Lecture Notes

“Model Reduction”

University of Hamburg
(summer term 2019)

Dr. Matthias Voigt
matthias.voigt@uni-hamburg.de

Preface

This document is based in large parts on the hand-written lecture notes of Christian Schröder who gave this course on model reduction at the TU Berlin in the winter term 2016/17. Special thanks go to Martijn Nagtegaal, Nora Heinrich, and Ines Ahrens for finding so many typos in the initial version of these lecture notes from winter term 2017/18. I believe that there are more errors and typos in this document, please send an email to `matthias.voigt@uni-hamburg.de` if you find any.

There are not many textbooks on model reduction, the most commonly known one has been written by A. C. Antoulas:

> A. C. Antoulas. Approximation of Large-Scale Dynamical Systems, volume 6 of Adv. Des. Control. SIAM Publications, Philadelphia, PA, 2005. doi:10.1137/1.9780898718713.

Most aspects discussed in this course have also been covered by Peter Benner on the Gene Golub SIAM Summer School 2013 at Fudan University in Shanghai, China. Some more applications and illustrative examples on model reduction can be found in his slides that you can download from the summer school's website[1]. Since this course is strongly based on control-theoretic basics, I recommend to read Chapters 3 and 4 of

> K. Zhou, J. C. Doyle, and K. Glover. Robust and Optimal Control, Prentice-Hall, Englewood Cliffs, NJ, 1996.

to look up these concepts. Further, more recent results discussed here will be cited throughout the lecture notes, so that you can read the original sources.

[1] http://g2s3.cs.ucdavis.edu/lecturers/Benner/Benner-lectures-online.pdf

Contents

CHAPTER 1

Introduction

1.1 What is Model Reduction?

Today, for the study of real-world processes, one usually sets up mathematical models usually consisting of differential (or differential-algebraic) equations that describe the behavior of the system under consideration. However, there is an ever-increasing need for higher accuracy which means that these models get more and more complex. The simulation, optimization, and control using such models then often leads to a very high demand in computational resources, both in terms of consumed time and memory – often even forbidding performing the desired task at all. Therefore, there is need for replacing the complex mathematical model by a much simpler model, that approximately behaves like the original model but which is computationally much less demanding. The process of finding this simpler representation is called *model reduction*. The typical set-up is depicted in Figure 1.1.

In this course we mainly consider *control systems* of the general form

$$\begin{aligned} \dot{x}(t) &= f(t, x(t), u(t)), \quad x(t_0) = x_0 \\ y(t) &= g(t, x(t), u(t)), \end{aligned} \tag{1.1}$$

where $\mathbb{I} = [t_0, t_\mathrm{f}]$ is a time interval of interest, $x : \mathbb{I} \to \mathbb{R}^n$ is the *state function* with initial value $x_0 \in \mathbb{R}^n$, $u : \mathbb{I} \to \mathbb{R}^m$ is the *input function*, $y : \mathbb{I} \to \mathbb{R}^p$ is the *output function*, and $f : \mathbb{I} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $g : \mathbb{I} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$. Usually, the input is a function that can be used to control the state of the system to

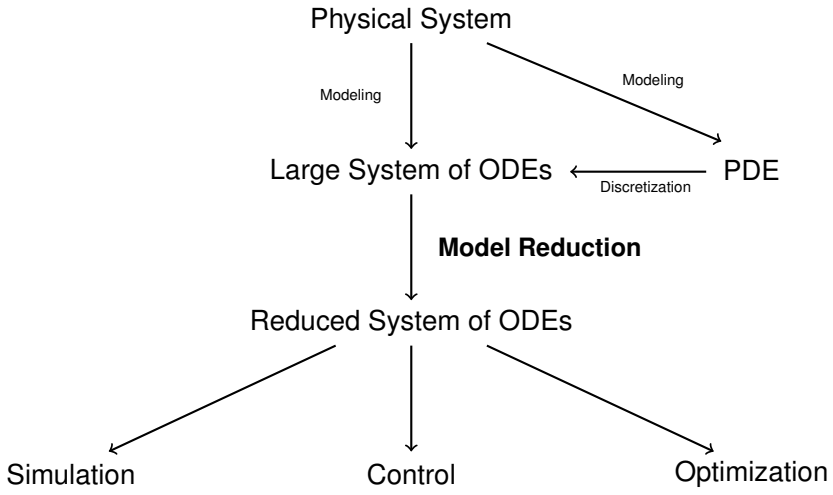

Figure 1.1: The broad setup of model reduction.

achieve a desired behavior. The output consists of "quantities of interest" that can often be measured in the real physical process.

The goal of model reduction is to replace the functions f and g in (1.1) by a *reduced-order model*

$$\begin{aligned} \dot{\widetilde{x}}(t) &= \widetilde{f}(t, \widetilde{x}(t), u(t)), \quad \widetilde{x}(t_0) = \widetilde{x}_0 \\ \widetilde{y}(t) &= \widetilde{g}(t, \widetilde{x}(t), u(t)), \end{aligned} \tag{1.2}$$

where $\widetilde{x} : \mathbb{I} \to \mathbb{R}^r$ with $r \ll n$ is the reduced state function with initial value $\widetilde{x}_0 \in \mathbb{R}^r$, and $\widetilde{f} : \mathbb{I} \times \mathbb{R}^r \times \mathbb{R}^m \to \mathbb{R}^n$, $\widetilde{g} : \mathbb{I} \times \mathbb{R}^r \times \mathbb{R}^m \to \mathbb{R}^p$. This model should be constructed such that $\|y - \widetilde{y}\|$ is "small" for all admissible inputs u. This will be made precise later. Note that we are only interested in the map from the input to the output, not in the evolution of the state itself.

In this course we focus on *linear time-invariant systems*, which are of the simpler form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{1.3}$$

for some matrices $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $C \in \mathbb{R}^{p \times n}$, and $D \in \mathbb{R}^{p \times m}$. Normally, we will also assume that $x(t_0) = x(0) = 0$.

We will discuss a rigorous mathematical theory for model reduction. We will discuss efficient numerical algorithms as well as theorems on the approximation quality, e. g., we state and prove error bounds. We will also touch on aspects of structure-preservation. This means, that if the original model has a certain structure, then also the reduced model should have this structure to account for physical properties that are encoded in the model.

1.2 Examples of Large-Scale Dynamical Systems

1.2.1 A Controlled Discretized Heat Equation

Consider the temperature distribution $T(t,\xi)$ of a one-dimensional beam of length $\ell = 1$. Here, $\xi \in [0,1]$ is the space variable and $t \geqslant 0$ denotes the time. On the right end of the beam we impose the boundary condition $T(t,1) = 0$ for all $t \geqslant 0$. On the left end we have a heat source that results in a controllable temperature flux

$$-\frac{\partial}{\partial \xi} T(t,0) = u(t).$$

The heat diffusion inside the beam is described by the heat equation

$$\frac{\partial}{\partial t} T(t,\xi) = k \cdot \frac{\partial^2}{\partial \xi^2} T(t,\xi) \quad \text{for all } \xi \in (0,1),\ t > 0.$$

Moreover, we are interested in the average beam temperature, i. e., our output is

$$y(t) = \int_0^1 T(t,\xi) \mathrm{d}\xi.$$

Finally, we need an initial condition which is given by

$$T(0,\xi) = 0 \quad \text{for all } \xi \in [0,1].$$

Now we discretize in space at n equidistant points and obtain

$$x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} := \begin{bmatrix} T(t,0) \\ T\left(t, \frac{1}{n}\right) \\ \vdots \\ T\left(t, \frac{n-1}{n}\right) \end{bmatrix}.$$

For $i = 2,\ 3,\ \ldots,\ n-1$ we find

$$\begin{aligned} \dot{x}_i(t) &= \frac{\partial}{\partial t} T\left(t, \tfrac{i-1}{n}\right) = k \cdot \frac{\partial^2}{\partial \xi^2} T\left(t, \tfrac{i-1}{n}\right) \\ &\approx k \cdot n^2 \left(T\left(t, \tfrac{i-2}{n}\right) - 2T\left(t, \tfrac{i-1}{n}\right) + T\left(t, \tfrac{i}{n}\right) \right) \\ &= k \cdot n^2 \left(x_{i-1}(t) - 2x_i(t) + x_{i+1}(t) \right). \end{aligned}$$

Analogously, we find

$$\dot{x}_n(t) \approx k \cdot n^2 (x_{n-1}(t) - 2x_n(t)),$$

since $x_{n+1}(t) := T(t, \frac{n}{n}) = 0$. Moreover, we have

$$\begin{aligned} \dot{x}_1(t) &= \frac{\partial}{\partial t} T\left(t,0\right) = k \cdot \frac{\partial^2}{\partial \xi^2} T\left(t,0\right) \\ &\approx k \cdot n \left(\frac{\partial}{\partial \xi} T\left(t, \tfrac{1}{n}\right) - \frac{\partial}{\partial \xi} T(t,0) \right) \\ &\approx k \cdot n \left(n(x_2(t) - x_1(t)) + u(t) \right). \end{aligned}$$

For the output we take a piecewise constant approximation, i. e.,

$$y(t) = \int_0^1 T(t,\xi)\mathrm{d}\xi \approx \frac{1}{n}\sum_{i=0}^{n-1} T(t,\tfrac{i}{n}) = \frac{1}{n}\sum_{i=1}^{n} x_i(t).$$

The zero initial conditions imply $x(0) = 0$. Our final controlled discretized heat equation now attains the form

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad x(0) = 0, \\ y(t) &= Cx(t), \end{aligned} \tag{1.4}$$

where

$$\begin{aligned} A &= kn^2 \begin{bmatrix} -1 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad B = kn \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{n\times 1}, \\ C &= \frac{1}{n}\begin{bmatrix} 1 & \dots & 1 \end{bmatrix} \in \mathbb{R}^{1\times n}. \end{aligned} \tag{1.5}$$

The larger n the better the solution of the PDE will be approximated, but the size of the system of ODEs in (1.4) and (1.5) will also grow and thus its evaluation will be more expensive.

Let $k = 1$ and $n = 1000$. Using the method of balanced truncation (discussed later), we can approximate the system by

$$\begin{aligned} \dot{\widetilde{x}}(t) &= \widetilde{A}\widetilde{x}(t) + \widetilde{B}u(t), \quad \widetilde{x}(0) = 0, \\ \widetilde{y}(t) &= \widetilde{C}\widetilde{x}(t), \end{aligned}$$

with

$$\begin{aligned} \widetilde{A} &= \begin{bmatrix} -2.256 & 1.775 & -0.6057 \\ -1.775 & -16.63 & 12.21 \\ -0.6057 & -12.21 & -40.66 \end{bmatrix}, \quad \widetilde{B} = \begin{bmatrix} -1.074 \\ -0.4136 \\ -0.1442 \end{bmatrix}, \\ \widetilde{C} &= \begin{bmatrix} -1.074 & 0.4136 & -0.1442 \end{bmatrix}. \end{aligned}$$

Simulation with various inputs shows that the outputs y and $\widetilde{y}$ are almost the same.

1.2.2 Further Examples

Here will briefly mention a few more examples to illustrate the importance of model reduction in practice.

Electrical Circuits. Electrical circuits containing only inductors, capacitors, and resistors can be modeled using modified modal analysis. This results in a linear system of the form

$$\begin{aligned} E\dot{x}(t) &= Ax(t) + Bu(t), \\ y(t) &= B^{\mathsf{T}}x(t), \end{aligned} \tag{1.6}$$

where the state $x(\cdot)$ contains the node potentials and currents through inductors and voltage sources. The input $u(\cdot)$ contains the currents of the current sources as well as the voltages of the voltage sources. The output $y(\cdot)$ contains the negative of the voltages of the current sources and the currents of the voltage sources. Here the matrices E, A, and B have the form

$$\begin{aligned} E &= \begin{bmatrix} A_{\mathcal{C}}\mathcal{C}A_{\mathcal{C}}^{\mathsf{T}} & 0 & 0 \\ 0 & \mathcal{L} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} -A_{\mathcal{R}}\mathcal{G}A_{\mathcal{R}}^{\mathsf{T}} & -A_{\mathcal{L}} & -A_{\mathcal{V}} \\ A_{\mathcal{L}}^{\mathsf{T}} & 0 & 0 \\ A_{\mathcal{V}}^{\mathsf{T}} & 0 & 0 \end{bmatrix}, \\ B &= \begin{bmatrix} -A_{\mathcal{I}} & 0 \\ 0 & 0 \\ 0 & -I \end{bmatrix}, \end{aligned} \tag{1.7}$$

where $\mathcal{G}$, $\mathcal{L}$, $\mathcal{C}$ are positive definite matrices containing the conductances, inductances, and capacities of the resistors, inductors, and capacitors, respectively. The matrices $A_{\mathcal{C}}$, $A_{\mathcal{R}}$, $A_{\mathcal{L}}$, $A_{\mathcal{V}}$, and $A_{\mathcal{I}}$ are incidence matrices that describe the network topology of the circuit. This model differs from (1.3), namely an additional matrix E is in front of $\dot{x}$ and moreover, E is singular. This means, that not all of the equations in (1.6) are differential equations, but there are also algebraic equations that result from Kirchhoff's laws. Therefore, such a system is called a *differential-algebraic system*. Moreover, the system (1.6) with (1.7) has certain symmetries that account for the physical properties of the circuit. For example, (1.6) with (1.7) is a *passive* system, meaning that

$$\int_0^T y(t)^{\mathsf{T}}u(t)\mathrm{d}t \geqslant 0$$

for all $T \geqslant 0$ and all smooth solution trajectories with $Ex(0) = 0$. This property must be reflected in the reduced-order model in order to get meaningful simulation results. In other words, structure-preserving methods are of great importance in applications.

Structural Mechanics. The goal of structural mechanics is the computation of mechanical deformations and internal forces and stresses within mechanical structures, such as buildings, bridges, machines, etc. Using the finite element method, the mechanical structure is decomposed into masses that are stiffly

connected. This leads to a large ordinary differential equation of second order of the form

$$\begin{aligned} M\ddot{x}(t) + D\dot{x}(t) + Kx(t) &= Bu(t), \\ y(t) &= C_1 x(t) + C_2 \dot{x}(t), \end{aligned} \tag{1.8}$$

where the state $x(\cdot)$ is the displacement of the masses from the equilibrium position and the input $u(\cdot)$ is an external force. Moreover, M and K are the positive definite mass and stiffness matrices and D is a positive definite damping matrix. Using a linearization, one can in principal rewrite (1.8) as a first-order system as in (1.3) and do model reduction on the first-order system. However, often it is important to have a reduced-order model of the form (1.8). It is often not possible to gain such a system when using methods for first-order systems. There are methods that work directly on (1.8), but there are still many open research problems.

CHAPTER 2

Basics of Systems and Control Theory

In this chapter we consider linear time-invariant (LTI) control systems

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), \quad x(t_0) = x_0, \\ y(t) &= Cx(t) + Du(t), \end{aligned} \tag{2.1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$, $x : [t_0, t_\mathrm{f}] \to \mathbb{R}^n$ is the *state* of the system, $u : [t_0, t_\mathrm{f}] \to \mathbb{R}^m$ denotes a *control input* and $y : [t_0, t_\mathrm{f}] \to \mathbb{R}^p$ is a *measurable output*. The set of LTI systems with state-space dimension n, m inputs, and p outputs is denoted by $\Sigma_{n,m,p}$ and we write $[A, B, C, D] \in \Sigma_{n,m,p}$. The goal of this chapter is to give a basic analysis and discussion of such systems in order to set the foundations for the model reduction methods we discuss later. Here we will rather skip the proofs or keep them short since this will mainly be the topic of the course on control theory. A more detailed introduction to the concepts presented here can be found in the textbook [ZDG96, Chapters 3 & 4].

2.1 Properties of LTI Systems

Next we discuss some fundamental properties of LTI dynamical systems. In the next definition we assume for simplicity that $t_\mathrm{f} = \infty$ and that $\mathcal{U}_\mathrm{ad} := \mathcal{PC}([t_0, t_\mathrm{f}], \mathbb{R}^m)$ is the set of admissible inputs, i. e., the set of all piecewise continuous functions mapping from $[t_0, t_\mathrm{f}]$ to $\mathbb{R}^m$, but in principal we could also take $\mathcal{U}_\mathrm{ad} = \mathcal{L}_2([t_0, t_\mathrm{f}], \mathbb{R}^m)$.

Definition 2.1: The LTI system $[A, B, C, D] \in \Sigma_{n,m,p}$ is called

a) *asymptotically stable*, if all solutions of the linear homogeneous ODE $\dot{x}(t) = Ax(t)$ satisfy $\lim_{t\to\infty} x(t) = 0$ for all initial conditions $x(t_0) = x_0$.

b) *controllable*, if for all initial conditions $x(t_0) = x_0$ and all $x_1 \in \mathbb{R}^n$, there exists a $t_1 > t_0$ and a control function $u \in \mathcal{U}_{\mathrm{ad}}$ such that $x(t_1) = x_1$.

c) *stabilizable*, if for all initial conditions $x(t_0) = x_0$ there exists a control function $u \in \mathcal{U}_{\mathrm{ad}}$ such that $\lim_{t\to\infty} x(t) = 0$.

d) *observable*, if for two solution trajectories (obtained with the same input $u \in \mathcal{U}_{\mathrm{ad}}$) $x(\cdot)$ and $\widetilde{x}(\cdot)$ it holds

$$Cx(t) = C\widetilde{x}(t) \quad \forall t \geqslant t_0 \Rightarrow x(t) = \widetilde{x}(t) \quad \forall t \geqslant t_0.$$

e) *detectable*, if for any solution $x(\cdot)$ of $\dot{x}(t) = Ax(t)$ with $Cx(t) \equiv 0$ it follows that $\lim_{t\to\infty} x(t) = 0$.

The following lemma characterizes these properties algebraically.

Lemma 2.2: The LTI system $[A, B, C, D] \in \Sigma_{n,m,p}$ is

a) asymptotically stable $\Leftrightarrow \Lambda(A) \subset \mathbb{C}^- := \{\lambda \in \mathbb{C} : \mathrm{Re}\,(\lambda) < 0\}$,

b) controllable $\Leftrightarrow \operatorname{rank}\begin{bmatrix}\lambda I_n - A & B\end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}$
$\Leftrightarrow \operatorname{rank}\begin{bmatrix}B & AB & \dots & A^{n-1}B\end{bmatrix} = n,$

c) stabilizable $\Leftrightarrow \operatorname{rank}\begin{bmatrix}\lambda I_n - A & B\end{bmatrix} = n \quad \forall \lambda \in \overline{\mathbb{C}^+} := \{\lambda \in \mathbb{C} : \mathrm{Re}\,(\lambda) \geqslant 0\}$
$\Leftrightarrow \exists F \in \mathbb{R}^{m\times n}$ such that $\Lambda(A + BF) \subset \mathbb{C}^-$,

d) observable $\Leftrightarrow \operatorname{rank}\begin{bmatrix}\lambda I_n - A \\ C\end{bmatrix} = n \quad \forall \lambda \in \mathbb{C}$

$$\Leftrightarrow \operatorname{rank}\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1}\end{bmatrix} = n,$$

e) detectable $\Leftrightarrow \operatorname{rank}\begin{bmatrix}\lambda I_n - A \\ C\end{bmatrix} = n \quad \forall \lambda \in \overline{\mathbb{C}^+}$
$\Leftrightarrow \exists G \in \mathbb{R}^{n\times p}$ such that $\Lambda(A + GC) \subset \mathbb{C}^-$.

Remark 2.3: a) Stabilizability weakens the concept of controllability in the sense that not all possible states are reachable, but uncontrollable parts tend to zero.

b) Detectability weakens observability in the same sense as stabilizability weakens controllability: not all of x can be observed but unobserved parts are asymptotically stable, i. e., deviations vanish over time.

c) The above concepts are *dual* in the sense that an LTI system is observable (detectable) if and only if the dual system

$$\dot{z}(t) = A^{\mathsf{T}} z(t) + C^{\mathsf{T}} v(t)$$

is controllable (stabilizable).

The following considerations motivate the Gramians that we define next. First we consider the input-to-state map

$$\zeta(t) = e^{At} B,$$

which is motivated by the fact that for $x(0) = x_0 = 0$ and impulsive inputs $u = u_0 \cdot \delta$ (where $u_0 \in \mathbb{R}^m$ and δ denotes the Dirac delta distribution), we obtain

$$\begin{aligned} x(t) &= \mathrm{e}^{At} x_0 + \int_0^t \mathrm{e}^{A(t-\tau)} B u(\tau) \mathrm{d}\tau \\ &= \int_0^t \mathrm{e}^{A(t-\tau)} B u_0 \delta(\tau) \mathrm{d}\tau \\ &= \mathrm{e}^{At} B u_0 = \zeta(t) u_0. \end{aligned}$$

Note that the above is formally *not correct*, since δ is not a function mapping from $\mathbb{R}$ to $\mathbb{R}$, but a distribution (often called generalized function) that is defined by

$$\delta : \mathcal{C}^\infty(\mathbb{R}, \mathbb{R}^n) \to \mathbb{R}^n, \quad f \mapsto f(0).$$

So actually we have more correctly

$$x(t) := \delta\left(\mathrm{e}^{A(t-\cdot)} B u_0\right) = \mathrm{e}^{At} B u_0.$$

Consider on the other hand the state-to-output map

$$\eta(t) = C e^{At},$$

which is motivated by the fact that for $x(0) = x_0$ and $u(t) \equiv 0$, we obtain

$$\begin{aligned} y(t) &= C\mathrm{e}^{At} x_0 + C \int_0^t \mathrm{e}^{A(t-\tau)} B u(\tau) \mathrm{d}\tau \\ &= C\mathrm{e}^{At} x_0 = \eta(t) x_0. \end{aligned}$$

For the analysis of LTI control systems we now make use of the following Gramians.

Definition 2.4: The matrix

$$P(T) = \int_0^T \mathrm{e}^{At} B B^{\mathsf{T}} \mathrm{e}^{A^{\mathsf{T}} t} \mathrm{d}t$$

is called the $(0, T)$-*controllability Gramian* of the system (2.1).

The matrix

$$Q(T) = \int_0^T \mathrm{e}^{A^{\mathsf{T}} t} C^{\mathsf{T}} C \mathrm{e}^{At} \mathrm{d}t$$

is called the $(0, T)$-*observability Gramian* of the system (2.1).

Remark 2.5: For a system (2.1), $P(T)$ and $Q(T)$ can be used to identify states of the system that are easily reachable and easily observable in the interval $(0, T)$ in the following sense:

a) For a reachable state x_* of the system (2.1), one can show that $\widehat{u}(t) = B^{\mathsf{T}} \mathrm{e}^{A^{\mathsf{T}}(t_* - t)} P(t_*)^{\dagger} x_*$, where $P(t_*)^{\dagger}$ is the Moore-Penrose inverse of $P(t_*)$, controls the system from $x(0) = 0$ to $x(t_*) = x_*$. Moreover, among all such controls, $\widehat{u}(t)$ is the one with minimal $\mathcal{L}_2$-norm.

b) For any $t_* > 0$ and $x_0 \in \mathbb{R}^n$, we have

$$x_0^{\mathsf{T}} Q(t_*) x_0 = \int_0^{t_*} x_0^{\mathsf{T}} \mathrm{e}^{A^{\mathsf{T}} t} C^{\mathsf{T}} C \mathrm{e}^{At} x_0 \mathrm{d}t = \int_0^{t_*} \left\| C \mathrm{e}^{At} x_0 \right\|_2^2 \mathrm{d}t = \left\| y_{x_0}(\cdot) \right\|_{\mathcal{L}_2}^2 .$$

Now we consider the above Gramians for $T \to \infty$.

Lemma 2.6: If A in (2.1) is asymptotically stable, then

a) the infinite controllability and observability Gramians

$$P = \lim_{T \to \infty} P(T) = \int_0^{\infty} \mathrm{e}^{At} B B^{\mathsf{T}} \mathrm{e}^{A^{\mathsf{T}} t} \mathrm{d}t$$

and

$$Q = \lim_{T \to \infty} Q(T) = \int_0^{\infty} \mathrm{e}^{A^{\mathsf{T}} t} C^{\mathsf{T}} C \mathrm{e}^{At} \mathrm{d}t$$

exist,

b) they solve the two *Lyapunov equations*

$$AP + PA^{\mathsf{T}} = -BB^{\mathsf{T}},$$
$$A^{\mathsf{T}} Q + QA = -C^{\mathsf{T}} C.$$

c) If (A, B) is controllable and (A, C) is observable, it moreover holds that $P = P^{\mathsf{T}} > 0$ and $Q = Q^{\mathsf{T}} > 0$. (Otherwise we just have $P = P^{\mathsf{T}} \geqslant 0$ and $Q = Q^{\mathsf{T}} \geqslant 0$.)

Proof. Exercise or lecture “Control Theory”. □

2.2 Laplace Transformation and Transfer Functions

In applications it is often useful to consider a dynamical system in the frequency domain. When doing so, the system can be treated using tools from linear algebra instead of from differential equations. A function $f : [0, \infty) \to \mathbb{R}^n$ is called *exponentially bounded*, if there exist numbers M and α such that $\|f(t)\|_2 \leqslant M\mathrm{e}^{\alpha t}$ for all $t \geqslant 0$. The value α is called a *bounding exponent*.

Definition 2.7: Let $f : [0, \infty) \to \mathbb{R}^n$ be exponentially bounded with bounding exponent α. Then

$$\mathcal{L}\{f\}(s) := \int_0^\infty f(\tau)\mathrm{e}^{-s\tau}\mathrm{d}\tau$$

for $\mathrm{Re}\,(s) > \alpha$ is called the *Laplace transform* of f. The process of forming the Laplace transform is called *Laplace transformation*.

It can be shown that the integral converges uniformly in a domain of the form $\mathrm{Re}\,(s) \geqslant \beta$ for all $\beta > \alpha$.

Moreover, the following two fundamental properties hold.

Theorem 2.8: Let f, g, $h : [0, \infty) \to \mathbb{R}^n$ be given. Then the following two statements hold true:

a) The Laplace transformation is linear, i. e., if f and g are exponentially bounded, then $h := \gamma f + \delta g$ is also exponentially bounded and

$$\mathcal{L}\{h\} = \gamma\mathcal{L}\{f\} + \delta\mathcal{L}\{g\}$$

holds for all γ, $\delta \in \mathbb{C}$.

b) If $f \in \mathcal{PC}^1([0, \infty), \mathbb{R}^n)$ and $\dot{f}$ is exponentially bounded, then f is exponentially bounded and

$$\mathcal{L}\{\dot{f}\}(s) = s\mathcal{L}\{f\}(s) - f(0).$$

Now we apply the Laplace transformation to the system $[A, B, C, D] \in \Sigma_{n,m,p}$. Assume that each of the Laplace transforms $X(s) := \mathcal{L}\{x\}(s)$, $U(s) := \mathcal{L}\{u\}(s)$, and $Y(s) := \mathcal{L}\{y\}(s)$ exist. By using Theorem 2.8, we obtain the Laplace transformed system

$$\begin{aligned} sX(s) - x(0) &= AX(s) + BU(s), \\ Y(s) &= CX(s) + DU(s). \end{aligned}$$

Under the assumption that $x(0) = 0$, we obtain the relation

$$Y(s) = \big(C(sI_n - A)^{-1}B + D\big)\, U(s).$$

This leads to the following definition.

Definition 2.9: The function

$$G(s) := C(sI_n - A)^{-1}B + D \in \mathbb{R}(s)^{p\times m}$$

is called the *transfer function* of the system $[A, B, C, D] \in \Sigma_{n,m,p}$. Here, $\mathbb{R}(s)^{p\times m}$ denotes the set of all $p \times m$ matrices that have real-rational functions as entries.

The following properties of rational functions will play an important role in the characterization of transfer functions.

Definition 2.10 (Properness)**:** Let $G(s) \in \mathbb{R}(s)^{p\times m}$ be given. We call $G(s)$

a) *strictly proper*, if $\lim_{\omega\to\infty} \|G(\mathrm{i}\omega)\|_2 = 0$;

b) *proper*, if $\lim_{\omega\to\infty} \|G(\mathrm{i}\omega)\|_2 < \infty$;

c) *improper*, if $\lim_{\omega\to\infty} \|G(\mathrm{i}\omega)\|_2 = \infty$.

Since $\lim_{\omega\to\infty}(\mathrm{i}\omega I_n - A)^{-1} = 0$, it easy to see that transfer function of systems $[A, B, C, D] \in \Sigma_{n,m,p}$ are always proper, improper transfer functions can only be realized by DAE systems. Furthermore, the transfer function of a system $[A, B, C, D] \in \Sigma_{n,m,p}$ is strictly proper if and only if $D = 0$. Now we define the notions of poles and zeros of rational matrices. For this we need the following terms.

Definition 2.11 (Unimodular matrix, monic/coprime polynomials)**:** Let $\mathbb{R}[s]$ denote the set of polynomials with real coefficients.

a) A polynomial matrix $U(s) \in \mathbb{R}[s]^{n\times n}$ is called *unimodular*, if its determinant is a nonzero constant in $\mathbb{R}$.

b) A polynomial $p(s) \in \mathbb{R}[s]$ is called *monic*, if its leading coefficient is one.

c) Two polynomials $p(s),\ q(s) \in \mathbb{R}[s]$ are called *coprime*, if their greatest common divisor is 1.

Matrices with rational entries can, via multiplication with suitable unimodular matrices, be transformed to *Smith-McMillan* form, described in the next theorem.

Theorem 2.12 (Smith-McMillan form)**:** For $G(s) \in \mathbb{R}(s)^{p\times m}$ there exist unimodular matrices $U(s) \in \mathbb{R}[s]^{p\times p}$ and $V(s) \in \mathbb{R}[s]^{m\times m}$, such that

$$U^{-1}(s)G(s)V^{-1}(s) = \begin{bmatrix} \widetilde{G}(s) & 0 \\ 0 & 0 \end{bmatrix} \quad \text{with} \quad \widetilde{G}(s) = \operatorname{diag}\left(\frac{\varepsilon_1(s)}{\psi_1(s)}, \ldots, \frac{\varepsilon_r(s)}{\psi_r(s)}\right) \tag{2.2}$$

for some monic and coprime polynomials $\varepsilon_j(s),\ \psi_j(s) \in \mathbb{R}[s]$ such that $\varepsilon_j(s)$ divides $\varepsilon_{j+1}(s)$ and $\psi_{j+1}(s)$ divides $\psi_j(s)$ for $j = 1, \ldots, r-1$.

The Smith-McMillan form can now be utilized to define poles and zeros of rational matrices.

Definition 2.13 (Poles and zeros)**:** Let $G(s) \in \mathbb{R}(s)^{p\times m}$ with Smith-McMillan form (2.2) be given. Then $\lambda \in \mathbb{C}$ is called

a) a *zero* of $G(s)$ if $\varepsilon_r(\lambda) = 0$;

b) a *pole* of $G(s)$ if $\psi_1(\lambda) = 0$.

Roughly speaking, the poles of $G(s)$ are the points $\lambda_0 \in \mathbb{C}$ where we have $\lim_{\lambda\to\lambda_0} \|G(\lambda)\| = \infty$. The zeros are the points $\lambda_0 \in \mathbb{C}$ where a rank drop occurs, i. e., those points where the rank of $G(\lambda_0)$ is strictly less than the rank for all other matrices $G(\lambda)$, where λ is in some neighborhood of λ_0.

2.3 Realizations

It is also possible to assign a dynamical system $[A, B, C, D] \in \Sigma_{n,m,p}$ to a given proper transfer function $G(s) \in \mathbb{R}(s)^{p\times m}$ which is, however, not unique. This leads to the following definitions.

Definition 2.14: Assume that the system $[A, B, C, D] \in \Sigma_{n,m,p}$ has the proper transfer function $G(s) \in \mathbb{R}(s)^{p\times m}$. Then we say that $[A, B, C, D]$ is a *realization* of $G(s)$. The smallest $n \geqslant 0$ such that $[A, B, C, D] \in \Sigma_{n,m,p}$ is a realization of $G(s)$ is called the *McMillan degree* of $G(s)$. A realization $[A, B, C, D] \in \Sigma_{n,m,p}$

of $G(s)$ is called *minimal*, if n is the McMillan degree of $G(s)$.

Remark 2.15: a) Realizations are not unique. If $[A, B, C, D] \in \Sigma_{n,m,p}$ is a realization of $G(s)$, then for any nonsingular matrix $T \in \mathbb{R}^{n\times n}$, the system

$$\left[T^{-1}AT, T^{-1}B, CT, D\right] \in \Sigma_{n,m,p}$$

is also a realization of $G(s)$. Transformations of the above kind are also called *state-space transformations*.

b) A realization is minimal, if and only if it is both controllable and observable.

If a realization is not minimal, we can obtain a minimal realization by using *Kalman decompositions*. There is a *controllability Kalman decomposition*, meaning that for $[A, B, C, D] \in \Sigma_{n,m,p}$ there exists an orthogonal matrix $Q \in \mathbb{R}^{n\times n}$ such that

$$Q^{\mathsf{T}}AQ = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \quad Q^{\mathsf{T}}B = \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \quad CQ = \begin{bmatrix} C_1 & C_2 \end{bmatrix}$$

where the system $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$ is controllable. In the above decomposition we have $\Lambda(A) = \Lambda(A_{11}) \cup \Lambda(A_{22})$. Here, the eigenvalues $\lambda \in \Lambda(A_{22})$ are called *uncontrollable modes* of the system $[A, B, C, D]$ since $B^{\mathsf{T}}v = 0$ holds for all eigenvectors $v \in \mathbb{C}^n\backslash\{0\}$ of A^{T} associated with eigenvalues in $\Lambda(A_{22})$.

On the other hand, there is the *observability Kalman decomposition*, i. e., there exists an orthogonal matrix $\widetilde{Q} \in \mathbb{R}^{n\times n}$ such that

$$\widetilde{Q}^{\mathsf{T}}A\widetilde{Q} = \begin{bmatrix} \widetilde{A}_{11} & 0 \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{bmatrix}, \quad \widetilde{Q}^{\mathsf{T}}B = \begin{bmatrix} \widetilde{B}_1 \\ \widetilde{B}_2 \end{bmatrix}, \quad C\widetilde{Q} = \begin{bmatrix} \widetilde{C}_1 & 0 \end{bmatrix},$$

where the system $[\widetilde{A}_{11}, \widetilde{B}_1, \widetilde{C}_1, D] \in \Sigma_{\widetilde{r},m,p}$ is observable. Similarly as above, eigenvalues $\lambda \in \Lambda(\widetilde{A}_{22})$ are called unobservable modes, since it holds $Cv = 0$ for all eigenvectors $v \in \mathbb{C}^n\backslash\{0\}$ of A associated with eigenvalues in $\Lambda(\widetilde{A}_{22})$.

A minimal realization is then obtained by first computing a controllability Kalman decomposition and applying an observability Kalman decomposition to the resulting controllable subsystem.

2.4 Hardy Spaces

In this section we consider linear spaces of rational functions in $\mathbb{R}(s)^{p\times m}$. These spaces are normed spaces or even inner product spaces that allow for

geometric concepts such as length of transfer functions or distances and angles between them. Later this will be useful to measure the approximation quality of reduced-order models in terms of distances between the transfer functions of the original model and the reduced one.

2.4.1 The Hilbert Space $\mathcal{H}_2^{p\times m}$

The space $\mathcal{H}_2^{p\times m}$ is defined by

$$\mathcal{H}_2^{p\times m} := \Bigg\{ G : \mathbb{C}^+ \to \mathbb{C}^{p\times m} \ : \ G \text{ is analytic in } \mathbb{C}^+ \text{ and } \int_{-\infty}^{\infty} \|G(\mathrm{i}\omega)\|_{\mathrm{F}}^2 \,\mathrm{d}\omega < \infty \Bigg\}.$$

Since every $G \in \mathcal{H}_2^{p\times m}$ is analytic, there exists a unique continuation to the imaginary axis. The space $\mathcal{H}_2^{p\times m}$ is a Hilbert space with the inner product

$$\langle F, G\rangle_{\mathcal{H}_2} := \frac{1}{2\pi}\int_{-\infty}^{\infty} \operatorname{tr}\left(F(\mathrm{i}\omega)^{\mathsf{H}} G(\mathrm{i}\omega)\right) \mathrm{d}\omega.$$

This inner product induces the *$\mathcal{H}_2$-norm*

$$\|G\|_{\mathcal{H}_2} := \langle G, G\rangle_{\mathcal{H}_2}^{1/2} = \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \|G(\mathrm{i}\omega)\|_{\mathrm{F}}^2 \,\mathrm{d}\omega\right)^{1/2}.$$

We are now interested in *rational* functions, i. e., in functions that are in $\mathcal{RH}_2^{p\times m} := \mathcal{H}_2^{p\times m} \cap \mathbb{R}(s)^{p\times m}$. First we have the following.

Lemma 2.16: The following statements are equivalent:

a) The function G is an element of $\mathcal{RH}_2^{p\times m}$.

b) The function G is strictly proper and all its poles are in $\mathbb{C}^-$.

c) The function G can be realized by a system $[A, B, C, D]$ with $\Lambda(A) \subset \mathbb{C}^-$ and $D = 0$.

The $\mathcal{H}_2$-norm of a transfer function can be utilized to bound the norm of the output by the norm of the input as follows. For this we will make use of the following result. It basically says that the $\mathcal{L}_2$-norm of a function on $\mathbb{R}$ is equal to the $\mathcal{L}_2$-norm of its Fourier transform on $\mathrm{i}\mathbb{R}$ (scaled by a constant).

Theorem 2.17 (Plancherel's Theorem)**:** Let $f \in \mathcal{L}_1(\mathbb{R}, \mathbb{R}^n) \cap \mathcal{L}_2(\mathbb{R}, \mathbb{R}^n)$. Then the *Fourier transform* of f, given by

$$F(\mathrm{i}\omega) := \mathcal{F}\{f\}(\mathrm{i}\omega) := \int_{-\infty}^{\infty} f(t)\mathrm{e}^{-\mathrm{i}\omega t}\mathrm{d}t$$

exists, it satisfies $F \in \mathcal{L}_2(\mathrm{i}\mathbb{R}, \mathbb{C}^n)$ and, moreover, it holds

$$\|f\|_{\mathcal{L}_2}^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \|F(\mathrm{i}\omega)\|_2^2 \, \mathrm{d}\omega.$$

When we consider functions f with $f(t) = 0$ for all $t < 0$, then the Fourier transform of f coincides with the Laplace transform of f restricted to the imaginary axis.

In fact, it can even be shown that the Laplace transform of $f \in \mathcal{L}_2([0, \infty), \mathbb{C}^n)$ will always give a result that is in $\mathcal{H}_2^n$. Conversely, applying the inverse Laplace transform to $F \in \mathcal{H}_2^n$ will return a function in $\mathcal{L}_2([0, \infty), \mathbb{C}^n)$. Summarizing, we can write

$$\mathcal{L}\left\{\mathcal{L}_2([0, \infty), \mathbb{C}^n)\right\} = \mathcal{H}_2^n.$$

Now we show that the $\mathcal{H}_2$-norm bounds the $\mathcal{L}_\infty$-norm of the output by the $\mathcal{L}_2$-norm of the input.

Theorem 2.18: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ with a transfer function $G \in \mathcal{RH}_2^{p \times m}$ be given. Then it holds

$$\|G\|_{\mathcal{H}_2} \geqslant \sup_{\substack{u \in \mathcal{L}_2([0,\infty), \mathbb{R}^m) \\ u \neq 0}} \frac{\|y\|_{\mathcal{L}_\infty}}{\|u\|_{\mathcal{L}_2}}.$$

Proof. Since $G \in \mathcal{RH}_2^{p \times m}$, we have $D = 0$ and therefore, it holds

$$y(t) = \int_0^t C\mathrm{e}^{A(t-\tau)} B u(\tau) \mathrm{d}\tau.$$

Set

$$g(t) := \begin{cases} C\mathrm{e}^{At}B, & t \geqslant 0, \\ 0, & t < 0. \end{cases}, \quad \widetilde{u}(t) := \begin{cases} u(t), & t \geqslant 0, \\ 0, & t < 0. \end{cases}$$

Taking norms, we obtain

$$\begin{aligned}\|y(t)\|_2 &= \left\| \int_{-\infty}^{t} g(t-\tau)\widetilde{u}(\tau)\mathrm{d}\tau \right\|_2 \\ &\leqslant \int_{-\infty}^{t} \|g(t-\tau)\|_{\mathrm{F}} \, \|\widetilde{u}(\tau)\|_2 \, \mathrm{d}\tau \\ &\leqslant \left(\int_{-\infty}^{t} \|g(t-\tau)\|_{\mathrm{F}}^2 \, \mathrm{d}\tau \right)^{1/2} \left(\int_{-\infty}^{t} \|\widetilde{u}(\tau)\|_2^2 \, \mathrm{d}\tau \right)^{1/2} \\ &\leqslant \left(\int_{-\infty}^{\infty} \|g(t-\tau)\|_{\mathrm{F}}^2 \, \mathrm{d}\tau \right)^{1/2} \left(\int_{-\infty}^{\infty} \|\widetilde{u}(\tau)\|_2^2 \, \mathrm{d}\tau \right)^{1/2},\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality. It can be shown that (exercise!)

$$\mathcal{F}\{g\}(\mathrm{i}\omega) = C(\mathrm{i}\omega I_n - A)^{-1}B.$$

Using Plancherel's Theorem we obtain

$$\int_{-\infty}^{\infty} \|g(t-\tau)\|_{\mathrm{F}}^2 \, \mathrm{d}\tau = \frac{1}{2\pi} \int_{-\infty}^{\infty} \left\|C(\mathrm{i}\omega I_n - A)^{-1}B\right\|_{\mathrm{F}}^2 \mathrm{d}\omega = \|G\|_{\mathcal{H}_2}^2 .$$

Therefore, we obtain $\|y(t)\|_2 \leqslant \|G\|_{\mathcal{H}_2} \|u\|_{\mathcal{L}_2}$. Since this inequality holds for all $t \geqslant 0$, we can take the supremum on the left-hand side and obtain the result. □

For SISO (single-input single-output) systems, it even holds

$$\|G\|_{\mathcal{H}_2} = \sup_{\substack{u \in \mathcal{L}_2([0,\infty),\mathbb{R}^m) \\ u \neq 0}} \frac{\|y\|_{\mathcal{L}_\infty}}{\|u\|_{\mathcal{L}_2}},$$

i. e., the $\mathcal{H}_2$-norm is the $\mathcal{L}_2$–$\mathcal{L}_\infty$-induced norm of the system. For general MIMO (multi-input multi-output) systems, the interpretation of the $\mathcal{H}_2$-norm is more involved. The $\mathcal{H}_2$-norm can be computed by using Plancherel's Theorem noticing that

$$\begin{aligned}\|G\|_{\mathcal{H}_2} &= \left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \|G(\mathrm{i}\omega)\|_{\mathrm{F}}^2 \, \mathrm{d}\omega \right)^{1/2} \\ &= \left(\int_0^{\infty} \left\|C\mathrm{e}^{At}B\right\|_{\mathrm{F}}^2 \mathrm{d}t \right)^{1/2} \\ &= \left(\int_0^{\infty} \mathrm{tr}\left(C\mathrm{e}^{At}BB^{\mathsf{T}}\mathrm{e}^{A^{\mathsf{T}}t}C^{\mathsf{T}} \right) \mathrm{d}t \right)^{1/2} \\ &= \mathrm{tr}\left(CPC^{\mathsf{T}} \right)^{1/2},\end{aligned}$$

where P is the controllability Gramian of the system. A similar expression can be obtained using the observability Gramian.

2.4.2 The Banach Space $\mathcal{H}_\infty^{p\times m}$

The space $\mathcal{H}_\infty^{p\times m}$ is defined by

$$\mathcal{H}_\infty^{p\times m} := \left\{ G : \mathbb{C}^+ \to \mathbb{C}^{p\times m} \ : \ G \text{ is analytic in } \mathbb{C}^+ \text{ and } \sup_{\omega\in\mathbb{R}} \|G(\mathrm{i}\omega)\|_2 < \infty \right\}.$$

Again, since every $G \in \mathcal{H}_\infty^{p\times m}$ is analytic, there exists a unique continuation to the imaginary axis. The space $\mathcal{H}_\infty^{p\times m}$ is a Banach space equipped with the $\mathcal{H}_\infty$*-norm*

$$\|G\|_{\mathcal{H}_\infty} := \sup_{\omega\in\mathbb{R}} \|G(\mathrm{i}\omega)\|_2 .$$

Again, we focus on *rational* functions, i. e., in functions that are in $\mathcal{RH}_\infty^{p\times m} := \mathcal{H}_\infty^{p\times m} \cap \mathbb{R}(s)^{p\times m}$. First we have the following.

Lemma 2.19: The following statements are equivalent:

a) The function G is an element of $\mathcal{RH}_\infty^{p\times m}$.

b) The function G is proper and all its poles are in $\mathbb{C}^-$.

c) The function G can be realized by a system $[A, B, C, D]$ with $\Lambda(A) \subset \mathbb{C}^-$.

Now we show that the $\mathcal{H}_\infty$-norm bounds the $\mathcal{L}_2$-norm of the output by the $\mathcal{L}_2$-norm of the input..

Theorem 2.20: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ with a transfer function $G \in \mathcal{RH}_\infty^{p\times m}$ be given. Then it holds

$$\|G\|_{\mathcal{H}_\infty} \geqslant \sup_{\substack{u\in\mathcal{L}_2([0,\infty),\mathbb{R}^m)\\ u\neq 0}} \frac{\|y\|_{\mathcal{L}_2}}{\|u\|_{\mathcal{L}_2}}.$$

Proof. It can be shown that an asymptotically stable system with an input $u \in \mathcal{L}_2([0,\infty),\mathbb{R}^m)$ results in an output $y \in \mathcal{L}_2([0,\infty),\mathbb{R}^p)$. (This can be proven using Young's convolution inequality [Bog07, Theorem 3.9.4].)

With $U(s) := \mathcal{L}\{u\}(s)$ and $Y(s) := \mathcal{L}\{y\}(s)$ and using Plancherel's Theorem

we obtain

$$\begin{aligned}
\|y\|_{\mathcal{L}_2}^2 &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \|Y(\mathrm{i}\omega)\|_2^2 \,\mathrm{d}\omega \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \|G(\mathrm{i}\omega)U(\mathrm{i}\omega)\|_2^2 \,\mathrm{d}\omega \\
&\leqslant \frac{1}{2\pi}\int_{-\infty}^{\infty} \|G(\mathrm{i}\omega)\|_2^2 \,\|U(\mathrm{i}\omega)\|_2^2 \,\mathrm{d}\omega \\
&\leqslant \sup_{\omega\in\mathbb{R}} \|G(\mathrm{i}\omega)\|_2^2 \cdot \frac{1}{2\pi}\int_{-\infty}^{\infty} \|U(\mathrm{i}\omega)\|_2^2 \,\mathrm{d}\omega \\
&= \sup_{\omega\in\mathbb{R}} \|G(\mathrm{i}\omega)\|_2^2 \cdot \|u\|_{\mathcal{L}_2}^2 .
\end{aligned}$$

□

It can also be shown that

$$\|G\|_{\mathcal{H}_\infty} = \sup_{\substack{u\in\mathcal{L}_2([0,\infty),\mathbb{R}^m)\\ u\neq 0}} \frac{\|y\|_{\mathcal{L}_2}}{\|u\|_{\mathcal{L}_2}},$$

i. e., the bound is tight. The proof of this is more lengthy, and therefore, it is omitted. There are also several algorithms for computing the $\mathcal{H}_\infty$-norm. The most established ones are based on an iteration on structured matrices or pencils. They are too involved to be discussed at this point.

CHAPTER 3

Eigenvalue-Based Approaches

Consider a linear system $[A, B, C, D] \in \Sigma_{n,m,p}$ with transfer function $G(s) \in \mathbb{R}(s)^{p\times m}$. Assume that we have a partition of the system as

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

where $A_{ij} \in \mathbb{R}^{n_i\times n_j}$, $B_i \in \mathbb{R}^{n_i\times m}$, and $C_i \in \mathbb{R}^{p\times n_i}$ for $i, j = 1, 2$. Then the system $[A_{11}, B_1, C_1, D] \in \Sigma_{n_1,m,p}$ is called a *truncation* of the original system $[A, B, C, D] \in \Sigma_{n,m,p}$. Assume that it has the transfer function $G_1(s) \in \mathbb{R}(s)^{p\times m}$. The goal is to find a good truncation in the following sense:

a) The state-space dimension n_1 is small compared to n.

b) The output y_1 of $[A_{11}, B_1, C_1, D]$ is similar to output y of $[A, B, C, D]$ for the same input u, i. e., $\|y - y_1\|$ is small in some suitable norm. This norm can be often estimated using the norm of $G(s) - G_1(s)$ such as the $\mathcal{H}_2$-norm or $\mathcal{H}_\infty$-norm.

c) If the original model is asymptotically stable, then also the reduced one should be asymptotically stable. In particular, both transfer functions should be in $\mathcal{RH}_\infty^{p\times m}$.

It is important to note that without any further assumptions, nothing can be said about asymptotic stability, controllability, or observability of the reduced-order system, even if the original system is asymptotically stable, controllable, or observable.

Example: Consider the system $[A, B, C, D] \in \Sigma_{2,1,1}$ with

$$A = \begin{bmatrix} 1 & \frac{5}{4} \\ -\frac{7}{4} & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad D = 5.$$

Then we have $\Lambda(A) = \left\{-\frac{1}{4}, -\frac{3}{4}\right\}$, i. e., the system is asymptotically stable. Moreover, we have

$$\operatorname{rank} \begin{bmatrix} B & AB \end{bmatrix} = \operatorname{rank} \begin{bmatrix} 0 & \frac{5}{4} \\ 1 & -2 \end{bmatrix} = 2,$$
$$\operatorname{rank} \begin{bmatrix} C \\ CA \end{bmatrix} = \operatorname{rank} \begin{bmatrix} 0 & 1 \\ -\frac{7}{4} & -2 \end{bmatrix} = 2,$$

this means that the system is controllable and observable.

Taking the truncation for $n_1 = 1$, we obtain the reduced-order model $[1, 0, 0, 5] \in \Sigma_{1,1,1}$ which is unstable, uncontrollable, and unobservable.

Most often, good truncations are achieved by performing a state-space transformation (Note that this does not change the transfer function!). Let $T = \begin{bmatrix} T_1 & T_2 \end{bmatrix} \in \mathbb{R}^{n \times n}$ be an invertible matrix with $T_1 \in \mathbb{R}^{n \times n_1}$ and $T_2 \in \mathbb{R}^{n \times n_2}$ be given and define $T^{-1} := \begin{bmatrix} W_1 & W_2 \end{bmatrix}^{\mathsf{T}}$ with $W_1 \in \mathbb{R}^{n \times n_1}$ and $W_2 \in \mathbb{R}^{n \times n_2}$. Then we consider the transformed system $[T^{-1}AT, T^{-1}B, CT, D] \in \Sigma_{n,m,p}$ and obtain the truncation (keeping the first n_1 rows and columns) by setting $[A_{11}, B_1, C_1, D] = \begin{bmatrix} W_1^{\mathsf{T}}AT_1, W_1^{\mathsf{T}}B, CT_1, D \end{bmatrix} \in \Sigma_{n_1,m,p}$.

Note that the above is *"model reduction by projection"*: We assume that the state $x(\cdot)$ lives approximately in low-dimensional subspace $\operatorname{im} T_1$. With $x(t) \approx T_1 x_1(t)$ we obtain

$$T_1 \dot{x}_1(t) \approx AT_1 x_1(t) + Bu(t),$$
$$y_1(t) = CT_1 x_1(t) + Du(t).$$

Next we "make the state equation square" again by imposing a *Petrov-Galerkin condition*

$$\operatorname{im} W_1 \perp \left(T_1 \dot{x}_1(t) - \left(AT_1 x_1(t) + Bu(t)\right)\right).$$

This results in

$$W_1^{\mathsf{T}} T_1 \dot{x}_1(t) = W_1^{\mathsf{T}} AT_1 x_1(t) + W_1^{\mathsf{T}} Bu(t).$$

By choosing T_1 and W_1 bi-orthogonal, i. e., $W_1^{\mathsf{T}} T_1 = I_{n_1}$, we obtain an ODE as state equation. This bi-orthogonality is automatically fulfilled by the above construction of the truncation. It remains to choose good projection matrices T_1 and W_1. This principle can also be generalized to non-linear systems.

3.1 Modal Truncation

In this chapter we discuss eigenvalue-based methods for model reduction. Assume that we have given a system $[A, B, C, D] \in \Sigma_{n,m,p}$. Assume that we have given a state-space transformation $T \in \mathbb{R}^{n\times n}$ such that

$$T^{-1}AT = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \quad T^{-1}B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad CT = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}. \tag{3.1}$$

Then for the transfer function we obtain

$$\begin{aligned} G(s) = C(sI_n - A)^{-1}B + D = \big(C_1(sI_{n_1} - A_{11})^{-1}B_1 + D\big) \\ + \big(C_2(sI_{n_2} - A_{22})^{-1}B_2\big) =: G_1(s) + G_2(s). \end{aligned}$$

If we can determine the above decomposition such that $n_1 \ll n$ and $\|G_2\|$ is small, we get

$$\|G - G_1\| = \|G_2\|$$

and therefore, $[A_{11}, B_1, C_1, D] \in \Sigma_{n_1,m,p}$ is a good reduced-order model. This process is called *modal truncation (or modal approximation, modal reduction)*. Here we discuss the computation of such reduced-order models.

Theorem 3.1: Assume that the system $[A, B, C, D] \in \Sigma_{n,m,p}$ is asymptotically stable (controllable, stabilizable, observable, detectable). Then the reduced-order model $[A_{11}, B_1, C_1, D] \in \Sigma_{n_1,m,p}$ in (3.1) is asymptotically stable (controllable, stabilizable, observable, detectable).

Proof. Exercise. □

Theorem 3.2: Let the system $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable with transfer function $G \in \mathcal{RH}_\infty^{p\times m}$ and assume that A is diagonalizable. Assume that there is an invertible matrix $T \in \mathbb{C}^{n\times n}$ such that

$$T^{-1}AT = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}, \quad T^{-1}B = \begin{bmatrix} \widehat{b}_1^{\mathsf{T}} \\ \vdots \\ \widehat{b}_n^{\mathsf{T}} \end{bmatrix}, \quad CT = \begin{bmatrix} \widehat{c}_1 & \dots & \widehat{c}_n \end{bmatrix}$$

and let the reduced-order model be $\big[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\big] \in \Sigma_{r,m,p}$ with

$$\widetilde{A} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_r \end{bmatrix}, \quad \widetilde{B} = \begin{bmatrix} \widehat{b}_1^{\mathsf{T}} \\ \vdots \\ \widehat{b}_r^{\mathsf{T}} \end{bmatrix}, \quad \widetilde{C} = \begin{bmatrix} \widehat{c}_1 & \dots & \widehat{c}_r \end{bmatrix}, \quad \widetilde{D} = D,$$

and with the transfer function $\widetilde{G} \in \mathcal{RH}_\infty^{p\times m}$. Then it holds

$$\|G - \widetilde{G}\|_{\mathcal{H}_\infty} \leqslant \sum_{j=r+1}^{n} \frac{\|\widehat{c}_j\|_2 \cdot \|\widehat{b}_j\|_2}{|\operatorname{Re}(\lambda_j)|}.$$

Proof. We have

$$G(s) = \sum_{j=1}^{n} \frac{1}{s-\lambda_j} \widehat{c}_j \widehat{b}_j^{\mathsf{T}} + D$$

and

$$\widetilde{G}(s) = \sum_{j=1}^{r} \frac{1}{s-\lambda_j} \widehat{c}_j \widehat{b}_j^{\mathsf{T}} + D.$$

Therefore, we have

$$\|G - \widetilde{G}\|_{\mathcal{H}_\infty} = \left\| \sum_{j=r+1}^{n} \frac{1}{\cdot - \lambda_j} \widehat{c}_j \widehat{b}_j^{\mathsf{T}} \right\|_{\mathcal{H}_\infty} \leqslant \sum_{j=r+1}^{n} \left\| \frac{1}{\cdot - \lambda_j} \widehat{c}_j \widehat{b}_j^{\mathsf{T}} \right\|_{\mathcal{H}_\infty}.$$

Moreover, it holds

$$\begin{aligned} \left\| \frac{1}{\cdot - \lambda_j} \widehat{c}_j \widehat{b}_j^{\mathsf{T}} \right\|_{\mathcal{H}_\infty} &= \sup_{\omega\in\mathbb{R}} \left\| \frac{1}{\mathrm{i}\omega - \lambda_j} \widehat{c}_j \widehat{b}_j^{\mathsf{T}} \right\|_2 \\ &= \left\| \widehat{c}_j \widehat{b}_j^{\mathsf{T}} \right\|_2 \cdot \sup_{\omega\in\mathbb{R}} \left| \frac{1}{\mathrm{i}\omega - \lambda_j} \right| \\ &= \|\widehat{c}_j\|_2 \cdot \|\widehat{b}_j\|_2 \cdot \left| \frac{1}{\operatorname{Re}(\lambda_j)} \right|, \end{aligned}$$

where the latter equality follows from the fact that $\mathrm{i}\omega - \lambda_j$ is minimized for $\omega = \operatorname{Im}(\lambda_j)$. □

Remark 3.3: a) In classical modal truncation, the eigenvalues are ordered with respect to distance to the imaginary axis, i. e.,

$$0 > \operatorname{Re}(\lambda_1) \geqslant \operatorname{Re}(\lambda_2) \geqslant \ldots \geqslant \operatorname{Re}(\lambda_n).$$

There a good numerical algorithms for approximating eigenvalues closest to the imaginary axis. However, the error bound suggests to order the eigenvalues such that

$$\frac{\|\widehat{c}_1\|_2 \cdot \|\widehat{b}_1\|_2}{|\operatorname{Re}(\lambda_1)|} \geqslant \frac{\|\widehat{c}_2\|_2 \cdot \|\widehat{b}_2\|_2}{|\operatorname{Re}(\lambda_2)|} \geqslant \ldots \geqslant \frac{\|\widehat{c}_n\|_2 \cdot \|\widehat{b}_n\|_2}{|\operatorname{Re}(\lambda_n)|}.$$

There are also algorithms that handle this sorting of the eigenvalues (see Section 3.2).

b) Modal truncation generates good local approximations of the transfer function. This means that the reduced-order model has a good approximation quality near those values on the imaginary axis that are close to some $\lambda \in \Lambda(A_{11})$ and can have a worse approximation quality near an eigenvalue $\lambda \in \Lambda(A)$, if λ is close to the imaginary axis, but $\lambda \notin \Lambda(A_{11})$.

c) There are problems if A is non-diagonalizable or if T is ill-conditioned, i. e., the condition number $\kappa(T) := \|T\|_2 \cdot \|T^{-1}\|_2$ is large. Then $\|\widehat{b}_j\|_2$ and $\|\widehat{c}_j\|_2$ can be large, even if B and C are of moderate norm. In this case, the state-space dimension of the reduced-order model often has to be increased to achieve a good approximation error.

d) The transformation matrix T can be chosen to be real by treating complex conjugate eigenvalues as pairs. This results in a real reduced-order model.

3.2 The Dominant Pole Algorithm

As mentioned above it is desirable to order the eigenvalues such that

$$\frac{\|\widehat{c}_1\|_2 \cdot \|\widehat{b}_1\|_2}{|\operatorname{Re}(\lambda_1)|} \geqslant \frac{\|\widehat{c}_2\|_2 \cdot \|\widehat{b}_2\|_2}{|\operatorname{Re}(\lambda_2)|} \geqslant \ldots \geqslant \frac{\|\widehat{c}_n\|_2 \cdot \|\widehat{b}_n\|_2}{|\operatorname{Re}(\lambda_n)|}.$$

The *dominant pole algorithm* that we will discuss now is doing exactly this. First we show that the vectors $\widehat{b}_j$ and $\widehat{c}_j$ have a special structure.

Lemma 3.4: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be an asymptotically stable system with transfer function $G \in \mathcal{RH}_\infty^{p\times m}$. Assume that A is diagonalizable. Then it holds

$$G(s) = \sum_{j=1}^{n} \frac{R_j}{s - \lambda_j} + D$$

with the *residues* $R_j = (Cx_j)(v_j^{\mathsf{H}} B)$, where x_j, $v_j \in \mathbb{C}^n$ denote the right and left eigenvectors of A associated with the eigenvalue λ_j for $j = 1, \ldots, n$. Moreover, here we assume the normalization condition $v_j^{\mathsf{H}} x_j = 1$ for $j = 1, \ldots, n$.

Proof. Let $T \in \mathbb{C}^{n\times n}$ be such that $T^{-1}AT = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Then we have that

$$T = \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}, \quad T^{-1} = \begin{bmatrix} v_1^{\mathsf{H}} \\ \vdots \\ v_n^{\mathsf{H}} \end{bmatrix}.$$

Using the notation of Theorem 3.2, we obtain

$$\begin{bmatrix} \widehat{b}_1 \\ \vdots \\ \widehat{b}_n \end{bmatrix} := T^{-1}B = \begin{bmatrix} v_1^{\mathsf{H}} B \\ \vdots \\ v_n^{\mathsf{H}} B \end{bmatrix}, \quad \begin{bmatrix} \widehat{c}_1 & \dots & \widehat{c}_n \end{bmatrix} := CT = \begin{bmatrix} Cx_1 & \dots & Cx_n \end{bmatrix},$$

which gives the result. □

We now derive the *dominant pole algorithm* for SISO systems [RM06b]. The case of MIMO systems is conceptually slightly different, see also Remark 3.5. So, assume that we have given an asymptotically stable system $\left[A, b, c^{\mathsf{T}}, 0\right] \in \Sigma_{n,1,1}$ with the transfer function $G(s) \in \mathbb{R}(s)$. Here we set the feedthrough term $D = 0$ to simplify the presentation, but it is no problem to include it as well. Then we have

$$\begin{aligned} G(s) &= \frac{Y(s)}{U(s)} = c^{\mathsf{T}}(sI_n - A)^{-1}b, \\ G^{\mathsf{H}}(s) &= \frac{\overline{Y}(s)}{\overline{U}(s)} = b^{\mathsf{T}}(sI_n - A)^{-\mathsf{H}}c, \end{aligned}$$

where $U(s)$ and $Y(s)$ are the Laplace transforms of u and y, respectively. This can be reformulated

$$\begin{aligned} \begin{bmatrix} sI_n - A & -b \\ c^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} X(s) \\ U(s) \end{bmatrix} &= \begin{bmatrix} 0 \\ Y(s) \end{bmatrix}, \\ \begin{bmatrix} (sI_n - A)^{\mathsf{H}} & c \\ -b^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} V(s) \\ \overline{U}(s) \end{bmatrix} &= \begin{bmatrix} 0 \\ \overline{Y}(s) \end{bmatrix}, \end{aligned} \tag{3.2}$$

with auxiliary vectors $X(s)$ and $V(s)$ (where $X(s)$ is the Laplace transform of the state of $\left[A, b, c^{\mathsf{T}}, 0\right] \in \Sigma_{n,1,1}$). If $\lambda \in \mathbb{C}$ is a pole of $G(s)$, then $\lim_{s\to\lambda} |G(s)| = \infty$ and one can choose $\lim_{s\to\lambda} U(s) = 0$, while $Y(s) \equiv 1$. This yields that $\lim_{s\to\lambda} X(s) = x$ and $\lim_{s\to\lambda} V(s) = v$ are right and left eigenvectors of A associated with the eigenvalue λ and the normalization conditions $c^{\mathsf{T}}x = 1$ and $-b^{\mathsf{T}}v = 1$.

We want to determine the most *dominant poles* of $G(s)$, i. e., those $\lambda_j \in \Lambda(A)$, where $|R_j|/|\operatorname{Re}(\lambda_j)|$ is the largest. We do this iteratively in a search in possibly growing subspaces. Assume that we have subspaces spanned by $\widehat{X} \in \mathbb{C}^{n\times k}$ and $\widehat{V} \in \mathbb{C}^{n\times k}$ for some $k \ll n$. Then we can project the eigenvalue problem for the matrix pencil $sI_n - A \in \mathbb{R}[s]^{n\times n}$ to a small eigenvalue problem for the matrix pencil

$$s\widehat{V}^{\mathsf{H}}\widehat{X} - \widehat{V}^{\mathsf{H}}A\widehat{X} \in \mathbb{C}[s]^{k\times k}.$$

Assume that this pencil has only semi-simple eigenvalues and that $\widehat{V}^{\mathsf{H}}\widehat{X}$ is invertible. Then one could alternatively consider the eigenvalue problem for the projected matrix $(\widehat{V}^{\mathsf{H}}\widehat{X})^{-1}\widehat{V}^{\mathsf{H}}A\widehat{X} \in \mathbb{C}^{k\times k}$. For this matrix pencil, we can easily

determine all eigenvalues $\widetilde{\lambda}_j \in \mathbb{C}$ and the associated right and left eigenvectors $\widetilde{x}_j \in \mathbb{C}^k$ and $\widetilde{v}_j \in \mathbb{C}^k$ for $j = 1, \ldots, k$. Then we obtain the eigenvalue and eigenvector approximations for the original problem as

$$\widehat{\lambda}_j = \widetilde{\lambda}_j, \quad \widehat{x}_j = \widehat{X}\widetilde{x}_j, \quad \widehat{v}_j = \widehat{V}\widetilde{v}_j, \quad j = 1, \ldots, k.$$

These approximations can now be sorted according to our dominance measure, i. e., we sort the eigenvalues such that

$$\frac{\left|c^{\mathsf{T}}\widehat{x}_1\widehat{v}_1^{\mathsf{H}}b\right|}{\left|\operatorname{Re}\left(\widehat{\lambda}_1\right)\right|} \geqslant \frac{\left|c^{\mathsf{T}}\widehat{x}_2\widehat{v}_2^{\mathsf{H}}b\right|}{\left|\operatorname{Re}\left(\widehat{\lambda}_2\right)\right|} \geqslant \ldots \geqslant \frac{\left|c^{\mathsf{T}}\widehat{x}_k\widehat{v}_k^{\mathsf{H}}b\right|}{\left|\operatorname{Re}\left(\widehat{\lambda}_k\right)\right|}.$$

So $\widehat{\lambda}_1$ is our current approximation for the most dominant pole. If $\left\|A\widehat{x}_1 - \widehat{\lambda}_1\widehat{x}_1\right\|_2 < \varepsilon$ (or $\left\|\widehat{v}_1^{\mathsf{H}}A - \widehat{\lambda}_1\widehat{v}_1^{\mathsf{H}}\right\|_2 < \varepsilon$) for some small tolerance $\varepsilon > 0$, then we assume that the eigenvalue $\widehat{\lambda}_1$ and the corresponding eigenvectors have converged.

If this is not the case, we expand the matrices $\widehat{X}$ and $\widehat{V}$ in order to enrich the spaces $\operatorname{im}\widehat{X}$ and $\operatorname{im}\widehat{V}$ in which we search for the eigenvectors. This is done by plugging in our current dominant pole approximation $\widehat{\lambda}_1$ into (3.2) and compute $\widehat{x} := X\left(\widehat{\lambda}_1\right)$ and $\widehat{v} := V\left(\widehat{\lambda}_1\right)$ (with $Y(s) := 1$). Then the expanded projection matrices are $\begin{bmatrix}\widehat{X} & \widehat{x}\end{bmatrix}$ and $\begin{bmatrix}\widehat{V} & \widehat{v}\end{bmatrix}$. For numerical stability, it advised to orthogonalize their columns afterwards.

On the other hand, if $\widehat{\lambda}_1$ has converged to an eigenvalue λ_1 with right and left eigenvectors $x_1, v_1 \in \mathbb{C}^n$, then we want to ensure that we do not not find it again in the next iterations. So we want to deflate this eigenvalue and its eigenvectors. This is done by projecting the system $[A, b, c^{\mathsf{T}}, 0]$, namely we replace it by $\left[A, \widetilde{b}, \widetilde{c}^{\mathsf{T}}, 0\right]$ with

$$\widetilde{b} := \left(I_n - \frac{x_1 v_1^{\mathsf{H}}}{v_1^{\mathsf{H}}x_1}\right) b, \quad \widetilde{c}^{\mathsf{T}} := c^{\mathsf{T}}\left(I_n - \frac{x_1 v_1^{\mathsf{H}}}{v_1^{\mathsf{H}}x_1}\right).$$

First of all, note that the matrix $I_n - \frac{x_1 v_1^{\mathsf{H}}}{v_1^{\mathsf{H}}x_1}$ is a projector. Projecting the system like this has the effect that the residue of the deflated eigenvalue is zero, since $\widetilde{c}^{\mathsf{T}}x_1 = 0$ and $v_1^{\mathsf{H}}\widetilde{b} = 0$ and the residues of the other eigenvalues remain unchanged (exercise!). Therefore, the already converged eigenvalues are not found again, since there dominance value is set to zero. To summarize this section we formulate the above results as Algorithm 3.1.

Algorithm 3.1 Dominant pole algorithm

Input: Asymptotically stable system $[A, b, c^{\mathsf{T}}, 0] \in \Sigma_{n,1,1}$ with transfer function $G(s) \in \mathbb{R}(s)$; an initial pole estimate $\lambda \in \mathbb{C}$, tolerance $\varepsilon > 0$, number of desired dominant poles k.

Output: k dominant poles $\Lambda = \{\lambda_1, \ldots, \lambda_k\} \subset \mathbb{C}$ of $G(s)$ with the associated right and left eigenvectors of A stored in R, $L \in \mathbb{C}^{n\times k}$.

1: Initialize $k_{\text{found}} := 0$, $\Lambda = \{\}$, $R = [\,]$, $L = [\,]$, $\widehat{X} = [\,]$, $\widehat{V} = [\,]$.

2: **while** $k > k_{\text{found}}$ **do**

3: Solve the linear system

$$\begin{bmatrix} \lambda I_n - A & -b \\ c^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} \widehat{x} \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

for $\widehat{x} \in \mathbb{C}^n$.

4: Solve the linear system

$$\begin{bmatrix} (\lambda I_n - A)^{\mathsf{H}} & c \\ -b^{\mathsf{T}} & 0 \end{bmatrix} \begin{bmatrix} \widehat{v} \\ u \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

for $\widehat{v} \in \mathbb{C}^n$.

5: Expand the search spaces: Set $\widehat{X} := \begin{bmatrix} \widehat{X} & \widehat{x} \end{bmatrix}$ and $\widehat{V} := \begin{bmatrix} \widehat{V} & \widehat{v} \end{bmatrix}$ and orthogonalize.

6: Compute the eigenvalues and eigenvectors of the matrix pencil $s\widehat{V}^{\mathsf{H}}\widehat{X} - \widehat{V}^{\mathsf{H}}A\widehat{X} \in \mathbb{C}[s]^{\ell\times\ell}$ and compute eigenvalue and eigenvector approximations, sort them according to the dominance measure and store them as $\widehat{\Lambda} := \{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_\ell\}$, $\widehat{X} := \begin{bmatrix} \widehat{x}_1 & \ldots & \widehat{x}_\ell \end{bmatrix}$, $\widehat{V} := \begin{bmatrix} \widehat{v}_1 & \ldots & \widehat{v}_\ell \end{bmatrix}$.

7: **while** $\left\| A\widehat{x}_1 - \widehat{\lambda}_1\widehat{x}_1 \right\| < \varepsilon$ **do**

8: Deflate the found eigenvalue: Set

$$\begin{aligned} k_{\text{found}} &:= k_{\text{found}} + 1, \quad \lambda_{k_{\text{found}}} := \widehat{\lambda}_1, \\ \Lambda &:= \Lambda \cup \{\lambda_{k_{\text{found}}}\}, \quad R := \begin{bmatrix} R & \widehat{x}_1 \end{bmatrix}, \quad L := \begin{bmatrix} L & \widehat{v}_1 \end{bmatrix}, \\ b &:= \left(I_n - \frac{\widehat{x}_1\widehat{v}_1^{\mathsf{H}}}{\widehat{v}_1^{\mathsf{H}}\widehat{x}_1} \right) b, \quad c^{\mathsf{T}} := c^{\mathsf{T}} \left(I_n - \frac{\widehat{x}_1\widehat{v}_1^{\mathsf{H}}}{\widehat{v}_1^{\mathsf{H}}\widehat{x}_1} \right). \end{aligned}$$

9: Set

$$\begin{aligned} \widehat{\Lambda} &:= \{\widehat{\lambda}_2, \ldots, \widehat{\lambda}_\ell\} =: \{\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{\ell-1}\}, \\ \widehat{X} &:= \begin{bmatrix} \widehat{x}_2 & \ldots & \widehat{x}_\ell \end{bmatrix} =: \begin{bmatrix} \widehat{x}_1 & \ldots & \widehat{x}_{\ell-1} \end{bmatrix}, \\ \widehat{V} &:= \begin{bmatrix} \widehat{v}_2 & \ldots & \widehat{v}_\ell \end{bmatrix} =: \begin{bmatrix} \widehat{v}_1 & \ldots & \widehat{v}_{\ell-1} \end{bmatrix}. \end{aligned}$$

10: **end while**

11: Set the new pole estimate $\lambda = \widehat{\lambda}_1$.

12: **end while**

Remark 3.5: a) The dominant pole algorithm presented here is a subspace accelerated version of an algorithm that was originally designed as a Newton method to find roots of $G^{-1}(s)$.

b) It is not guaranteed that the method finds the most dominant poles, but it often works well in practice (in particular, if there are only a few very dominant poles). Convergence of poles can be enhanced by using a Newton scheme or a Rayleigh quotient iteration to update the pole estimates.

c) The projections in (3.2) should not be constructed explicitly. It is rather advised to compute the action of the projection on a vector $z \in \mathbb{C}^n$ if needed. This means that we we compute

$$\left(I_n - \frac{xv^{\mathsf{H}}}{v^{\mathsf{H}}x}\right) z = z - \frac{v^{\mathsf{H}}z}{v^{\mathsf{H}}x} \cdot x$$

using two inner products and one scaled vector addition.

d) The algorithm can be modified to deal with MIMO systems [RM06a]. The most drastic changes are in lines 3 and 4 of Algorithm 3.1, where we replace the linear systems by

$$(\lambda I_n - A)\widehat{x} = Bu, \quad (\lambda I_n - A)^{\mathsf{H}}\widehat{v} = Cw,$$

where $u \in \mathbb{C}^m$ and $w \in \mathbb{C}^p$ are chosen to be the right and left singular vectors of $G(\lambda)$ corresponding to its largest singular value.

e) The algorithm can also be modified to output real R and L in order to obtain a real reduced-order model. For this, pairs of complex conjugate eigenvalues have to be deflated together.

CHAPTER 4

Balancing-Based Approaches

In this chapter we discuss another kind of transformation that simultaneously transforms the controllability and observability Gramians to diagonal form. Then we can sort the transformed states according to their input or output energy and truncate those which are hard to control or hard to observe.

4.1 Input and Output Energy

Consider a system $[A, B, C, D] \in \Sigma_{n,m,p}$. Here we consider the system for $t \in \mathbb{R}$ and assume that $x(-\infty) = 0$. Assume that we have an input $u \in \mathcal{L}_2((-\infty, 0], \mathbb{R}^m)$ steering the state to $x(0) = x_0 \in \mathbb{R}^n$. Then

$$E_u := \left(\int_{-\infty}^{0} \|u(\tau)\|_2^2 \, \mathrm{d}\tau \right)^{1/2} = \|u\|_{\mathcal{L}_2((-\infty,0],\mathbb{R}^m)}$$

is called the *input energy* and if $y \in \mathcal{L}_2([0, \infty), \mathbb{R}^p)$, then

$$E_y := \left(\int_{0}^{\infty} \|y(\tau)\|_2^2 \, \mathrm{d}\tau \right)^{1/2} = \|y\|_{\mathcal{L}_2([0,\infty),\mathbb{R}^p)}$$

is called the *output energy*. In many applications these can be interpreted as actual physical energies of the system.

For the initial state $x(0) = x_0 \in \mathbb{R}^n$ we define

$$E_u(x_0) := \inf_{\substack{u \in \mathcal{L}_2((-\infty,0],\mathbb{R}^m) \\ x(-\infty)=0,\, x(0)=x_0}} \|u\|_{\mathcal{L}_2((-\infty,0],\mathbb{R}^m)}, \tag{4.1}$$

which is the minimal energy needed to steer the system from the state zero state to the state x_0 in an arbitrary time. If $E_u(x_0)$ is small, then the state x_0 is easy to reach, otherwise it is hard to reach. Note that $E_u(x_0) = \infty$ is also possible. Then the state x_0 is unreachable and the system is uncontrollable.

Now assume that $x(0) = x_0$ and that $u|_{[0,\infty)} = 0$. Then we have $y(t) = C\mathrm{e}^{At}x_0$. We define

$$E_y(x_0) := \|y\|_{\mathcal{L}_2([0,\infty),\mathbb{R}^p)} = \left\|C\mathrm{e}^{A\cdot}x_0\right\|_{\mathcal{L}_2([0,\infty),\mathbb{R}^p)},$$

which is the output energy gained from the state x_0. If $E_y(x_0)$ is large, then x_0 is easy to observe, otherwise it is hard to observe. If $E_y(x_0) = 0$, the the state x_0 is unobservable, and therefore the system is unobservable.

The next theorem shows that the $E_u(x_0)$ and $E_y(x_0)$ can be expressed by the controllability and observability Gramians, respectively, which make them feasible for numerical computations.

Theorem 4.1: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable and controllable. Then the following two statements are satisfied:

a) It holds
$$E_u(x_0)^2 = x_0^{\mathsf{T}} P^{-1} x_0,$$
where P is the controllability Gramian of the system. Moreover, $u_*(t) := B^{\mathsf{T}}\mathrm{e}^{-A^{\mathsf{T}}t}P^{-1}x_0$ is a trajectory for which the infimum in (4.1) is attained.

b) It holds
$$E_y(x_0)^2 = x_0^{\mathsf{T}} Q x_0,$$
where Q is the observability Gramian of the system.

Proof. a) Let (x, u) be an arbitrary solution trajectory with $x(-\infty) = 0$, $x(0) = x_0$, and $E_u < \infty$. Then we have

$$x(0) = \int_{-\infty}^{0} \mathrm{e}^{-A\tau} B u(\tau) \mathrm{d}\tau.$$

We show that $E_u \geqslant E_{u_*}$ for the above defined u_*. Define $v := u - u_*$. Then we have

$$\begin{aligned}\int_{-\infty}^{0} u_*(\tau)^{\mathsf{T}} v(\tau)\mathrm{d}\tau &= x_0^{\mathsf{T}} P^{-1} \left(\int_{-\infty}^{0} \mathrm{e}^{-A\tau} B u(\tau)\mathrm{d}\tau \right. \\ &\qquad \left. - \underbrace{\int_{-\infty}^{0} \mathrm{e}^{-A\tau} B B^{\mathsf{T}} \mathrm{e}^{-A^{\mathsf{T}}\tau}\mathrm{d}\tau}_{=P} P^{-1} x_0 \right) \\ &= x_0^{\mathsf{T}} P^{-1}(x_0 - x_0) = 0.\end{aligned}$$

Hence we obtain

$$\begin{aligned} E_u^2 &= \int_{-\infty}^{0} u(\tau)^{\mathsf{T}} u(\tau) \mathrm{d}\tau \\ &= \int_{-\infty}^{0} (v(\tau) + u_*(\tau))^{\mathsf{T}} (v(\tau) + u_*(\tau)) \mathrm{d}\tau \\ &= \underbrace{\int_{-\infty}^{0} v(\tau)^{\mathsf{T}} v(\tau) \mathrm{d}\tau}_{\geqslant 0} + 2 \underbrace{\int_{-\infty}^{0} u_*(\tau)^{\mathsf{T}} v(\tau) \mathrm{d}\tau}_{=0} + \underbrace{\int_{-\infty}^{0} u_*(\tau)^{\mathsf{T}} u_*(\tau) \mathrm{d}\tau}_{\geqslant 0} \\ &\geqslant E_{u_*}^2. \end{aligned}$$

Moreover, we have

$$\begin{aligned} E_{u_*}^2 = x_0^{\mathsf{T}} P^{-1} \int_{-\infty}^{0} \mathrm{e}^{-A\tau} B B^{\mathsf{T}} \mathrm{e}^{-A^{\mathsf{T}}\tau} \mathrm{d}\tau P^{-1} x_0 &= x_0^{\mathsf{T}} P^{-1} P P^{-1} x_0 \\ &= x_0^{\mathsf{T}} P^{-1} x_0. \end{aligned}$$

b) We have

$$E_y(x_0)^2 = \int_0^{\infty} y(\tau)^{\mathsf{T}} y(\tau) \mathrm{d}\tau = x_0^{\mathsf{T}} \int_0^{\infty} \mathrm{e}^{A^{\mathsf{T}}\tau} C^{\mathsf{T}} C \mathrm{e}^{A\tau} \mathrm{d}\tau x_0 = x_0^{\mathsf{T}} Q x_0.$$

This completes the proof.

□

Now consider an eigendecomposition of P, i. e., $P = U\Sigma U^{\mathsf{T}}$ with orthogonal $U = \begin{bmatrix} u_1 & \dots & u_n \end{bmatrix}$ and $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_n)$, where $\sigma_1 \geqslant \sigma_2 \geqslant \dots \geqslant \sigma_n$. Then the energy needed to reach the state $x_0 = u_i$ from $x(-\infty) = 0$ is given by $E_u(u_i)^2 = u_i^{\mathsf{T}} P^{-1} u_i = 1/\sigma_i$. Thus eigenvectors of P corresponding to large eigenvalues are easy to reach and eigenvectors of P corresponding to small eigenvalues are hard to reach. The eigenvectors corresponding to zero eigenvalues are unreachable. Analogously, the eigenvectors corresponding to large eigenvalues of Q are easy to observe, the ones corresponding to small eigenvalues are hard to observe and those corresponding to zero eigenvalues are unobservable.

4.2 Balancing Transformations and Balanced Truncation

We motivate the concept of balancing transformations by means of an example. The application of theses transformations then leads to the method of balanced truncation that was discussed first in [Moo81].

Example: Consider the parameter-dependent system $[A(\alpha), B(\alpha), C(\alpha), D] \in \Sigma_{2,1,1}$ with

$$A(\alpha) = \begin{bmatrix} -1 & -\frac{4}{\alpha} \\ 4\alpha & -2 \end{bmatrix}, \quad B(\alpha) = \begin{bmatrix} 1 \\ 2\alpha \end{bmatrix}, \quad C(\alpha) = \begin{bmatrix} -1 & \frac{2}{\alpha} \end{bmatrix}, \quad D = 0$$

for $\alpha > 0$. This system is asymptotically stable, controllable, and observable. Thus the controllability and observability Gramians $P(\alpha)$ and $Q(\alpha)$ are symmetric positive definite, where

$$P(\alpha) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \alpha^2 \end{bmatrix}, \quad Q(\alpha) = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{\alpha^2} \end{bmatrix}.$$

The eigenvectors of P and Q are $e_1 = \left[\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right]$ and $e_2 = \left[\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}\right]$.

Assume that the state function $x(\cdot) = \begin{bmatrix} x_1(\cdot) \\ x_2(\cdot) \end{bmatrix}$ is expressed as a (time-dependent) linear combination of the eigenvectors of $P(\alpha)$, in our case we get

$$x(t) = \beta_1(\alpha, t)e_1 + \beta_2(\alpha, t)e_2.$$

Intuitively, if a state e_i is hard to reach, then its coefficient β_i is negligible, so truncating it should not change the system's dynamics drastically. In our example we have two cases:

a) $\alpha \ll 1$: In this case, e_2 is much harder to reach than e_1. Thus we truncate x_2 and obtain the reduced-order model $[-1, 1, -1, 0] \in \Sigma_{1,1,1}$.

b) $\alpha \gg 1$: In this case, e_1 is much harder to reach than e_2. Thus we truncate x_1 and obtain the reduced-order model $\left[-2, 2\alpha, \frac{2}{\alpha}, 0\right] \in \Sigma_{1,1,1}$.

Alternatively, we could express $x(\cdot)$ as a (time-dependent) linear combination of the eigenvectors of $Q(\alpha)$, in our case we get

$$x(t) = \gamma_1(\alpha, t)e_1 + \gamma_2(\alpha, t)e_2.$$

Similarly as above, if a state e_i is hard to observe, then its coefficient γ_i is negligible, so truncating it should not change the system's dynamics too much. Again we have two cases in our example:

a) $\alpha \ll 1$: In this case, e_1 is much harder to observe than e_2. Thus we truncate x_1 and obtain the reduced-order model $\left[-2, 2\alpha, \frac{2}{\alpha}, 0\right] \in \Sigma_{1,1,1}$.

b) $\alpha \gg 1$: In this case, e_2 is much harder to observe than e_1. Thus we truncate x_2 and obtain the reduced-order model $[-1, 1, -1, 0] \in \Sigma_{1,1,1}$.

Remark 4.2: a) Both approaches lead to different reduced-order models. In general, this would be OK, but in the above example this leads to contradictory reduced-order models.

b) The behavior in the example can be explained as follows: If $\alpha \ll 1$, then e_2 is very hard to reach, but at the same time it is also very easy to observe and thus has a considerable influence on the output function.

c) The transfer function of the system is $G(s) = \frac{3s+8}{s^2+3s+18}$, but the reduced-order model depends on α, which should not be the case.

The solution of the above problems is to truncate states that are simultaneously hard to reach *and* hard to oberserve. In general, finding these states is difficult, but it is easy if $P = Q$.

Definition 4.3: An asymptotically stable system $[A, B, C, D] \in \Sigma_{n,m,p}$ with controllability Gramian P and observability Gramian Q is called *balanced*, if $P = Q = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$.

If a system is not balanced, then we can find a state-space transformation that balances the system. Before, we have to check how state-space transformations affect the Gramians.

Lemma 4.4: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable. Let $T \in \mathbb{R}^{n\times n}$ be invertible and define $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right] := [T^{-1}AT, T^{-1}B, CT, D]$. Then

a) P is the controllability Gramian of $[A, B, C, D]$, if and only if $\widetilde{P} := T^{-1}PT^{-\mathsf{T}}$ is the controllability Gramian of $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right]$;

b) Q is the observability Gramian of $[A, B, C, D]$, if and only if $\widetilde{Q} := T^{\mathsf{T}}QT$ is the observability Gramian of $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right]$.

Proof. Exercise. □

Now we show how to balance a system using so-called *balancing transformations*.

Theorem 4.5: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable, controllable, and observable. Then there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$ such that $[T^{-1}AT, T^{-1}B, CT, D]$ is balanced.

Proof. By assumption, for the Gramians P and Q we have $P > 0$ and $Q > 0$. Thus, there exist Cholesky decompositions $P = RR^{\mathsf{T}}$ and $Q = LL^{\mathsf{T}}$,

where R and L are lower triangular and invertible. Now consider the singular value decomposition $L^{\mathsf{T}}R = U\Sigma V^{\mathsf{T}}$ with orthogonal $U, V \in \mathbb{R}^{n\times n}$ and $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n \geqslant 0$. Since L and R are invertible, so is $L^{\mathsf{T}}R$ and therefore, we have $\sigma_n > 0$.

Set $T := RV\Sigma^{-\frac{1}{2}}$. Since $I_n = \Sigma^{-\frac{1}{2}}U^{\mathsf{T}}L^{\mathsf{T}}RV\Sigma^{-\frac{1}{2}}$ we find $T^{-1} = \Sigma^{-\frac{1}{2}}U^{\mathsf{T}}L^{\mathsf{T}}$. For the controllability Gramian $\widetilde{P}$ of the transformed system we have

$$\begin{aligned}\widetilde{P} &= T^{-1}PT^{-\mathsf{T}}\\ &= \Sigma^{-\frac{1}{2}}U^{\mathsf{T}}L^{\mathsf{T}}RR^{\mathsf{T}}LU\Sigma^{-\frac{1}{2}}\\ &= \Sigma^{-\frac{1}{2}}U^{\mathsf{T}}U\Sigma V^{\mathsf{T}}V\Sigma U^{\mathsf{T}}U\Sigma^{-\frac{1}{2}} = \Sigma.\end{aligned}$$

Analogously, for the transformed observability Gramian $\widetilde{Q}$ we obtain

$$\begin{aligned}\widetilde{Q} &= T^{\mathsf{T}}QT\\ &= \Sigma^{-\frac{1}{2}}V^{\mathsf{T}}R^{\mathsf{T}}LL^{\mathsf{T}}RV\Sigma^{-\frac{1}{2}}\\ &= \Sigma^{-\frac{1}{2}}V^{\mathsf{T}}V\Sigma U^{\mathsf{T}}U\Sigma V^{\mathsf{T}}V\Sigma^{-\frac{1}{2}} = \Sigma = \widetilde{P}.\end{aligned}$$

i. e., the transformed system is balanced. □

Example: We revisit the above example. We have

$$R = \begin{bmatrix}\frac{1}{\sqrt{2}} & 0\\ 0 & \alpha\end{bmatrix}, \quad L = \begin{bmatrix}\frac{1}{\sqrt{2}} & 0\\ 0 & \frac{1}{\alpha}\end{bmatrix}.$$

Moreover, it holds

$$L^{\mathsf{T}}R = \begin{bmatrix}\frac{1}{2} & 0\\ 0 & 1\end{bmatrix} = \begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & \frac{1}{2}\end{bmatrix}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}.$$

The balancing transformation is given by

$$T = RV\Sigma^{-\frac{1}{2}} = \begin{bmatrix}\frac{1}{\sqrt{2}} & 0\\ 0 & \alpha\end{bmatrix}\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & \sqrt{2}\end{bmatrix} = \begin{bmatrix}0 & 1\\ \alpha & 0\end{bmatrix}.$$

Now the balanced system $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right] \in \Sigma_{2,1,1}$ is given by

$$\begin{aligned}\widetilde{A} &= \begin{bmatrix}0 & \frac{1}{\alpha}\\ 1 & 0\end{bmatrix}\begin{bmatrix}-1 & -\frac{4}{\alpha}\\ 4\alpha & -2\end{bmatrix}\begin{bmatrix}0 & 1\\ \alpha & 0\end{bmatrix} = \begin{bmatrix}-2 & 4\\ -4 & -1\end{bmatrix},\\ \widetilde{B} &= \begin{bmatrix}0 & \frac{1}{\alpha}\\ 1 & 0\end{bmatrix}\begin{bmatrix}1\\ 2\alpha\end{bmatrix} = \begin{bmatrix}2\\ 1\end{bmatrix},\\ \widetilde{C} &= \begin{bmatrix}-1 & \frac{2}{\alpha}\end{bmatrix}\begin{bmatrix}0 & 1\\ \alpha & 0\end{bmatrix} = \begin{bmatrix}2 & -1\end{bmatrix},\\ \widetilde{D} &= 0.\end{aligned}$$

The transformed system does not depend on α. The eigenvector e_2 of $\widetilde{P} = \widetilde{Q}$ (e_1 in the old coordinates) is harder to to reach and harder to observe than e_1.

This leads to Algorithm 4.1 for model reduction that is called *balanced truncation*.

Algorithm 4.1 Balanced truncation (basic version)

Input: Asymptotically stable and minimal system $[A, B, C, D] \in \Sigma_{n,m,p}$, desired reduced order r.

Output: Reduced-order model $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$.

1: Solve the Lyapunov equations

$$AP + PA^{\mathsf{T}} = -BB^{\mathsf{T}}, \quad A^{\mathsf{T}}Q + QA = -C^{\mathsf{T}}C$$

for $P > 0$ and $Q > 0$.

2: Compute Cholesky factorization $P = RR^{\mathsf{T}}$ and $Q = LL^{\mathsf{T}}$.

3: Compute the singular value decomposition $L^{\mathsf{T}}R = U\Sigma V^{\mathsf{T}}$.

4: Set $T := RV\Sigma^{-\frac{1}{2}}$ (and $T^{-1} = \Sigma^{-\frac{1}{2}}U^{\mathsf{T}}L^{\mathsf{T}}$).

5: Do the balancing transformation

$$\left[T^{-1}AT, T^{-1}B, CT, D\right] = \left[\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \begin{bmatrix} C_1 & C_2 \end{bmatrix}, D\right]$$

and set the reduced-order model as $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$.

4.3 Hankel Operator and Hankel Singular Values

In this section we want to discuss the foundation for the analysis of the balanced truncation algorithm introduced above. For this, we need the Hankel operator and the Hankel singular values [Ant05, Sec. 5.4]. Consider the state equation $\dot{x}(t) = Ax(t) + Bu(t)$ with $x(-\infty) = 0$ and an input $u \in \mathcal{L}_2((-\infty, 0], \mathbb{R}^m)$ that acts on the negative time-horizon leading to $x(0) = x_0$. By switching off the input at $t = 0$, the output equation $y(t) = Cx(t) + Du(t)$ gives an output signal $y \in \mathcal{L}_2([0, \infty), \mathbb{R}^p)$ on the positive time-horizon. This defines an operator

$$\mathcal{H} : \mathcal{L}_2((-\infty, 0], \mathbb{R}^m) \to \mathcal{L}_2([0, \infty), \mathbb{R}^p), \quad u \mapsto y,$$

which is called the *Hankel operator* of the system $[A, B, C, D] \in \Sigma_{n,m,p}$. We have

$$x_0 = \int_{-\infty}^{0} \mathrm{e}^{-A\tau}Bu(\tau)\mathrm{d}\tau, \quad y(t) = C\mathrm{e}^{At}x_0,$$

and thus we obtain

$$(\mathcal{H}u)(t) = y(t) = \int_{-\infty}^{0} C\mathrm{e}^{A(t-\tau)}Bu(\tau)\mathrm{d}\tau.$$

Lemma 4.6: If A is asymptotically stable, then $\mathcal{H}$ is a bounded linear operator.

Proof. Exercise. □

Definition 4.7: Let $\mathcal{V}$, $\mathcal{W}$ be two linear spaces with the inner products $\langle\cdot,\cdot\rangle_{\mathcal{V}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, respectively. Furthermore, let $\mathcal{L} : \mathcal{V} \to \mathcal{W}$ be a linear operator. Then $\mathcal{L}^* : \mathcal{W} \to \mathcal{V}$ is called the *adjoint* of $\mathcal{L}$, if

$$\langle \mathcal{L}v, w\rangle_{\mathcal{W}} = \langle v, \mathcal{L}^* w\rangle_{\mathcal{V}} \quad \text{for all } v \in \mathcal{V},\ w \in \mathcal{W}.$$

For $\mathcal{H}$ as above and $u \in \mathcal{L}_2((-\infty, 0], \mathbb{R}^m)$, $y \in \mathcal{L}_2([0, \infty), \mathbb{R}^p)$ we obtain

$$\begin{aligned}
\langle \mathcal{H}u, y\rangle_{\mathcal{L}_2([0,\infty),\mathbb{R}^p)} &= \int_0^\infty ((\mathcal{H}u)(t))^{\mathsf{T}} y(t)\mathrm{d}t \\
&= \int_0^\infty \int_{-\infty}^0 u(\tau)^{\mathsf{T}} B^{\mathsf{T}} \mathrm{e}^{A^{\mathsf{T}}(t-\tau)} C^{\mathsf{T}} y(t)\mathrm{d}\tau\mathrm{d}t \\
&= \int_{-\infty}^0 u(\tau)^{\mathsf{T}} \int_0^\infty B^{\mathsf{T}} \mathrm{e}^{A^{\mathsf{T}}(t-\tau)} C^{\mathsf{T}} y(t)\mathrm{d}t\mathrm{d}\tau \\
&= \langle u, \mathcal{H}^* y\rangle_{\mathcal{L}_2((-\infty,0],\mathbb{R}^m)}.
\end{aligned}$$

Therefore, we have

$$\mathcal{H}^* : \mathcal{L}_2([0, \infty), \mathbb{R}^p) \to \mathcal{L}_2((-\infty, 0], \mathbb{R}^m), \quad y \mapsto \int_0^\infty B^{\mathsf{T}} \mathrm{e}^{A^{\mathsf{T}}(t-\cdot)} C^{\mathsf{T}} y(t)\mathrm{d}t.$$

Definition 4.8: Let $\mathcal{V}$, $\mathcal{W}$ be two linear spaces with the inner products $\langle\cdot,\cdot\rangle_{\mathcal{V}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{W}}$, respectively. Let $\mathcal{L} : \mathcal{V} \to \mathcal{W}$ be a linear operator with adjoint $\mathcal{L}^* : \mathcal{W} \to \mathcal{V}$. Then $\sigma \in \mathbb{R}^+$ is called a *singular value* of $\mathcal{L}$, if σ^2 is an eigenvalue of $\mathcal{L}^*\mathcal{L}$, i. e., there exists a $v \in \mathcal{V}\backslash\{0\}$ such that $\mathcal{L}^*\mathcal{L}v = \sigma^2 v$.

Note that, if $\mathcal{L}^*\mathcal{L}v = \lambda v$, then

$$\lambda \|v\|_{\mathcal{V}}^2 = \lambda\langle v, v\rangle_{\mathcal{V}} = \langle v, \mathcal{L}^*\mathcal{L}v\rangle_{\mathcal{V}} = \langle \mathcal{L}v, \mathcal{L}v\rangle_{\mathcal{W}} = \|\mathcal{L}v\|_{\mathcal{W}}^2,$$

i. e., λ is real and nonnegative.

Definition 4.9: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable and let $\mathcal{H}$ be its Hankel operator. Then the positive singular values of $\mathcal{H}$ are called *Hankel singular values*.

We want to compute the Hankel singular values using the state-space matrices A, B, C, D only. This will be the goal of the following considerations.

Theorem 4.10: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable. Let P and Q its controllability and observability Gramians and $\mathcal{H}$ its Hankel operator. Then the Hankel singular values are exactly the (positive) square-roots of the eigenvalues of PQ.

Proof. We have

$$y(t) := (\mathcal{H}u)(t) = \int_{-\infty}^{0} Ce^{A(t-\tau)}Bu(\tau)d\tau = Ce^{At}z$$

for

$$z := \int_{-\infty}^{0} e^{-A\tau}Bu(\tau)d\tau. \tag{4.2}$$

Then we get

$$(\mathcal{H}^*y)(t) = \int_{0}^{\infty} B^{\mathsf{T}}e^{A^{\mathsf{T}}(\tau-t)}C^{\mathsf{T}}y(\tau)d\tau = B^{\mathsf{T}}e^{-A^{\mathsf{T}}t}\int_{0}^{\infty} e^{A^{\mathsf{T}}\tau}C^{\mathsf{T}}y(\tau)d\tau.$$

This leads to

$$(\mathcal{H}^*\mathcal{H}u)(t) = (\mathcal{H}^*y)(t) = B^{\mathsf{T}}e^{-A^{\mathsf{T}}t}\int_{0}^{\infty} e^{A^{\mathsf{T}}\tau}C^{\mathsf{T}}Ce^{A\tau}zd\tau = B^{\mathsf{T}}e^{-A^{\mathsf{T}}t}Qz.$$

Assume that $\sigma > 0$ is a singular value of $\mathcal{H}$. Then there exists an eigenfunction $u \in \mathcal{L}_2((-\infty, 0], \mathbb{R}^m)$ of $\mathcal{H}^*\mathcal{H}$ corresponding to an eigenvalue $\sigma^2 > 0$, i. e.,

$$(\mathcal{H}^*\mathcal{H}u)(t) = B^{\mathsf{T}}e^{-A^{\mathsf{T}}t}Qz = \sigma^2 u(t).$$

This gives

$$u(t) = \frac{1}{\sigma^2}B^{\mathsf{T}}e^{-A^{\mathsf{T}}t}Qz. \tag{4.3}$$

With (4.2) we get

$$z = \int_{-\infty}^{0} e^{-A\tau}B\frac{1}{\sigma^2}B^{\mathsf{T}}e^{-A^{\mathsf{T}}\tau}Qzd\tau = \frac{1}{\sigma^2}PQz,$$

i. e., σ^2 is an eigenvalue of PQ.

Now assume that σ^2 is an eigenvalue of PQ with an eigenvector $z \in \mathbb{R}^n \backslash \{0\}$. Define $u \in \mathcal{L}_2((-\infty, 0], \mathbb{R}^m)$ as in (4.3). Then we have

$$\begin{aligned}
(\mathcal{H}^*\mathcal{H}u)(t) &= B^\mathsf{T} \mathrm{e}^{-A^\mathsf{T} t} \int_0^\infty \mathrm{e}^{A^\mathsf{T}\tau} C^\mathsf{T} \int_{-\infty}^0 C\mathrm{e}^{A(\tau-s)} B u(s) \mathrm{d}s \mathrm{d}\tau \\
&= B^\mathsf{T} \mathrm{e}^{-A^\mathsf{T} t} \int_0^\infty \mathrm{e}^{A^\mathsf{T}\tau} C^\mathsf{T} \int_{-\infty}^0 C\mathrm{e}^{A(\tau-s)} B \frac{1}{\sigma^2} B^\mathsf{T} \mathrm{e}^{-A^\mathsf{T} s} Q z \mathrm{d}s \mathrm{d}\tau \\
&= B^\mathsf{T} \mathrm{e}^{-A^\mathsf{T} t} \int_0^\infty \mathrm{e}^{A^\mathsf{T}\tau} C^\mathsf{T} C \mathrm{e}^{A\tau} \frac{1}{\sigma^2} \int_{-\infty}^0 \mathrm{e}^{-As} B B^\mathsf{T} \mathrm{e}^{-A^\mathsf{T} s} Q z \mathrm{d}s \mathrm{d}\tau \\
&= B^\mathsf{T} \mathrm{e}^{-A^\mathsf{T} t} \int_0^\infty \mathrm{e}^{A^\mathsf{T}\tau} C^\mathsf{T} C \mathrm{e}^{A\tau} \underbrace{\frac{1}{\sigma^2} P Q z}_{=z} \mathrm{d}\tau \\
&= B^\mathsf{T} \mathrm{e}^{-A^\mathsf{T} t} Q z = \sigma^2 u(t),
\end{aligned}$$

i. e., σ is a singular value of $\mathcal{H}$. □

Note that for a minimal system, there exist the Cholesky factorizations $P = RR^\mathsf{T}$ and $Q = LL^\mathsf{T}$. Thus, if σ^2 is an eigenvalue of PQ, then we have

$$PQz = \left(RR^\mathsf{T}\right)\left(LL^\mathsf{T}\right)z = \sigma^2 z$$

This is equivalent to

$$\left(L^\mathsf{T}R\right)\left(R^\mathsf{T}L\right)L^\mathsf{T}z = \sigma^2 L^\mathsf{T}z,$$

which implies that σ is a singular value of $L^\mathsf{T}R$. Therefore, we obtain the Hankel singular values as a side product when computing the balancing transformation in Algorithm 4.1.

4.4 Properties of Balanced Truncation

In this section we analyze properties of Algorithm 4.1, see also [Ant05, Sec. 7.2]. In particular, we will derive an error bound using the Hankel singular values.

Theorem 4.11: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable and minimal. Apply Algorithm 4.1 to obtain the reduced-order model $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$. Assume $\sigma_r > \sigma_{r+1}$ for the Hankel singular values σ_i, $i = 1, \ldots, n$. Then the reduced-order model $[A_{11}, B_1, C_1, D]$ is asymptotically stable, minimal, and balanced with the Gramians $P_{11} = Q_{11} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r) =: \Sigma_1$.

Proof. Since the system $[A, B, C, D] \in \Sigma_{n,m,p}$ is minimal, the balancing transformation with T leads to the transformed Gramians

$$\widetilde{P} = \widetilde{Q} = \mathrm{diag}(\sigma_1, \ldots, \sigma_n) =: \mathrm{diag}(\Sigma_1, \Sigma_2) > 0.$$

Then (in balanced coordinates), the Lyapunov equations

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} A_{11}^{\mathsf{T}} & A_{21}^{\mathsf{T}} \\ A_{12}^{\mathsf{T}} & A_{22}^{\mathsf{T}} \end{bmatrix} = - \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \begin{bmatrix} B_1^{\mathsf{T}} & B_2^{\mathsf{T}} \end{bmatrix}, \tag{4.4}$$

$$\begin{bmatrix} A_{11}^{\mathsf{T}} & A_{21}^{\mathsf{T}} \\ A_{12}^{\mathsf{T}} & A_{22}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = - \begin{bmatrix} C_1^{\mathsf{T}} \\ C_2^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} C_1 & C_2 \end{bmatrix} \tag{4.5}$$

are satisfied. If the reduced-order model is asymptotically stable, i. e., $\Lambda(A_{11}) \subset \mathbb{C}^-$, then $\Sigma_1 > 0$ is the controllability and observability Gramian of the reduced-order model, i. e., the reduced-order model is minimal and balanced. Now we show that we indeed have $\Lambda(A_{11}) \subset \mathbb{C}^-$. Let $\lambda \in \mathbb{C}$ be an eigenvalue of A_{11}^{T} with eigenvector $v \in \mathbb{C}^r$. Then we obtain

$$\underbrace{-\left\|B_1^{\mathsf{T}} v\right\|_2^2}_{\leqslant 0} = v^{\mathsf{H}} A_{11} \Sigma_1 v + v^{\mathsf{H}} \Sigma_1 A_{11}^{\mathsf{T}} v = 2 \operatorname{Re}(\lambda) \underbrace{v^{\mathsf{H}} \Sigma_1 v}_{>0}.$$

This implies $\operatorname{Re}(\lambda) \leqslant 0$. It remains to show that A_{11} has no eigenvalues on the imaginary axis. Therefore, assume that there exist imaginary eigenvalues. Let $\mathrm{i}\omega \in \mathrm{i}\mathbb{R}$ be an imaginary eigenvalue and $\{v_1, \ldots, v_q\} \subset \mathbb{C}^r$ be an orthonormal basis of $\ker(A_{11} - \mathrm{i}\omega I_r)$ and define $V = \begin{bmatrix} v_1 & \ldots & v_q \end{bmatrix}$. Then we have

$$(A_{11} - \mathrm{i}\omega I_r) V = 0, \quad V^{\mathsf{H}} \left(A_{11}^{\mathsf{T}} + \mathrm{i}\omega I_r\right) = 0.$$

Moreover, we have

$$(A_{11} - \mathrm{i}\omega I_r)\Sigma_1 + \Sigma_1 \left(A_{11}^{\mathsf{T}} + \mathrm{i}\omega I_r\right) = -B_1 B_1^{\mathsf{T}}, \tag{4.6}$$

$$\left(A_{11}^{\mathsf{T}} + \mathrm{i}\omega I_r\right)\Sigma_1 + \Sigma_1 (A_{11} - \mathrm{i}\omega I_r) = -C_1^{\mathsf{T}} C_1. \tag{4.7}$$

Multiplying (4.7) with V^{H} from the left and with V from the right gives

$$\underbrace{V^{\mathsf{H}} \left(A_{11}^{\mathsf{T}} + \mathrm{i}\omega I_r\right)\Sigma_1 V}_{=0} + \underbrace{V^{\mathsf{H}} \Sigma_1 (A_{11} - \mathrm{i}\omega I_r) V}_{=0} = -V^{\mathsf{H}} C_1^{\mathsf{T}} C_1 V,$$

resulting in $C_1 V = 0$. Multiplying (4.7) with V from the right yields

$$\left(A_{11}^{\mathsf{T}} + \mathrm{i}\omega I_r\right)\Sigma_1 V + \underbrace{\Sigma_1 (A_{11} - \mathrm{i}\omega I_r) V}_{=0} = -C_1^{\mathsf{T}} \underbrace{C_1 V}_{=0},$$

and thus $\left(A_{11}^{\mathsf{T}} + \mathrm{i}\omega I_r\right)\Sigma_1 V = 0$. Now multiplying (4.6) with $V^{\mathsf{H}}\Sigma_1$ from the left and with $\Sigma_1 V$ from the right results in

$$\underbrace{V^{\mathsf{H}} \Sigma_1 (A_{11} - \mathrm{i}\omega I_r)\Sigma_1^2 V}_{=0} + \underbrace{V^{\mathsf{H}} \Sigma_1^2 \left(A_{11}^{\mathsf{T}} + \mathrm{i}\omega I_r\right)\Sigma_1 V}_{=0} = -V^{\mathsf{H}} \Sigma_1 B_1 B_1^{\mathsf{T}} \Sigma_1 V,$$

giving $B_1^\mathsf{T}\Sigma_1 V = 0$. By multiplying (4.6) with $\Sigma_1 V$ from the right, we obtain

$$(A_{11} - \mathrm{i}\omega I_r)\Sigma_1^2 V + \underbrace{\Sigma_1\big(A_{11}^\mathsf{T} + \mathrm{i}\omega I_r\big)\Sigma_1 V}_{=0} = -B_1 \underbrace{B_1^\mathsf{T}\Sigma_1 V}_{=0},$$

so we have $(A_{11} - \mathrm{i}\omega I_r)\Sigma_1^2 V = 0$. Since V spans $\ker(A_{11} - \mathrm{i}\omega I_r)$, we have

$$\Sigma_1^2 V = V\Xi \quad \text{for } \Xi \in \mathbb{C}^{q\times q} \text{ with } \Lambda(\Xi) \subseteq \Lambda\big(\Sigma_1^2\big). \tag{4.8}$$

Multiplying the (2,1) block of (4.4) by $\Sigma_1 V$ from the right yields

$$A_{21}\Sigma_1^2 V + \Sigma_2 A_{12}^\mathsf{T}\Sigma_1 V = -B_2 B_1^\mathsf{T}\Sigma_1 V = 0.$$

On the other hand, multiplying the (2,1) block of (4.5) by V from the right results in

$$A_{12}^\mathsf{T}\Sigma_1 V + \Sigma_2 A_{21} V = -C_2^\mathsf{T} C_1 V = 0.$$

Using (4.8) and both of the last two equations we get

$$A_{21}V\Xi = A_{21}\Sigma_1^2 V = -\Sigma_2 A_{12}^\mathsf{T}\Sigma_1 V = \Sigma_2^2 A_{21} V,$$

hence

$$(A_{21}V)\Xi - \Sigma_2^2(A_{21}V) = 0.$$

This is a *Sylvester matrix equation* with the unknown $A_{21}V$. Since by (4.8), $\Lambda(\Xi) \cap \Lambda(\Sigma_2^2) = \varnothing$, it is uniquely solvable (see exercise!) and thus we have $A_{21}V = 0$. Now we have

$$\widetilde{A}\begin{bmatrix} V \\ 0 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} V \\ 0 \end{bmatrix} = \begin{bmatrix} A_{11}V \\ A_{21}V \end{bmatrix} = \mathrm{i}\omega \begin{bmatrix} V \\ 0 \end{bmatrix}.$$

Thus, $\mathrm{i}\omega$ is an imaginary eigenvalue of A, contradicting its asymptotic stability. □

In the next theorem we will move towards an error bound for balanced truncation.

Theorem 4.12: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ with transfer function $G \in \mathcal{RH}_\infty^{p\times m}$ be asymptotically stable and balanced with the controllability and observability Gramians $P = Q = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$. Let $\sigma_r > \sigma_{r+1} = \ldots = \sigma_n$. Let $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$ be the reduced-order model of order r obtained by Algorithm 4.1 with transfer function $\widetilde{G} \in \mathcal{RH}_\infty^{p\times m}$. Then it holds that

$$\big\|G - \widetilde{G}\big\|_{\mathcal{H}_\infty} \leqslant 2\sigma_{r+1}$$

(independently of the multiplicity of σ_{r+1}).

Proof. Define

$$\Sigma_1 := \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r), \quad \Sigma_2 := \operatorname{diag}(\sigma_{r+1}, \sigma_{r+2}, \ldots, \sigma_n),$$

and the error transfer function

$$E(s) = G(s) - \widetilde{G}(s) = C(sI_n - A)^{-1}B + D - \left(C_1(sI_r - A_{11})^{-1}B_1 + D\right),$$

and consider its realization $\left[\widehat{A}, \widehat{B}, \widehat{C}, \widehat{D}\right] \in \Sigma_{n+r,m,p}$ with

$$\widehat{A} = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & A_{11} & A_{12} \\ 0 & A_{21} & A_{22} \end{bmatrix}, \quad \widehat{B} = \begin{bmatrix} B_1 \\ B_1 \\ B_2 \end{bmatrix}, \quad \widehat{C} = \begin{bmatrix} -C_1 & C_1 & C_2 \end{bmatrix}, \quad \widehat{D} = 0.$$

Using the state-space transformation

$$T := \begin{bmatrix} I_r & I_r & 0 \\ I_r & -I_r & 0 \\ 0 & 0 & I_{n-r} \end{bmatrix} \quad \text{with} \quad T^{-1} = \frac{1}{2}\begin{bmatrix} I_r & I_r & 0 \\ I_r & -I_r & 0 \\ 0 & 0 & 2I_{n-r} \end{bmatrix},$$

we obtain the alternative realization

$$\left[\widehat{A}_1, \widehat{B}_1, \widehat{C}_1, \widehat{D}_1\right] := \left[T^{-1}\widehat{A}T, T^{-1}\widehat{B}, \widehat{C}T, \widehat{D}\right] \in \Sigma_{n+r,m,p}$$

with

$$\widehat{A}_1 = \begin{bmatrix} A_{11} & 0 & \frac{1}{2}A_{12} \\ 0 & A_{11} & -\frac{1}{2}A_{12} \\ A_{21} & -A_{21} & A_{22} \end{bmatrix}, \quad \widehat{B}_1 = \begin{bmatrix} B_1 \\ 0 \\ B_2 \end{bmatrix},$$

$$\widehat{C}_1 = \begin{bmatrix} 0 & -2C_1 & C_2 \end{bmatrix}, \quad \widehat{D}_1 = 0.$$

Now define

$$A_0 := \widehat{A}_1, \quad B_0 := \begin{bmatrix} \widehat{B}_1 & \widehat{B}_2 \end{bmatrix} := \begin{bmatrix} B_1 & 0 \\ 0 & \sigma_{r+1}\Sigma_1^{-1}C_1^{\mathsf{T}} \\ B_2 & -C_2^{\mathsf{T}} \end{bmatrix},$$

$$C_0 := \begin{bmatrix} \widehat{C}_1 \\ \widehat{C}_2 \end{bmatrix} := \begin{bmatrix} 0 & -2C_1 & C_2 \\ -2\sigma_{r+1}B_1^{\mathsf{T}}\Sigma_1^{-1} & 0 & -B_2^{\mathsf{T}} \end{bmatrix},$$

$$D_0 := \begin{bmatrix} 0 & 2\sigma_{r+1}I_p \\ 2\sigma_{r+1}I_m & 0 \end{bmatrix}.$$

Then the transfer function of $[A_0, B_0, C_0, D_0] \in \Sigma_{r+n,m+p,m+p}$ is given by

$$\begin{aligned} E_0(s) &:= C_0(sI_{r+n} - A_0)^{-1}B_0 + D_0 \\ &= \begin{bmatrix} \widehat{C}_1\left(sI_{r+n} - \widehat{A}_1\right)^{-1}\widehat{B}_1 & \widehat{C}_1\left(sI_{r+n} - \widehat{A}_1\right)^{-1}\widehat{B}_2 + 2\sigma_{r+1}I_p \\ \widehat{C}_2\left(sI_{r+n} - \widehat{A}_1\right)^{-1}\widehat{B}_1 + 2\sigma_{r+1}I_m & \widehat{C}_2\left(sI_{r+n} - \widehat{A}_1\right)^{-1}\widehat{B}_2 \end{bmatrix} \end{aligned}$$

Since A and A_{11} are both asymptotically stable, then also the matrices $\widehat{A}$, $\widehat{A}_1$, and A_0 are asymptotically stable by construction. Moreover, $[A_0, B_0, C_0, D_0]$ has the controllability Gramian $P_0 = \operatorname{diag}\left(\Sigma_1,\, \sigma_{r+1}^2\Sigma_1^{-1},\, 2\Sigma_2\right)$, because

$$\begin{aligned} A_0P_0 + P_0A_0^{\mathsf{T}} + B_0B_0^{\mathsf{T}} &= \begin{bmatrix} A_{11}\Sigma_1 & 0 & A_{12}\Sigma_2 \\ 0 & \sigma_{r+1}^2 A_{11}\Sigma_1^{-1} & -A_{12}\Sigma_2 \\ A_{21}\Sigma_1 & -\sigma_{r+1}^2 A_{21}\Sigma_1^{-1} & 2A_{22}\Sigma_2 \end{bmatrix} \\ &\quad + \begin{bmatrix} \Sigma_1 A_{11}^{\mathsf{T}} & 0 & \Sigma_1 A_{21}^{\mathsf{T}} \\ 0 & \sigma_{r+1}^2\Sigma_1^{-1}A_{11}^{\mathsf{T}} & -\sigma_{r+1}^2\Sigma_1^{-1}A_{21}^{\mathsf{T}} \\ \Sigma_2 A_{12}^{\mathsf{T}} & -\Sigma_2 A_{12}^{\mathsf{T}} & 2\Sigma_2 A_{22}^{\mathsf{T}} \end{bmatrix} \\ &\quad + \begin{bmatrix} B_1B_1^{\mathsf{T}} & 0 & B_1B_2^{\mathsf{T}} \\ 0 & \sigma_{r+1}^2\Sigma_1^{-1}C_1^{\mathsf{T}}C_1\Sigma_1^{-1} & -\sigma_{r+1}\Sigma_1^{-1}C_1^{\mathsf{T}}C_2 \\ B_2B_1^{\mathsf{T}} & -\sigma_{r+1}C_2^{\mathsf{T}}C_1\Sigma_1^{-1} & B_2B_2^{\mathsf{T}} + C_2^{\mathsf{T}}C_2 \end{bmatrix} = 0, \end{aligned} \tag{4.9}$$

where the latter equality follows from combining (4.4) and (4.5). Moreover, we have

$$B_0D_0^{\mathsf{T}} = \begin{bmatrix} 0 & 2\sigma_{r+1}B_1 \\ 2\sigma_{r+1}^2\Sigma_1^{-1}C_1^{\mathsf{T}} & 0 \\ -2\sigma_{r+1}C_2^{\mathsf{T}} & 2\sigma_{r+1}B_2 \end{bmatrix} = -P_0C_0^{\mathsf{T}}. \tag{4.10}$$

We have

$$\|E\|_{\mathcal{H}_\infty} = \sup_{\omega\in\mathbb{R}} \|E(\mathrm{i}\omega)\|_2 \leqslant \sup_{\omega\in\mathbb{R}} \|E_0(\mathrm{i}\omega)\|_2 = \sup_{\omega\in\mathbb{R}} \left(\lambda_{\max}\left(E_0(\mathrm{i}\omega)E_0(\mathrm{i}\omega)^{\mathsf{H}}\right)\right)^{1/2}.$$

Define the conjugated transfer function $E_0^{\sim}(s) := E_0(-\overline{s})^{\mathsf{H}}$ which is realized by $\left[-A_0^{\mathsf{T}}, C_0^{\mathsf{T}}, -B_0^{\mathsf{T}}, D_0^{\mathsf{T}}\right] \in \Sigma_{r+n,m+p,m+p}$. Then a realization of $E_0(s)E_0^{\sim}(s)$ is given by $\left[\widetilde{A}_0, \widetilde{B}_0, \widetilde{C}_0, \widetilde{D}_0\right] \in \Sigma_{2(r+n),m+p,m+p}$ (see homework!) with

$$\widetilde{A}_0 = \begin{bmatrix} A_0 & -B_0B_0^{\mathsf{T}} \\ 0 & -A_0^{\mathsf{T}} \end{bmatrix}, \quad \widetilde{B}_0 = \begin{bmatrix} B_0D_0^{\mathsf{T}} \\ C_0^{\mathsf{T}} \end{bmatrix}, \quad \widetilde{C}_0 = \begin{bmatrix} C_0 & -D_0B_0^{\mathsf{T}} \end{bmatrix},$$
$$\widetilde{D}_0 = D_0D_0^{\mathsf{T}}.$$

Using the state-space transformation

$$\widetilde{T} := \begin{bmatrix} I_{r+n} & -P_0 \\ 0 & I_{r+n} \end{bmatrix} \quad \text{with} \quad \widetilde{T}^{-1} := \begin{bmatrix} I_{r+n} & P_0 \\ 0 & I_{r+n} \end{bmatrix},$$

we obtain an equivalent realization by $\left[\widehat{A}_0, \widehat{B}_0, \widehat{C}_0, \widehat{D}_0\right] \in \Sigma_{2(r+n),m+p,m+p}$

$$\begin{aligned} \widehat{A}_0 &:= \widetilde{T}^{-1}\widetilde{A}_0\widetilde{T} = \begin{bmatrix} A_0 & -A_0P_0 - P_0A_0^{\mathsf{T}} - B_0B_0^{\mathsf{T}} \\ 0 & -A_0^{\mathsf{T}} \end{bmatrix} \overset{(4.9)}{=} \begin{bmatrix} A_0 & 0 \\ 0 & -A_0^{\mathsf{T}} \end{bmatrix}, \\ \widehat{B}_0 &:= \widetilde{T}^{-1}\widetilde{B}_0 = \begin{bmatrix} B_0D_0^{\mathsf{T}} + P_0C_0^{\mathsf{T}} \\ C_0^{\mathsf{T}} \end{bmatrix} \overset{(4.10)}{=} \begin{bmatrix} 0 \\ C_0^{\mathsf{T}} \end{bmatrix}, \\ \widehat{C}_0 &:= \widetilde{C}_0\widetilde{T} = \begin{bmatrix} C_0 & -C_0P_0 - D_0B_0^{\mathsf{T}} \end{bmatrix} \overset{(4.10)}{=} \begin{bmatrix} C_0 & 0 \end{bmatrix}, \\ \widehat{D}_0 &:= \widetilde{D}_0. \end{aligned}$$

Finally we obtain

$$\begin{aligned} E_0(s)E_0^{\sim}(s) &= \widehat{C}_0\big(sI_{2(r+n)} - \widehat{A}_0\big)^{-1}\widehat{B}_0 + \widehat{D}_0 \\ &= \begin{bmatrix} C_0 & 0 \end{bmatrix} \begin{bmatrix} sI_{r+n} - A_0 & 0 \\ 0 & sI_{r+n} + A_0^{\mathsf{T}} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ C_0^{\mathsf{T}} \end{bmatrix} + D_0 D_0^{\mathsf{T}} \\ &= \begin{bmatrix} 0 & 2\sigma_{r+1} I_p \\ 2\sigma_{r+1} I_m & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 2\sigma_{r+1} I_m \\ 2\sigma_{r+1} I_p & 0 \end{bmatrix} = 4\sigma_{r+1}^2 I_{m+p}. \end{aligned}$$

This implies

$$\begin{aligned} \|E\|_{\mathcal{H}_\infty} &\leqslant \sup_{\omega\in\mathbb{R}} \Big(\lambda_{\max}\big(E_0(\mathrm{i}\omega)E_0(\mathrm{i}\omega)^{\mathsf{H}}\big)\Big)^{1/2} \\ &= \sup_{\omega\in\mathbb{R}} \big(\lambda_{\max}\big(E_0(\mathrm{i}\omega)E_0^{\sim}(\mathrm{i}\omega)\big)\big)^{1/2} = \sqrt{4\sigma_{r+1}^2} = 2\sigma_{r+1}. \end{aligned}$$

□

Now we can conclude an $\mathcal{H}_\infty$ error bound for the general case.

Corollary 4.13: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ with transfer function $G \in \mathcal{RH}_\infty^{p\times m}$ be asymptotically stable and balanced with the controllability and observability Gramians $P = Q = \operatorname{diag}(\sigma_1 I_{s_1}, \sigma_2 I_{s_2}, \ldots, \sigma_k I_{s_k})$, where $\sigma_1 > \sigma_2 > \ldots > \sigma_k \geqslant 0$. Let $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$ be the reduced-order model of order r obtained by Algorithm 4.1 with $r = s_1 + s_2 + \ldots + s_\ell$ for some $\ell \leqslant k$ and with transfer function $\widetilde{G} \in \mathcal{RH}_\infty^{p\times m}$. Then it holds that

$$\|G - \widetilde{G}\|_{\mathcal{H}_\infty} \leqslant \sum_{j=\ell+1}^{k} 2\sigma_j.$$

Proof. Denote by $G_j(s)$ the transfer function of the reduced-order model obtained by Algorithm 4.1 by truncating only the Hankel singular values $\sigma_{j+1}, \ldots, \sigma_k$. So we have $G(s) = G_k(s)$ and $\widetilde{G}(s) = G_\ell(s)$. Now it holds that

$$\begin{aligned} G(s) - G_\ell(s) = (G_k(s) - G_{k-1}(s)) + (G_{k-1}(s) - G_{k-2}(s)) + \ldots \\ + (G_{\ell+1}(s) - G_\ell(s)), \end{aligned}$$

which implies

$$\|G - \widetilde{G}\|_{\mathcal{H}_\infty} \leqslant \sum_{j=\ell+1}^{k} \|G_j - G_{j-1}\|_{\mathcal{H}_\infty} \leqslant 2 \sum_{j=\ell+1}^{k} \sigma_j.$$

□

There are also other error bounds for balanced truncation. An *a-posteriori* error bound is given as follows: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ with transfer function $G \in \mathcal{RH}_\infty^{p\times m}$ be asymptotically stable and balanced. Let $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$ be the reduced-order model of order r obtained by Algorithm 4.1 with transfer function $\widetilde{G} \in \mathcal{RH}_\infty^{p\times m}$. Moreover, assume that $Y_1 \in \mathbb{R}^{r\times r}$ and $Y_2 \in \mathbb{R}^{(n-r)\times r}$ solve the Sylvester equation

$$\begin{bmatrix} A_{11}^\mathsf{T} & A_{21}^\mathsf{T} \\ A_{12}^\mathsf{T} & A_{22}^\mathsf{T} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} + \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} A_{11} = -C^\mathsf{T} C_1.$$

Then we have

$$\|G - \widetilde{G}\|_{\mathcal{H}_2}^2 \leqslant \operatorname{tr}\left(\left(B_2 B_2^\mathsf{T} + 2 Y_2 A_{12} \right) \Sigma_2 \right).$$

4.5 Numerical Solution of Large-Scale Lyapunov Equations

In this section, we discuss the numerical solution of large-scale Lyapunov equations. Since a Lyapunov equation is a special Sylvester equation, the same conditions for unique solvability apply. This means, that a Lyapunov equation

$$AX + XA^\mathsf{T} = -W$$

has a unique solution, if and only if $\Lambda(A) \cap \Lambda(-A) = \varnothing$. Since for balanced truncation, A is assumed to be asymptotically stable, this condition is fulfilled a-priori. In the following we will the derive the alternating directions implicit (ADI) iteration, that was introduced to solve partial differential equations in [PR55]. We will see soon that this method is also suitable for large-scale Lyapunov equations that appear in model reduction. There are many other methods, in particular Krylov subspace methods [Sim07], that are often equally good. For sake of brevity, we will not discuss these here in detail.

4.5.1 Derivation of the ADI Iteration

Consider the discrete-time Lyapunov equation

$$X = AXA^\mathsf{T} + W, \quad A \in \mathbb{R}^{n\times n}, \quad W = W^\mathsf{T} \in \mathbb{R}^{n\times n}. \tag{4.11}$$

The existence of a unique solution is ensured if $|\lambda| < 1$ for all $\lambda \in \Lambda(A)$ (see exercise). This motivates the basic iteration

$$X_k = AX_{k-1}A^\mathsf{T} + W, \quad k \geqslant 1, \quad X_0 \in \mathbb{R}^{n\times n}. \tag{4.12}$$

Let A be diagonalizable, i.e., there exists a nonsingular matrix $V \in \mathbb{C}^{n\times n}$ such that $A = V\Lambda V^{-1}$. Let $\rho(A) := \max_{\lambda\in\Lambda(A)} |\lambda|$ denote the spectral radius of A. Since

$$\begin{aligned}\|X_k - X\|_2 &= \left\|A(X_{k-1} - X)A^{\mathsf{T}}\right\|_2 = \ldots = \left\|A^k(X_0 - X)(A^{\mathsf{T}})^k\right\|_2 \\ &\leqslant \|A^k\|_2^2\|X_0 - X\|_2 \leqslant \|V\|_2^2\|V^{-1}\|_2^2\rho(A)^{2k}\|X_0 - X\|_2, \end{aligned} \tag{4.13}$$

this iteration converges because $\rho(A) < 1$ (fixed point argumentation).

Continuous-time Lyapunov equations can be treated similarly, however, we must first transform the data as pointed out in the next lemma.

Lemma 4.14: The continuous-times Lyapunov equation

$$AX + XA^{\mathsf{T}} = -W, \quad \Lambda(A) \subset \mathbb{C}^-$$

is equivalent to the discrete-time Lyapunov equation

$$\begin{aligned} X =& C(p)XC(p)^{\mathsf{H}} + \widetilde{W}(p), \quad C(p) := (A - \overline{p}I_n)(A + pI_n)^{-1}, \\ \widetilde{W}(p) :=& -2\operatorname{Re}(p)(A + pI_n)^{-1}W(A + pI_n)^{-\mathsf{H}} \end{aligned} \tag{4.14}$$

for $p \in \mathbb{C}^-$.

Proof. Exercise. □

Moreover, $\rho(C(p)) < 1$ (see exercise). Applying (4.12) to (4.14) gives the *Smith iteration*

$$X_k = C(p)X_{k-1}C(p)^{\mathsf{H}} + \widetilde{W}(p), \quad k \geqslant 1, \quad X_0 \in \mathbb{R}^{n\times n}. \tag{4.15}$$

Similarly as in (4.13), we have

$$\|X_k - X\|_2 \leqslant \|V\|_2^2\|V^{-1}\|_2^2\rho(C(p))^{2k}\|X_0 - X\|_2.$$

This means that we obtain fast convergence by choosing p such that $\rho(C(p)) < 1$ is as small as possible. We will discuss this later in more detail.

By varying the shifts p in (4.15) in every step, we obtain the *ADI iteration for Lyapunov equations*

$$X_k = C(p_k)X_{k-1}C(p_k)^{\mathsf{H}} + \widetilde{W}(p_k), \quad k \geqslant 1, \quad X_0 \in \mathbb{R}^{n\times n}, \quad p_k \in \mathbb{C}^-. \tag{4.16}$$

4.5.2 The ADI Shift Parameter Problem

One can show, similarly to (4.13), that

$$\|X_k - X\|_2 \leqslant \|V\|_2^2 \left\|V^{-1}\right\|_2^2 \rho(M_k)^2 \|X_0 - X\|_2, \quad M_k := \prod_{i=1}^{k} C(p_i), \tag{4.17}$$

where V is a transformation matrix diagonalizing A (assuming it is diagonalizable). The eigenvalues of the product of the Cayley transformations M_k are

$$\Lambda(M_k) = \left\{ \prod_{i=1}^{k} \frac{\lambda - \overline{p_i}}{\lambda + p_i} \;\middle|\; \lambda \in \Lambda(A) \right\}.$$

Good shifts $p_1^*, \ldots, p_k^*$ should make $\rho(M_k) < 1$ as small as possible. This motivates the ADI shift parameter problem

$$[p_1^*, \ldots, p_k^*] = \operatorname{argmin}_{[p_1,\ldots,p_k] \in (\mathbb{C}^-)^k} \max_{\lambda \in \Lambda(A)} \left| \prod_{i=1}^{k} \frac{\lambda - \overline{p_i}}{\lambda + p_i} \right|. \tag{4.18}$$

In general, this is very hard to solve. For instance, in general, $\rho(C(p))$ is not differentiable and the problem is very expensive, if A is a large matrix. However, there are some procedures that work well in practice:

- **Wachspress shifts [Wac13]:** Embed $\Lambda(A)$ in an elliptic function region that depends on the the parameters $\max_{\lambda \in \Lambda(A)} \operatorname{Re}(\lambda)$, $\min_{\lambda \in \Lambda(A)} \operatorname{Re}(\lambda)$, and $\arctan \max_{\lambda \in \Lambda(A)} \left| \frac{\operatorname{Im}(\lambda)}{\operatorname{Re}(\lambda)} \right|$ (or approximations thereof). Then, (4.18) can be solved by employing elliptic integral.
- **Heuristic Penzl shifts [Pen00]:** If A is a large and sparse matrix, $\Lambda(A)$ is replaced by a small number of approximate eigenvalues (e.g., Ritz values). Then (4.18) is solved heuristically.
- **Self-generating shifts [BKS15]:** If A is large and sparse, these shifts are based on projections of A with the data obtained by previous iterations. These shifts also make use of the right-hand side W.

4.5.3 The Low-Rank Phenomenon

Now we consider

$$AX + XA^{\mathsf{T}} = -BB^{\mathsf{T}}, \tag{4.19}$$

where $A \in \mathbb{R}^{n \times n}$ and n is 'large', but A is sparse, i. e., only a few entries in A are non-zero. Therefore, multiplication with A can be performed in $\mathcal{O}(n)$ rather than $\mathcal{O}(n^2)$ FLOPS. Also solves with A or $A + pI$ can be performed efficiently.

However, $X \in \mathbb{R}^{n\times n}$ is usually dense and thus X cannot be stored for large n since we would need $\mathcal{O}(n^2)$ memory. Thus the question arises whether it is possible to store the solution X more efficiently. In practice we often have $B \in \mathbb{R}^{n\times m}$, where $m \ll n$, i. e., the right-hand side BB^{T} has a low rank. Recall that if (A, B) is controllable then $X = X^{\mathsf{T}} \succcurlyeq 0$ and $\operatorname{rank}(X) = n$.

It is a very common observation in practice that the eigenvalues of X solving (4.19) decay very rapidly towards zero, and fall early below the machine precision. The following theorem explains this eigenvalue decay [SZ02].

Theorem 4.15: Let A be diagonalizable, i. e., there exists an invertible matrix $V \in \mathbb{C}^{n\times n}$ such that $A = V\Lambda V^{-1}$. Then the eigenvalues of X solving (4.19) satisfy

$$\frac{\lambda_{km+1}(X)}{\lambda_1(X)} \leqslant \|V\|_2^2 \left\|V^{-1}\right\|_2^2 \rho(M_k)^2$$

for any choice of shift parameters p used to construct M_k (in particular, the optimal ones).

Remark 4.16:

- If the eigenvalues of A cluster in the complex plane, only a few p_k in the clusters suffice to get a small $\rho(M_k)$ and thus $\lambda_i(X)$ decay fast.
- If A is normal, then $\|V\|_2 \left\|V^{-1}\right\|_2 = 1$ and the bound gives a good explanation for the decay. The nonnormal case is much harder to understand.
- This bound (and most others) does not precisely incorporate the eigenvectors of A as well as the precise influence of B.

Consequence: If there is a fast decay of $\lambda_i(X)$, then X can be well approximated as $X = X^{\mathsf{T}} \approx ZZ^{\mathsf{H}}$, where $Z \in \mathbb{C}^{n\times r}$ with $r \ll n$ is a *low-rank solution factor*. Hence, only nr memory is required. Thus, in the next subsection we consider algorithms for computing the factor Z without explicitly forming X.

4.5.4 The Low-Rank Cholesky Factor ADI Iteration

The idea [Pen00] consists of considering one step of the dense ADI iteration (4.16) and inserting $X_j = Z_j Z_j^{\mathsf{H}}$. This leads to

$$\begin{aligned} X_j &= C(p_j) X_{j-1} C(p_j)^{\mathsf{H}} + \widetilde{W}(p_j) \\ &= (A - \overline{p}_j I_n)(A + p_j I_n)^{-1} Z_{j-1} Z_{j-1}^{\mathsf{H}} (A + p_j I_n)^{-\mathsf{H}} (A - \overline{p}_j I_n)^{\mathsf{H}} \\ &\quad - 2\operatorname{Re}(p_j)(A + p_j I_n)^{-1} BB^{\mathsf{T}} (A + p_j I_n)^{-\mathsf{H}}. \end{aligned}$$

Note that if $p_j \in \mathbb{R}_-$, then $X_j \in \mathbb{R}^{n\times n}$. Furthermore, if $p_j \in \mathbb{C}^-$ and $p_{j+1} = \overline{p_j}$, then $X_{j+1} \in \mathbb{R}^{n\times n}$. Obviously, we have $X_j = Z_j Z_j^{\mathsf{H}}$ with

$$Z_j = \left[\sqrt{-2\operatorname{Re}(p_j)}(A + p_j I_n)^{-1}B \quad (A - \overline{p}_j I_n)(A + p_j I_n)^{-1} Z_{j-1}\right].$$

With $Z_0 = 0$ we find a low rank variant the ADI iteration (4.16) forming Z_j successively (grows by m columns in each step).

The drawback is that all columns are processed in every step which leads to quickly growing costs (in total jm linear systems have to be solved to get Z_j).

However, there is a remedy to this problem. We have observed that

$$S_i = (A + p_i I_n)^{-1} \text{ and } T_j = (A - \overline{p}_j I_n)$$

commute for all i, j with each other and themselves.

Now consider Z_j being the iterate after iteration step j

$$Z_j = \left[\alpha_j S_j B \quad (T_j S_j)\alpha_{j-1} S_{j-1} B \quad \ldots \quad (T_j S_j)\cdots(T_2 S_2)\alpha_1 S_1 B\right]$$

with $\alpha_i = \sqrt{-2\operatorname{Re}(p_i)}$. Due to the commutativity, the order of application of the shifts is not important, and we reverse their application to obtain the following alternative iterate

$$\begin{aligned}
\widetilde{Z}_j &= \left[\alpha_1 S_1 B \quad \alpha_2 (T_1 S_1) S_2 B \quad \ldots \quad \alpha_j (T_1 S_1)\cdots(T_{j-1} S_{j-1}) S_j B\right] \\
&= \left[\alpha_1 S_1 B \quad \alpha_2 (T_1 S_2) S_1 B \quad \ldots \quad \alpha_j (T_{j-1} S_j)(T_{j-2} S_{j-1})\cdots(T_1 S_2) S_1 B\right] \\
&= \left[\alpha_1 V_1 \quad \alpha_2 V_2 \quad \ldots \quad \alpha_j V_j\right], \\
V_1 &= S_1 B, \quad V_i = T_{i-1} S_i V_{i-1}, \quad i = 1, \ldots, j.
\end{aligned}$$

We have $X_j = \widetilde{Z}_j \widetilde{Z}_j^{\mathsf{H}}$, but in this formulation only the new columns are processed. Even more structure is revealed by the Lyapunov residual.

Theorem 4.17: The residual at step j of (4.16), started with $X_0 = 0$, is of rank at most m and given by

$$\begin{aligned}
R_j :=& A\widetilde{Z}_j \widetilde{Z}_j^{\mathsf{H}} + \widetilde{Z}_j \widetilde{Z}_j^{\mathsf{H}} A^{\mathsf{T}} + BB^{\mathsf{T}} = W_j W_j^{\mathsf{H}}, \\
W_j =& M_j B = C(p_j) W_{j-1} = W_{j-1} - 2\operatorname{Re}(p_j) V_j, \quad W_0 := B,
\end{aligned}$$

where $M_j := \prod_{i=1}^{j} C(p_i)$. Moreover, it holds $V_j = (A + p_j I_n)^{-1} W_{j-1}$.

Proof. We have

$$\begin{aligned}
R_j &= AX_j + X_j A^{\mathsf{T}} + BB^{\mathsf{T}} = A(X_j - X) + (X_j - X)A^{\mathsf{T}} \quad \text{(by (4.19))} \\
&= AM_j(X_0 - X)M_j^{\mathsf{H}} + M_j(X_0 - X)M_j^{\mathsf{H}} A^{\mathsf{T}} \\
&= -M_j A X M_j^{\mathsf{H}} - M_j X A^{\mathsf{T}} M_j^{\mathsf{H}} \\
&= -M_j(AX + XA^{\mathsf{T}})M_j^{\mathsf{H}} = M_j BB^{\mathsf{T}} M_j^{\mathsf{H}}.
\end{aligned}$$

Algorithm 4.2 Low-rank ADI (LRCF-ADI) iteration for Lyapunov equations

Input: A, B from (4.19), shifts $P = \{p_1, \ldots, p_{\text{maxiter}}\} \subset \mathbb{C}^-$, residual tolerance tol.

Output: Z_k such that $X = Z_k Z_k^{\mathsf{H}}$ (approx.) solves (4.19).

1: Initialize $j = 1$, $W_0 := B$, $Z_0 := [\,]$.
2: **while** $\|W_{j-1}\|_2 \geqslant \text{tol}$ **do**
3: Set $V_j := (A + p_j I_n)^{-1} W_{j-1}$.
4: Set $W_j := W_{j-1} - 2\operatorname{Re}(p_j) V_j$.
5: Set $Z_j := \begin{bmatrix} Z_{j-1} & \sqrt{-\operatorname{Re}(p_j)}\, V_j \end{bmatrix}$.
6: Set $j := j + 1$.
7: **end while**

Moreover, it holds

$$\begin{aligned} V_j &= T_{j-1} S_j V_{j-1} = T_{j-1} S_j T_{j-2} S_{j-1} V_{j-2} = \ldots = \\ &= S_j \left(\prod_{k=1}^{j-1} T_k S_k \right) B = S_j M_{j-1} B = (A + p_j I_n)^{-1} W_{j-1}, \end{aligned} \tag{4.20}$$

and

$$W_j = M_j B = S_j T_j W_{j-1} = W_{j-1} - 2\operatorname{Re}(p_j) S_j W_{j-1} = W_{j-1} - 2\operatorname{Re}(p_j) V_j.$$

□

Thanks to the above theorem, the norm of the Lyapunov residual can be cheaply computed via $\|R_j\|_2 = \|W_j W_j^{\mathsf{H}}\|_2 = \|W_j\|_2^2$. All this leads to Algorithm 4.2 which is also often referred to as low-rank Cholesky-factor ADI (LRCF-ADI) iteration. Algorithm 4.2 produces complex low-rank factors, if some of the shifts are complex, which might be required for problems with nonsymmetric A.

However, it is still possible to ensure that $Z_j \in \mathbb{R}^{n \times n_j}$, see [BKS13].

Definition 4.18: A set of shift parameters P is called proper if for all $p \in P$, also $\overline{p} \in P$.

Theorem 4.19: Assume $P = \{p_1, \ldots, p_k\}$ to be a set of proper shifts and assume w. l. o. g. that $p_{j+1} = \overline{p}_j \notin \mathbb{R}$. Then for V_j, V_{j+1} it holds

$$\begin{aligned} V_{j+1} &= \overline{V_j} + 2\beta_j \operatorname{Im}(V_j), \\ W_{j+1} &= W_{j-1} - 4\operatorname{Re}(p_j)\left(\operatorname{Re}(V_j) + \beta_j \operatorname{Im}(V_j)\right), \end{aligned}$$

with $\beta_j = \frac{\operatorname{Re}(p_j)}{\operatorname{Im}(p_j)}$.

Question: Why does that help? V_j, V_{j+1} are still complex. Consider

$$Z_{j+1} = \begin{bmatrix} Z_{j-1} & \alpha_j V_j & \alpha_j V_{j+1} \end{bmatrix}.$$

This gives

$$X_{j+1} = Z_{j+1} Z_{j+1}^{\mathsf{H}} = Z_{j-1} Z_{j-1}^{\mathsf{H}} + \alpha_j^2 \widehat{Z} \widehat{Z}^{\mathsf{H}},$$

with

$$\begin{aligned} \widehat{Z} &= \begin{bmatrix} V_j & V_{j+1} \end{bmatrix} = \begin{bmatrix} \operatorname{Re}(V_j) + \mathrm{i} \operatorname{Im}(V_j) & \operatorname{Re}(V_j) + 2\beta_j \operatorname{Im}(V_j) - \mathrm{i} \operatorname{Im}(V_j) \end{bmatrix} \\ &= \begin{bmatrix} \operatorname{Re}(V_j) & \operatorname{Im}(V_j) \end{bmatrix} \underbrace{\begin{bmatrix} I_m & I_m \\ \mathrm{i} I_m & (2\beta_j - \mathrm{i}) I_m \end{bmatrix}}_{=:N}. \end{aligned}$$

This yields

$$\widehat{Z} \widehat{Z}^{\mathsf{H}} = \begin{bmatrix} \operatorname{Re}(V_j) & \operatorname{Im}(V_j) \end{bmatrix} N N^{\mathsf{H}} \begin{bmatrix} \operatorname{Re}(V_j) & \operatorname{Im}(V_j) \end{bmatrix}^{\mathsf{H}}$$

and

$$\begin{aligned} 0 < N N^{\mathsf{H}} &= \begin{bmatrix} 2I_m & 2\beta_j I_m \\ 2\beta_j I_m & (4\beta_j^2 + 1) I_m \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} I_m & 0 \\ \beta_j I_m & I_m \end{bmatrix}}_{=:L} \underbrace{\begin{bmatrix} 2I_m & 0 \\ 0 & 2(\beta_j^2 + 1) I_m \end{bmatrix}}_{=:\Gamma > 0} \begin{bmatrix} I_m & \beta_j I_m \\ 0 & I_m \end{bmatrix}. \end{aligned}$$

Therefore, we can alternatively choose the following $\breve{Z}$ instead of $\widehat{Z}$ to obtain the same Z_{j+1}, namely

$$\begin{aligned} \breve{Z} &:= \begin{bmatrix} \operatorname{Re}(V_j) & \operatorname{Im}(V_j) \end{bmatrix} L \Gamma^{\frac{1}{2}} \\ &= \sqrt{2} \begin{bmatrix} \operatorname{Re}(V_j) + \beta_j \operatorname{Im}(V_j) & \sqrt{(\beta_j^2 + 1)} \operatorname{Im}(V_j) \end{bmatrix} \in \mathbb{R}^{n \times 2m}, \end{aligned}$$

in other words, Z_{j+1} is constructed to be real. This leads to the real version of the LRCF-ADI iteration.

4.5.5 Balanced Truncation Using the LRCF-ADI Method

Algorithm 4.1 can now be modified by including low-rank methods for solving the Lyapunov equations. The result is Algorithm 4.3.

Remark 4.20: The $\mathcal{H}_\infty$ error bound for balanced truncation does not necessarily hold anymore. First of all, we do not compute all Hankel singular values. Hence, the ones which have not been computed, can only be estimated using

Algorithm 4.3 Balanced truncation (pro version)

Input: Asymptotically stable system $[A, B, C, D] \in \Sigma_{n,m,p}$, desired maximum reduced order $r_{\max}$, ADI residual tolerance tol.

Output: Reduced-order model $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right] \in \Sigma_{r,m,p}$ with $r \leqslant r_{\max}$.

1: Solve the Lyapunov equations

$$AP + PA^{\mathsf{T}} = -BB^{\mathsf{T}}, \quad A^{\mathsf{T}}Q + QA = -C^{\mathsf{T}}C$$

using (the real version of) Algorithm 4.2 to determine two low-rank factors $R \in \mathbb{R}^{n \times r_P}$ and $L \in \mathbb{R}^{n \times r_Q}$ such that $P \approx RR^{\mathsf{T}}$ and $Q \approx LL^{\mathsf{T}}$ and with residuals less than tol.

2: Set $r := \min\{r_{\max}, r_P, r_Q\}$.

3: Compute the SVD of $L(:, 1:r)^{\mathsf{T}} R(:, 1:r) = U\Sigma V^{\mathsf{T}}$.

4: Set $T := RV\Sigma^{-\frac{1}{2}}$ and $W := LU\Sigma^{-\frac{1}{2}}$.

5: Balance and truncate to obtain the reduced-order model

$$\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right] := \left[W^{\mathsf{T}}AT, W^{\mathsf{T}}B, CT, D\right].$$

the smallest singular value of Σ in Step 3 of Algorithm 4.3. Moreover, the singular values contained in Σ may be corrupted by the approximation errors done when computing R and L. Therefore, the reduction error can only be estimated in practice.

CHAPTER 5

Passivity-Preserving Balancing-Based Model Reduction

In this chapter we will focus on some aspects of structure-preservation. In this chapter we consider *passive* systems which often appear in the modeling of electrical circuits, power network, mechanical systems, and many more. See Chapter 1 for some examples. Thus, when doing model reduction, one would like to obtain a passive reduced-order model in order to preserve the physical properties in the model. In this chapter, we will first define passivity and show that each passive system admits a positive real transfer function. Thereafter we will discuss a passivity-preserving model reduction scheme using alternative energy functionals. This will lead to the method of positive real balanced truncation which we will analyze afterwards. Since positive real balanced truncation relies on algebraic Riccati equations rather than Lyapunov equations, we will also treat the numerical solution of large-scale algebraic Riccati equations. Many of the results presented here can be found in the famous works by Jan C. Willems [Wil71, Wil72a, Wil72b].

5.1 Passivity and Positive Real Transfer Functions

First we define passivity for a LTI systems.

Definition 5.1: Let $[A, B, C, D] \in \Sigma_{n,m,m}$ be given. Then the system is called

passive, if

$$\int_0^T y(\tau)^\mathsf{T} u(\tau) \mathrm{d}\tau \geqslant 0 \tag{5.1}$$

holds for all $T \geqslant 0$ and all solution trajectories $(x, u, y) \in \mathcal{L}_2([0, T], \mathbb{R}^{n+2m})$ of the system with $x(0) = 0$.

The expression on the left-hand side of (5.1) can be understood as the energy that is supplied to the system in the time interval $[0, T]$. Therefore, passivity of a dynamical system is the property that for each solution, more energy has to be supplied than energy that can be extracted from the system. So the system cannot internally produce energy. Passivity is connected to two energy functionals:

a) the *(virtual) available storage* $V^- : \mathbb{R}^n \to \mathbb{R}$ with

$$V^-(x_0) := \sup \left\{ -\int_0^\infty 2y(\tau)^\mathsf{T} u(\tau) \mathrm{d}\tau \;\middle|\; (x, u, y) \in \mathcal{L}_2([0, \infty), \mathbb{R}^{n+2m}) \right.$$
$$\left. \text{is a solution of } [A, B, C, D] \text{ with } x(0) = x_0 \right\};$$

b) the *required supply* $V^+ : \mathbb{R}^n \to \mathbb{R}$ with

$$V^+(x_0) := \inf \left\{ \int_{-\infty}^0 2y(\tau)^\mathsf{T} u(\tau) \mathrm{d}\tau \;\middle|\; (x, u, y) \in \mathcal{L}_2((-\infty, 0], \mathbb{R}^{n+2m}) \right.$$
$$\left. \text{is a solution of } [A, B, C, D] \text{ with } x(0) = x_0 \right\}.$$

The value of $V^+(x_0)$ is the least amount of energy that has to be supplied to the system to reach the state x_0. On the other hand, the value of $V^-(x_0)$ is the maximum amount of energy that can be extracted from the system by stabilizing solution trajectories.

We will later see that under some conditions, the functionals V^+, V^- are so-called *storage functions*. A storage function is a function $V : \mathbb{R}^n \to \mathbb{R}^+$ with $V(0) = 0$ that fulfills the *dissipation inequality*

$$V(x_1) - V(x_0) \leqslant \int_{t_0}^{t_1} 2y(\tau)^\mathsf{T} u(\tau) \mathrm{d}\tau, \tag{5.2}$$

where $(x, u, y) \in \mathcal{L}_2([t_0, t_1], \mathbb{R}^{n+2m})$ is a solution trajectory of the system with $x(t_0) = x_0$ and $x(t_1) = x_1$. If V is differentiable, then the dissipation inequality can be formulated in its differential form

$$V'(x(t)) \cdot \dot{x}(t) \leqslant 2y(t)^\mathsf{T} u(t), \tag{5.3}$$

where $V' : \mathbb{R}^n \to \mathbb{R}^{1\times n}$ is the Jacobian of V.

Now we consider the special case of *quadratic* storage functions, i. e., $V(x) = x^\mathsf{T} P x$ for some symmetric positive semi-definite matrix $P \in \mathbb{R}^{n\times n}$. For such, $V'(x) = 2x^\mathsf{T} P$ and (5.3) gives

$$\begin{aligned} V'(x(t)) \cdot \dot{x}(t) &= 2x(t)^\mathsf{T} P \dot{x}(t) \\ &= x(t)^\mathsf{T} P(Ax(t) + Bu(t)) + (Ax(t) + Bu(t))^\mathsf{T} P x(t) \\ &= \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^\mathsf{T} \begin{bmatrix} A^\mathsf{T} P + PA & PB \\ B^\mathsf{T} P & 0 \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \\ &\leqslant 2y(t)^\mathsf{T} u(t) \\ &= (Cx(t) + Du(t))^\mathsf{T} u(t) + u(t)^\mathsf{T}(Cx(t) + Du(t)) \\ &= \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}^\mathsf{T} \begin{bmatrix} 0 & C^\mathsf{T} \\ C & D + D^\mathsf{T} \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}. \end{aligned}$$

Therefore, each quadratic storage function can be expressed by a solution $P \geqslant 0$ of the *linear matrix inequality (LMI)*

$$\begin{bmatrix} A^\mathsf{T} P + PA & PB - C^\mathsf{T} \\ B^\mathsf{T} P - C & -D - D^\mathsf{T} \end{bmatrix} \leqslant 0, \quad P = P^\mathsf{T}. \tag{5.4}$$

It can be shown that that if the system is controllable, then the LMI (5.4) has two extremal solutions $P^+ \in \mathbb{R}^{n\times n}$ and $P^- \in \mathbb{R}^{n\times n}$ such that $P^- \leqslant P \leqslant P^+$ for each solution $P \in \mathbb{R}^{n\times n}$ of the LMI and with the properties

$$V^-(x_0) = x_0^\mathsf{T} P^- x_0, \quad V^+(x_0) = x_0^\mathsf{T} P^+ x_0.$$

We have the following theorem connecting all these concepts.

Theorem 5.2: Let $[A, B, C, D] \in \Sigma_{n,m,m}$ be controllable. Then the following statements are equivalent:

a) The system $[A, B, C, D]$ is passive.

b) It holds that $V^+(x_0) \geqslant 0$ for all $x_0 \in \mathbb{R}^n$.

c) There exists a function $V : \mathbb{R}^n \to \mathbb{R}^+$ that satisfies the dissipation inequality (5.2).

Moreover, whenever one of the above conditions is fulfilled, then we have

$$-\infty < V^-(x_0) \leqslant V^+(x_0) < \infty.$$

Proof. We show "a) ⇒ b)": Assume that b) is not satisfied, i. e., there exist an $\varepsilon > 0$ and a solution trajectory $(x, u, y) \in \mathcal{L}_2((-\infty, 0], \mathbb{R}^{n+2m})$ with $x(0) = x_0$

(and $\lim_{t\to-\infty} x(t) = 0$) such that

$$\int_{-\infty}^{0} 2y(\tau)^{\mathsf{T}} u(\tau)\mathrm{d}\tau < -\varepsilon.$$

It can be shown that for each $\varepsilon > 0$ there exist a $T > 0$ and a solution trajectory $(\widetilde{x}, \widetilde{u}, \widetilde{y}) \in \mathcal{L}_2((-\infty, 0], \mathbb{R}^{n+2m})$ with compact support in the interval $[-T, 0]$ and with $\widetilde{x}(0) = x_0$ such that

$$\left| \int_{-\infty}^{0} 2y(\tau)^{\mathsf{T}} u(\tau)\mathrm{d}\tau - \int_{-\infty}^{0} 2\widetilde{y}(\tau)^{\mathsf{T}} \widetilde{u}(\tau)\mathrm{d}\tau \right| < \frac{\varepsilon}{2}.$$

This gives

$$\begin{aligned} \int_{-\infty}^{0} 2\widetilde{y}(\tau)^{\mathsf{T}} \widetilde{u}(\tau)\mathrm{d}\tau &= \int_{-T}^{0} 2\widetilde{y}(\tau)^{\mathsf{T}} \widetilde{u}(\tau)\mathrm{d}\tau \\ &= \int_{0}^{T} 2\widetilde{y}(\tau - T)^{\mathsf{T}} \widetilde{u}(\tau - T)\mathrm{d}\tau < -\frac{\varepsilon}{2}. \end{aligned}$$

Therefore, condition a) is violated for the solution trajectory $(\widetilde{x}(\cdot - T), \widetilde{u}(\cdot - T), \widetilde{y}(\cdot - T)) \in \mathcal{L}_2([0, T], \mathbb{R}^{n+2m})$.

The statement "b) $\Rightarrow$ c)" follows from the fact that V^+ is a storage function.

Now we show "c) $\Rightarrow$ a)": From the dissipation inequality with $t_0 = 0$, $x(t_0) = 0$, $t_1 = T$, and the condition $V(0) = 0$ we obtain

$$0 \leqslant V(x_1) \leqslant \int_{0}^{T} 2y(\tau)^{\mathsf{T}} u(\tau)\mathrm{d}\tau,$$

which gives the result.

The last inequality follows from b) since for all $T > 0$ and all solution trajectories $(x, u, y) \in \mathcal{L}_2((-\infty, T], \mathbb{R}^{n+2m})$ with $x(0) = x_0$ we obtain

$$-\int_{0}^{T} 2y(\tau)^{\mathsf{T}} u(\tau)\mathrm{d}\tau \leqslant \int_{-\infty}^{0} 2y(\tau)^{\mathsf{T}} u(\tau)\mathrm{d}\tau.$$

With $T \to \infty$, taking the supremum on the left-hand side 9+ and the infimum on the right-hand side gives $V^-(x_0) \leqslant V^+(x_0)$ for all $x_0 \in \mathbb{R}^n$. The finiteness of both functionals then follows from controllability, since every point x_0 can be reached by a solution trajectory. □

The passivity property is equivalent to a structural property of its transfer function, namely, they are positive real.

Definition 5.3: Let $[A, B, C, D] \in \Sigma_{n,m,m}$ be given with the transfer function $G(s) \in \mathbb{R}(s)^{m\times m}$. Then $G(s)$ is called *positive real*, if $G(s)$ has no poles in $\mathbb{C}^+$ and

$$\Psi(\lambda) := G(\lambda) + G(\lambda)^{\mathsf{H}} \geqslant 0 \quad \forall\, \lambda \in \mathbb{C}^+. \tag{5.5}$$

The following famous theorem (called the positive real lemma, sometimes also the Kalman-Yakubovich-Popov(-Anderson) lemma) makes a connection between solvability of the LMI (5.4) and positive realness.

Theorem 5.4: Let $[A, B, C, D] \in \Sigma_{n,m,m}$ be given with the transfer function $G(s) \in \mathbb{R}(s)^{m\times m}$ and let Ψ be as in (5.5). Then the following statements are satisfied:

a) If the LMI (5.4) has a solution $P > 0$, then $G(s)$ is positive real.

b) If the system $[A, B, C, D]$ is minimal and $G(s)$ is positive real, then there exists a solution $P > 0$ of the LMI (5.4).

Proof. We prove statement a): Let $P > 0$ be a solution of the LMI (5.4). Let $v \in \mathbb{C}^n$ be an eigenvector corresponding to an eigenvalue $\lambda \in \mathbb{C}$ of A. Then we have

$$v^{\mathsf{H}}A^{\mathsf{T}}Pv + v^{\mathsf{H}}PAv = \overline{\lambda}v^{\mathsf{H}}Pv + \lambda v^{\mathsf{H}}Pv = 2\operatorname{Re}(\lambda)\underbrace{v^{\mathsf{H}}Pv}_{>0} \leqslant 0.$$

Therefore, we have $\operatorname{Re}(\lambda) \leqslant 0$ and thus, $G(s)$ has no poles in $\mathbb{C}^+$.

Moreover, we have that

$$\begin{aligned}\begin{bmatrix} A & B \end{bmatrix}\begin{bmatrix}(\lambda I_n - A)^{-1}B \\ I_m\end{bmatrix} &= A(\lambda I_n - A)^{-1}B + B \\ &= (A + \lambda I_n - A)(\lambda I_n - A)^{-1}B \\ &= \lambda(\lambda I_n - A)^{-1}B.\end{aligned}$$

With $\lambda \in \mathbb{C}^+$ and using the above identity we obtain

$$\begin{aligned}&\begin{bmatrix}(\lambda I_n - A)^{-1}B \\ I_m\end{bmatrix}^{\mathsf{H}}\begin{bmatrix}A^{\mathsf{T}}P + PA & PB \\ B^{\mathsf{T}}P & 0\end{bmatrix}\begin{bmatrix}(\lambda I_n - A)^{-1}B \\ I_m\end{bmatrix} \\ =&\begin{bmatrix}(\lambda I_n - A)^{-1}B \\ I_m\end{bmatrix}^{\mathsf{H}}\left(\begin{bmatrix}PA & PB \\ 0 & 0\end{bmatrix} + \begin{bmatrix}A^{\mathsf{T}}P & 0 \\ B^{\mathsf{T}}P & 0\end{bmatrix}\right)\begin{bmatrix}(\lambda I_n - A)^{-1}B \\ I_m\end{bmatrix} \\ =&(\lambda + \overline{\lambda})B^{\mathsf{T}}(\lambda I_n - A)^{-\mathsf{H}}P(\lambda I_n - A)^{-1}B \geqslant 0.\end{aligned}$$

Now we get

$$\begin{aligned}\Psi(\lambda) &= \begin{bmatrix}(\lambda I_n - A)^{-1}B \\ I_m\end{bmatrix}^{\mathsf{H}} \begin{bmatrix}0 & C^{\mathsf{T}} \\ C & D + D^{\mathsf{T}}\end{bmatrix} \begin{bmatrix}(\lambda I_n - A)^{-1}B \\ I_m\end{bmatrix} \\ &\geqslant \begin{bmatrix}(\lambda I_n - A)^{-1}B \\ I_m\end{bmatrix}^{\mathsf{H}} \begin{bmatrix}-A^{\mathsf{T}}P - PA & -PB + C^{\mathsf{T}} \\ -B^{\mathsf{T}}P + C & D + D^{\mathsf{T}}\end{bmatrix} \begin{bmatrix}(\lambda I_n - A)^{-1}B \\ I_m\end{bmatrix} \\ &\geqslant 0,\end{aligned}$$

where the latter inequality follows from the fact, that P solves (5.4).

The proof of statement b) is quite complicated and technical and therefore, we omit it here. The proof can be found in [Ran96]. □

Remark 5.5: There are many relaxations of the assumptions of the positive real lemma. For instance, it can be shown that $G(s)$ is already positive real, if there exists a solution $P \geqslant 0$ of the LMI (5.4). However, the techniques for the proof get more involved, see, e. g., [AV73].

From the above theorem, the following corollary is immediate.

Corollary 5.6: The system $[A, B, C, D] \in \Sigma_{n,m,p}$ is passive if and only if its transfer function $G(s)$ is positive real.

Proof. Let $[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}] \in \Sigma_{\tilde{n},m,p}$ be a minimal realization of $G(s)$. Then, $[A, B, C, D]$ is passive if and only if $[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}]$ is passive, since both generate the same input/output pairs. From Theorem 5.2 and controllability, this is equivalent to the existence of a matrix $\widetilde{P} \geqslant 0$ such that the LMI

$$\begin{bmatrix}\widetilde{A}^{\mathsf{T}}\widetilde{P} + \widetilde{P}\widetilde{A} & \widetilde{P}\widetilde{B} - \widetilde{C}^{\mathsf{T}} \\ \widetilde{B}^{\mathsf{T}}\widetilde{P} - \widetilde{C} & -\widetilde{D} - \widetilde{D}^{\mathsf{T}}\end{bmatrix} \leqslant 0, \quad \widetilde{P} = \widetilde{P}^{\mathsf{T}}$$

is satisfied. From Theorem 5.4 and Remark 5.5, this is equivalent to positive realness of $G(s)$. □

5.2 Positive Real Balanced Truncation

Now we want to derive a balancing-type algorithm for passivity-preserving model reduction. Assume that $P \geqslant 0$ is a solution of the LMI (5.4). Then there exist matrices $K \in \mathbb{R}^{q\times n}$ and $L \in \mathbb{R}^{q\times m}$ such that

$$\begin{bmatrix}A^{\mathsf{T}}P + PA & PB - C^{\mathsf{T}} \\ B^{\mathsf{T}}P - C & -D - D^{\mathsf{T}}\end{bmatrix} = -\begin{bmatrix}K^{\mathsf{T}} \\ L^{\mathsf{T}}\end{bmatrix}\begin{bmatrix}K & L\end{bmatrix}.$$

Under the assumption that $D + D^\mathsf{T}$ is invertible (then $q \geqslant m$), we can apply the Schur complement on both sides and obtain

$$A^\mathsf{T}P + PA + \big(PB - C^\mathsf{T}\big)\big(D + D^\mathsf{T}\big)^{-1}\big(B^\mathsf{T}P - C\big) \\ = -K^\mathsf{T}K + K^\mathsf{T}L\big(L^\mathsf{T}L\big)^{-1}L^\mathsf{T}K.$$

Let $L = U\left[\begin{smallmatrix}\Sigma\\0\end{smallmatrix}\right]V^\mathsf{T}$ with orthogonal matrices $U \in \mathbb{R}^{q\times q}$, $V \in \mathbb{R}^{m\times m}$ and an invertible diagonal matrix $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_m)$ with $\sigma_1 \geqslant \ldots \geqslant \sigma_m > 0$ be given. Then we have

$$\begin{aligned} &- K^\mathsf{T}K + K^\mathsf{T}L\big(L^\mathsf{T}L\big)^{-1}L^\mathsf{T}K \\ &= -K^\mathsf{T}K + K^\mathsf{T}U\begin{bmatrix}\Sigma\\0\end{bmatrix}V^\mathsf{T}\left(V\begin{bmatrix}\Sigma & 0\end{bmatrix}U^\mathsf{T}U\begin{bmatrix}\Sigma\\0\end{bmatrix}V^\mathsf{T}\right)^{-1}V\begin{bmatrix}\Sigma & 0\end{bmatrix}U^\mathsf{T}K \\ &\qquad - K^\mathsf{T}K + K^\mathsf{T}U\begin{bmatrix}I_m & 0\\0 & 0\end{bmatrix}U^\mathsf{T}K \leqslant 0. \end{aligned}$$

Therefore, each solution P of the LMI (5.4) satisfies the *algebraic Riccati inequality*

$$A^\mathsf{T}P + PA + \big(PB - C^\mathsf{T}\big)\big(D + D^\mathsf{T}\big)^{-1}\big(B^\mathsf{T}P - C\big) \leqslant 0, \quad P = P^\mathsf{T}.$$

It can be seen that the extremal elements of its solution set, P^+ and P^-, even satisfy the *algebraic Riccati equation (ARE)*

$$A^\mathsf{T}P + PA + \big(PB - C^\mathsf{T}\big)\big(D + D^\mathsf{T}\big)^{-1}\big(B^\mathsf{T}P - C\big) = 0, \quad P = P^\mathsf{T}. \tag{5.6}$$

Recall that if the system $[A, B, C, D]$ is controllable and passive, then the extremal solutions P^+ and P^- exist and $P^+ \geqslant 0$. It can be shown (see [Obe91, Sect. 6]) that if the system is also observable, then we even have

$$0 < P^- \leqslant P^+. \tag{5.7}$$

Furthermore, it can be shown (exercise!) that if $P > 0$ is a solution of the ARE, then $Q = P^{-1}$ is a solution of the *dual ARE*

$$AQ + QA^\mathsf{T} + \big(QC^\mathsf{T} - B\big)\big(D + D^\mathsf{T}\big)^{-1}\big(CQ - B^\mathsf{T}\big) = 0, \quad Q = Q^\mathsf{T}. \tag{5.8}$$

Therefore, if we have (5.7), then there exist a minimal solution $Q^- \in \mathbb{R}^{n\times n}$ and a maximal solution $Q^+ \in \mathbb{R}^{n\times n}$ with

$$0 < Q^- \leqslant Q \leqslant Q^+ \quad \text{for all solutions } Q \text{ of (5.8)},$$

and with $Q^- = (P^+)^{-1}$ and $Q^+ = (P^-)^{-1}$. The minimal solutions P^- and Q^- are called the *positive real (observability and controllability) Gramians* and they are now subject to our balancing procedure. These will attain the role of the controllability and observability Gramian from the previous chapter.

Definition 5.7: Let $[A, B, C, D] \in \Sigma_{n,m,m}$ be minimal and passive with positive real Gramians $P^- > 0$ and $Q^- > 0$. Then the system is called *positive real balanced*, if $P^- = Q^- = \Sigma$. In this case, the eigenvalues of the matrix Σ are called the *positive real characteristic values*.

Now we discuss the associated balancing transformations.

Theorem 5.8: Let $[A, B, C, D] \in \Sigma_{n,m,m}$ be minimal and passive with positive real Gramians $P^- > 0$ and $Q^- > 0$. Then there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$ such that $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right] := [T^{-1}AT, T^{-1}B, CT, D]$ with the positive real Gramians $\widetilde{P}^- > 0$ and $\widetilde{Q}^- > 0$ is positive real balanced.

Proof. Since $P^- > 0$ and $Q^- > 0$, there exist Cholesky decompositions $P^- = RR^\mathsf{T}$ and $Q^- = LL^\mathsf{T}$, where R and L are lower triangular and invertible. Now consider the singular value decomposition $L^\mathsf{T}R = U\Sigma V^\mathsf{T}$ with orthogonal $U, V \in \mathbb{R}^{n\times n}$ and $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_n \geqslant 0$. Since L and R are invertible, so is $L^\mathsf{T}R$ and therefore, we have $\sigma_n > 0$.

It is easy to check that as for the case of standard balancing, we have that

$$\widetilde{P}^- = T^\mathsf{T}P^-T, \quad \widetilde{Q}^- = T^{-1}Q^-T^{-\mathsf{T}}.$$

Now the rest of the proof is similar to the proof of Theorem 4.5. With $T := LU\Sigma^{-\frac{1}{2}}$ we find $T^{-1} = \Sigma^{-\frac{1}{2}}V^\mathsf{T}R^\mathsf{T}$ which make the system positive real balanced (exercise!). □

This leads to Algorithm 5.1 for model reduction that is called *positive real balanced truncation* and which has first been considered in [Obe91, Sect. 6].

Remark 5.9: a) If the passive system $[A, B, C, D] \in \Sigma_{n,m,m}$ is not minimal, then the algebraic Riccati equations (5.6) and (5.8) may not have (minimal) solutions. However, for many problems the existence of minimal solutions can be derived from the structure of the models, such as for electrical circuit models.

b) If the matrix $D + D^\mathsf{T}$ is not invertible, then the algebraic Riccati equations cannot be formed. In this case, one has to resort to *Lur'e equations* such as

$$\begin{aligned} A^\mathsf{T}P + PA &= -K^\mathsf{T}K, \quad P = P^\mathsf{T}, \\ PB - C^\mathsf{T} &= -K^\mathsf{T}L, \\ D + D^\mathsf{T} &= L^\mathsf{T}L, \end{aligned} \tag{5.9}$$

Algorithm 5.1 Positive real balanced truncation (basic version)

Input: Minimal and passive system $[A, B, C, D] \in \Sigma_{n,m,m}$ with invertible $D + D^{\mathsf{T}}$, desired reduced order r.

Output: Passive reduced-order model $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,m}$.

1: Compute the minimal (and positive definite) solutions P^- and Q^- of the algebraic Riccati equations

$$A^{\mathsf{T}}P + PA + \left(PB - C^{\mathsf{T}}\right)\left(D + D^{\mathsf{T}}\right)^{-1}\left(B^{\mathsf{T}}P - C\right) = 0, \quad P = P^{\mathsf{T}},$$
$$AQ + QA^{\mathsf{T}} + \left(QC^{\mathsf{T}} - B\right)\left(D + D^{\mathsf{T}}\right)^{-1}\left(CQ - B^{\mathsf{T}}\right) = 0, \quad Q = Q^{\mathsf{T}}.$$

2: Compute Cholesky factorization $P^- = RR^{\mathsf{T}}$ and $Q^- = LL^{\mathsf{T}}$.

3: Compute the singular value decomposition $L^{\mathsf{T}}R = U\Sigma V^{\mathsf{T}}$.

4: Set $T := LU\Sigma^{-\frac{1}{2}}$ (and $T^{-1} = \Sigma^{-\frac{1}{2}}V^{\mathsf{T}}R^{\mathsf{T}}$).

5: Do the balancing transformation

$$\left[T^{-1}AT, T^{-1}B, CT, D\right] = \left[\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \begin{bmatrix} C_1 & C_2 \end{bmatrix}, D\right]$$

and set the reduced-order model as $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,m}$.

which has to be solved for the triple $(P, K, L) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{q\times n} \times \mathbb{R}^{q\times m}$, where q is as small as possible among all such triples solving (5.9). This minimal rank property is motivated by the fact, that the algebraic Riccati inequality turns to an equation, if and only if $L \in \mathbb{R}^{q\times m}$ is invertible, i. e., $q = m$. This is the smallest possible rank, since $D + D^{\mathsf{T}}$ was assumed to be invertible in this case. In the general case, the solutions of the LMI which have this minimal rank property are called *rank-minimizing solutions*. It can be shown that this minimal rank is

$$q = \operatorname{rank}_{\mathbb{R}(s)}(G(s) + G^{\sim}(s)).$$

Extremal solutions of (5.9) are always rank-minimizing.

5.3 Analysis of the Method

In this section, we give a brief analysis of the properties of positive real balanced truncation (mainly without the proofs). First of all, we see that the reduced-order model is again passive.

Theorem 5.10: Let $[A, B, C, D] \in \Sigma_{n,m,m}$ with invertible $D + D^\mathsf{T}$ be asymptotically stable, minimal, and passive. Apply Algorithm 5.1 to obtain the reduced-order model $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,m}$. Assume further, that for the positive real characteristic values $\sigma_1, \sigma_2, \ldots, \sigma_n > 0$ sorted in decreasing order, it holds that $\sigma_r > \sigma_{r+1}$. Then the reduced-order model $[A_{11}, B_1, C_1, D]$ is asymptotically stable, minimal, passive, and positive real balanced with the positive real Gramians $P_{11} = Q_{11} = \Sigma_1 := \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$.

Proof. The preservation of passivity is easy to see, since it holds that $\Sigma_1 > 0$ solves the reduced ARE

$$A_{11}^\mathsf{T}\tilde{P} + \tilde{P}A_{11} + \left(\tilde{P}B_1 - C_1^\mathsf{T}\right)\left(D + D^\mathsf{T}\right)^{-1}\left(B_1^\mathsf{T}\tilde{P} - C_1\right) = 0, \quad \tilde{P} = \tilde{P}^\mathsf{T}.$$

Therefore, the corresponding LMI (5.4) has a positive definite solution and thus by Theorem 5.4, the reduced transfer function $G(s) = C_1(sI_r - A_{11})^{-1}B_1 + D$ is positive real. By Corollary 5.6, the reduced-order model is passive. The proof of asymptotic stability and minimality is quite involved and therefore, it is omitted here. □

Next we want to address error bounds. In contrast to standard balancing, there are no a priori error bounds in the $\mathcal{H}_\infty$-norm, even if the original and the reduced-order model both have transfer functions in $\mathcal{RH}_\infty^{m\times m}$. Instead, one has to resort to the so-called *gap metric*. The following has been taken from [GO13].

Definition 5.11: Let $\mathcal{V}_1$ and $\mathcal{V}_2$ be two closed subspaces ("closed" means that any sequence of elements of the subspace has its limit in this subspace) of a Hilbert space $\mathcal{H}$ with induced norm $\|\cdot\|_\mathcal{H}$. Then the *gap* between $\mathcal{V}_1$ and $\mathcal{V}_2$ is defined by

$$g(\mathcal{V}_1, \mathcal{V}_2) := \|\Pi_{\mathcal{V}_1} - \Pi_{\mathcal{V}_2}\|_{\mathcal{L}(\mathcal{H},\mathcal{H})},$$

where $\Pi_{\mathcal{V}_1}, \Pi_{\mathcal{V}_2} : \mathcal{H} \to \mathcal{H}$ denote the orthogonal projections onto the spaces $\mathcal{V}_1, \mathcal{V}_2$ (which exist by the closedness assumption).

It can be shown that g makes the set of all closed subspaces of $\mathcal{H}$ to a (complete) metric space. Moreover, it can be shown that

$$g(\mathcal{V}_1, \mathcal{V}_2) = \max\left\{\vec{g}(\mathcal{V}_1, \mathcal{V}_2),\ \vec{g}(\mathcal{V}_2, \mathcal{V}_1)\right\},$$

where

$$\vec{g}(\mathcal{V}_1, \mathcal{V}_2) = \|(I - \Pi_{\mathcal{V}_2})\Pi_{\mathcal{V}_1}\|_{\mathcal{L}(\mathcal{H},\mathcal{H})} = \sup_{v\in\mathcal{V}_1,\, \|v\|_\mathcal{H}=1} \operatorname{dist}(v, \mathcal{V}_2).$$

is the *directed gap*. Now we apply these concepts to spaces related to linear systems.

Definition 5.12: Let $\Sigma := [A, B, C, D] \in \Sigma_{n,m,p}$ be a given asymptotically stable system with transfer function $G \in \mathcal{RH}_\infty^{p\times m}$. Let $u \in \mathcal{L}_2([0,\infty), \mathbb{R}^m)$ be given and define $U \in \mathcal{H}_2^m$ by $U(s) := \mathcal{L}\{u\}(s)$. Furthermore, define the *multiplication operator*

$$M_G : \mathcal{H}_2^m \to \mathcal{H}_2^p, \quad (M_G U)(s) = G(s)U(s) \quad \forall\, s \in \mathbb{C}^+.$$

The *graph* of the dynamical system Σ is then defined by

$$\mathcal{G}(\Sigma) := \left\{ \begin{bmatrix} I_m \\ M_G \end{bmatrix} U \;\middle|\; U \in \mathcal{H}_2^m \right\} = \operatorname{im} \begin{bmatrix} I_m \\ M_G \end{bmatrix},$$

which is a closed subspace of the Hilbert space $\mathcal{H}_2^{p+m}$.

The *gap metric* between two asymptotically stable systems Σ_1, Σ_2 is defined by

$$\delta(\Sigma_1, \Sigma_2) := g(\mathcal{G}(\Sigma_1), \mathcal{G}(\Sigma_2)).$$

The gap metric can be interpreted as the distance of the subspaces of input/output trajectories generated by two dynamical systems in frequency domain. With the gap metric, we can now obtain the following error bounds.

Theorem 5.13: Let $\Sigma := [A, B, C, D] \in \Sigma_{n,m,m}$ be an asymptotically stable, minimal, and passive system with transfer function $G \in \mathcal{RH}_\infty^{m\times m}$ and let $\widetilde{\Sigma} := [A_{11}, B_1, C_1, D] \in \Sigma_{r,m,m}$ be the reduced-order model obtained by positive real balanced truncation with transfer function $\widetilde{G} \in \mathcal{RH}_\infty^{m\times m}$. Assume further, that for the positive real characteristic values $\sigma_1, \sigma_2, \ldots, \sigma_n > 0$ sorted in decreasing order, it holds that $\sigma_r > \sigma_{r+1}$. Then we have the (a priori) gap metric error bound

$$\delta\big(\Sigma, \widetilde{\Sigma}\big) \leqslant \sum_{j=r+1}^{n} \sigma_j.$$

Moreover, there is the (a posteriori) $\mathcal{H}_\infty$ error bound

$$\begin{aligned} \big\|G - \widetilde{G}\big\|_{\mathcal{H}_\infty} \leqslant 2 \min \Big\{ &\Big(1 + \|G\|_{\mathcal{H}_\infty}^2\Big)\Big(1 + \big\|\widetilde{G}\big\|_{\mathcal{H}_\infty}\Big), \\ &\big(1 + \|G\|_{\mathcal{H}_\infty}\big)\Big(1 + \big\|\widetilde{G}\big\|_{\mathcal{H}_\infty}^2\Big)\Big\} \sum_{j=r+1}^{n} \sigma_j. \end{aligned}$$

Proof. Omitted. □

Remark 5.14: a) The $\mathcal{H}_\infty$ error bound is only an a posteriori error bound, since it requires the knowledge of the reduced-order model. Therefore, its use in practice is limited.

b) The gap metric can be expressed by the *normalized coprime factorizations* of $G(s)$ and $\widetilde{G}(s)$, see [Geo88]. A (right) normalized coprime factorization is given by $G(s) = N(s)M(s)^{-1}$, where $\left[\begin{smallmatrix} M \\ N \end{smallmatrix}\right] \in \mathcal{RH}_\infty^{(p+m)\times m}$ and there exist $Y \in \mathcal{RH}_\infty^{m\times m}$, $Z \in \mathcal{RH}_\infty^{m\times p}$ such that the *Bézout identity*

$$Y(s)M(s) + Z(s)N(s) = I_m$$

with the normalization condition

$$M^\sim(s)M(s) + N^\sim(s)N(s) = I_m$$

is satisfied. If $G(s) = N(s)M(s)^{-1}$ and $\widetilde{G}(s) = \widetilde{N}(s)\widetilde{M}(s)^{-1}$ are normalized coprime factorizations of $G(s)$ and $\widetilde{G}(s)$, respectively, then for the directed gap we have

$$\vec{g}\left(\mathcal{G}(\Sigma), \mathcal{G}\left(\widetilde{\Sigma}\right)\right) = \inf_{H\in\mathcal{H}_\infty^{m\times m}} \left\| \begin{bmatrix} M \\ N \end{bmatrix} - \begin{bmatrix} \widetilde{M} \\ \widetilde{N} \end{bmatrix} H \right\|_{\mathcal{H}_\infty}.$$

From this, some bounds for the gap metric can be derived. Efficient methods for its computation however, seem to be widely unexplored, except for [Geo88].

c) The gap metric error bound can also be expressed by the gap of subspaces of $\mathcal{L}_2([0,\infty), \mathbb{R}^{p+m})$ in the time domain. This analysis makes use of the *behavior approach of systems theory* which was developed by Jan C. Willems in the early 90s.

5.4 Numerical Solution of Large-Scale Algebraic Riccati Equations

5.4.1 Newton's Method for Solving Algebraic Riccati Equations

In this section we discuss the numerical solution of algebraic Riccati equations of the form

$$A^\mathsf{T}P + PA + \left(PB - C^\mathsf{T}\right)\left(D + D^\mathsf{T}\right)^{-1}\left(B^\mathsf{T}P - C\right) = 0, \quad P = P^\mathsf{T}.$$

This ARE can be rewritten as

$$\begin{aligned}\mathcal{R}(P) &:= F + \widehat{A}^{\mathsf{T}}P + P\widehat{A} + PGP = 0, \quad P = P^{\mathsf{T}}, \quad \text{where}\\ \widehat{A} &:= A - B\big(D + D^{\mathsf{T}}\big)^{-1}C,\\ F &:= C^{\mathsf{T}}\big(D + D^{\mathsf{T}}\big)^{-1}C \geqslant 0,\\ G &:= B\big(D + D^{\mathsf{T}}\big)^{-1}B^{\mathsf{T}} \geqslant 0.\end{aligned} \tag{5.10}$$

Remember that we want to compute the minimal solution P^- of (5.10). Under the assumption that the pair $(A,\, B)$ is stabilizable (equivalently, the pair $\big(\widehat{A},\, B\big)$ is stabilizable), it can even be shown that P^- is the unique *stabilizing solution* of (5.10), that is

$$\Lambda\big(\widehat{A} + GP^-\big) = \Lambda\left(A - B\big(D + D^{\mathsf{T}}\big)^{-1}\big(C - B^{\mathsf{T}}P^-\big)\right) \subset \mathbb{C}^-.$$

We consider (5.10) as a nonlinear system of equations and apply Newton's method which has first been considered in [Kle68]. For this, we need to evaluate the (Fréchet) derivative of $\mathcal{R}(P)$ with respect to P.

Definition 5.15 (Fréchet differentiability, Fréchet derivative)**:** Let $(\mathcal{X},\, \|\cdot\|_{\mathcal{X}})$ and $(\mathcal{Y},\, \|\cdot\|_{\mathcal{Y}})$ be two normed linear spaces and let $\mathcal{U} \subset \mathcal{X}$ be an open subset. An operator $\mathcal{F} : \mathcal{U} \to \mathcal{Y}$ is called *Fréchet differentiable* at $X \in \mathcal{U}$ if there exists a bounded linear operator $\mathcal{F}'(X) : \mathcal{X} \to \mathcal{Y}$ such that

$$\lim_{\|N\|_{\mathcal{X}} \to 0} \frac{1}{\|N\|_{\mathcal{X}}} \big\|\mathcal{F}(X+N) - \mathcal{F}(X) - (\mathcal{F}'(X))(N)\big\|_{\mathcal{Y}} = 0.$$

The operator $\mathcal{F}'(X)$ is called *Fréchet derivative* of $\mathcal{F}$ at X. The map $\mathcal{F}' : \mathcal{U} \to \mathcal{L}(\mathcal{X}, \mathcal{Y})$ with $X \mapsto \mathcal{F}'(X)$ is called *Fréchet derivative* of $\mathcal{F}$ on $\mathcal{U}$.

Let us see whether $\mathcal{R}(\cdot)$ is Fréchet differentiable and (if yes) determine its Fréchet derivative. If the Fréchet derivative exists it is given by

$$\begin{aligned}(\mathcal{R}'(P))(N) &= \lim_{h\to 0} \frac{1}{h}(\mathcal{R}(P+hN) - \mathcal{R}(P))\\ &= \lim_{h\to 0} \frac{1}{h}\Big(F + \widehat{A}^{\mathsf{T}}(P+hN) + (P+hN)\widehat{A}\\ &\qquad +(P+hN)G(P+hN) - \big(F + \widehat{A}^{\mathsf{T}}P + P\widehat{A} + PGP\big)\Big)\\ &= \lim_{h\to 0} \frac{1}{h}\big(h\widehat{A}^{\mathsf{T}}N + hN\widehat{A} + hPGN + hNGP + h^2NGN\big)\\ &= \lim_{h\to 0} \big(\widehat{A}^{\mathsf{T}}N + N\widehat{A} + PGN + NGP + hNGN\big)\\ &= \widehat{A}^{\mathsf{T}}N + N\widehat{A} + PGN + NGP\\ &= \big(\widehat{A} + GP\big)^{\mathsf{T}}N + N\big(\widehat{A} + GP\big).\end{aligned}$$

Algorithm 5.2 Newton's method for the algebraic Riccati equation

Input: $\widehat{A}$, F, G as in (5.10) and initial value P_0 such that $\Lambda\big(\widehat{A} + GP_0\big) \subset \mathbb{C}^-$.
Output: Stabilizing (and minimal) solution P^- solving (5.10).

1: **for** $j = 1, 2, \ldots$ **do**
2: Set $\widehat{A}_j := \widehat{A} + GP_{j-1}$.
3: Solve $\widehat{A}_j^{\mathsf{T}} N_{j-1} + N_{j-1}\widehat{A}_j = -\mathcal{R}(P_{j-1})$ for N_{j-1}.
4: Set $P_j := P_{j-1} + N_{j-1}$.
5: **end for**
6: Set $P^- := P_j$.

In other words, the Fréchet derivative of a Riccati operator is a Lyapunov operator. Now the Newton iteration is given by

$$(\mathcal{R}'(P_{j-1}))(N_{j-1}) = -\mathcal{R}(P_{j-1}), \quad P_j = P_{j-1} + N_{j-1}, \quad j = 1, 2, \ldots$$

and the iteration is summarized in Algorithm 5.2. This formulation of the algorithm has the disadvantage that $\mathcal{R}(P_{j-1})$ is evaluated in every iteration. Therefore, let us revisit the computation of the update N_{j-1}. We know that

$$\begin{aligned}\big(\widehat{A} + GP_{j-1}\big)^{\mathsf{T}} N_{j-1} + N_{j-1}\big(\widehat{A} + GP_{j-1}\big) \\ = -F - \widehat{A}^{\mathsf{T}} P_{j-1} - P_{j-1}\widehat{A} - P_{j-1}GP_{j-1}.\end{aligned} \tag{5.11}$$

Plugging in $N_{j-1} = P_j - P_{j-1}$ then gives

$$\begin{aligned}\big(\widehat{A} + GP_{j-1}\big)^{\mathsf{T}} (P_j - P_{j-1}) + (P_j - P_{j-1})\big(\widehat{A} + GP_{j-1}\big) \\ = -F - \widehat{A}^{\mathsf{T}} P_{j-1} - P_{j-1}\widehat{A} - P_{j-1}GP_{j-1}.\end{aligned}$$

Some manipulations and rearrangements of the terms finally lead to

$$\big(\widehat{A} + GP_{j-1}\big)^{\mathsf{T}} P_j + P_j\big(\widehat{A} + GP_{j-1}\big) = -F + P_{j-1}GP_{j-1}. \tag{5.12}$$

This leads to Kleinman's formulation of the Newton iteration which is given in Algorithm 5.3. The question arises whether Algorithm 5.3 converges to the right solution. The following theorem makes this clear, see also [LR95] for a proof.

Theorem 5.16: Consider the ARE (5.10) with stabilizable (A, B). Let P^- be its unique stabilizing solution. Let further $P_0 \in \mathbb{R}^{n\times n}$ be stabilizing, i. e., it holds that $\Lambda\big(\widehat{A} + GP_0\big) \subset \mathbb{C}^-$. Then the iterates P_j, $j = 1, 2, \ldots$ fulfill the following statements:

a) The matrix P_j is stabilizing.

b) It holds that $P_1 \leqslant \cdots \leqslant P_j \leqslant P_{j+1} \leqslant \cdots \leqslant P^-$.

Algorithm 5.3 Newton-Kleinman iteration for the algebraic Riccati equation

Input: $\widehat{A}$, F, G as in (5.10) and initial value P_0 such that $\Lambda\big(\widehat{A} + GP_0\big) \subset \mathbb{C}^-$.
Output: Stabilizing (and minimal) solution P^- solving (5.10).

1: **for** $j = 1, 2, \ldots$ **do**
2: Set $\widehat{A}_j := \widehat{A} + GP_{j-1}$ and $F_j := -F + P_{j-1}GP_{j-1}$.
3: Solve $\widehat{A}_j^{\mathsf{T}} P_j + P_j \widehat{A}_j = F_j$.
4: **end for**
5: Set $P^- := P_j$.

c) It holds that $\lim_{j\to\infty} P_j = P^-$.

d) The convergence is globally quadratic, i. e., there exists a constant $\gamma > 0$ such that

$$\left\|P^- - P_j\right\| \leqslant \gamma \left\|P^- - P_{j-1}\right\|^2, \quad j = 1, 2, \ldots.$$

Proof. The proof of this theorem needs a few technical results from the solution theory of Lyapunov and Riccati equations. Therefore, we omit it here. □

Remark 5.17: a) If $\widehat{A}$ is not asymptotically stable (otherwise $P_0 = 0$ is stabilizing), then the computation of a stabilizing P_0 usually costs as much as another iteration step since this requires the solution of one additional Lyapunov equation (e. g., in Bass' algorithm).

b) It can be proven that if $\Lambda\big(\widehat{A} + GP_{j-1}\big) \subset \mathbb{C}^-$, then it holds that $\Lambda\big(\widehat{A} + G(P_{j-1} + tN_{j-1})\big) \subset \mathbb{C}^-$ for all $t \in [0, 2]$. This motivates line search algorithms to optimize the step length after computing the direction N_{j-1} in Algorithm 5.2. This means that we set $P_j := P_{j-1} + tN_{j-1}$ where t is chosen as

$$t = \operatorname{argmin}_{\tau\in[0,2]} \left\|\mathcal{R}(P_{j-1} + \tau N_{j-1})\right\|_{\mathrm{F}}.$$

The computation of t is usually much cheaper than the actual Newton step which can drastically accelerate the iteration [BB98].

5.4.2 The Low-Rank Newton-Kleinman Method

We now aim at applying the Newton-Kleinman iteration for large-scale AREs and derive a low-rank formulation. Recall that the ARE attains the form

$$\begin{aligned}\mathcal{R}(P) := C^{\mathsf{T}}\big(D+D^{\mathsf{T}}\big)^{-1}C + \Big(A - B\big(D+D^{\mathsf{T}}\big)^{-1}C\Big)^{\mathsf{T}} P \\ + P\Big(A - B\big(D+D^{\mathsf{T}}\big)^{-1}C\Big) + PB\big(D+D^{\mathsf{T}}\big)^{-1}B^{\mathsf{T}}P = 0,\end{aligned}$$

where $A \in \mathbb{R}^{n\times n}$ is large and sparse, B, $C^{\mathsf{T}} \in \mathbb{R}^{n\times m}$, $D + D^{\mathsf{T}} > 0$, and $m \ll n$. Thus, the constant and the quadratic term admit low-rank factorizations $C^{\mathsf{T}}\big(D+D^{\mathsf{T}}\big)^{-1}C = \widetilde{C}^{\mathsf{T}}\widetilde{C}$ and $B\big(D+D^{\mathsf{T}}\big)^{-1}B^{\mathsf{T}} = \widetilde{B}\widetilde{B}^{\mathsf{T}}$. Inserting this into (5.12) gives the iteration scheme

$$\begin{aligned}&\Big(A - B\big(D+D^{\mathsf{T}}\big)^{-1}\big(C - B^{\mathsf{T}}P_{j-1}\big)\Big)^{\mathsf{T}} P_j + \\ &P_j \underbrace{\Big(A - B\big(D+D^{\mathsf{T}}\big)^{-1}\big(C - B^{\mathsf{T}}P_{j-1}\big)\Big)}_{=:\widehat{A}_j} = -\widetilde{C}^{\mathsf{T}}\widetilde{C} + P_{j-1}\widetilde{B}\widetilde{B}^{\mathsf{T}}P_{j-1}\end{aligned} \tag{5.13}$$

Unfortunately, we cannot solve the Lyapunov equation (5.13) directly with the low-rank ADI method, since its right-hand side may be indefinite. However, we can split it into two Lyapunov equations

$$\begin{aligned}\widehat{A}_j^{\mathsf{T}}P_{1,j} + P_{1,j}\widehat{A}_j &= -\widetilde{C}^{\mathsf{T}}\widetilde{C},\\ \widehat{A}_j^{\mathsf{T}}P_{2,j} + P_{2,j}\widehat{A}_j &= -P_{j-1}\widetilde{B}\widetilde{B}^{\mathsf{T}}P_{j-1},\end{aligned}$$

where we start with the initial values $P_{1,0} = P_0$ and $P_{2,0} = 0$ [RS10]. Then we obtain the iterate $P_j = P_{1,j} - P_{2,j}$ by linearity of the Lyapunov operator. This results in the *low-rank Newton-Kleinman method for algebraic Riccati equations, see also [BS13].*

One problem remains: even if A is sparse and B, C^{T} are thin, the feedback

$$\widehat{A}_j := A - B\big(D+D^{\mathsf{T}}\big)^{-1}\underbrace{\big(C - B^{\mathsf{T}}P_{j-1}\big)}_{=:K_j} \tag{5.14}$$

is usually dense. This means that we should never explicitly form (5.14).

There are several ways to solve linear systems with the system matrix $\widehat{A}_j - p_j I_n$ efficiently in the low-rank ADI method:

a) **Application of an iterative solver:** This option only requires multiplications with $\widehat{A}_j$. Since B and K_j have only a few columns and rows, respectively, these can be carried out efficiently. On the other hand, the convergence of iterative solvers is often slow, as long as no good preconditioner is available.

b) **Application of the Sherman-Morrison-Woodbury identity:** It holds that

$$\left(A - B\big(D + D^{\mathsf{T}}\big)^{-1} K_j\right)^{-1} = A^{-1} + A^{-1}B\big(D + D^{\mathsf{T}} - K_j A^{-1} B\big)^{-1} K_j A^{-1}.$$

Then a linear system solve with $\widehat{A}_j$ only requires two sparse (block) solves with A and one small dense solve with the matrix $D + D^{\mathsf{T}} - K_j A^{-1} B$.

CHAPTER 6

Interpolatory Model Reduction

In this section we will discuss interpolatory methods for model reduction. Here we take a more transfer function point of view and try to find reduced representations of it. This reduction is then normally done by evaluating the original transfer function (and its derivatives) in a number of points in the complex plane and then to construct a *rational interpolant* that match this information. Mostly, this rational interpolant has a realization of low order which will be our reduced-order model.

6.1 Moment Matching

6.1.1 Moments

Consider the LTI system $[A, B, C, D] \in \Sigma_{n,m,p}$ with transfer function $G(s) \in \mathbb{R}(s)^{p\times m}$. Since G is rational and proper, its poles are contained in $\Lambda(A)$. Therefore, G is analytic in a neighborhood of all $s_0 \in \mathbb{C}\backslash\Lambda(A)$. Hence it can be locally expanded into a Taylor series at the expansion point s_0. Thus, for finite s_0 we obtain

$$G(s) = \sum_{k=0}^{\infty} (s - s_0)^k M_k(s_0)$$

for some neighborhood of s_0 and some matrices $M_0(s_0)$, $M_1(s_0)$, $M_2(s_0)$, ... On the other hand, for $s_0 = \infty$ we obtain the Laurent series

$$G(s) = \sum_{k=0}^{\infty} s^{-k} M_k(\infty).$$

The matrices $M_k(s_0)$ are called the (k-th) *moments* of G at s_0. For the case $s_0 = \infty$ they are also called the *Markov parameters* of the transfer function. Now we want to determine the moments. For this we need the following lemma.

Lemma 6.1 (Neumann series)**:** Let $A \in \mathbb{C}^{n\times n}$ with spectral radius $\rho(A) < 1$ be given. Then $I_n - A$ is invertible and it holds that

$$(I_n - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

For finite s_0 we have

$$\begin{aligned} G(s) &= C((s-s_0)I_n - A + s_0 I_n)^{-1}B + D \\ &= C\left((s-s_0)(s_0 I_n - A)^{-1} + I_n\right)^{-1}(s_0 I_n - A)^{-1}B + D \\ &= C\left(I_n - (s_0 - s)(s_0 I_n - A)^{-1}\right)^{-1}(s_0 I_n - A)^{-1}B + D \\ &= \sum_{k=0}^{\infty} C(s_0 I_n - A)^{-k-1}B(s_0 - s)^k + D, \end{aligned}$$

where we have used, that by Lemma 6.1 and s sufficiently close to s_0 it holds that

$$\left(I_n - (s_0 - s)(s_0 I_n - A)^{-1}\right)^{-1} = \sum_{k=0}^{\infty} (s_0 - s)^k (s_0 I_n - A)^{-k}.$$

Therefore, we have

$$M_k(s_0) = \begin{cases} C(s_0 I_n - A)^{-1}B + D, & \text{if } k = 0, \\ (-1)^k C(s_0 I_n - A)^{-k-1}B, & \text{if } k \geqslant 1. \end{cases}$$

Moreover, for $s_0 = \infty$ and sufficiently large s, we have

$$\begin{aligned} G(s) &= C(sI_n - A)^{-1}B + D \\ &= \tfrac{1}{s} C \underbrace{\left(I_n - \tfrac{1}{s}A\right)^{-1}}_{=\sum_{k=0}^{\infty} \frac{1}{s^k} A^k} B + D \\ &= \sum_{k=1}^{\infty} C A^{k-1} B \tfrac{1}{s^k} + D. \end{aligned}$$

Therefore, we have

$$M_k(\infty) = \begin{cases} D, & \text{if } k = 0, \\ CA^{k-1}B, & \text{if } k \geqslant 1. \end{cases}$$

Model reduction by moment matching consists of finding a reduced-order model $[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}] \in \Sigma_{r,m,p}$ with $r \ll n$ such that for a given $s_0 \in (\mathbb{C} \cup \{\infty\}) \setminus \Lambda(A)$ the corresponding moments $\widetilde{M}_k(s_0)$, $k = 1, 2, \ldots$ fulfill

$$\widetilde{M}_k(s_0) = M_k(s_0), \quad k = 0, 1, \ldots, \ell$$

for ℓ as large as possible. The moment matching problem for $s_0 = 0$ is also called *Padé approximation problem*, for $s_0 = \infty$ it is also called the *partial realization problem*. Since for finite s_0 we have

$$G(s) = \sum_{k=0}^{\infty} (s - s_0)^k M_k(s_0) = \sum_{k=0}^{\ell} (s - s_0)^k M_k(s_0) + \mathcal{O}\big((s - s_0)^{\ell+1}\big),$$
$$\widetilde{G}(s) = \sum_{k=0}^{\infty} (s - s_0)^k \widetilde{M}_k(s_0) = \sum_{k=0}^{\ell} (s - s_0)^k M_k(s_0) + \mathcal{O}\big((s - s_0)^{\ell+1}\big),$$

we obtain

$$G(s) - \widetilde{G}(s) = \mathcal{O}\big((s - s_0)^{\ell+1}\big).$$

Similarly, for $s_0 = \infty$ we get

$$G(s) - \widetilde{G}(s) = \mathcal{O}\big(s^{-(\ell+1)}\big).$$

6.1.2 One-Sided Moment Matching

For ease of notation we will now have a look at moment matching for SISO systems. In particular we discuss efficient ways to generate the reduced-order model without explicitly forming the moments.

Definition 6.2: Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be given. The *generalized controllability matrices* of $[A, B, C, D]$ for $s_0 \in (\mathbb{C} \cup \{\infty\}) \setminus \Lambda(A)$ are given by

$$\mathcal{C}_k(s_0) = \begin{bmatrix} (s_0 I_n - A)^{-1}B & (s_0 I_n - A)^{-2}B & \ldots & (s_0 I_n - A)^{-k}B \end{bmatrix}, \quad k = 1, \ldots, n,$$

if $s_0 \in \mathbb{C} \setminus \Lambda(A)$, and by

$$\mathcal{C}_k(\infty) = \begin{bmatrix} B & AB & \ldots & A^{k-1}B \end{bmatrix}, \quad k = 1, \ldots, n,$$

if $s_0 = \infty$.

Note that $\mathcal{C}_n(\infty)$ is the Kalman controllability matrix that can be used for checking controllability of a linear system. We have the following lemma.

Lemma 6.3: Let $s_0 \in (\mathbb{C} \cup \{\infty\})\backslash\Lambda(A)$ be given. Then the pair $(A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times 1}$ is controllable if and only if it holds that $\operatorname{rank}\mathcal{C}_k(s_0) = k$ for all $k = 1, 2, \ldots, n$.

Proof. Consider the case that $s_0 = \infty$. Then controllability of (A, B) is equivalent to

$$\operatorname{rank}\mathcal{C}_n(\infty) = \operatorname{rank}\begin{bmatrix} B & AB & \ldots & A^{n-1}B \end{bmatrix} = n.$$

This is furthermore equivalent to

$$\operatorname{rank}\mathcal{C}_k(\infty) = k, \quad k = 1, \ldots, n.$$

On the other hand, for $s_0 \in \mathbb{C}\backslash\Lambda(A)$, the result follows as above by noticing that (A, B) is controllable, if and only if

$$\begin{aligned}
\mathbb{C}^n &= \operatorname{span}\left\{B, AB, \ldots, A^{n-1}B\right\} \\
&= \operatorname{span}\left\{B, (s_0 I_n - A)B, \ldots, (s_0 I_n - A)^{n-1}B\right\} \\
&= (s_0 I_n - A)^{-n} \operatorname{span}\left\{B, (s_0 I_n - A)B, \ldots, (s_0 I_n - A)^{n-1}B\right\} \\
&= \operatorname{span}\left\{(s_0 I_n - A)^{-1}B, (s_0 I_n - A)^{-2}B, \ldots, (s_0 I_n - A)^{-n}B\right\}.
\end{aligned}$$

□

These matrices determine projection matrices that result in reduced-order models with matched moments. In particular, we have the following theorem.

Theorem 6.4: Let $[A, B, C, D] \in \Sigma_{n,1,1}$ be controllable with the moments $M_k(s_0)$, $k = 0, 1, \ldots$ for some given $s_0 \in (\mathbb{C} \cup \{\infty\})\backslash\Lambda(A)$. Assume that $\ell \in \{1, 2, \ldots, n\}$ and that $\widetilde{T} \in \mathbb{C}^{\ell\times\ell}$ is invertible. Set $T := \mathcal{C}_\ell(s_0)\widetilde{T} \in \mathbb{C}^{n\times\ell}$ and choose $W \in \mathbb{C}^{n\times\ell}$ such that $W^{\mathsf{H}}T = I_\ell$. Define $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right] := \left[W^{\mathsf{H}}AT, W^{\mathsf{H}}B, CT, D\right] \in \Sigma_{\ell,1,1}$ with the moments $\widetilde{M}_k(s_0)$, $k = 0, 1, \ldots$. Then the first ℓ moments are matched, i. e., it holds that

$$M_k(s_0) = \widetilde{M}_k(s_0), \quad k = 0, 1, \ldots, \ell - 1.$$

Proof. We can assume w. l. o. g. that $\widetilde{T} = I_\ell$, because the moments $\widetilde{M}_k(s_0)$, $k = 1, 2, \ldots$ are invariant under state-space transformations. Since the system is controllable, we have that $\operatorname{rank} T = \ell$ and there exists a matrix $W \in \mathbb{C}^{n\times\ell}$

such that $W^{\mathsf{H}}T = I_\ell$. Here we will only do the proof for finite s_0. The proof for $s_0 = \infty$ is similar, and therefore, we omit it here. We have that

$$\begin{aligned} s_0 I_\ell - \widetilde{A} &= W^{\mathsf{H}}(s_0 I_n - A)T \\ &= W^{\mathsf{H}}(s_0 I_n - A)\begin{bmatrix}(s_0 I_n - A)^{-1}B & \dots & (s_0 I_n - A)^{-\ell}B\end{bmatrix} \\ &= \begin{bmatrix}W^{\mathsf{H}}B & W^{\mathsf{H}}(s_0 I_n - A)^{-1}B & \dots & W^{\mathsf{H}}(s_0 I_n - A)^{-\ell+1}B\end{bmatrix} \\ &=: \begin{bmatrix}\widetilde{B} & e_1 & \dots & e_{\ell-1}\end{bmatrix}. \end{aligned}$$

Therefore, for $k \in \{0, 1, \dots, \ell - 1\}$ we have

$$\left(s_0 I_\ell - \widetilde{A}\right)^{k+1} = \begin{bmatrix} * & \dots & * & \widetilde{B} & e_1 & \dots & e_{\ell-1-k}\end{bmatrix},$$

which gives $\left(s_0 I_\ell - \widetilde{A}\right)^{-(k+1)}\widetilde{B} = e_{k+1}$.

So, for the moments we have,

$$\begin{aligned} \widetilde{M}_0(s_0) &= \widetilde{C}\left(s_0 I_\ell - \widetilde{A}\right)^{-1}\widetilde{B} + D \\ &= CTe_1 + D \\ &= C\mathcal{C}_\ell(s_0)e_1 + D \\ &= C(s_0 I_n - A)^{-1}B + D = M_0(s_0), \end{aligned}$$

and for $k = 1, \dots, \ell - 1$ we obtain

$$\begin{aligned} \widetilde{M}_k(s_0) &= (-1)^k\widetilde{C}\left(s_0 I_\ell - \widetilde{A}\right)^{-(k+1)}\widetilde{B} \\ &= (-1)^k CTe_{k+1} \\ &= (-1)^k C\mathcal{C}_\ell(s_0)e_{k+1} \\ &= (-1)^k C(s_0 I_n - A)^{-(k+1)}B = M_k(s_0). \end{aligned}$$

□

Remark 6.5: a) The assumption that $[A, B, C, D]$ is controllable can be weakened in the sense that it is only required that ℓ is small enough such that $\operatorname{rank}\mathcal{C}_\ell(s_0) = \ell$.

b) On the other hand, if $\ell = \operatorname{rank}\mathcal{C}_n(s_0)$, then the reduced system matches all moments.

We are free to choose the matrix $\widetilde{T}$ in Theorem 6.4. The simplest choice $\widetilde{T} = I_\ell$ unfortunately usually leads to very ill-conditioned projection matrices T, implying that the numerical computation of T is very sensitive to round-off errors. It is better to choose T such that it has orthonormal columns. Since $\operatorname{im}\mathcal{C}(s_0)$ is a Krylov subspace, i. e., a space of the general form

$$\mathcal{K}_\ell(F, v) := \operatorname{span}\left\{v, Fv, \dots, F^{\ell-1}v\right\},$$

T can be efficiently computed by the *(shift-and-invert) Arnoldi method* (see the course on numerical linear algebra).

Since T has orthonormal columns, we can choose $W = T$. In the case $s_0 = \infty$, there are further simplifications. The Arnoldi iteration computes $T \in \mathbb{C}^{n\times\ell}$ with orthonormal columns and a matrix $H \in \mathbb{C}^{\ell\times\ell}$ in Hessenberg form such that

$$AT = TH + f_{\ell+1}e_\ell^{\mathsf{T}}$$

for some vector $f_{\ell+1} \in \mathbb{C}^n$ with $T^{\mathsf{H}} f_{\ell+1} = 0$. Thus we obtain

$$\widetilde{A} = W^{\mathsf{H}}AT = T^{\mathsf{H}}\big(TH + f_{\ell+1}e_\ell^{\mathsf{T}}\big) = H.$$

Moreover, since B is the initial vector of the Krylov space $\mathcal{K}_\ell(A, B)$, we have

$$\widetilde{B} = W^{\mathsf{H}}B = T^{\mathsf{H}}B = \|B\|_2\, e_1.$$

Thus, $\widetilde{A}$ and $\widetilde{B}$ are obtained at no extra cost, only the matrix $\widetilde{C}$ has to be computed.

Note that stability and observability may be lost by moment matching as the following example shows.

Example: Consider the system $[A, B, C, D] \in \Sigma_{2,1,1}$ with

$$A = \begin{bmatrix} -1 & 5 \\ 0 & -2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & -1 \end{bmatrix}. \tag{6.1}$$

It is easily checked that this system is stable and observable. For $s_0 = \infty$ and $\ell = 1$, we obtain $T = \frac{1}{\sqrt{2}}B = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix} 1 \\ 1 \end{smallmatrix}\right] = W$. This gives

$$\widetilde{A} = \frac{1}{2}\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} -1 & 5 \\ 0 & -2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1,$$
$$\widetilde{C} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = 0.$$

Therefore, the reduced model is neither stable nor observable.

Note that stability and observability can be shown for symmetric systems with $A = A^{\mathsf{T}}$ and $B = C^{\mathsf{T}}$.

By using the generalized observability matrices

$$\mathcal{O}_k(s_0) = \begin{bmatrix} C(s_0 I_n - A)^{-1} \\ C(s_0 I_n - A)^{-2} \\ \vdots \\ C(s_0 I_n - A)^{-k} \end{bmatrix}, \quad k = 1, \ldots, n,$$

if $s_0 \in \mathbb{C}\backslash\Lambda(A)$, and

$$\mathcal{O}_k(\infty) = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{k-1} \end{bmatrix}, \quad k = 1, \ldots, n,$$

if $s_0 = \infty$, we obtain the following result which is completely analogous to Theorem 6.4.

Theorem 6.6: Let $[A, B, C, D] \in \Sigma_{n,1,1}$ be observable with the moments $M_k(s_0)$, $k = 0, 1, \ldots$ for some given $s_0 \in (\mathbb{C} \cup \{\infty\})\backslash\Lambda(A)$. Assume that $\ell \in \{1, 2, \ldots, n\}$ and that $\widetilde{W} \in \mathbb{C}^{\ell\times\ell}$ is invertible. Set $W := \mathcal{O}_\ell(s_0)^{\mathsf{H}}\widetilde{W} \in \mathbb{C}^{n\times\ell}$ and choose $T \in \mathbb{C}^{n\times\ell}$ such that $W^{\mathsf{H}}T = I_\ell$. Define $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right] := \left[W^{\mathsf{H}}AT, W^{\mathsf{H}}B, CT, D\right] \in \Sigma_{\ell,1,1}$ with the moments $\widetilde{M}_k(s_0)$, $k = 0, 1, \ldots$. Then the first ℓ moments are matched, i. e., it holds that

$$M_k(s_0) = \widetilde{M}_k(s_0), \quad k = 0, 1, \ldots, \ell - 1.$$

The matrix W can be computed by the Arnoldi algorithm applied to the matrix pair $\left(A^{\mathsf{T}}, C^{\mathsf{T}}\right)$. There are also variants of moment matching for MIMO systems which make use of the block Arnoldi method and for matching moments at several interpolation points $s_0, s_1, \ldots$.

6.1.3 Two-Sided Moment Matching

We have just seen that choosing the projection matrices T and W such that $W^{\mathsf{H}}T = I_\ell$ and

i) $\operatorname{im} T = \operatorname{im} \mathcal{C}_\ell(s_0)$ or

ii) $\operatorname{im} W = \operatorname{im} \mathcal{O}_\ell(s_0)^{\mathsf{H}}$

guarantees that ℓ moment at s_0 are matched. So what happens if we do both? We will show soon that in this case we can match 2ℓ moments. However, this is not always possible since a certain regularity condition is needed which we will prove now.

Lemma 6.7: Let $[A, B, C, D] \in \Sigma_{n,1,1}$ be given with the generalized controllability and observability matrices $\mathcal{C}_\ell(s_0)$ and $\mathcal{O}_\ell(s_0)$ for some $\ell \in \{1, 2, \ldots, n\}$. Then there exist matrices $T, W \in \mathbb{C}^{n\times\ell}$ with $W^{\mathsf{H}}T = I_\ell$, $\operatorname{im} T = \operatorname{im} \mathcal{C}_\ell(s_0)$, and $\operatorname{im} W = \operatorname{im} \mathcal{O}_\ell(s_0)^{\mathsf{H}}$, if and only if the matrix $\mathcal{H}_\ell(s_0) := \mathcal{O}_\ell(s_0)\mathcal{C}_\ell(s_0)$ is invertible.

Proof. First we show "⇐": So assume that $\mathcal{H}_\ell(s_0)$ is invertible. By choosing $T := \mathcal{C}_\ell(s_0)$ and $W := \mathcal{O}_\ell(s_0)^{\mathsf{H}}\mathcal{H}_\ell(s_0)^{-\mathsf{H}}$, the conditions $\operatorname{im} T = \operatorname{im}\mathcal{C}_\ell(s_0)$ and $\operatorname{im} W = \operatorname{im}\mathcal{O}_\ell(s_0)^{\mathsf{H}}$ are obviously satisfied. Moreover, we have

$$W^{\mathsf{H}}T = \mathcal{H}_\ell(s_0)^{-1}\underbrace{\mathcal{O}_\ell(s_0)\mathcal{C}_\ell(s_0)}_{=\mathcal{H}_\ell(s_0)} = I_\ell.$$

Now we show "⇒": So let $T, W \in \mathbb{C}^{n\times\ell}$ such that $W^{\mathsf{H}}T = I_\ell$, $\operatorname{im} T = \operatorname{im}\mathcal{C}_\ell(s_0)$, and $\operatorname{im} W = \operatorname{im}\mathcal{O}_\ell(s_0)^{\mathsf{H}}$. Then there exist two invertible matrices $K_{\mathcal{C}}$, $K_{\mathcal{O}} \in \mathbb{C}^{\ell\times\ell}$ such that $T = \mathcal{C}_\ell(s_0)K_{\mathcal{C}}$ and $W = \mathcal{O}_\ell(s_0)^{\mathsf{H}}K_{\mathcal{O}}^{\mathsf{H}}$. Therefore, we get

$$I_\ell = W^{\mathsf{H}}T = K_{\mathcal{O}}\mathcal{O}_\ell(s_0)\mathcal{C}_\ell(s_0)K_{\mathcal{C}} = K_{\mathcal{O}}\mathcal{H}_\ell(s_0)K_{\mathcal{C}}.$$

So $\mathcal{H}_\ell(s_0) = K_{\mathcal{O}}^{-1}K_{\mathcal{C}}^{-1}$ is invertible. □

Theorem 6.8: Let $[A, B, C, D] \in \Sigma_{n,1,1}$ be given with the moments $M_k(s_0)$, $k = 0, 1, \ldots$ for some given $s_0 \in (\mathbb{C}\cup\{\infty\})\backslash\Lambda(A)$. Let $T, W \in \mathbb{C}^{n\times\ell}$ with $W^{\mathsf{H}}T = I_\ell$ and define the reduced-order model by $\big[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\big] := \big[W^{\mathsf{H}}AT, W^{\mathsf{H}}B, CT, D\big] \in \Sigma_{\ell,1,1}$ with the moments $\widetilde{M}_k(s_0)$, $k = 0, 1, \ldots$. If $\operatorname{im}\mathcal{C}_{\ell_1}(s_0) \subseteq \operatorname{im} T$ and $\operatorname{im}\mathcal{O}_{\ell_2}(s_0)^{\mathsf{H}} \subseteq \operatorname{im} W$ for some $\ell_1, \ell_2 \in \mathbb{N}_0$, then it holds that

$$M_k(s_0) = \widetilde{M}_k(s_0), \quad k = 0, 1, \ldots, \begin{cases} \ell_1 + \ell_2 - 1, & \text{if } s_0 \in \mathbb{C}\backslash\Lambda(A), \\ \ell_1 + \ell_2, & \text{if } s_0 = \infty. \end{cases}$$

Proof. Here we prove the theorem only for the case $s_0 \in \mathbb{C}\backslash\Lambda(A)$, the case $s_0 = \infty$ is analogous. We show the statement in several steps:

Step 1: We show that for any $F \in \mathbb{C}^{\ell\times n}$ with $FT = I_\ell$ we have $TFv = v$ for all $v \in \operatorname{im} T$. Indeed, if $v \in \operatorname{im} T$, there exists a $z \in \mathbb{C}^\ell$ such that $v = Tz$. This gives $TFv = TFTz = Tz = v$.

Step 2: We show that $(s_0 I_n - A)^{-k}B = T\big(s_0 I_\ell - \widetilde{A}\big)^{-k}\widetilde{B}$ for $k = 1, \ldots, \ell_1$. First note that with $F := \big(s_0 I_\ell - \widetilde{A}\big)^{-1}W^{\mathsf{H}}(s_0 I_n - A)$ we have $FT = \big(s_0 I_\ell - \widetilde{A}\big)^{-1}W^{\mathsf{H}}(s_0 I_n - A)T = I_\ell$, so we can use Step 1 here.

Now we prove the statement via induction. First we show the case $k = 1$: It

holds that

$$\begin{aligned} T\big(s_0 I_\ell - \widetilde{A}\big)^{-1}\widetilde{B} &= T\big(s_0 I_\ell - \widetilde{A}\big)^{-1} W^{\mathsf{H}} B \\ &= T\big(s_0 I_\ell - \widetilde{A}\big)^{-1} W^{\mathsf{H}} (s_0 I_n - A)(s_0 I_n - A)^{-1} B \\ &= TF \underbrace{(s_0 I_n - A)^{-1} B}_{\in \operatorname{im} \mathcal{C}_{\ell_1}(s_0) \subseteq \operatorname{im} T} \\ &= (s_0 I_n - A)^{-1} B. \end{aligned}$$

Now assume that $(s_0 I_n - A)^{-k} B = T\big(s_0 I_\ell - \widetilde{A}\big)^{-k}\widetilde{B}$. Then we have

$$\begin{aligned} T\big(s_0 I_\ell - \widetilde{A}\big)^{-(k+1)}\widetilde{B} &= T\big(s_0 I_\ell - \widetilde{A}\big)^{-1} W^{\mathsf{H}} \underbrace{T\big(s_0 I_\ell - \widetilde{A}\big)^{-k}\widetilde{B}}_{=(s_0 I_n - A)^{-k} B} \\ &= T\big(s_0 I_\ell - \widetilde{A}\big)^{-1} W^{\mathsf{H}} (s_0 I_n - A)(s_0 I_n - A)^{-(k+1)} B \\ &= TF \underbrace{(s_0 I_n - A)^{-(k+1)} B}_{\in \operatorname{im} \mathcal{C}_{\ell_1}(s_0) \subseteq \operatorname{im} T, \text{ if } k+1 \leqslant \ell_1} = (s_0 I_n - A)^{-(k+1)} B. \end{aligned}$$

Step 3: We show that $C(s_0 I_n - A)^{-k} = \widetilde{C}\big(s_0 I_\ell - \widetilde{A}\big)^{-k} W^{\mathsf{H}}$ for $k = 1, 2, \ldots, \ell_2$. This statement can be analogously proven as in Step 2 by replacing A by A^{T}, $\widetilde{A}$ by $\widetilde{A}^{\mathsf{H}}$, B by C^{T}, $\widetilde{B}$ by $\widetilde{C}^{\mathsf{H}}$, T by W, and W by T.

Step 4: We show that $M_k(s_0) = \widetilde{M}_k(s_0)$ for $k = 0, 1, \ldots, \ell_1 + \ell_2 - 1$. For $k = 0$ we obtain

$$\begin{aligned} \widetilde{M}_0(s_0) = \widetilde{C}\big(s_0 I_\ell - \widetilde{A}\big)^{-1}\widetilde{B} + D &= C \underbrace{T\big(s_0 I_\ell - \widetilde{A}\big)^{-1}\widetilde{B}}_{=(s_0 I_n - A)^{-1} B} + D \\ &= C(s_0 I_n - A)^{-1} B + D = M_0(s_0), \end{aligned}$$

where we have used Step 2. Now for $k \in \{1, 2, \ldots, \ell_1 + \ell_2 - 1\}$ let $k_1 \in \{1, 2, \ldots, \ell_1\}$ and $k_2 \in \{1, 2, \ldots, \ell_2\}$ be such that $k_1 + k_2 = k + 1$. Then we have

$$\begin{aligned} \widetilde{M}_k(s_0) &= (-1)^k \widetilde{C}\big(s_0 I_\ell - \widetilde{A}\big)^{-(k+1)}\widetilde{B} \\ &= (-1)^k \widetilde{C}\big(s_0 I_\ell - \widetilde{A}\big)^{-k_2} W^{\mathsf{H}} T\big(s_0 I_\ell - \widetilde{A}\big)^{-k_1}\widetilde{B} \\ &= (-1)^k C(s_0 I_n - A)^{-k_2}(s_0 I_n - A)^{-k_1} B = M_k(s_0). \end{aligned}$$

This completes the proof. □

Remark 6.9: a) Most often, the reduced-order model is computed for $\ell_1 = \ell_2 = \ell$. In case the matrix $\mathcal{H}_\ell(s_0)$ is invertible, then the corresponding projection matrices T and W can be computed by the nonsymmetric Lanczos process.

b) If the system $[A, B, C, D] \in \Sigma_{n,m,p}$ is a MIMO system, then the nonsymmetric block-Lanczos method can be used. Then we also have that

$$\operatorname{span}\left\{(s_0 I_n - A)^{-1} B, \ldots, (s_0 I_n - A)^{-\ell} B\right\} \subseteq \operatorname{im} T,$$
$$\operatorname{span}\left\{(s_0 I_n - A)^{-\mathsf{H}} C^{\mathsf{H}}, \ldots, \left((s_0 I_n - A)^{\ell}\right)^{-\mathsf{H}} C^{\mathsf{H}}\right\} \subseteq \operatorname{im} W$$

implies

$$M_k(s_0) = \widetilde{M}_k(s_0), \quad k = 0, 1, \ldots, 2\ell - 1,$$

but the projection spaces must have a larger dimension to match the same order of moments. Alternatively, *tangential interpolation* techniques can be employed which yield moment matching properties in certain tangential directions, but normally also give smaller projection matrices. This will be discussed in more detail later.

Algorithm 6.1 employs the nonsymmetric Lanczos algorithm [Saa82] for moment matching. In the literature it is known as the *Padé-via-Lanczos (PVL) algorithm*.

The quantities computed in Algorithm 6.1 have the following properties (exercise!):

a) The matrices W and T are biorthogonal, i. e., $W^{\mathsf{H}} T = I_\ell$.

b) It holds that

$$(s_0 I_n - A)^{-1} T = TH + \begin{bmatrix} 0 & \ldots & 0 & t_{\ell+1} \end{bmatrix} \beta_{\ell+1},$$
$$(s_0 I_n - A)^{-\mathsf{H}} W = WH^{\mathsf{H}} + \begin{bmatrix} 0 & \ldots & 0 & w_{\ell+1} \end{bmatrix} \overline{\gamma_{\ell+1}},$$

where

$$H = \begin{bmatrix} \alpha_1 & \gamma_2 & & \\ \beta_2 & \alpha_2 & \ddots & \\ & \ddots & \ddots & \gamma_\ell \\ & & \beta_\ell & \alpha_\ell \end{bmatrix}.$$

c) We have that

$$\operatorname{im} T = \operatorname{im} \mathcal{C}_\ell(s_0), \quad \operatorname{im} W = \operatorname{im} \mathcal{O}_\ell(s_0)^{\mathsf{H}}.$$

Remark 6.10: In Algorithm 6.1 breakdowns may occur. It can be shown that this is the case if and only if the matrix $\mathcal{H}_k(s_0)$ is singular for some $k \leqslant \ell$. With regard to Lemma 6.7 this means that we cannot construct matrices $W, T \in \mathbb{C}^{n \times k}$ with $W^{\mathsf{H}} T = I_k$ such that $\operatorname{im} T = \operatorname{im} \mathcal{C}_\ell(s_0)$ and $\operatorname{im} W = \operatorname{im} \mathcal{O}_\ell(s_0)^{\mathsf{H}}$.

Algorithm 6.1 Padé-via-Lanczos for moment matching

Input: $[A, B, C, D] \in \Sigma_{n,1,1}$, $s_0 \in \mathbb{C}\backslash\Lambda(A)$, reduced order ℓ.

Output: Reduced-order model $[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}] = \left[W^{\mathsf{H}}AT, W^{\mathsf{H}}B, CT, D\right] \in \Sigma_{\ell,1,1}$ with $M_k(s_0) = \widetilde{M}_k(s_0)$ for $k = 0, 1, \ldots, 2\ell - 1$.

1: **if** $C(s_0 I_n - A)^{-2}B = 0$ **then**
2: Breakdown.
3: **else**
4: Set $t_1 := \frac{1}{\|(s_0 I_n - A)^{-1}B\|_2}(s_0 I_n - A)^{-1}B$, $w_1 := \frac{\|(s_0 I_n - A)^{-1}B\|_2}{C(s_0 I_n - A)^{-2}B}(s_0 I_n - A)^{-\mathsf{H}}C^{\mathsf{H}}$.
5: Set $\alpha_1 := w_1^{\mathsf{H}}(s_0 I_n - A)^{-1}t_1$, $t_0 := 0$, $w_0 := 0$, $\beta_1 := 0$, $\gamma_1 := 0$.
6: **end if**
7: **for** $k = 1, 2, \ldots, \ell$ **do**
8: Set $r_k := (s_0 I_n - A)^{-1}t_k - \alpha_k t_k - \gamma_k t_{k-1}$.
9: Set $s_k := (s_0 I_n - A)^{-\mathsf{H}}w_k - \overline{\alpha_k} w_k - \beta_k w_{k-1}$.
10: **if** $s_k^{\mathsf{H}} r_k = 0$ **then**
11: Breakdown.
12: **else**
13: Set $\beta_{k+1} := \|r_k\|_2$.
14: Set $\gamma_{k+1} := \frac{1}{\beta_{k+1}} s_k^{\mathsf{H}} r_k$.
15: Set $t_{k+1} = \frac{1}{\beta_{k+1}} r_k$.
16: Set $w_{k+1} = \frac{1}{\overline{\gamma_{k+1}}} s_k$.
17: Set $\alpha_{k+1} := w_{k+1}^{\mathsf{H}}(s_0 I_n - A)^{-1}t_{k+1}$.
18: **end if**
19: **end for**
20: Construct the reduced-order model as

$$\widetilde{A} := W^{\mathsf{H}}AT, \quad \widetilde{B} := W^{\mathsf{H}}B, \quad \widetilde{C} := CT, \quad \widetilde{D} := D,$$

where

$$W := \begin{bmatrix} w_1 & \ldots & w_\ell \end{bmatrix}, \quad T := \begin{bmatrix} t_1 & \ldots & t_\ell \end{bmatrix}.$$

We conclude this subsection with some general remarks on the difficulties of moment matching model reduction.

Remark 6.11: a) Moment matching is a comparably cheap method for model reduction, since for each s_0, only one (sparse) LU factorization or one preconditioner has to be computed and stored.

b) Computable error estimates or bounds are often very pessimistic or expensive to evaluate. Moreover, there is only a good approximation quality close to the expansion point s_0.

c) The expansion point s_0 can often only be chosen heuristically by using a priori knowledge from the system under consideration.

d) Preservation of physical properties such as stability or passivity can only be guaranteed in special cases. Usually, a post-processing to restore these properties would be required, but this may destroy the moment matching properties.

In the following section we will discuss optimal choices for interpolation points which will cure many of the above mentioned problems.

6.2 $\mathcal{H}_2$-Optimal Interpolation: The Iterative Rational Krylov Algorithm

This section is based on [GAB08]. Consider an LTI system $[A, B, C, D] \in \Sigma_{n,1,1}$. From the results of the previous section it is known that choosing $W,\ T \in \mathbb{C}^{n\times\ell}$ such that $W^{\mathsf{H}}T = I_\ell$ with

$$\begin{aligned}\operatorname{im} T &= \operatorname{im}\begin{bmatrix}(s_1 I_n - A)^{-1}B & \dots & (s_\ell I_n - A)^{-1}B\end{bmatrix},\\ \operatorname{im} W &= \operatorname{im}\begin{bmatrix}(s_1 I_n - A)^{-\mathsf{H}}C^{\mathsf{T}} & \dots & (s_\ell I_n - A)^{-\mathsf{H}}C^{\mathsf{T}}\end{bmatrix}\end{aligned}$$

will lead to a reduced-order model $[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}] := \left[W^{\mathsf{H}}AT, W^{\mathsf{H}}B, CT, D\right] \in \Sigma_{\ell,1,1}$ with

$$G(s_k) = \widetilde{G}(s_k), \quad G'(s_k) = \widetilde{G}'(s_k), \quad k = 1,\, 2,\, \dots,\, \ell.$$

The question arises whether for fixed ℓ we can choose the interpolation points $s_1,\, \dots,\, s_\ell$ such that $\left\|G - \widetilde{G}\right\|$ can be minimized in a suitable system norm. It turns out that we can determine reduced-order models which fulfill such an optimality criterion locally in the $\mathcal{H}_2$-norm. Recall from Chapter 2 that for SISO systems we have that

$$\left\|G - \widetilde{G}\right\|_{\mathcal{H}_2} = \sup_{\substack{u\in\mathcal{L}_2([0,\infty),\mathbb{R})\\ u\neq 0}} \frac{\|y - \widetilde{y}\|_{\mathcal{L}_\infty}}{\|u\|_{\mathcal{L}_2}},$$

where $y,\, \widetilde{y} \in \mathcal{L}_\infty([0,\infty),\mathbb{R})$ are the outputs of the full and reduced-order models obtained by feeding in the input $u \in \mathcal{L}_2([0,\infty),\mathbb{R})$. Further recall that

$$\left\|G - \widetilde{G}\right\|_{\mathcal{H}_2} = \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \left|G(\mathrm{i}\omega) - \widetilde{G}(\mathrm{i}\omega)\right|^2 \mathrm{d}\omega\right)^{1/2}$$

for $G - \widetilde{G} \in \mathcal{RH}_2$ which is usually the case, since we have $D = \widetilde{D}$ in projection-based model reduction.

Finding a global optimum is hard, but we are satisfied with a local optimum which can be obtained by checking first-order (necessary) optimality conditions. For this we will need the $\mathcal{H}_2$ inner product, which for SISO systems $G, H \in \mathcal{RH}_2$ reduces to

$$\langle G, H \rangle_{\mathcal{H}_2} = \frac{1}{2\pi} \int_{-\infty}^{\infty} G(-\mathrm{i}\omega) H(\mathrm{i}\omega) \mathrm{d}\omega.$$

Now we look for a different way of evaluating this inner product. For this we need the residues of a function $G(s) \in \mathbb{R}(s)$ with possibly multiple poles. The residue of G at pole $\lambda \in \mathbb{C}$ of order m is given by

$$\operatorname{res}(G, \lambda) := \frac{1}{(m-1)!} \lim_{s \to \lambda} \frac{\mathrm{d}^{m-1}}{\mathrm{d}s^{m-1}} \left((s-\lambda)^m G(s) \right).$$

Note that if G is proper and the pole λ is simple, then the residues can be computed as in Lemma 3.4. Residues are important tools in complex analysis, for instance for the evaluation of integrals. One of the key results in complex analysis is the *residue theorem* which we state next (in a simplified version).

Theorem 6.12 (Residue theorem)**:** Let $f : \mathcal{D} \to \mathbb{C}$ be a meromorphic function (i. e., a function holomorphic in almost all points in $\mathcal{D}$), $\Gamma \subset \mathcal{D}$ be a simple closed contour in counterclockwise orientation that does not contain a pole of f, and let $\mu_1, \ldots, \mu_k$ be the poles of f inside the contour Γ. Then we have

$$\int_{\Gamma} f(s) \mathrm{d}s = 2\pi \mathrm{i} \sum_{j=1}^{k} \operatorname{res}(f, \mu_j).$$

Lemma 6.13: Let $G, H \in \mathcal{RH}_2$ and let $\lambda_1, \ldots, \lambda_p$ be the poles of G and $\mu_1, \ldots, \mu_q$ be the poles of H. Then it holds that

$$\langle G, H \rangle_{\mathcal{H}_2} = \sum_{j=1}^{q} \operatorname{res}(G(-\cdot)H(\cdot), \mu_j) = \sum_{j=1}^{p} \operatorname{res}(H(-\cdot)G(\cdot), \lambda_j).$$

If μ_j is a simple pole of H, then it holds that

$$\operatorname{res}(G(-\cdot)H(\cdot), \mu_j) = G(-\mu_j) \operatorname{res}(H, \mu_j).$$

Moreover, if μ_j is a double pole of H, then we have

$$\operatorname{res}(G(-\cdot)H(\cdot), \mu_j) = G(-\mu_j) \operatorname{res}(H, \mu_j) - G'(-\mu_j) h_0(\mu_j),$$

where $h_0(\mu_j) = \lim_{s \to \mu_j} \left((s-\mu_j)^2 H(s) \right)$.

Proof. First note that the function $G(-s)H(s)$ may have poles only in $\mu_1, \ldots, \mu_q$ and $-\lambda_1, \ldots, -\lambda_p$. For any $R > 0$ define the semicircular contour

$$\Gamma_R := \underbrace{\{z \mid z = \mathrm{i}\omega \text{ with } \omega \in [-R, R]\}}_{=:\Gamma_{1,R}} \cup \underbrace{\left\{z \;\middle|\; z = R\mathrm{e}^{\mathrm{i}\theta} \text{ with } \theta \in \left[\tfrac{\pi}{2}, \tfrac{3\pi}{2}\right]\right\}}_{=:\Gamma_{2,R}}.$$

For sufficiently large R, all poles of H are inside Γ_R. Since $G, H \in \mathcal{RH}_2$, it holds that $\lim_{|s|\to\infty} G(s) = \lim_{|s|\to\infty} H(s) = 0$, and thus we have $G(-s)H(s) = \mathcal{O}\big(|s|^{-2}\big)$ for $|s| \to \infty$. Therefore, $G(-s)H(s)$ decays faster then the length of the contour $\Gamma_{2,R}$ grows for $R \to \infty$ and thus we have

$$\lim_{R\to\infty} \int_{\Gamma_{2,R}} G(-s)H(s)\mathrm{d}s = 0.$$

By using the residue theorem, we obtain

$$\begin{aligned}
\langle G, H\rangle_{\mathcal{H}_2} &= \frac{1}{2\pi}\int_{-\infty}^{\infty} G(-\mathrm{i}\omega)H(\mathrm{i}\omega)\mathrm{d}\omega \\
&= \frac{1}{2\pi\mathrm{i}} \lim_{R\to\infty} \int_{\Gamma_R} G(-s)H(s)\mathrm{d}s \\
&= \sum_{j=1}^{q} \operatorname{res}(G(-\cdot)H(\cdot), \mu_j).
\end{aligned}$$

(Note that if a pole μ_j of H is canceled by a zero of $G(-\cdot)$, then the above formula is still correct, since then $\operatorname{res}(G(-\cdot)H(\cdot), \mu_j) = 0$.) If μ_j is a simple pole of H, then if $G(-\mu_j) \neq 0$, it is also a simple pole of $G(-\cdot)H(\cdot)$ and we get

$$\begin{aligned}
\operatorname{res}(G(-\cdot)H(\cdot), \mu_j) &= \lim_{s\to\mu_j} (s - \mu_j)G(-s)H(s) \\
&= G(-\mu_j) \lim_{s\to\mu_j} (s - \mu_j)H(s) = G(-\mu_j)\operatorname{res}(H, \mu_j).
\end{aligned}$$

On the other hand, if $G(-\mu_j) = 0$, then $G(-\cdot)H(\cdot)$ has no pole at μ_j and

$$0 = \operatorname{res}(G(-\cdot)H(\cdot), \mu_j) = G(-\mu_j)\operatorname{res}(H, \mu_j).$$

If μ_j is a double pole of H, then $G(-\cdot)H(\cdot)$ has no pole, a simple pole, or a double pole at μ_j. With

$$\begin{aligned}
H(s) &= \frac{H_{-2}}{(s - \mu_j)^2} + \frac{H_{-1}}{s - \mu_j} + \ldots, \\
G(-s) &= G_0 + G_1(s - \mu_j) + \ldots,
\end{aligned}$$

where $G_0 = G(-\mu_j)$ and $G_1 = -G'(-\mu_j)$, it holds that

$$\begin{aligned}
&\operatorname{res}(G(-\cdot)H(\cdot), \mu_j) \\
&= G_0 \cdot H_{-1} + G_1 \cdot H_{-2} \\
&= G(-\mu_j) \lim_{s \to \mu_j} \frac{\mathrm{d}}{\mathrm{d}s} \left((s-\mu_j)^2 H(s)\right) - G'(-\mu_j) \lim_{s \to \mu_j} (s-\mu_j)^2 H(s) \\
&= G(-\mu_j) \operatorname{res}(H, \mu_j) - G'(-\mu_j) h_0(\mu_j).
\end{aligned}$$

□

Now we get back to the optimization problem. Assume for simplicity that $G \in \mathcal{RH}_2$ (The case of non-strictly proper G can be handled similarly.) The question arises how we can check whether $\widetilde{G} \in \mathcal{RH}_2$ with McMillan degree ℓ locally minimizes $\|G - \widetilde{G}\|_{\mathcal{H}_2}$. First we attempt to answer an easier question: Is there a $\widehat{G} \in \mathcal{RH}_2$ of McMillan degree ℓ with the same poles as $\widetilde{G}$ that yields a smaller $\mathcal{H}_2$-error norm?

Theorem 6.14: Let $\boldsymbol{\mu} := \{\mu_1, \ldots, \mu_\ell\} \subset \mathbb{C}^-$ be closed under complex conjugation. We define $\mathcal{M}(\boldsymbol{\mu})$ to be the set of all functions in $\mathcal{RH}_2$ of McMillan degree at most ℓ whose poles are simple and contained in the set $\boldsymbol{\mu}$. Then there are the following facts:

a) If $H \in \mathcal{M}(\boldsymbol{\mu})$, then H is the transfer function of a stable and minimal LTI system with state-space dimension at most ℓ.

b) The set $\mathcal{M}(\boldsymbol{\mu})$ is an ℓ-dimensional closed subspace of $\mathcal{H}_2$.

c) The function $\widehat{G}$ solves the minimization problem

$$\|G - \widehat{G}\|_{\mathcal{H}_2} = \min_{H \in \mathcal{M}(\boldsymbol{\mu})} \|G - H\|_{\mathcal{H}_2}, \tag{6.2}$$

if and only if

$$\left\langle G - \widehat{G}, H \right\rangle_{\mathcal{H}_2} = 0 \quad \text{for all } H \in \mathcal{M}(\boldsymbol{\mu}).$$

Furthermore, the solution $\widehat{G}$ of the minimization problem (6.2) exists and is unique.

Proof. Statements a) and b) are easy to show (exercise!).

We only discuss statement c). Let $\widehat{G} \in \mathcal{M}(\boldsymbol{\mu})$ be such that $\left\langle G - \widehat{G}, H \right\rangle_{\mathcal{H}_2} = 0$

for all $H \in \mathcal{M}(\boldsymbol{\mu})$. Then for $H \in \mathcal{M}(\boldsymbol{\mu})$ we get

$$\begin{aligned}\|G-H\|_{\mathcal{H}_2}^2 &= \|(G-\widehat{G})+(\widehat{G}-H)\|_{\mathcal{H}_2}^2 \\ &= \|G-\widehat{G}\|_{\mathcal{H}_2}^2 + \|\widehat{G}-H\|_{\mathcal{H}_2}^2 + 2\langle G-\widehat{G}, \widehat{G}-H\rangle_{\mathcal{H}_2} \\ &= \|G-\widehat{G}\|_{\mathcal{H}_2}^2 + \|\widehat{G}-H\|_{\mathcal{H}_2}^2,\end{aligned}$$

where the latter equality follows from $\widehat{G}-H \in \mathcal{M}(\boldsymbol{\mu})$. This proves that $\widehat{G}$ is the unique minimizer.

Let $\widehat{G}$ be a minimizer of (6.2), which is in $\mathcal{M}(\boldsymbol{\mu})$ by the closedness property. Then we get

$$\|(\widehat{G}+\varepsilon H)-G\|_{\mathcal{H}_2}^2 - \|G-\widehat{G}\|_{\mathcal{H}_2}^2 = 2\varepsilon\langle \widehat{G}-G, H\rangle_{\mathcal{H}_2} + \varepsilon^2\|H\|_{\mathcal{H}_2}^2 \geqslant 0 \quad (6.3)$$

If $\langle \widehat{G}-G, H\rangle_{\mathcal{H}_2} \neq 0$ for some $H \in \mathcal{M}(\boldsymbol{\mu})$, then for sufficiently small $\varepsilon > 0$ we have $\left|2\varepsilon\langle \widehat{G}-G, H\rangle_{\mathcal{H}_2}\right| > \varepsilon^2\|H\|_{\mathcal{H}_2}^2$. However, this is a contradiction to the nonnegativity of (6.3). Therefore, we have $\langle \widehat{G}-G, H\rangle_{\mathcal{H}_2} = 0$ for all $H \in \mathcal{M}(\boldsymbol{\mu})$. □

By Lemma 6.13 we obtain

$$0 = \langle G-\widehat{G}, H\rangle_{\mathcal{H}_2} = \sum_{j=1}^{\ell} \big(G(-\mu_j)-\widehat{G}(-\mu_j)\big)\operatorname{res}(H,\mu_j)$$

for all $H \in \mathcal{M}(\boldsymbol{\mu})$, which is equivalent to

$$G(-\mu_j) = \widehat{G}(-\mu_j), \quad j = 1, 2, \ldots, \ell.$$

So in other words, $\widehat{G}$ has to interpolate G at the mirror images of its own poles. The next question is whether the poles $\mu_1, \mu_2, \ldots, \mu_\ell$ are optimal.

Definition 6.15: Let $G \in \mathcal{RH}_2$ be given and let $\delta(\cdot)$ denote the McMillan degree of a rational function. Then the function $\widetilde{G} \in \mathcal{RH}_2$ is a *local minimizer* of the minimization problem

$$\min_{H\in\mathcal{RH}_2,\,\delta(H)=\ell} \|G-H\|_{\mathcal{H}_2}, \quad (6.4)$$

if

$$\|G-\widetilde{G}\|_{\mathcal{H}_2} \leqslant \|G-H\|_{\mathcal{H}_2}$$

is satisfied for all $H \in \mathcal{RH}_2$ with $\delta(H) = \ell$ and $\|\widetilde{G}-H\|_{\mathcal{H}_2} \leqslant \varepsilon$ for some $\varepsilon > 0$.

We will now derive necessary optimality conditions for $\widetilde{G}$ being a local minimizer of the minimization problem (6.4).

Theorem 6.16: If $\widetilde{G} \in \mathcal{RH}_2$ is a local minimizer of (6.4) and $\widetilde{G}$ has only simple poles $\boldsymbol{\mu} := \{\mu_1, \ldots, \mu_\ell\}$ (closed under complex conjugation), then it holds that

$$\left\langle G - \widetilde{G}, \widetilde{G} \cdot H_1 + H_2 \right\rangle_{\mathcal{H}_2} = 0$$

for all $H_1, H_2 \in \mathcal{M}(\boldsymbol{\mu})$.

Proof. Suppose that $\widetilde{G}_\varepsilon \in \mathcal{RH}_2$ with $\delta(\widetilde{G}_\varepsilon) = \ell$ and $\|\widetilde{G} - \widetilde{G}_\varepsilon\|_{\mathcal{H}_2} \leqslant c \cdot \varepsilon$ for a sufficiently small $\varepsilon > 0$ and some constant $c > 0$. Then it holds that

$$\begin{aligned}
\|G - \widetilde{G}\|_{\mathcal{H}_2}^2 &\leqslant \|G - \widetilde{G}_\varepsilon\|_{\mathcal{H}_2}^2 \\
&= \|(G - \widetilde{G}) + (\widetilde{G} - \widetilde{G}_\varepsilon)\|_{\mathcal{H}_2}^2 \\
&= \|G - \widetilde{G}\|_{\mathcal{H}_2}^2 + 2\left\langle G - \widetilde{G}, \widetilde{G} - \widetilde{G}_\varepsilon \right\rangle_{\mathcal{H}_2} + \|\widetilde{G} - \widetilde{G}_\varepsilon\|_{\mathcal{H}_2}^2.
\end{aligned}$$

This implies that for all sufficiently small $\varepsilon > 0$ we have that

$$0 \leqslant 2\left\langle G - \widetilde{G}, \widetilde{G} - \widetilde{G}_\varepsilon \right\rangle_{\mathcal{H}_2} + \|\widetilde{G} - \widetilde{G}_\varepsilon\|_{\mathcal{H}_2}^2. \tag{6.5}$$

Let $\mu_1, \ldots, \mu_\ell$ be ordered such that $\mu_1, \mu_2, \ldots, \mu_{\ell_R}$ are the real poles and $\mu_j = \alpha_j \pm \mathrm{i}\beta_j$ with $\beta_j > 0$ for $j = \ell_R + 1, \ldots, \ell_R + \ell_C$ are the nonreal poles. Then any $H_1 \in \mathcal{M}(\boldsymbol{\mu})$ can be written as a partial fraction expansion as

$$H_1(s) = \sum_{j=1}^{\ell_R} \frac{\gamma_j}{s - \mu_j} + \sum_{j=\ell_R+1}^{\ell_R+\ell_C} \frac{\rho_j(s - \alpha_j) + \tau_j}{(s - \alpha_j)^2 + \beta_j^2}$$

for some $\gamma_j, \rho_j, \tau_j \in \mathbb{R}$. Thus we get

$$\begin{aligned}
\left\langle G - \widetilde{G}, \widetilde{G} \cdot H_1 + H_2 \right\rangle_{\mathcal{H}_2} = \sum_{j=1}^{\ell_R} \gamma_j \left\langle G - \widetilde{G}, \frac{\widetilde{G}}{\cdot - \mu_j} \right\rangle_{\mathcal{H}_2} \\
+ \sum_{j=\ell_R+1}^{\ell_R+\ell_C} \rho_j \left\langle G - \widetilde{G}, \frac{(\cdot - \alpha_j)\widetilde{G}}{(\cdot - \alpha_j)^2 + \beta_j^2} \right\rangle_{\mathcal{H}_2} \\
+ \sum_{j=\ell_R+1}^{\ell_R+\ell_C} \tau_j \left\langle G - \widetilde{G}, \frac{\widetilde{G}}{(\cdot - \alpha_j)^2 + \beta_j^2} \right\rangle_{\mathcal{H}_2} + \left\langle G - \widetilde{G}, H_2 \right\rangle_{\mathcal{H}_2}.
\end{aligned}$$

By Theorem 6.14, we have $\left\langle G - \widetilde{G}, H_2 \right\rangle_{\mathcal{H}_2} = 0$. Now we show that also the other summands vanish. Assume that for some $j \in \{1, 2, \ldots, \ell_R\}$ we have that

$$\left\langle G - \widetilde{G}, \frac{\widetilde{G}}{\cdot - \mu_j} \right\rangle_{\mathcal{H}_2} \neq 0.$$

We write $\widetilde{G}(s) = \frac{p(s)}{(s-\mu_j)q(s)}$ for some polynomials $p(s)$, $q(s) \in \mathbb{R}[s]$ and define

$$\widetilde{G}_\varepsilon(s) := \frac{p(s)}{(s-\mu_j-(\pm\varepsilon))q(s)},$$

where the sign of $\pm\varepsilon$ is chosen to match the one of $\left\langle G-\widetilde{G}, \frac{\widetilde{G}}{\cdot-\mu_j}\right\rangle_{\mathcal{H}_2}$. Then we have

$$\widetilde{G}_\varepsilon(s) = \widetilde{G}(s) \pm \varepsilon \frac{p(s)}{(s-\mu_j)^2 q(s)} + \mathcal{O}(\varepsilon^2),$$

which leads to

$$\widetilde{G}(s) - \widetilde{G}_\varepsilon(s) = \mp\varepsilon \frac{\widetilde{G}(s)}{s-\mu_j} + \mathcal{O}(\varepsilon^2)$$

and

$$\left\langle G-\widetilde{G}, \widetilde{G}-\widetilde{G}_\varepsilon\right\rangle_{\mathcal{H}_2} = -\varepsilon \underbrace{\left|\left\langle G-\widetilde{G}, \frac{\widetilde{G}}{\cdot-\mu_j}\right\rangle_{\mathcal{H}_2}\right|}_{=:\widetilde{c}>0} + \mathcal{O}(\varepsilon^2).$$

Together with (6.5) we get

$$0 \leqslant -2\widetilde{c}\varepsilon + c^2\varepsilon^2 + \mathcal{O}(\varepsilon^2) = -2\widetilde{c}\varepsilon + \mathcal{O}(\varepsilon^2).$$

But this is a contradiction for sufficiently small $\varepsilon > 0$.

Now write $\widetilde{G}(s) = \frac{p(s)}{((s-\alpha_j)^2+\beta_j^2)q(s)}$ for some $j \in \{\ell_R+1, \ell_R+2, \ldots, \ell_R+\ell_C\}$ and $p(s)$, $q(s) \in \mathbb{R}[s]$. In an analogous fashion, it can be shown that

$$\left\langle G-\widetilde{G}, \frac{(\cdot-\alpha_j)\widetilde{G}}{(\cdot-\alpha_j)^2+\beta_j^2}\right\rangle_{\mathcal{H}_2} = 0,$$
$$\left\langle G-\widetilde{G}, \frac{\widetilde{G}}{(\cdot-\alpha_j)^2+\beta_j^2}\right\rangle_{\mathcal{H}_2} = 0,$$

by using

$$\widetilde{G}_\varepsilon(s) = \frac{p(s)}{((s-\alpha_j-(\pm\varepsilon))^2+\beta_j^2)q(s)} \quad \text{and}$$
$$\widetilde{G}_\varepsilon(s) = \frac{p(s)}{((s-\alpha_j)^2+\beta_j^2-(\pm\varepsilon))q(s)},$$

respectively. From this we can conclude the claim of the theorem. □

The question arises whether we can interpret the above theorem as interpolation conditions. With Lemma 6.13 we obtain

$$0 = \left\langle G-\widetilde{G}, \widetilde{G}\cdot H_1\right\rangle_{\mathcal{H}_2} = \sum_{j=1}^{\ell} \operatorname{res}\left(\left(G(-\cdot)-\widetilde{G}(-\cdot)\right)\widetilde{G}(\cdot)H_1(\cdot), \mu_j\right).$$

If H_1 has exactly the same poles as $\widetilde{G}$, then μ_j is a double pole of $\widetilde{G} \cdot H_1$ (recall that $\widetilde{G},\ H_1 \in \mathcal{M}(\boldsymbol{\mu})$) and $G(-\mu_j) = \widetilde{G}(-\mu_j)$ for all $j = 1,\ 2,\ \ldots,\ \ell$. This gives

$$\begin{aligned} 0 &= -\sum_{j=1}^{\ell} \left(G'(-\mu_j) - \widetilde{G}'(-\mu_j)\right) \left(\lim_{s \to \mu_j} (s-\mu_j)^2 \widetilde{G}(s) H_1(s)\right) \\ &= -\sum_{j=1}^{\ell} \left(G'(-\mu_j) - \widetilde{G}'(-\mu_j)\right) \operatorname{res}\left(\widetilde{G}, \mu_j\right) \operatorname{res}(H_1, \mu_j). \end{aligned}$$

Since this is true for all $H_1 \in \mathcal{M}(\boldsymbol{\mu})$, we obtain that

$$G'(-\mu_j) = \widetilde{G}'(-\mu_j), \quad j = 1,\ 2,\ \ldots,\ \ell.$$

The analysis carried out above can also be generalized to deal with transfer functions $G(s) \in \mathbb{R}(s)^{p \times m}$ of MIMO systems. This will lead to *tangential interpolation conditions*, meaning that the original and the reduced transfer function moments only match in certain tangential directions. We will not discuss this in detail here, but summarize the result in the following theorem.

Theorem 6.17: Let an asymptotically stable system $[A, B, C, D] \in \Sigma_{n,m,p}$ with transfer function $G \in \mathcal{RH}_\infty^{p \times m}$ be given. Let

$$\widetilde{G}(s) = \sum_{j=1}^{\ell} \frac{1}{s - \mu_j} \widehat{c}_j \widehat{b}_j^{\mathsf{H}} + \widetilde{D}$$

with $\widehat{b}_j \in \mathbb{C}^m$ and $\widehat{c}_j \in \mathbb{C}^p$, $j = 1,\ 2,\ \ldots,\ \ell$ be a local minimizer of the minimization problem

$$\min_{H \in \mathcal{RH}_\infty^{p \times m},\, \delta(H) = \ell} \left\| G - H \right\|_{\mathcal{H}_2}.$$

Then the following statements are satisfied:

a) It holds that $D = \widetilde{D}$.

b) It holds that

$$G(-\mu_j)\widehat{b}_j = \widetilde{G}(-\mu_j)\widehat{b}_j, \quad \widehat{c}_j^{\mathsf{H}} G(-\mu_j) = \widehat{c}_j^{\mathsf{H}} \widetilde{G}(-\mu_j), \quad j = 1, \ldots, \ell.$$

c) It holds that

$$\widehat{c}_j^{\mathsf{H}} G'(-\mu_j)\widehat{b}_j = \widehat{c}_j^{\mathsf{H}} \widetilde{G}'(-\mu_j)\widehat{b}_j, \quad j = 1,\ \ldots,\ \ell.$$

Now we want to put this into an algorithm. We want to compute a state-space realization of the reduced transfer function $\widetilde{G}$ without explicitly computing it. Given the interpolation points $\widehat{s}_j$ and the tangential directions $\widehat{b}_j$ and $\widehat{c}_j$ for $j =$

$1, 2, \ldots, \ell$, we can construct a reduced-order model such that the tangential interpolation conditions

$$G(\hat{s}_j)\hat{b}_j = \tilde{G}(\hat{s}_j)\hat{b}_j, \quad \hat{c}_j^{\mathsf{H}} G(\hat{s}_j) = \hat{c}_j^{\mathsf{H}} \tilde{G}(\hat{s}_j), \quad j = 1, \ldots, \ell$$

and

$$\hat{c}_j^{\mathsf{H}} G'(\hat{s}_j)\hat{b}_j = \hat{c}_j^{\mathsf{H}} \tilde{G}'(\hat{s}_j)\hat{b}_j, \quad j = 1, \ldots, \ell.$$

are satisfied. We do this by two-sided interpolation with projection matrices $W, T \in \mathbb{C}^{n\times\ell}$ with $W^{\mathsf{H}}T = I_\ell$ and

$$\begin{aligned}
\operatorname{im} T &= \operatorname{im}\left[(\hat{s}_1 I_n - A)^{-1}B\hat{b}_1 \quad \ldots \quad (\hat{s}_\ell I_n - A)^{-1}B\hat{b}_\ell\right], \\
\operatorname{im} W &= \operatorname{im}\left[(\hat{s}_1 I_n - A)^{-\mathsf{H}}C^{\mathsf{H}}\hat{c}_1 \quad \ldots \quad (\hat{s}_\ell I_n - A)^{-\mathsf{H}}C^{\mathsf{H}}\hat{c}_\ell\right].
\end{aligned}$$

On the other hand, if a reduced-order system $\left[\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\right] \in \Sigma_{\ell,m,p}$ is given, we compute new interpolation data as follows: Assume that $\tilde{A}$ is diagonalizable, i. e., there exists an invertible matrix $X \in \mathbb{C}^{\ell\times\ell}$ such that

$$X^{-1}\tilde{A}X = M = \operatorname{diag}(\mu_1, \ldots, \mu_\ell).$$

Then we have

$$\tilde{G}(s) = \tilde{C}\big(sI_\ell - \tilde{A}\big)^{-1}\tilde{B} + \tilde{D} = \tilde{C}X(sI_\ell - M)^{-1}X^{-1}\tilde{B} + \tilde{D}.$$

With

$$\begin{bmatrix} \hat{b}_1^{\mathsf{H}} \\ \vdots \\ \hat{b}_\ell^{\mathsf{H}} \end{bmatrix} := X^{-1}\tilde{B}, \quad \begin{bmatrix} \hat{c}_1 & \ldots & \hat{c}_\ell \end{bmatrix} := \tilde{C}X$$

we then obtain

$$\tilde{G}(s) = \sum_{j=1}^{\ell} \frac{1}{s - \mu_j}\hat{c}_j\hat{b}_j^{\mathsf{H}} + \tilde{D}.$$

From this we can construct a new model as above that fulfills the tangential interpolation conditions at the data $\big(-\mu_j, \hat{b}_j, \hat{c}_j\big)$. This leads to an algorithm that alternates between the construction of a reduced-order model from interpolation data and the generation of new interpolation data from a given reduced-order model. If this iteration converges, then the tangential interpolation conditions are satisfied. This can be interpreted as a fixed-point iteration and results in Algorithm 6.2. The algorithm is usually terminated if the change in the pole locations between two consecutive iterations is below a given tolerance, or if the distance (e. g., measured in the $\mathcal{H}_2$-norm) of the reduced transfer functions between two consecutive iterations is sufficiently small.

Algorithm 6.2 Iterative rational Krylov algorithm (IRKA)

Input: Asymptotically stable $[A, B, C, D] \in \Sigma_{n,m,p}$, desired reduced order ℓ.

Output: Reduced-order model $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right] \in \Sigma_{\ell,m,p}$.

1: Choose initial interpolation data $\left(\mu_j, \widehat{b}_j, \widehat{c}_j\right)$, $j = 1, \ldots, \ell$.

2: **while** not converged **do**

3: Construct the model $\left[\widetilde{A}, \widetilde{B}, \widetilde{C}, \widetilde{D}\right] := [W^{\mathsf{H}}AT, W^{\mathsf{H}}B, CT, D] \in \Sigma_{\ell,m,p}$ where $W, T \in \mathbb{C}^{n\times\ell}$ with $W^{\mathsf{H}}T = I_\ell$ and

$$\operatorname{im} T = \operatorname{im}\left[(-\mu_1 I_n - A)^{-1}B\widehat{b}_1 \quad \ldots \quad (-\mu_\ell I_n - A)^{-1}B\widehat{b}_\ell\right],$$
$$\operatorname{im} W = \operatorname{im}\left[(-\mu_1 I_n - A)^{-\mathsf{H}}C^{\mathsf{H}}\widehat{c}_1 \quad \ldots \quad (-\mu_\ell I_n - A)^{-\mathsf{H}}C^{\mathsf{H}}\widehat{c}_\ell\right].$$

4: Compute new interpolation data $\left(\mu_j, \widehat{b}_j, \widehat{c}_j\right)$, $j = 1, \ldots, \ell$: Compute an invertible matrix $X \in \mathbb{C}^{\ell\times\ell}$ such that $X^{-1}\widetilde{A}X = \operatorname{diag}(\mu_1, \ldots, \mu_\ell)$ and set

$$\begin{bmatrix}\widehat{b}_1^{\mathsf{H}}\\ \vdots \\ \widehat{b}_\ell^{\mathsf{H}}\end{bmatrix} := X^{-1}\widetilde{B}, \quad \begin{bmatrix}\widehat{c}_1 & \ldots & \widehat{c}_\ell\end{bmatrix} := \widetilde{C}X, \quad \widetilde{D} := D.$$

5: **end while**

Remark 6.18: a) The initial interpolation data is either chosen randomly or by a cheaply computable reduced-order model such as a partial realization.

b) Stability of intermediate reduced-order models is not guaranteed. Unstable intermediate systems may occur, in particular, if the initial interpolation data is far away of the optimal one.

c) If the full-order model contains only real matrices, then also a real reduced-order model can be obtained by choosing the interpolation data such that it is closed under complex conjugation and by a realification of the projection matrices W and T by noting that

$$\operatorname{im}\left[(-\mu I_n - A)^{-1}B\widehat{b} \quad (-\overline{\mu} I_n - A)^{-1}B\overline{\widehat{b}}\right]$$
$$= \operatorname{im}\left[\operatorname{Re}\left((-\mu I_n - A)^{-1}B\widehat{b}\right) \quad \operatorname{Im}\left((-\mu I_n - A)^{-1}B\widehat{b}\right)\right].$$

d) Convergence of IRKA cannot be guaranteed, counter-examples are known.

e) In case of convergence, the resulting system may only be locally optimal, but in many cases, it even converges to a very good local or even global optimizer in only a few iterations. Unfortunately, a complete convergence

analysis of the method it still an open problem. Partial results are only known for symmetric systems.

f) The reduced-order models obtained by IRKA are usually competitive to those obtained by balanced truncation.

6.3 Interpolation from Data: The Loewner Framework

In this section we discuss another framework for interpolation that is solely based on data obtained from the system under consideration, see [MA07]. In many situations it may be the case that a model is not available, but only some input/output data may given. Then we seek for a model of low order that best reproduces this data. Note that this process is rather called *reduced-order modeling* instead of model reduction, since no model is given. The methods discussed here are related to the field of system identification, where it is the goal to determine a model or its parameters from measurements, to validate it and to assess its robustness with respect to uncertainties in the data via statistical considerations.

Assume that we have given *right* or *column data* $(\lambda_i,\, r_i,\, w_i) \in \mathbb{C} \times \mathbb{C}^m \times \mathbb{C}^p$, $i = 1,\, 2,\, \ldots,\, k$, and *left* or *row data* $(\mu_j,\, \ell_j,\, v_j) \in \mathbb{C} \times \mathbb{C}^p \times \mathbb{C}^m$, $j = 1,\, 2,\, \ldots,\, q$. We seek a function $\widetilde{G}(s) \in \mathbb{C}(s)^{p\times m}$ such that

$$\begin{aligned} \widetilde{G}(\lambda_i) r_i &= w_i, \quad i = 1,\, 2,\, \ldots,\, k, \\ \ell_j^{\mathsf{H}} \widetilde{G}(\mu_j) &= v_j^{\mathsf{H}}, \quad j = 1,\, 2,\, \ldots,\, q. \end{aligned}$$

For simplicity we assume that $\{\lambda_1,\, \ldots,\, \lambda_k\} \cap \{\mu_1,\, \ldots,\, \mu_q\} = \varnothing$. We reorganize the right data as

$$\begin{aligned} \Lambda &:= \operatorname{diag}(\lambda_1,\, \ldots,\, \lambda_k) \in \mathbb{C}^{k\times k}, \\ R &:= \begin{bmatrix} r_1 & r_2 & \ldots & r_k \end{bmatrix} \in \mathbb{C}^{m\times k}, \\ W &:= \begin{bmatrix} w_1 & w_2 & \ldots & w_k \end{bmatrix} \in \mathbb{C}^{p\times k}, \end{aligned} \tag{6.6}$$

and the left data as

$$\begin{aligned} M &:= \operatorname{diag}(\mu_1,\, \ldots,\, \mu_q) \in \mathbb{C}^{q\times q}, \\ L^{\mathsf{H}} &:= \begin{bmatrix} \ell_1 & \ell_2 & \ldots & \ell_q \end{bmatrix} \in \mathbb{C}^{p\times q}, \\ V^{\mathsf{H}} &:= \begin{bmatrix} v_1 & v_2 & \ldots & v_q \end{bmatrix} \in \mathbb{C}^{m\times q}. \end{aligned} \tag{6.7}$$

From the data we construct the *Loewner matrix*

$$\mathbb{L} := \begin{bmatrix} \frac{v_1^{\mathsf{H}} r_1 - \ell_1^{\mathsf{H}} w_1}{\mu_1 - \lambda_1} & \cdots & \frac{v_1^{\mathsf{H}} r_k - \ell_1^{\mathsf{H}} w_k}{\mu_1 - \lambda_k} \\ \vdots & \ddots & \vdots \\ \frac{v_q^{\mathsf{H}} r_1 - \ell_q^{\mathsf{H}} w_1}{\mu_q - \lambda_1} & \cdots & \frac{v_q^{\mathsf{H}} r_k - \ell_q^{\mathsf{H}} w_k}{\mu_q - \lambda_k} \end{bmatrix} \in \mathbb{C}^{q \times k}, \tag{6.8}$$

and the *shifted Loewner matrix*

$$\mathbb{L}_\sigma := \begin{bmatrix} \frac{\mu_1 v_1^{\mathsf{H}} r_1 - \ell_1^{\mathsf{H}} w_1 \lambda_1}{\mu_1 - \lambda_1} & \cdots & \frac{\mu_1 v_1^{\mathsf{H}} r_k - \ell_1^{\mathsf{H}} w_k \lambda_k}{\mu_1 - \lambda_k} \\ \vdots & \ddots & \vdots \\ \frac{\mu_q v_q^{\mathsf{H}} r_1 - \ell_q^{\mathsf{H}} w_1 \lambda_1}{\mu_q - \lambda_1} & \cdots & \frac{\mu_q v_q^{\mathsf{H}} r_k - \ell_q^{\mathsf{H}} w_k \lambda_k}{\mu_q - \lambda_k} \end{bmatrix} \in \mathbb{C}^{q \times k}. \tag{6.9}$$

The matrix pencil $s\mathbb{L} - \mathbb{L}_\sigma \in \mathbb{C}[s]^{q \times k}$ is also called the *Loewner pencil*.

Lemma 6.19: With the notation introduced in (6.6), (6.7), (6.8), and (6.9), the two Sylvester equations

$$\mathbb{L}\Lambda - M\mathbb{L} = LW - VR \tag{6.10}$$

and

$$\mathbb{L}_\sigma\Lambda - M\mathbb{L}_\sigma = LW\Lambda - MVR \tag{6.11}$$

are satisfied.

Proof. By denoting with $[\cdot]_{ij}$, the (i,j)-th entry of the matrix in the brackets, we obtain that

$$\begin{aligned} [\mathbb{L}\Lambda - M\mathbb{L}]_{ij} &= \frac{v_i^{\mathsf{H}} r_j - \ell_i^{\mathsf{H}} w_j}{\mu_i - \lambda_j}\lambda_j - \mu_i \frac{v_i^{\mathsf{H}} r_j - \ell_i^{\mathsf{H}} w_j}{\mu_i - \lambda_j} \\ &= \frac{v_i^{\mathsf{H}} r_j \lambda_j - \ell_i^{\mathsf{H}} w_j \lambda_j - \mu_i v_i^{\mathsf{H}} r_j + \mu_i \ell_i^{\mathsf{H}} w_j}{\mu_i - \lambda_j} \\ &= \ell_i^{\mathsf{H}} w_j - v_i^{\mathsf{H}} r_j = [LW - VR]_{ij}. \end{aligned}$$

Moreover, we have

$$\begin{aligned} [\mathbb{L}_\sigma\Lambda - M\mathbb{L}_\sigma]_{ij} &= \frac{\mu_i v_i^{\mathsf{H}} r_j - \ell_i^{\mathsf{H}} w_j \lambda_j}{\mu_i - \lambda_j}\lambda_j - \mu_i \frac{\mu_i v_i^{\mathsf{H}} r_j - \ell_i^{\mathsf{H}} w_j \lambda_j}{\mu_i - \lambda_j} \\ &= \frac{\mu_i v_i^{\mathsf{H}} r_j \lambda_j - \ell_i^{\mathsf{H}} w_j \lambda_j^2 - \mu_i^2 v_i^{\mathsf{H}} r_j + \mu_i \ell_i^{\mathsf{H}} w_j \lambda_j}{\mu_i - \lambda_j} \\ &= \ell_i^{\mathsf{H}} w_j \lambda_j - \mu_i v_i^{\mathsf{H}} r_j = [LW\Lambda - MVR]_{ij}. \end{aligned}$$

□

Remark 6.20: Let the data be sampled from a (regular) linear descriptor system

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}Ex(t) &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) \end{aligned}$$

with transfer function $G(s) = C(sE - A)^{-1}B$ and define the matrices

$$\begin{aligned} \mathcal{C}_k &:= \begin{bmatrix} (\lambda_1 E - A)^{-1}Br_1 & \dots & (\lambda_k E - A)^{-1}Br_k \end{bmatrix} \in \mathbb{C}^{n\times k}, \\ \mathcal{O}_q &:= \begin{bmatrix} \ell_1^{\mathsf{H}} C(\mu_1 E - A)^{-1} \\ \vdots \\ \ell_q^{\mathsf{H}} C(\mu_q E - A)^{-1} \end{bmatrix} \in \mathbb{C}^{q\times n}, \end{aligned}$$

which are called *generalized tangential controllability matrices* and *generalized tangential observability matrices*, respectively. With these, it follows that

$$\begin{aligned} [\mathbb{L}]_{ij} &= \frac{v_i^{\mathsf{H}} r_j - \ell_i^{\mathsf{H}} w_j}{\mu_i - \lambda_j} \\ &= \frac{\ell_i^{\mathsf{H}} G(\mu_i) r_j - \ell_i^{\mathsf{H}} G(\lambda_j) r_j}{\mu_i - \lambda_j} \\ &= \frac{\ell_i^{\mathsf{H}} C\left((\mu_i E - A)^{-1} - (\lambda_j E - A)^{-1}\right) Br_j}{\mu_i - \lambda_j} \\ &= \frac{\ell_i^{\mathsf{H}} C(\mu_i E - A)^{-1}((\lambda_j E - A) - (\mu_i E - A))(\lambda_j E - A)^{-1} Br_j}{\mu_i - \lambda_j} \\ &= -\ell_i^{\mathsf{H}} C(\mu_i E - A)^{-1} E(\lambda_j E - A)^{-1} Br_j, \end{aligned}$$

and similarly

$$\begin{aligned} [\mathbb{L}_\sigma]_{ij} &= \frac{\mu_i v_i^{\mathsf{H}} r_j - \lambda_j \ell_i^{\mathsf{H}} w_j}{\mu_i - \lambda_j} \\ &= \frac{\mu_i \ell_i^{\mathsf{H}} G(\mu_i) r_j - \lambda_j \ell_i^{\mathsf{H}} G(\lambda_j) r_j}{\mu_i - \lambda_j} \\ &= -\ell_i^{\mathsf{H}} C(\mu_i E - A)^{-1} A(\lambda_j E - A)^{-1} Br_j. \end{aligned}$$

This gives

$$\mathbb{L} = -\mathcal{O}_q E \mathcal{C}_k, \quad \mathbb{L}_\sigma = -\mathcal{O}_q A \mathcal{C}_k.$$

Now we state and proof a result that gives us the structure of the dynamical system that interpolates the data.

Theorem 6.21: Assume that there is given data as in (6.6), (6.7) with $k = q$. Let the associated Loewner pencil $s\mathbb{L} - \mathbb{L}_\sigma$ be regular (i. e., $\det(s\mathbb{L} - \mathbb{L}_\sigma)$ is not the zero polynomial) and assume that no λ_i, $i = 1, 2, \ldots, k$ and μ_j, $j = 1, 2, \ldots, q$ is an eigenvalue of the pencil $s\mathbb{L} - \mathbb{L}_\sigma$. Then

$$\widetilde{E} = -\mathbb{L}, \quad \widetilde{A} = -\mathbb{L}_\sigma, \quad \widetilde{B} = V, \quad \widetilde{C} = W$$

is an interpolating descriptor system realization, i. e., the function $\widetilde{G}(s) := \widetilde{C}\big(s\widetilde{E} - \widetilde{A}\big)^{-1}\widetilde{B}$ interpolates the given data.

Proof. For the proof we make use of the Sylvester equations (6.10) and (6.11). By multiplying the first equation by s and subtracting it from the second one, we obtain

$$(\mathbb{L}_\sigma - s\mathbb{L})\Lambda - M(\mathbb{L}_\sigma - s\mathbb{L}) = LW(\Lambda - sI_k) - (M - sI_q)VR. \tag{6.12}$$

Multiplying this equation by e_i from the right and setting $s = \lambda_i$, we obtain

$$(\lambda_i I_q - M)(\mathbb{L}_\sigma - \lambda_i\mathbb{L})e_i = (\lambda_i I_q - M)Vr_i,$$

which is equivalent to

$$(\mathbb{L}_\sigma - \lambda_i\mathbb{L})e_i = Vr_i$$

and yields

$$w_i = We_i = W(\mathbb{L}_\sigma - \lambda_i\mathbb{L})^{-1}Vr_i.$$

Therefore, we obtain $w_i = \widetilde{G}(\lambda_i)r_i$, $i = 1, \ldots, k$, i. e., the right data is interpolated. To prove that also the left data is interpolated we multiply (6.12) by e_j^{T} from the left and take $s = \mu_j$. This gives

$$e_j^{\mathsf{T}}(\mathbb{L}_\sigma - \mu_j\mathbb{L})(\Lambda - \mu_j I_k) = e_j^{\mathsf{T}}LW(\Lambda - \mu_j I_k),$$

which is equivalent to

$$e_j^{\mathsf{T}}(\mathbb{L}_\sigma - \mu_j\mathbb{L}) = \ell_j^{\mathsf{H}}W.$$

Therefore, we obtain

$$v_j^{\mathsf{H}} = e_j^{\mathsf{T}}V = \ell_j^{\mathsf{H}}W(\mathbb{L}_\sigma - \mu_j\mathbb{L})^{-1}V,$$

which yields $v_j^{\mathsf{H}} = \ell_j^{\mathsf{H}}\widetilde{G}(\mu_j)$, $j = 1, \ldots, q$. This completes the proof. □

A common situation that arises in practice is redundant data, i. e., the case where we have to much data. In this case, the Loewner pencil $s\mathbb{L} - \mathbb{L}_\sigma$ is singular and thus the transfer function $\widetilde{G}$ does not exist. The question that arises is how we can treat this situation in the Loewner framework.

Theorem 6.22: Let data as in (6.6), (6.7) and the associated Loewner matrices (6.8) and (6.9) be given. Assume that

$$\operatorname{rank}(\xi\mathbb{L} - \mathbb{L}_\sigma) = \operatorname{rank}\begin{bmatrix}\mathbb{L} & \mathbb{L}_\sigma\end{bmatrix} = \operatorname{rank}\begin{bmatrix}\mathbb{L} \\ \mathbb{L}_\sigma\end{bmatrix} = r \qquad \forall\, \xi \in \{\lambda_1, \ldots, \lambda_k\} \cup \{\mu_1, \ldots, \mu_q\}.$$

Consider the economic SVDs

$$\begin{bmatrix}\mathbb{L} & \mathbb{L}_\sigma\end{bmatrix} = Y\Sigma_{\mathrm{l}}\widetilde{X}^{\mathsf{H}}, \quad \begin{bmatrix}\mathbb{L} \\ \mathbb{L}_\sigma\end{bmatrix} = \widetilde{Y}\Sigma_{\mathrm{r}}X^{\mathsf{H}} \tag{6.13}$$

with Σ_{l}, $\Sigma_{\mathrm{r}} \in \mathbb{R}^{r\times r}$, and $X \in \mathbb{C}^{k\times r}$ and $Y \in \mathbb{C}^{q\times r}$. If R and L^{H} both have full column rank, then

$$\widetilde{E} = -Y^{\mathsf{H}}\mathbb{L}X, \quad \widetilde{A} = -Y^{\mathsf{H}}\mathbb{L}_\sigma X, \quad \widetilde{B} = Y^{\mathsf{H}}V, \quad \widetilde{C} = WX$$

is an interpolating descriptor system realization, i. e., the function $\widetilde{G}(s) := \widetilde{C}\big(s\widetilde{E} - \widetilde{A}\big)^{-1}\widetilde{B}$ interpolates the given data.

Proof. Here we show only show that the right interpolation conditions are satisfied, the proof for the left interpolation conditions is analogous. From (6.13) and since X, Y have orthonormal columns, we have that $\mathbb{L} = -\widetilde{Y}_1\Sigma_{\mathrm{r}}X^{\mathsf{H}}$ and $\mathbb{L}_\sigma = -\widetilde{Y}_2\Sigma_{\mathrm{r}}X^{\mathsf{H}}$. Therefore, we obtain

$$\begin{aligned}\mathbb{L}XX^{\mathsf{H}} &= -\widetilde{Y}_1\Sigma_{\mathrm{r}}X^{\mathsf{H}}XX^{\mathsf{H}} = -\widetilde{Y}_1\Sigma_{\mathrm{r}}X^{\mathsf{H}} = \mathbb{L},\\ \mathbb{L}_\sigma XX^{\mathsf{H}} &= -\widetilde{Y}_2\Sigma_{\mathrm{r}}X^{\mathsf{H}}XX^{\mathsf{H}} = -\widetilde{Y}_2\Sigma_{\mathrm{r}}X^{\mathsf{H}} = \mathbb{L}_\sigma.\end{aligned}$$

Moreover, we have

$$\mathbb{L}_\sigma - \mathbb{L}\Lambda = VR,$$

which follows from

$$[\mathbb{L}_\sigma - \mathbb{L}\Lambda]_{ij} = \frac{\mu_i v_i^{\mathsf{H}}r_j - \lambda_j\ell_i^{\mathsf{H}}w_j}{\mu_i - \lambda_j} - \frac{v_i^{\mathsf{H}}r_j - \ell_i^{\mathsf{H}}w_j}{\mu_i - \lambda_j}\lambda_j = v_i^{\mathsf{H}}r_j = [VR]_{ij}, \qquad i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, q.$$

Similarly we have

$$\mathbb{L}_\sigma - M\mathbb{L} = LW.$$

This gives

$$LWXX^{\mathsf{H}} = (\mathbb{L}_\sigma - M\mathbb{L})XX^{\mathsf{H}} = \mathbb{L}_\sigma - M\mathbb{L} = LW.$$

Sine L^{H} has full column rank, this implies $WXX^{\mathsf{H}} = W$. With the above identities we get

$$\begin{aligned} -\widetilde{A}X^{\mathsf{H}} + \widetilde{E}X^{\mathsf{H}}\Lambda &= Y^{\mathsf{H}}\mathbb{L}_\sigma XX^{\mathsf{H}} - Y^{\mathsf{H}}\mathbb{L}XX^{\mathsf{H}}\Lambda \\ &= Y^{\mathsf{H}}(\mathbb{L}_\sigma - \mathbb{L}\Lambda) \\ &= Y^{\mathsf{H}}VR. \end{aligned}$$

Then for $i = 1, 2, \ldots, k$ we get

$$\begin{aligned} \widetilde{G}(\lambda_i)r_i &= \widetilde{C}\big(\lambda_i\widetilde{E} - \widetilde{A}\big)^{-1}\widetilde{B}r_i \\ &= \widetilde{C}\big(\lambda_i\widetilde{E} - \widetilde{A}\big)^{-1}Y^{\mathsf{H}}Vr_i \\ &= \widetilde{C}\big(\lambda_i\widetilde{E} - \widetilde{A}\big)^{-1}\big(-\widetilde{A}X^{\mathsf{H}} + \widetilde{E}X^{\mathsf{H}}\Lambda\big)e_i \\ &= \widetilde{C}\big(\lambda_i\widetilde{E} - \widetilde{A}\big)^{-1}\big(\lambda_i\widetilde{E} - \widetilde{A}\big)X^{\mathsf{H}}e_i \\ &= \widetilde{C}X^{\mathsf{H}}e_i = WXX^{\mathsf{H}}e_i = We_i = w_i. \end{aligned}$$

□

Remark 6.23: In general there are many projections leading to the same transfer function. To make this precise, assume that $\Phi \in \mathbb{C}^{k\times r}$, $\Psi \in \mathbb{C}^{q\times r}$ are given such that $X^{\mathsf{H}}\Phi$ and $\Psi^{\mathsf{H}}Y$ are both nonsingular. Then the model given by

$$\widehat{E} = -\Phi^{\mathsf{H}}\mathbb{L}\Psi, \quad \widehat{A} = -\Phi^{\mathsf{H}}\mathbb{L}_\sigma\Psi, \quad \widehat{B} = \Phi^{\mathsf{H}}V, \quad \widehat{C} = W\Psi$$

has the transfer function $\widetilde{G}(s) := \widetilde{C}\big(s\widetilde{E} - \widetilde{A}\big)^{-1}\widetilde{B}$ with the notation as in Theorem 6.22.

To illustrate the method we give a simple example.

Example: Consider the function $G(s) = \frac{1}{s^2+1}$. We sample this function at the points

$$\begin{aligned} &\lambda_1 = 1, \quad \lambda_2 = 2, \quad \lambda_3 = 3, \\ &\mu_1 = -1, \quad \mu_2 = -2, \quad \mu_3 = -3, \end{aligned}$$

and obtain the data

$$\begin{aligned} &\Lambda = \operatorname{diag}(1,\,2,\,3), \quad R = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \quad W = \begin{bmatrix} \frac{1}{2} & \frac{1}{5} & \frac{1}{10} \end{bmatrix}, \\ &M = \operatorname{diag}(-1,\,-2,\,-3), \quad L^{\mathsf{H}} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \quad V^{\mathsf{H}} = \begin{bmatrix} \frac{1}{2} & \frac{1}{5} & \frac{1}{10} \end{bmatrix}. \end{aligned}$$

Then the Loewner pencil $s\mathbb{L} - \mathbb{L}_\sigma$ is given by

$$\mathbb{L} = \begin{bmatrix} 0 & -\frac{1}{10} & -\frac{1}{10} \\ \frac{1}{10} & 0 & -\frac{1}{50} \\ \frac{1}{10} & \frac{1}{50} & 0 \end{bmatrix}, \quad \mathbb{L}_\sigma = \begin{bmatrix} \frac{1}{2} & \frac{3}{10} & \frac{1}{5} \\ \frac{3}{10} & \frac{1}{5} & \frac{7}{50} \\ \frac{1}{5} & \frac{7}{50} & \frac{1}{10} \end{bmatrix}.$$

It is easily checked by some SVDs that

$$\operatorname{rank}(\xi\mathbb{L}-\mathbb{L}_\sigma)=\operatorname{rank}\begin{bmatrix}\mathbb{L} & \mathbb{L}_\sigma\end{bmatrix}=\operatorname{rank}\begin{bmatrix}\mathbb{L}\\ \mathbb{L}_\sigma\end{bmatrix}=2 \qquad \forall\,\xi\in\{-3,\,-2,\,-1,\,1,\,2,\,3\}\,.$$

Therefore, we choose $\Phi,\ \Psi\in\mathbb{C}^{3\times 2}$. We take

$$\Phi=\Psi=\begin{bmatrix}5 & -5\\ 1 & 0\\ 0 & 1\end{bmatrix}.$$

This gives a reduced-order

$$\widehat{E}=-\Phi^{\mathsf{H}}\mathbb{L}\Psi=\begin{bmatrix}0 & \frac{51}{50}\\ -\frac{51}{50} & 0\end{bmatrix},\quad \widehat{A}=-\Phi^{\mathsf{H}}\mathbb{L}_\sigma\Psi=\begin{bmatrix}-\frac{157}{10} & \frac{643}{50}\\ \frac{643}{50} & -\frac{53}{5}\end{bmatrix},$$
$$\widehat{B}=\Phi^{\mathsf{H}}V=\begin{bmatrix}\frac{27}{10}\\ -\frac{12}{5}\end{bmatrix},\quad \widehat{C}=W\Psi=\begin{bmatrix}\frac{27}{10} & -\frac{12}{5}\end{bmatrix}.$$

Now a simple calculation gives

$$\widehat{C}\big(s\widehat{E}-\widehat{A}\big)^{-1}\widehat{B}=\frac{1}{s^2+1}=G(s),$$

so we have reconstructed our original model (which can be realized by a system of state-space dimension two) from the data.

CHAPTER 7

Outlook

In this course on model reduction we have mainly considered system-theoretic methods, i. e., methods that mainly try to approximate the input/output behavior of a dynamical system. In this chapter we will have a brief look onto some further aspects of model reduction that could not be covered in this course in full detail.

7.1 Parametric Model Reduction

In industrial applications one often considers dynamical systems that depend on parameters. Then one is often interested in optimizing these parameters which often requires a lot of simulations of the dynamical system for many parameters. In this context one is particularly interested in reduced representations of the model to make these simulations feasible. The great challenge consists of finding a reduced-order model that is a good approximation of the original one for all parameters. A good survey on such methods is [BGW15]. To make this precise, consider a *parametric LTI system*

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

where $A(p) \in \mathbb{R}^{n\times n}$, $B(p) \in \mathbb{R}^{n\times m}$, $C(p) \in \mathbb{R}^{q\times n}$, and $D(p) \in \mathbb{R}^{q\times m}$ for all $p \in \Omega \subset \mathbb{R}^d$. In this setting it is desirable to have a parameter-affine representation

of the model as

$$\begin{aligned} A(p) &= A_0 + a_1(p)A_1 + \ldots + a_{\kappa_A}(p)A_{\kappa_A}, \\ B(p) &= B_0 + b_1(p)B_1 + \ldots + b_{\kappa_B}(p)B_{\kappa_B}, \\ C(p) &= C_0 + c_1(p)C_1 + \ldots + c_{\kappa_C}(p)C_{\kappa_C} \end{aligned} \tag{7.1}$$

for fixed matrices $A_0, \ldots, A_{\kappa_A} \in \mathbb{R}^{n\times n}$, $B_0, \ldots, B_{\kappa_B} \in \mathbb{R}^{n\times m}$, $C_0, \ldots, C_{\kappa_C} \in \mathbb{R}^{q\times n}$ and functions $a_1, \ldots, a_{\kappa_A}, b_1, \ldots, b_{\kappa_B}, c_1, \ldots, c_{\kappa_C} : \Omega \to \mathbb{C}$. Note that any system can be written in this form, but for computational efficiency it is desirable to have $\kappa_A, \kappa_B, \kappa_C \ll n$. This representation easily allows the construction of reduced-order models via projection, i. e., projection matrices $W, T \in \mathbb{R}^{n\times r}$ are constructed such that $W^{\mathsf{T}}T = I_r$ and such that the reduced-order model is given by

$$\begin{aligned} \dot{\widetilde{x}}(t;p) &= \widetilde{A}(p)\widetilde{x}(t;p) + \widetilde{B}(p)u(t), \\ \widetilde{y}(t;p) &= \widetilde{C}(p)\widetilde{x}(t;p) + \widetilde{D}(p)u(t), \end{aligned}$$

where

$$\begin{aligned} \widetilde{A}(p) &= W^{\mathsf{T}}A(p)T = W^{\mathsf{T}}A_0T + a_1(p)W^{\mathsf{T}}A_1T + \ldots + a_{\kappa_A}(p)W^{\mathsf{T}}A_{\kappa_A}T, \\ \widetilde{B}(p) &= W^{\mathsf{T}}B(p) = W^{\mathsf{T}}B_0 + b_1(p)W^{\mathsf{T}}B_1 + \ldots + b_{\kappa_B}(p)W^{\mathsf{T}}B_{\kappa_B}, \\ \widetilde{C}(p) &= C(p)T = C_0T + c_1(p)C_1T + \ldots + c_{\kappa_C}(p)C_{\kappa_C}T, \\ \widetilde{D}(p) &= D(p). \end{aligned}$$

The main questions that have to be faced here, are the following:

a) Do we want to determine *global projection matrices* W and T that are good for all $p \in \Omega$ or do we rather determine *local projection matrices* $W_1, \ldots, W_k$, $T_1, \ldots, T_k$ from models for particular parameters $p_1, \ldots, p_k$?

b) If we determine local projection matrices, how do we choose good parameters $p_1, \ldots, p_k$ and how can we use the information to get a reduced-order model for $p \notin \{p_1, \ldots, p_k\}$?

There are several ways to attack these questions. If the local projection matrices $W_1, \ldots, W_k, T_1, \ldots, T_k$ from models for particular parameters $p_1, \ldots, p_k$ are known, then one can obtain a reduced-order model for some $p \notin \{p_1, \ldots, p_k\}$ by

a) *transfer function interpolation*: If $\widetilde{G}_j(s) = \widetilde{C}(p_j)\big(sI_n - \widetilde{A}(p_j)\big)^{-1}\widetilde{B}(p_j) + \widetilde{D}(p_j)$, then one can construct

$$\widetilde{G}(s,p) = \sum_{j=1}^{k} g_j(p)\widetilde{G}_j(s)$$

for an arbitrary $p \in \Omega$, where $g_j(\cdot)$ are some interpolation functions (such as Lagrange polynomials). Further, rational interpolation techniques have been developed to find good global projection matrices. These lead to (tangential) interpolation conditions that do not only lead to moment matching at interpolation points $s_1, \ldots, s_\ell$, but also at the parameters $p_1, p_2, \ldots, p_k$.

b) *matrix interpolation*: Instead of interpolating the transfer functions, the reduced functions $\widetilde{A}(\cdot)$, $\widetilde{B}(\cdot)$, $\widetilde{C}(\cdot)$, $\widetilde{D}(\cdot)$ can be obtained via interpolation at the parameters $p_1, \ldots, p_k$.

Note that the above techniques do not need the affine representation (7.1). On the other hand, the affine representation is explicitly used in determining good global projection matrices by parametric balanced truncation. This method computes global low-rank factorizations of parametric Lyapunov equations, but it still has limitations since it can only be applied to special classes of systems such as systems with pointwise positive definite $A(p)$ that allow the development of error estimators. These error estimators are important to find good sampling parameters p in the algorithm.

7.2 Sampling-Based Methods

A further class of methods that has not been discussed in this course are sampling-based methods. These methods mainly consist of sampling the solution of the system under consideration for several initial conditions or parameter values and approximating the space in which the solutions live. If a basis for this space for some initial conditions or parameters is known, then the hope is that also the solutions for the other values can be well represented in this basis. Sampling-based methods play a great role in model reduction of PDEs, see, e. g., [Vol13]. Consider for example 1D Burgers' equation

$$\frac{\partial}{\partial t}y(x,t) - \nu\frac{\partial^2}{\partial x^2}y(x,t) + \frac{1}{2}\frac{\partial}{\partial x}y(x,t)^2 = 0, \quad x \in [0,1],\ t \geqslant 0$$

with the boundary and initial conditions

$$\begin{aligned} y(0,t) = y(1,t) &= 0, \quad t \geqslant 0, \\ y(x,0) &= y_0(x), \quad x \in [0,1]. \end{aligned}$$

A finite element discretization with the finite element basis $\{\varphi_1, \ldots, \varphi_N\}$ leads to the spatially discretized model

$$\widehat{M}\frac{\mathrm{d}}{\mathrm{d}t}\widehat{y}(t) - \nu\widehat{K}\widehat{y}(t) + \widehat{L}\big(\widehat{y}(t)\big) = 0.$$

Then an approximate solution of the original problem is of the form

$$\widehat{y}_h(x,t) = \sum_{i=1}^{N} \varphi_i(x)\widehat{y}_i(t),$$

where $\widehat{y}_i(t)$ is the i-th component of $\widehat{y}(t)$. The goal is to replace this model by a reduced-order model

$$\widetilde{M}\frac{\mathrm{d}}{\mathrm{d}t}\widetilde{y}(t) - \nu\widetilde{K}\widetilde{y}(t) + \widetilde{L}\big(\widetilde{y}(t)\big) = 0$$

that can be described by an orthonormal basis $\{\psi_1, \ldots, \psi_k\}$ with $k \ll N$. Then an approximate solution of the original problem is of the form

$$\widetilde{y}_h(x,t) = \sum_{i=1}^{k} \psi_i(x)\widetilde{y}_i(t),$$

where $\widetilde{y}_i(t)$ is the i-th component of $\widetilde{y}(t)$. The question is how to determine the basis $\{\psi_1, \ldots, \psi_k\}$. Assume that we have given *"snapshots"* of the full-order solution $\widehat{y}_h(x,t)$, i. e., we know $\{\widehat{y}_1(\cdot), \ldots, \widehat{y}_n(\cdot)\} := \{\widehat{y}_h(\cdot,t_1), \ldots, \widehat{y}_h(\cdot,t_n)\} \subset X$, where X is assumed to be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_X$ and induced norm $\|\cdot\|_X := \langle\cdot,\cdot\rangle_X^{1/2}$. For given $k \leqslant n$, the method of *proper orthogonal decomposition (POD)* computes an optimal solution of the optimization problem

$$\min_{\{\psi_1,\ldots,\psi_k\}} \sum_{i=1}^{n} \|\widetilde{y}_i - \widehat{y}_i\|_X^2, \quad \langle\psi_j,\psi_\ell\rangle_X = \delta_{j\ell} \quad \forall\, j,\ell \in \{1,\ldots,k\},$$

where $\widetilde{y}_i(x) = \sum_{j=1}^{k}\langle\widehat{y}_i,\psi_j\rangle_X\,\psi_j(x)$ is an approximation of $\widehat{y}_i$ using the basis $\{\psi_1, \ldots, \psi_k\}$. Assume w. l. o. g. that $X = \mathbb{R}^N$. (Note that if X is any finite dimensional Hilbert space, then there exists an isometric isomorphism $\Phi : (X, \langle\cdot,\cdot\rangle_X) \to \big(\mathbb{R}^N, \langle\cdot,\cdot\rangle_{\mathbb{R}^N}\big)$, where $\langle\cdot,\cdot\rangle_{\mathbb{R}^N}$ is the standard inner product in $\mathbb{R}^N$ – so we can exploit the notation and simply denote $\Phi(y)$ by y.) So we get an equivalent optimization problem

$$\min_{\{\psi_1,\ldots,\psi_k\}} \sum_{i=1}^{n} \|\widetilde{y}_i - \widehat{y}_i\|_2^2, \quad \langle\psi_j,\psi_\ell\rangle_{\mathbb{R}^N} = \delta_{j\ell} \quad \forall\, j,\ell \in \{1,\ldots,k\},$$

where $\widetilde{y}_i = \sum_{j=1}^{k}\langle\widehat{y}_i,\psi_j\rangle_{\mathbb{R}^N}\,\psi_j$. Let $Y = \begin{bmatrix}\widehat{y}_1 & \ldots & \widehat{y}_n\end{bmatrix} \in \mathbb{R}^{N\times n}$ and

$$YY^{\mathsf{T}}\widetilde{\psi}_j = \lambda_j\widetilde{\psi}_j \quad \text{for } j = 1, 2, \ldots, n \quad \text{with } \lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n.$$

Then the POD basis is given by $\{\psi_1, \ldots, \psi_k\} = \big\{\widetilde{\psi}_1, \ldots, \widetilde{\psi}_k\big\}$. In other words, the POD basis consists of the k dominant left singular vectors of the snapshot matrix Y. Furthermore, the error is given by

$$\sum_{i=1}^{n}\left\|\sum_{j=1}^{k}\langle\widehat{y}_i,\psi_j\rangle_{\mathbb{R}^N}\,\psi_j - \widehat{y}_i\right\|_2^2 = \sum_{j=k+1}^{n}\lambda_j.$$

In the literature several further aspects are discussed such as optimal choices of snapshots for parameter-dependent problems (by developing error estimators) and the application to optimal control problems.

Bibliography

[Ant05] A. C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. SIAM Publications, Philadelphia, PA, USA, 2005.

[AV73] B. D. O. Anderson and S. Vongpanitlerd. *Network Analysis and Synthesis – A Modern Systems Theory Approach*. Prentice-Hall, Englewood Cliffs, NJ, USA, 1973.

[BB98] P. Benner and R. Byers. An exact line search method for solving generalized continuous-time algebraic Riccati equations. *IEEE Trans. Automat. Control*, 43(1):101–107, 1998.

[BGW15] P. Benner, S. Gugercin, and K. Willcox. A survey on projection-based model reduction methods for parametric dynamical systems. *SIAM Rev.*, 57(4):483–531, 2015.

[BKS13] P. Benner, P. Kürschner, and J. Saak. Efficient handling of complex shift parameters in the low-rank cholesky factor ADI method. *Numer. Algorithms*, 62(2):225–251, 2013.

[BKS15] P. Benner, P. Kürschner, and J. Saak. Self-generating and efficient shift parameters in ADI methods for large Lyapunov and Sylvester equations. *Electron. Trans. Numer. Anal.*, 43:142–162, 2014–2015.

[Bog07] V. I. Bogachev. *Measure Theory*. Springer-Verlag, Berlin, Heidelberg, 1st edition, 2007.

[BS13] P. Benner and J. Saak. Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: A state of the art survey. *GAMM-Mitt.*, 6(1):32–52, 2013.

[GAB08] 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.

[Geo88] T. T. Georgiou. On the computation of the gap metric. *Systems Control Lett.*, 11(4):253–257, 1988.

[GO13] C. Guiver and M. R. Opmeer. Error bounds in the gap metric for dissipative balanced approximations. *Linear Algebra Appl.*, 439(12):3659–3698, 2013.

[Kle68] D. Kleinman. On an iterative technique for Riccati equation computations. *IEEE Trans. Automat. Control*, 13(1):114–115, 1968.

[LR95] P. Lancester and L. Rodman. *Algebraic Riccati Equations*. Oxford University Press, New York, 1995.

[MA07] A. J. Mayo and A. C. Antoulas. A framework for the solution of the generalized realization problem. *Linear Algebra Appl.*, 425(2–3):634–662, 2007.

[Moo81] B. C. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Trans. Automat. Control*, AC-26(1):17–32, 1981.

[Obe91] R. Ober. Balanced parametrization of classes of linear systems. *SIAM J. Control Optim.*, 29(6):1251–1287, 1991.

[Pen00] T. Penzl. A cyclic low rank Smith method for large sparse Lyapunov equations. *SIAM J. Sci. Comput.*, 21(4):1401–1418, 2000.

[PR55] D. Peaceman and H. Rachford. The numerical solution of elliptic and parabolic differential equations. *J. Soc. Indust. Appl. Math.*, 3(1):28–41, 1955.

[Ran96] A. Rantzer. On the Kalman-Yakubovich-Popov lemma. *Systems Control Lett.*, 28(1):7–10, 1996.

[RM06a] J. Rommes and N. Martins. Efficient computation of multivariate transfer function dominant poles using subspace acceleration. *IEEE Trans. Power Syst.*, 21(4):1471–1483, 2006.

[RM06b] J. Rommes and N. Martins. Efficient computation of transfer function dominant poles using subspace acceleration. *IEEE Trans. Power Syst.*, 21(3):1218–1226, 2006.

[RS10] T. Reis and T. Stykel. Positive real and bounded real balancing for model reduction of descriptor systems. *Internat. J. Control*, 83(1):74–88, 2010.

[Saa82] Y. Saad. The Lanczos biorthogonalization algorithm and other oblique projection projection methods for solving large unsymmetric systems. *SIAM J. Numer. Anal.*, 19(3):485–506, 1982.

[Sim07] V. Simoncini. A new iterative method for solving large-scale Lyapunov equations. *SIAM J. Sci. Comput.*, 29(3):1268–1288, 2007.

[SZ02] D. C. Sorenson and Y. Zhou. Bounds on eigenvalue decay rates and sensitivity of solutions to Lyapunov equations. CAAM Technical Reports, Rice University, 2002.

[Vol13] S. Volkwein. Proper orthogonal decomposition: Theory and reduced-order modeling, August 2013. Available from http://www.math.uni-konstanz.de/numerik/personen/volkwein/teaching/POD-Book.pdf.

[Wac13] E. Wachspress. *The ADI Model Problem*. Springer-Verlag, New York, NY, USA, 1st edition, 2013.

[Wil71] J. C. Willems. Least squares stationary optimal control and the algebraic Riccati equation. *IEEE Trans. Automat. Control*, AC-16(6):621–634, 1971.

[Wil72a] J. C. Willems. Dissipative dynamical systems part I: General theory. *Arch. Ration. Mech. Anal.*, 45(5):321–351, 1972.

[Wil72b] J. C. Willems. Dissipative dynamical systems part II: Linear systems with quadratic supply rates. *Arch. Ration. Mech. Anal.*, 45(5):352–393, 1972.

[ZDG96] K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice-Hall, Englewood Cliffs, NJ, 1996.