are biorthogonal with the columns of V. For a system with multiple inputs and multiple outputs, B and C are matrices, then the block Lanczos algorithm in [11] can be used to compute V and W in (10).

If only the matrix V is used to compute the reduced model, i.e. $W = V$, then the Arnoldi process can be applied to compute V in (10) for a SISO system, and the Band Arnoldi process in [12] can be applied to compute V for a MIMO system. For more discussions on the algorithms of computing V, W in (10), see [4, 12]. In [16], algorithms of computing V and W in (11) are discussed in detail.

Nowadays, more and more concerns are on automatic generation of the reduced model. For the methods based on rational interpolation, the question is how to adaptively select the interpolation points s_j, $j = 0, \ldots, k$. Many techniques have been proposed so far, though most of them are more or less heuristic. The algorithm IRKA proposed in [18] iteratively selects the interpolation points s_j, so that a necessary condition for a locally optimal reduced model is satisfied, and it is applicable to SISO systems. The algorithm is then extended to MIMO systems [18].

In the following sections, we only consider the standard state space system in (1).

4 Modal truncation method

The modal truncation method [9] is based on the eigendecomposition of the system matrix A in (1). Assume that A is diagonalizable, i.e. $T^{-1}AT = D_A$ (T can be a complex matrix), then the matrices V, W for the reduced model are constructed as

$$V = T(:, 1:r) = [t_1, \ldots, t_r],$$
$$\tilde{W}^* = T^{-1}(1:r, :), W = \tilde{W}(V^*\tilde{W})^{-1}.$$

Here, the columns in $T = [t_1, \ldots, t_n]$ are eigenvectors of A, $D_A = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ includes the eigenvalues of A. The matrix V is composed of the first r dominant eigenvectors of A, which corresponds to the eigenvalues closest to the imaginary axis. The eigendecomposition of A can be computed by, e.g. Krylov subspace methods, Jacobi-Davidson method.

The reduced model is given by $\hat{A} = W^*AV = \mathrm{diag}(\lambda_1, \ldots, \lambda_r)$, $\hat{B} = W^*B$, $\hat{C} = CV$. This is equivalent to doing truncation for the following matrices,

$$T^{-1}AT = \begin{bmatrix} \hat{A} & \\ & \hat{A}_2 \end{bmatrix}, T^{-1}B = \begin{bmatrix} \hat{B} \\ \hat{B}_2 \end{bmatrix}, CT = [\hat{C}, \hat{C}_2].$$

The error bound for the transfer function of the reduced model is

$$||G - \hat{G}||_\infty \leq ||C_2||||B_2|| \frac{1}{\min_{\lambda \in \Lambda(\hat{A}_2)} |Re(\lambda)|}.$$

The error bound is not computable for very large-scale systems, since the whole spectrum of A needs to be computed in principle.

The modal truncation method only uses information from A, the information from B and C is not taken use of, which might cause big errors. The performance of the method can be improved by the dominant pole algorithm [28], where A, B and C are used to measure the dominant poles. The

left and right eigenvectors corresponding to the dominant poles are used to construct the reduced model.

5 Balanced truncation method

The balanced truncation method was proposed in [25]. The basic principle of balanced truncation method is as follows. Firstly, the Gramian matrices P and Q are computed by solving the Lyapunov equations in (4). Secondly, a balancing matrix $T = T_b$ (see Theorem 3) is used to obtain a balanced system by state space transformation, $\tilde{A} = TAT^{-1}$, $\tilde{B} = TB$, $\tilde{C} = CT^{-1}$. It can be readily verified that the Gramians of the transformed system are diagonal matrices, i.e. $TPT^T = \Sigma$, $T^{-T}QT^{-1} = \Sigma$.

If $\sigma_{r+1} \ll \sigma_r$, $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ can be divided into two parts $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, $\Sigma_2 = diag(\sigma_{r+1}, \ldots, \sigma_n)$. According to this separation, $\tilde{A}$, $\tilde{B}$ ad $\tilde{C}$ can be divided as

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

where $A_{11} \in \mathbb{R}^{r\times r}$, $B_1 \in \mathbb{R}^{r\times m}$, $C_1 \in \mathbb{R}^{p\times r}$ correspond to Σ_1. The reduced model is constructed as $\hat{A} = A_{11}$, $\hat{B} = B_1$, $\hat{C} = C_1$, $\hat{D} = D$.

The motivation of balanced truncation is that the HSVs are the invariants of the system, which means HSVs do not change under state space transformation. Once a system is balanced, the smallest HSVs can be easily distinguished from the diagonalized Gramian Σ, and the system can be truncated according to the separation of Σ. With the deletion of the smallest HSVs, the unimportant states which are difficult to observe and difficult to control are truncated from the system [1], so that only important information of the original system is retained in the reduced model.

In practice, the reduced model is obtained not by explicitly forming the balanced system, instead, the square root (SR) method is used to compute the balanced reduced model. The basic idea is to use the SVD decomposition of SR^T,

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

The two matrices V and W are computed as $W = R^T V_1 \Sigma_1^{-\frac{1}{2}}$, $V = S^T U_1 \Sigma_1^{-\frac{1}{2}}$. The reduced system matrices are $\hat{A} = W^T AV$, $\hat{B} = W^T B$, $\hat{C} = CV$. It is easily verified that $W^T V = I$, so that VW^T is an oblique projector, hence balanced truncation method is a Petrov-Galerkin projection method.

An important property of balanced truncation method is the computable error bound,

$$||G - \hat{G}||_\infty \leq 2 \sum_{j=r+1}^{n} \sigma_j, \tag{12}$$

then from (8), we get

$$||y - \hat{y}||_2 \leq \left(2 \sum_{j=r+1}^{n} \sigma_j \right) ||u||_2.$$

From the error bound, the reduced model can be automatically obtained by adaptively choosing r according to the desired accuracy. The properties of the reduced model computed by balanced truncation are summarized in the following theorem.

Theorem 10. *Let the reduced-order system $\hat{\Sigma} : (\hat{A}, \hat{B}, \hat{C}, \hat{D})$ with $r \leq n$ be computed by balanced truncation. Then the reduced-order model $\hat{\Sigma}$ is balanced, stable, minimal, and its HSVs are $\sigma_1, \ldots, \sigma_r$.*

6 Balancing related methods

The balancing related methods were developed for different purposes of model reduction. The linear-quadratic Gaussian balanced truncation (LQGBT) method in [20] can be used as a model reduction method for unstable systems, and it also provides a closed-loop balancing technique. Compared with the standard balanced truncation method in Section 5, the only difference is that the controllability and the observability Gramians are replaced by the solutions P, Q of the dual algebraic Riccati equations (AREs)

$$\begin{aligned} AP + PA^T - PC^TCP + BB^T &= 0, \\ A^TQ + QA - QBB^TQ + C^TC &= 0. \end{aligned}$$

The stochastic balancing method (BST) firstly appeared in [8] for balancing stochastic systems, and was generalized in [14], where a relative error bound for the reduced model is proposed. Instead of solving two Lyapunov equations required by the standard balanced truncation method, one Laypunov equation and one ARE must be solved to get the Gramians P and Q,

$$\begin{aligned} AP + PA^T + BB^T &= 0, \\ \bar{A}^TQ + Q\bar{A} + QB_W(DD^T)^{-1}B_W^TQ + C^T(DD^T)^{-1}C &= 0, \end{aligned}$$

where $\bar{A} := A - B_W(DD^T)^{-1}C, B_W := BD^T + PC^T$.

The positive real balanced truncation method [8, 15] is applicable for positive real systems, also called passive systems. The method is based on positive-real equations, related to positive real (Kalman-Yakubovich-Popov-Anderson) lemma. The following two AREs need to be solved,

$$\bar{A}P + P\bar{A}^T + PC^T\bar{R}^{-1}CP + B\bar{R}^{-1}B^T = 0,$$
$$\bar{A}^TQ + Q\bar{A} + QB\bar{R}^{-1}B^TQ + C^T\bar{R}^{-1}C = 0,$$

where $\bar{A} := A - B\bar{R}^{-1}C, \bar{R} := D + D^T$.

In contrast to the error bound for the standard balanced truncation method in (12), the computable error bounds for the LQGBT method and the BST method are

$$\text{LQGBT} : ||G - \hat{G}||_\infty \leq 2\sum_{j=r+1}^{n} \frac{\sigma_j^{LQG}}{\sqrt{1+(\sigma_j^{LQG})^2}},$$

$$\text{BST} : ||G - \hat{G}||_\infty \leq \left(\prod_{j=r+1}^{n} \frac{1+\sigma_j^{BST}}{1-(\sigma_j^{BST})} - 1\right) ||G||_\infty,$$

Actually, the error bound for the BST method is an error bound for the relative error.

Other balancing-based methods include bounded-real balanced truncation method [26], H_∞ balanced truncation method [24], as well as frequency-weighted versions of the above approaches. A good textbook for learning the balanced truncation methods is [1], where the mathematical basics required for model reduction are also provided. For a restudy of modal truncation method and details of dominant pole method, please refer to the thesis [28]. In the thesis [16], methods based on Padé approximation are reviewed, and method based on rational interpolation are proposed.

7 Solving matrix equations

The major computational part of the balanced truncation methods or the balancing related methods is solving the large-scale matrix equations. The efficiency of these model order reduction methods depends on fast numerical algorithms of solving the matrix equations.

Solvability and complexity issues

Consider the Sylvester equation $AX + XB + W = 0$, $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$, $X \in \mathbb{R}^{n\times m}$, $W \in \mathbb{R}^{n\times m}$, using the Kronecker (tensor) product, $AX + XB + W = 0$ is equivalent to

$$\big((I_m \otimes A) + \big(B^T \otimes I_n\big)\big)\,\mathrm{vec}(X) = \mathrm{vec}(-W). \tag{13}$$

Observing that

$M := (I_m \otimes A) + \big(B^T \otimes I_n\big)$ is invertible $\Longleftrightarrow$
$0 \notin \Lambda(M) = \Lambda\big((I_m \otimes A) + (B^T \otimes I_n)\big) = \{\lambda_j + \mu_k, \mid \lambda_j \in \Lambda(A),\ \mu_k \in \Lambda(B)\}$
$\Longleftrightarrow \Lambda(A) \cap \Lambda(-B) = \emptyset$,

we have the following corollary,

Corollary 1. *If A, B are Hurwitz matrices, then the Sylvester equation $AX + XB + W = 0$ has unique solution.*

Note that when $B = A^T$, we get the Lyapunov equation

$$AX + XA^T + W = 0. \tag{14}$$

A straightforward way of solving the Sylvester equation is via the equivalent linear system of equations in (13). This requires LU factorization of a $nm \times nm$ matrix; for $n \approx m$, the computational complexity is $\frac{2}{3}n^6$. The storage memory is also unacceptable, since we need n^4 data for X.

Traditional methods of solving the matrix equations include the Bartels-Stewart method for Sylvester and Lyapunov equations, the Hessenberg-Schur method for Sylvester equations, and Hammarling's method for Lyapunov equations with A Hurwitz.

All methods are based on the fact that if A, B^T are in Schur form, then

$$M = (I_m \otimes A) + \big(B^T \otimes I_n\big)$$

is block-upper triangular. Hence, $Mx = b$ can be solved by back-substitution.

However, clever implementation of the back-substitution process still requires $nm(n+m)$ flops. All methods require Schur decomposition of A and/or Schur or Hessenberg decomposition of B, which requires $25n^3$ flops for Schur decomposition. Therefore, these methods are not feasible for large-scale problems with $n > 10,000$.

The sign function method

The sign function method is used to solve the Lyapunov equation in (14).

Definition 12. For $Z \in \mathbb{R}^{n\times n}$ with $\Lambda(Z) \cap \imath\mathbb{R} = \emptyset$ and Jordan canonical form

$$Z = S \begin{bmatrix} J^+ & 0 \\ 0 & J^- \end{bmatrix} S^{-1}$$

the matrix sign function is

$$\operatorname{sign}(Z) := S \begin{bmatrix} I_k & 0 \\ 0 & -I_{n-k} \end{bmatrix} S^{-1}.$$

Lemma 5. *Let $T \in \mathbb{R}^{n\times n}$ be nonsingular and Z as above, then*

$$\operatorname{sign}(TZT^{-1}) = T\operatorname{sign}(Z)T^{-1}.$$

Since $\operatorname{sign}(Z)$ is the square root of I_n, i.e. $(\operatorname{sign}(Z))^2 - I_n = 0$, one can use Newton's method to get $\operatorname{sign}(Z)$ by solving $f(\tilde{Z}) := \tilde{Z}^2 - I_n = 0$:

$$\tilde{Z}_0 \leftarrow Z, \qquad \tilde{Z}_{j+1} \leftarrow \frac{1}{2}\left(c_j\tilde{Z}_j + \frac{1}{c_j}\tilde{Z}_j^{-1}\right), \qquad j = 1, 2, \ldots, \tag{15}$$

finally, $\operatorname{sign}(Z) = \lim_{j\to\infty}\tilde{Z}_j$ [19]. The variable $c_j > 0$ is a scaling parameter for convergence acceleration and rounding error minimization, e.g.

$$c_j = \sqrt{\frac{||\tilde{Z}_j^{-1}||_F}{||\tilde{Z}_j||_F}},$$

based on "equilibrating" the norms of the two summands.

Solving the Lyapunov equation in (14) with the matrix sign function method is based on the following observation. If $X \in \mathbb{R}^{n\times n}$ is a solution of (14), then

$$\underbrace{\begin{bmatrix} I_n & -X \\ 0 & I_n \end{bmatrix}}_{=T^{-1}}\underbrace{\begin{bmatrix} A & W \\ 0 & -A^T \end{bmatrix}}_{=:H}\underbrace{\begin{bmatrix} I_n & X \\ 0 & I_n \end{bmatrix}}_{=:T} = \begin{bmatrix} A & 0 \\ 0 & -A^T \end{bmatrix}.$$

Hence, if A is Hurwitz (i.e., asymptotically stable), then

$$\begin{aligned}\operatorname{sign}(H) &= \operatorname{sign}\left(T\begin{bmatrix} A & 0 \\ 0 & -A^T \end{bmatrix}T^{-1}\right) = T\operatorname{sign}\left(\begin{bmatrix} A & 0 \\ 0 & -A^T \end{bmatrix}\right)T^{-1} \\ &= \begin{bmatrix} -I_n & 2X \\ 0 & I_n \end{bmatrix}.\end{aligned}$$

Apply the sign function iteration in (15): $\tilde{Z} \leftarrow \frac{1}{2}(\tilde{Z} + \tilde{Z}^{-1})$ ($c_j = 1$) using $\tilde{Z}_0 = H = \begin{bmatrix} A & W \\ 0 & -A^T \end{bmatrix}$, and observe that

$$H + H^{-1} = \begin{bmatrix} A & W \\ 0 & -A^T \end{bmatrix} + \begin{bmatrix} A^{-1} & A^{-1}WA^{-T} \\ 0 & -A^{-T} \end{bmatrix},$$

we get the sign function iteration for the Lyapunov equation:

$$\begin{aligned} A_0 &\leftarrow A, & A_{j+1} &\leftarrow \tfrac{1}{2}\left(A_j + A_j^{-1}\right), \\ W_0 &\leftarrow W, & W_{j+1} &\leftarrow \tfrac{1}{2}\left(W_j + A_j^{-1}W_jA_j^{-T}\right), \end{aligned} \qquad j = 0, 1, 2, \ldots. \tag{16}$$

Define $A_\infty := \lim_{j\to\infty} A_j$, $W_\infty := \lim_{j\to\infty} W_j$, we immediately get the following theorem.

Theorem 11. *If A is Hurwitz, then*

$$A_\infty = -I_n \qquad \textit{and} \qquad X = \frac{1}{2}W_\infty.$$

Now consider the second iteration in (16) for $W_j = B_j B_j^T$, starting with $W_0 = BB^T =: B_0 B_0^T$, one can see that

$$\begin{aligned} \tfrac{1}{2}\left(W_j + A_j^{-1} W_j A_j^{-T}\right) &= \tfrac{1}{2}\left(B_j B_j^T + A_j^{-1} B_j B_j^T A_j^{-T}\right) \\ &= \tfrac{1}{2}\left[B_j \; A_j^{-1} B_j\right]\left[B_j \; A_j^{-1} B_j\right]^T. \end{aligned}$$

Hence, the factored iteration for the sign function method is [7],

$$B_{j+1} \;\leftarrow\; \frac{1}{\sqrt{2}}\left[B_j \; A_j^{-1} B_j\right] \tag{17}$$

with $S := \frac{1}{\sqrt{2}} \lim_{j\to\infty} B_j$ and $X = SS^T$. From Theorem 11, a simple stopping criterion is taken as $||A_j + I_n||_F \leq tol$. It is clear that the iteration in (17) can be used to solve the Lyapunov equations in (4) to get the controllability and observability Gramians P and Q.

The alternating direction implicit (ADI) method

The Peaceman Rachford ADI method was originally used to solve the linear system $Au = b$ where $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite and $b \in \mathbb{R}^n$. The idea is to decompose $A = H + V$ with $H, V \in \mathbb{R}^{n\times n}$ such that

$$\begin{aligned} (H + pI)v &= r \\ (V + pI)w &= t \end{aligned}$$

can be solved easily or efficiently. The standard ADI iteration for solving $Au = b$ is as follows. If H, V are symmetric positive definite matrices, then $\exists p_k$, $k = 1, 2, \ldots$ such that

$$\begin{aligned} u_0 &= 0 \\ (H + p_k I)u_{k-\frac{1}{2}} &= (p_k I - V)u_{k-1} + b \\ (V + p_k I)u_k &= (p_k I - H)u_{k-\frac{1}{2}} + b \end{aligned}$$

converges to $u \in \mathbb{R}^n$ solving $Au = b$.

Notice that the Lyapunov operator $\mathcal{L} : P \mapsto AX + XA^T$ can be decomposed into the linear operators,

$$\mathcal{L}_H : X \mapsto AX, \qquad \mathcal{L}_V : X \mapsto XA^T.$$

In analogy to the standard ADI method, we find the ADI iteration for the Lyapunov equation $AX + XA^T + W = 0$ [30],

$$\begin{aligned} X_0 &= 0 \\ (A + p_k I)X_{k-\frac{1}{2}} &= -W - X_{k-1}(A^T - p_k I) \\ (A + p_k I)X_k^T &= -W - X_{k-\frac{1}{2}}^T(A^T - p_k I). \end{aligned}$$

Consider applying the above ADI iteration to the Lyapunov equation $AX + XA^T + BB^T = 0$ for a stable matrix $A \in \mathbb{R}^{n\times n}$, with $B \in \mathbb{R}^{n\times m}$, $m \ll n$. The two step ADI iteration can be rewritten into one step by removing $X_{k-\frac{1}{2}}$,

$$\begin{aligned} Z_0 Z_0^T &= 0, \\ Z_k Z_k^T &= -2p_k(A + p_k I)^{-1}BB^T(A + p_k I)^{-T} \\ &\quad + (A + p_k I)^{-1}(A - p_k I)Z_{k-1}Z_{k-1}^T(A - p_k I)^T(A + p_k I)^{-T}, \end{aligned}$$

with the low-rank factorization of X_k, $X_k = Z_k Z_k^T$, $k = 0, \ldots, k_{\max}$, $Z_k \in \mathbb{R}^{n\times r_k}$, $r_k \ll n$. This is the scheme of low-rank (vector) ADI method [5, 21, 17, 27]. From the above iteration for $Z_k Z_k^T$, it is easily known that the low-rank factor Z_k of X_k can be iteratively computed as

$$Z_k = [\sqrt{-2p_k}(A + p_k I)^{-1}B,\ (A + p_k I)^{-1}(A - p_k I)Z_{k-1}],$$

so that in practical implementations only Z_k is iterated.

It is noticed that at each iteration step k, the number of vectors needing to be updated in Z_k increases by m. A more efficient algorithm of computing Z_k is proposed in [21], which keeps the number of updated vectors constant at each iteration step.

Assuming $k_{\max}$ is the maximal number of iterations, and observing that $(A - p_i I)$, $(A + p_k I)^{-1}$ commute, then at the last step k_{max}, $Z_{k_{\max}-1}$ can be rewritten as [21],

$$Z_{k_{\max}-1} = \left[z_{k_{\max}},\ P_{k_{\max}-1}z_{k_{\max}},\ P_{k_{\max}-2}(P_{k_{\max}-1}z_{k_{\max}}),\ \ldots,\ P_1(P_2 \cdots P_{k_{\max}-1}z_{k_{\max}})\right]. \tag{18}$$

$$z_{k_{\max}} = \sqrt{-2p_{k_{\max}}}(A + p_{k_{\max}} I)^{-1}B$$

and

$$P_i := \frac{\sqrt{-2p_i}}{\sqrt{-2p_{i+1}}}\left[I - (p_i + p_{i+1})(A + p_i I)^{-1}\right].$$

From (18), we derive the iteration for Z_k, $k = 0, 1, \ldots, k_{max} - 1$,

$$\begin{aligned} Z_0 &= z_{k_{max}}, \\ Z_k &= \left[z_{k_{max}}, P_{k_{\max}-1}z_{k_{\max}}, \ldots, P_{k_{\max}-k}\ldots(P_{k_{\max}-1}z_{k_{\max}})\right], \end{aligned} \tag{19}$$

where the number of updated vectors at each step is always m.

Factored Galerkin-ADI iteration method

Factored Galerkin-ADI iteration method is a projection-based method for solving Lyapunov equations with $A + A^T < 0$. The basic steps are

1. Compute orthonormal basis $Z = [z_1, \ldots, z_r] \in \mathbb{R}^{n\times r}$ of subspace $\mathcal{Z} \subset \mathbb{R}^n$ ($\dim \mathcal{Z} = r$), i.e. range$(Z) = \mathcal{Z}$.
2. Set $A_r := Z^T AZ$, $B_r := Z^T B$.
3. Solve small-size Lyapunov equation $A_r\hat{X} + \hat{X}A_r^T + B_r B_r^T = 0$.
4. Use $X \approx Z\hat{X}Z^T$.

The subspace $\mathcal{Z}$ can be taken as, e.g.

$$\mathcal{Z} = \mathcal{K}_r(A, B) = \text{span}\{B, AB, A^2B, \ldots, A^{r-1}B\},$$

which corresponds to the Krylov subspace methods.

The K-PIK method uses the combined subspace

$$\mathcal{Z} = \mathcal{K}_r(A, B) \cup \mathcal{K}_r(A^{-1}, A^{-1}B).$$

The rational Krylov subspace method uses

$$\mathcal{Z} = \text{colspan}\{(A - s_1)^{-1}B, \ldots, (A - s_r)^{-1}B\}.$$

$\mathcal{Z}$ can also be taken as the ADI subspace

$$\mathcal{Z} = \text{colspan}\{z_{k_{max}}, P_{k_{\max}-1}z_{k_{\max}}, \ldots, P_{k_{\max}-r+1}\ldots(P_{k_{\max}-1}z_{k_{\max}})\}.$$

The ADI subspace is proved to be a rational Krylov subspace in [21]. In the following subsections, we discuss the numerical methods for solving large-scale algebraic Riccati equation (ARE)

$$A^T X + XA - XBB^T X + C^T C = 0.$$

Newton's method for AREs

Consider the ARE,

$$0 = \mathcal{R}(X) := A^T X + XA - XBB^T X + C^T C, \tag{20}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$. The Frechét derivative of $\mathcal{R}(X)$ at X is $\mathcal{R}'_X : Z \mapsto (A - BB^T X)^T Z + Z(A - BB^T X)$.

Newton-Kantorovich method follows the iteration $X_{j+1} = X_j - \left(\mathcal{R}'_{X_j}\right)^{-1}\mathcal{R}(X_j)$, $j = 0, 1, 2, \ldots$, and can be described by Algorithm 1.

Algorithm 1 Newton's method (with line search) for AREs

1: FOR $j = 0, 1, \ldots$
2: $A_j \leftarrow A - BB^T X_j =: A - BK_j$.
3: Solve the Lyapunov equation $A_j^T N_j + N_j A_j = -\mathcal{R}(X_j)$.
4: $X_{j+1} \leftarrow X_j + t_j N_j$.
5: ENDFOR j

If $A_j = A - BK_j = A - BB^T X_j$ is stable $\forall\ j \geq 0$, then $\mathcal{R}(X_j)$ converges to zero, $\lim_{j\to\infty} ||\mathcal{R}(X_j)||_F = 0$, so X_j converges to the solution of ARE, $\lim_{j\to\infty} X_j = X_* \geq 0$. It is seen that during the algorithm, large-scale Lyapunov equations need to be efficiently solved, where the algorithms discussed in the above two subsections can be applied.

Low-Rank Newton-ADI for AREs

If we re-write Newton's method for AREs, in particular Step 3 in Algorithm 1, we get

$$A_j^T \underbrace{(X_j + N_j)}_{=X_{j+1}} + \underbrace{(X_j + N_j)}_{=X_{j+1}} A_j = \underbrace{-C^T C - X_j B B^T X_j}_{=:-W_j W_j^T}$$

Set $X_j = Z_j Z_j^T$ for $\operatorname{rank}(Z_j) \ll n$, we have

$$A_j^T \left(Z_{j+1} Z_{j+1}^T\right) + \left(Z_{j+1} Z_{j+1}^T\right) A_j + W_j W_j^T = 0 \tag{21}$$

Then $Z_{j+1}, j = 0, 1, \ldots$ can be obtained by solving Lyapunov equations in (21) with the factored ADI iteration in (19), so that Algorithm1 is combined with the low-rank ADI methods.

Software

In the toolbox LYAPACK, there are MATLAB routines for solving large, sparse Lyapunov equations, and AREs equations. The main methods used are low-rank ADI and Newton-ADI iterations. It can be downloaded from [32].

The Matrix Equations and Sparse Solvers library(MESS) [33], is the extended and revised version of the LYAPACK Toolbox. It includes solvers for large-scale differential Riccati equations.There are many algorithmic improvements, for example, new ADI parameter selection, column compression based on RRQR algorithm, a more efficient use of direct solvers, treatment of generalized systems without factorization of the mass matrix, new ADI versions avoiding complex arithmetic etc. It is available as a MATLAB toolbox,

as well as a C-library. The C version provides a large set of axillary subroutines for sparse matrix computations and efficient usage of modern multicore workstations.

8 Conclusion

In this lecture, popular model order reduction methods applicable for non-parametric LTI systems are discussed. The numerical algorithms for solving large-scale matrix equations are explored.

Most of the above discussed methods can be either directly applied, or extended to treating descriptor systems $E\dot{x} = Ax + Bu$, E singular. Some methods are generalized for bilinear and stochastic systems. Methods based on Padé approximation/rational interpolation are also extended for nonlinear systems. The well-known proper orthogonal decomposition (POD) method for nonlinear systems is not covered in the lecture. Parametric model order reduction (PMOR) for parametric systems, such as

$$\dot{x} = A(p)x + B(p)u, \quad y = C(p)x,$$

where $p \in \mathbb{R}^d$ is a free parameter vector, is a huge topic, and cannot be included in the lecture either. For a survey of PMOR methods, see e.g. [6]. For a wide view about various MOR methods for different complex systems, please visit the MORwiki [34], where one is also free to share the benchmarks for testing and verifying MOR algorithms.

References

1. A.C. Antoulas. Approximation of Large-Scale Dynamical Systems. *SIAM Publications*, Philadelphia, PA, 2005.
2. A.C. Antoulas, C.A. Beattie, and S. Gugercin. Interpolatory model reduction of large-scale dynamical systems. *in Efficient Modeling and Control of Large-Scale Systems*, Javad Mohammadpour and Karolos M. Grigoriadis, eds., Springer US, 3–58, 2010.
3. G.A. Baker, Jr. and P. Graves-Morris. Padé Approximants, Second Edition. *Cambridge University Press*, New York, 1996.
4. Z. Bai. Krylov subspace techniques for reduced-order modeling of large-scale dynamical systems. *Applied Numerical Mathematics*, 43(1–2):9–44, 2002.
5. P. Benner and J.-R. Li and T. Penzl. Numerical Solution of Large Lyapunov Equations, Riccati Equations, and Linear-Quadratic Control Problems. *Numerical Algorithms*, 15(9):755–777, 2008.
6. P. Benner, S. Gugercin, and K. Willcox. A survey of model reduction methods for parametric systems, *MPI Magdeburg Preprints*, 2013.
7. P. Benner and E. S. Quintana-Orti. Solving stable generalized Lyapunov equations with the matrix sign function. *Numerical Algorithms*, 20:75–100, 1999.
8. U. B. Desai and D. Pal. A transformation approach to stochastic model reduction. *IEEE Transactions on Automatic Control*, AC-29(12):1097–1100, 1984.

9. E. J. Davison. A method for simplifying linear dynamic systems. *IEEE Transactions on Automatic Control*, AC-11(1):93–101, 1966.
10. P. Feldmann, R. W. Freund. Efficient linear circuit analysis by Padé approximation via the Lanczos process. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.*, 14(5):639–649, 1995.
11. P. Feldmann, R.W Freund. Reduced-order modeling of large linear subcircuits via a block Lanczos algorithm. *Proc. 32nd ACM/IEEE Design Automation Conf.*, , 474-479, 1995.
12. R. W. Freund. Model reduction methods based on Krylov subspaces. *Acta Numerica*, 12:267–319, 2003.
13. R.W. Freund. Krylov Subspace methods for reduced-order modeling in circuit simulation. *JCAP.*, 123:395-421, 2000.
14. M. Green. A relative error bound for balanced stochastic truncation. *IEEE Transactions on Automatic Control*, AC-33:961-965, 1988.
15. M. Green. Balanced stochastic realizations. *Linear Algebra and its Applications*, 98:211-247, 1988.
16. E. J. Grimme. Krylov projection methods for model reduction. PhD thesis, Univ. Illinois, Urbana-Champaign, 1997.
17. S. Gugercin, D.C. Sorensen, and A. C. Antoulas. A modified low-rank Smith method for large-scale Lyapunov equations. *Numerical Algorithms*, 32(1):27–55, 2003.
18. S. Gugercin, A. C. Antoulas, and C. A. Beattie. $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems. *SIAM J. Matrix Anal. Appl.*, 30(2):609–638, 2008.
19. N. J. Higham. Functions of Matrices: Theory and Computation. *SIAM*, 2008.
20. E. Jonckheere and L. Silverman. A new set of invariants for linear systems–application to reduced order compensator. *IEEE Trans Automat. Control*, AC-28:953-964, 1983.
21. J.-R. Li and J. White. Low Rank Solution of Lyapunov Equations. *SIAM J. Matrix Anal. Appl.*, 24(1):260–280, 2002.
22. A. J. Laub, M. T. Heat, C. C. Paige and R. C. Ward. Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms. *IEEE Transactions on Automatic Control*, AC-32(2):115–122, 1987.
23. Maximum-modulus principle. *Encyclopedia of Mathematics.* http://www.encyclopediaofmath.org/index.php/Maximum-modulus_principle
24. D. Mustafa and K. Glover. Controller reduction by $\mathcal{H}_\infty$-balanced truncation. *IEEE Transactions on Automatic Control*, 36:668–682, 1991.
25. B. C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, AC-26:17–32, 1981.
26. P. C. Opdenacker and E. A. Jonckheere. A contraction mapping preserving balanced reduction scheme and its infinity norm error bounds. *IEEE Transactions on Circuits and Systems*, CAS 35:184–189, 1988.
27. T. Penzl. A cyclic low-rank Smith method for large sparse Lyapunov equations. *SIAM J. Sci. Comput.*, 21(4):1401-1418, 2000.
28. J. Rommes. Methods for eigenvalue problems with applications in model order reduction. PhD thesis, Utrecht University (Netherlands), 2007.
29. M. Tombs and I. Postlethwaite. Truncated balanced realization of a stable non-minimal state-space system. *International Journal of Control*, 46:1319–1330, 1987.
30. E. L. Wachspress. Iterative solution of the Lyapunov matrix equation. *Applied Mathematics Letters* 1(1), 87–90, 1988.
31. E.L. Wachspress. The ADI Model Problem, Windsor, CA, 1995
32. LYAPACK, http://www.tu-chemnitz.de/sfb393/lyapack/.
33. MESS, http://svncsc.mpi-magdeburg.mpg.de/trac/messtrac.
34. http://www.modelreduction.org.