Model Order Reduction

Dr. Dominik Garmatter

Chemnitz University of Technology
Institut of Mathematics

Wintersemester 2019/2020

https://www.tu-chemnitz.de/mathematik/wire/

Preliminary version, will be adjusted and corrected throughout the lecture period.
Last update: February 3, 2020

Preface

The first Chapter of this script is largely based on the lecture notes of the lecture "Reduzierte Basis Methoden" held by from Prof. Dr. Bernard Haasdonk at the University of Stuttgart in the summer term 2011, where I attended the lecture myself as a student, as well as the summer terms 2015 and 2019[1].

I thank Prof. Haasdonk for the allowance of using his lecture notes as the foundation of this script.

The second Chapter of this script is based on Chapters 2 & 4 of the lecture notes of the lecture "Model Reduction" held by Dr. Matthias Voigt at the University of Hamburg in the summer term 2019[2]. Basic concepts of control theory have been added from slides of the lecture "Control Theory" regularly held by Prof. Dr. Carsten Scherer at the University of Stuttgart.

I thank both Dr. Voigt and Prof. Scherer for the allowance of using their lecture notes for this script.

Dominik Garmatter, Chemnitz, February 3, 2020.

[1] `http://www2.ians.uni-stuttgart.de/am/Haasdonk/data/skripte/`
[2] `https://www.math.uni-hamburg.de/home/voigt/modellreduktion.html`

Contents

Chapter 1

Parametric Model Order Reduction

1.1 Introduction

We begin with the description of two basic parametric problems.

Example 1.1 (Parametric Partial Differential align)
Let $\Omega \subseteq \mathbb{R}^d$ be an open polygonal domain and $\mathcal{P} \subset \mathbb{R}^p$ be a set of 'admissible' parameters. For a parameter vector *$\mu \in \mathcal{P}$, we seek a function $u(x;\mu) : \Omega \to \mathbb{R}$, the 'temperature', satisfying*

$$\begin{aligned} -\nabla \cdot (\sigma(x;\mu)\nabla u(x;\mu)) &= h(x;\mu), \quad && x \in \Omega, \\ u(x;\mu) &= 0, && x \in \partial\Omega, \end{aligned}$$

with 'heat conductivity' $\sigma(x;\mu) : \Omega \to \mathbb{R}$ and 'heat source' $h(x;\mu) : \Omega \to \mathbb{R}$, for example

$$h(x;\mu) := \begin{cases} 1, & x \in \Omega_{source}, \\ 0, & otherwise \end{cases},$$

with a subdomain $\Omega_{source} \subset \Omega$. Furthermore, an output *might be desired, for example the 'average temperature' over a subdomain $\Omega_{av} \subset \Omega$*

$$s(\mu) := \frac{1}{|\Omega_{av}|} \int_{\Omega_{av}} u(x;\mu)\,\mathrm{d}x.$$

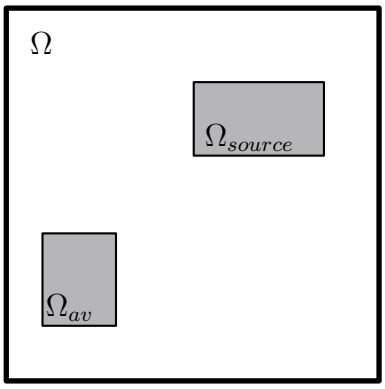


Figure 1.1: Exemplary domain and subdomains for $d = 2$.

Example 1.2 (Parametric Stationary align System)
For a parameter vector $\mu \in \mathcal{P}$, *find a* state vector $u(\mu) \in \mathbb{R}^n$ *and an* output $s(\mu) \in \mathbb{R}^k$ *satisfying*

$$\begin{aligned} 0 &= A(\mu)u(\mu) + B(\mu)w(\mu), \\ s(\mu) &= C(\mu)u(\mu), \end{aligned}$$

with parameter dependent system matrices $A(\mu) \in \mathbb{R}^{n\times n}$, $B(\mu) \in \mathbb{R}^{n\times m}$, $C(\mu) \in \mathbb{R}^{k\times n}$ *and input vector* $w(\mu) \in \mathbb{R}^m$.

In this chapter, we want to tackle problems that can be described via a *weak formulation in Hilbert spaces* as follows. Let X be a real Hilbert space, $a(\cdot,\cdot;\mu) : X \times X \to \mathbb{R}$ be a bilinear form and $f(\cdot;\mu) : X \to \mathbb{R}$, $l(\cdot;\mu) : X \to \mathbb{R}$ be linear forms. For a given $\mu \in \mathcal{P}$ we then seek a $u(\mu) \in X$ and output $s(\mu) \in \mathbb{R}$ satisfying

$$\begin{aligned} a(u(\mu), v; \mu) &= f(v;\mu), \quad \forall v \in X, \\ s(\mu) &= l(u(\mu);\mu). \end{aligned}$$

Both examples so far can be formulated this way.

- For Example 1.1: with the Hilbert space

  $$X = H_0^1(\Omega) := \{g \in L^2(\Omega) \mid D^\alpha g \in L^2(\Omega) \text{ for all } |\alpha| \leq 1 \text{ and } g\,|_{\partial\Omega} = 0\}$$

we have

$$\underbrace{\int_\Omega \sigma(x;\mu)\nabla u(x;\mu)\cdot\nabla v(x)\,\mathrm{d}x}_{=:a(u,v;\mu)} = \underbrace{\int_\Omega h(x;\mu)v(x)\,\mathrm{d}x}_{=:f(v;\mu)}, \quad \forall v\in X,$$

$$s(\mu) = \underbrace{\frac{1}{|\Omega_{av}|}\int_{\Omega_{av}} u(x;\mu)\,\mathrm{d}x}_{=:l(u(\mu);\mu)}.$$

- For Example 1.2 (with $k=1$): with the Hilbert space $X=\mathbb{R}^n$ we have

$$\underbrace{v^\mathsf{T}A(\mu)u(\mu)}_{=:a(u,v;\mu)} = \underbrace{-v^\mathsf{T}B(\mu)w(\mu)}_{=:f(v;\mu)}, \quad \forall v\in X,$$

$$s(\mu) = \underbrace{C(\mu)u(\mu)}_{=:l(u(\mu);\mu)}.$$

The Motivation/basic idea for parametric model order reduction is the following.

- $\mathcal{M} := \{u(\mu) \mid \mu\in\mathcal{P}\subset\mathbb{R}^p\}$ is the solution manifold parametrized by μ.
- The solution space X is in general ∞-dimensional (Sobolev space) or finite but high-dimensional (FEM, FD, FV space for Example 1.1; large n in Example 1.2). But $\mathcal{M}$ can often be approximated by a low-dimensional subspace $X_N\subset X$ such that $u_N(\mu)\in X_N$ is a good approximation of $u(\mu)$ 'for many' $\mu\in\mathcal{P}$.
- In Reduced Basis Methods (RB-Methods) the space X_N is formed by solution samples, so-called *snapshots*, such that

$$X_N \subset \operatorname{span}(\{u(\mu_1),\dots,u(\mu_n)\}) \subset X$$

for suitably chosen parameter samples $\mu_1,\dots,\mu_n\in\mathcal{P}$. A further goal of RB-Methods is to control the approximation error via error bounds such that

$$\|u(\mu)-u_N(\mu)\|_X \le \Delta_N(\mu).$$

We illustrate this approach (approximating $\mathcal{M}$ via snapshots) by a simple example.

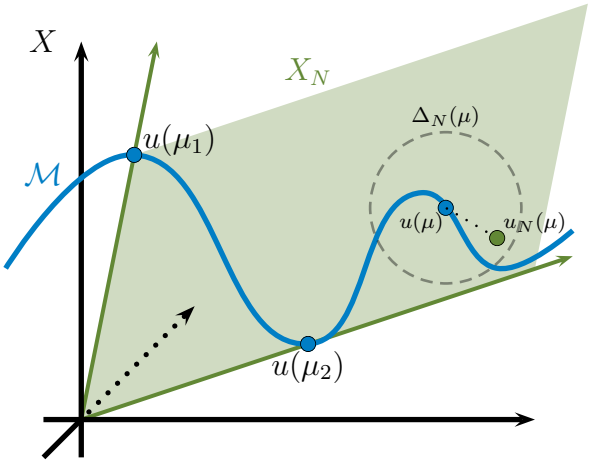


Figure 1.2: Illustration of the Reduced Basis Method.

Example 1.3 (Parametrized Boundary Value Problem)
For $\mu \in \mathcal{P} := [0,1] \subset \mathbb{R}$, find $u(x;\mu) \in C^2([0,1])$ satisfying

$$(1+\mu)u''(x) = 1, \quad x \in (0,1), \quad u(0) = u(1) = 1. \tag{BVP}$$

We construct the reduced space X_N using two snapshots *(special solutions).*

- *For $\mu = 0$ we obtain $u_0(x) := u(x;0) = \frac{1}{2}x^2 - \frac{1}{2}x + 1$.*
- *For $\mu = 1$ we obtain $u_1(x) := u(x;1) = \frac{1}{4}x^2 - \frac{1}{4}x + 1$.*
- *Define $X_N := \operatorname{span}(\{u_0, u_1\})$.*

The reduced solution is then a linear combination

$$u_N(x;\mu) := c_0(\mu)u_0(x) + c_1(\mu)u_1(x)$$

and by inserting this into (BVP) we obtain for $c_0(\mu)$ and $c_1(\mu)$

$$u_N(0;\mu) = u_N(1;\mu) = 1 \quad \Leftrightarrow \quad c_0(\mu) = 1 - c_1(\mu) \tag{$\diamond$}$$

as well as

$$\begin{aligned} (1+\mu)u_N''(x;\mu) = 1 \quad &\Leftrightarrow \quad c_0(\mu) + \frac{c_1(\mu)}{2} = \frac{1}{\mu+1} \\ \overset{(\diamond)}{\Leftrightarrow} -\frac{c_1(\mu)}{2} = \frac{1}{\mu+1} - 1 \quad &\Leftrightarrow \quad c_1(\mu) = 2 - \frac{2}{\mu+1} \end{aligned}$$

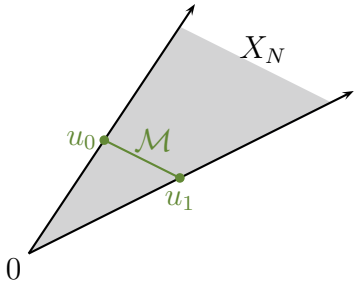


Figure 1.3: Reduced Basis approach for (BVP).

and thus $c_0(\mu) = \frac{2}{\mu+1} - 1$. *Thus,* $u_N(x;\mu)$ *is a solution of (BVP) for all* $\mu \in \mathcal{P}$ *and as the solution of (BVP) is unique, we have the error statement*

$$\|u(x;\mu) - u_N(x;\mu)\|_\infty = \sup_{x\in[0,1]} |u(x;\mu) - u_N(x;\mu)| = 0$$

for all $\mu \in \mathcal{P}$. *Therefore, the solution manifold* $\mathcal{M} = \{u(\mu) \mid \mu \in \mathcal{P}\}$ *is contained in a 2-dimensional space* X_N. *More precisely, due to* $0 \leq c_0(\mu)$, $c_1(\mu) \leq 1$ *and* $c_0(\mu) + c_1(\mu) = 1$ *for all* $\mu \in \mathcal{P}$, $\mathcal{M}$ *is the set of convex combinations of* $u_0(x)$ *and* $u_1(x)$, *hence a 1-dimensional affine subspace.*

We introduce some Notation/Terminology (here tailored around partial differential aligns (PDEs)).

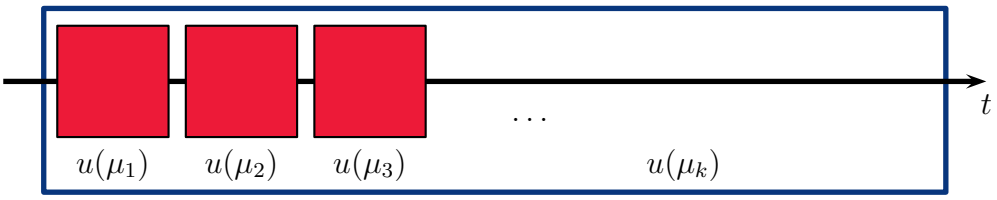
- A PDE is an *analytical model* that characterizes the *exact solution* $u(\mu) \in X$ residing in a typically ∞-dimensional function space.
- A *detailed model* is a computational procedure that characterizes an approximation $u(\mu) \in X$ in a high-dimensional function space with quite general approximation properties (e.g. a FEM, FD, or FV space, with $\dim(X) = 10^3 - 10^8$). Thus, $u(\mu)$ and X can correspond to either an analytical or a detailed model in this chapter.
- A *reduced model* is a computational procedure that characterizes a *reduced solution* $u_N(\mu) \in X_N$ in a very problem adapted and typically low-dimensional space ($\dim(X_N) = 1 - 100$).

- Model Order Reduction deals with the construction of reduced models and the investigation of their properties.

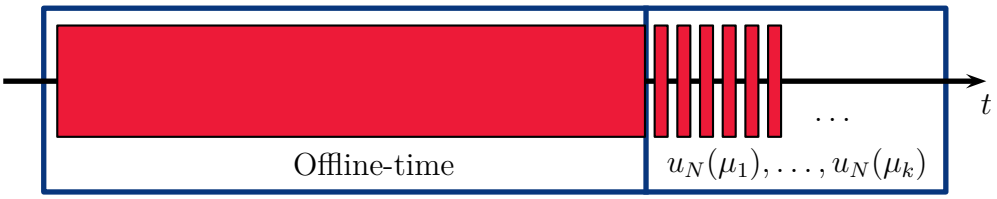
Since the evaluation of a parametric reduced model requires less CPU time and the model itself requires less memory storage, such models have applications in various fields.

- *Multi-query contexts:* repeated model evaluations under a varying parameter; occurring in design, optimization, inverse problems, parameter studies, etc.
- *Real-time contexts:* applications that require a fast simulation response; occurring in real-time control of technical processes, interactive user interfaces, or web forms.
- *Cool-computing contexts:* applications where only 'simple' hardware is available; occurring in smartphones, electronic controllers.

The quick evaluation of the reduced model can be achieved via an *Offline/Online decomposition* of the procedure. A general consensus in model order reduction is that the construction of the reduced model, which is done in the *Offline-Phase*, can be computationally *expensive* as this is done only *once*. The reduced model can then be rapidly evaluated in the *Online-Phase*. Due to the multi-query context, the high (one-time) Offline-costs are then amortized via sufficiently many online requests.



(a) Multiple solution queries with detailed model.



(b) Multiple solution queries with reduced model.

In this chapter we will deal with the following questions.

- *Reduced model:* how can we compute the reduced solution $u_N(\mu)$?
- *Stability:* can we guarantee stability of the reduced model for growing $\dim(X_N) = N$?
- *Error estimator:* can the error $\|u(\mu) - u_N(\mu)\|_X$ be bounded by an error estimator $\Delta_N(\mu)$? Are such estimators cheap to evaluate?
- *Effectivity of error estimator:* is there an upper bound on the overestimation of the error estimator to the error?
- *Computational efficiency:* how can $u_N(\mu)$ be computed rapidly for many different μ?
- *Reduced basis:* how can we construct a subspace that is small but well-approximating? Can we prove approximation qualities?

1.2 Theoretical Background

1.2.1 Linear functional analysis in Hilbert spaces

Definition 1.4 (Hilbert Space)
Let X be a real vector space, $\langle \cdot, \cdot \rangle_X : X \times X \to \mathbb{R}$ a scalar product with induced norm $\|\cdot\|_X := \sqrt{\langle \cdot, \cdot \rangle_X}$. If X is complete with respect to $\|\cdot\|_X$ we call X a real Hilbert space.

Throughout this lecture, we will only deal with real Hilbert spaces.

Example 1.5 (Hilbert Spaces)
Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$ be an open and bounded set. The following spaces are Hilbert spaces:

(a) $X := \mathbb{R}^d$, *with scalar product* $\langle x, x' \rangle_X := \sum_{i_1}^d x_i x_i'$ *with* $x, x' \in X$,

(b) $X := H_0^1(\Omega)$, *with scalar product* $\langle f, g \rangle_X := \int_\Omega \nabla f(x) \cdot \nabla g(x) \, \mathrm{d}x$ *for* $f, g \in X$,

(c) $X := H^1(\Omega)$, *with scalar product*

$$\langle f, g \rangle_X := \int_\Omega f(x) g(x) \, \mathrm{d}x + \int_\Omega \nabla f(x) \cdot \nabla g(x) \, \mathrm{d}x$$

for $f, g \in X$.

The space $X := C^0([0,1])$ *with scalar product* $\langle f, g\rangle_X := \int_0^1 f(x)g(x)\,\mathrm{d}x$ *for* $f, g \in X$ *is not a Hilbert space (see Exercise 1.1).*

Definition 1.6 (Linear Operators)
Let X, Y *be Hilbert spaces.*

(a) A linear mapping $A : X \to Y$ *is a* continuous linear Operator*, if*

$$\|A\| := \sup_{x \in X\setminus\{0\}} \frac{\|Ax\|_Y}{\|x\|_X} < \infty.$$

(b) The space of continuous linear Operators *mapping from* X *to* Y *is denoted* $\mathcal{L}(X, Y)$*. Furthermore,* $\mathcal{L}(X) := \mathcal{L}(X, X)$*.*

(c) The dual space *of* X *is denoted* $X' := \mathcal{L}(X, \mathbb{R})$*. Elements of* X' *are* continuous linear functionals *and for* $l \in X'$ *we have*

$$\|l\|_{X'} := \sup_{x \in X\setminus\{0\}} \frac{|l(x)|}{\|x\|_X}.$$

(d) For $A \in \mathcal{L}(X, Y)$ *we call*

- $\mathcal{R}(A) := \{Ax \mid x \in X\} \subset Y$ *the* Range *of* A *and*
- $\mathcal{N}(A) := \{x \in X \mid Ax = 0\} \subset X$ *the* Kernel *of* A*.*

(e) $A \in \mathcal{L}(X, Y)$ *is a* compact operator *if* $\overline{A(B)}$ *is compact for all bounded sets* $B \subset X$*.* $\mathcal{K}(X, Y)$ *denotes the space of compact operators from* X *to* Y*.*

(f) $A \in \mathcal{L}(X, Y)$ *is called an* Isometry *if* $\|Ax\|_Y = \|x\|_X$ *for all* $x \in X$*.*

Theorem 1.7 (Orthogonal Projection)
Let X *be a Hilbert space and* $Y \subset X$ *a closed subspace. Then, there exists a unique linear mapping* $P : X \to Y : x \mapsto Px$ *satisfying*

$$\|x - Px\|_X = \inf_{y \in Y} \|x - y\|_X$$

and we call this map the orthogonal Projection *of* X *onto* Y*. Furthermore, we have:*

(a) $Px \in Y$ *for* $x \in X$ *is equivalently characterized via*

$$\langle x - Px, y\rangle_X = 0, \quad \forall y \in Y.$$

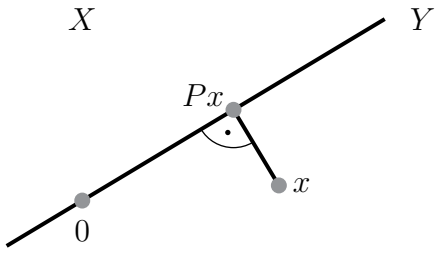


Figure 1.5: Illustration of Orthogonal Projection.

(b) If $\dim(Y) = n < \infty$ *and* $\{\varphi_i\}_{i=1}^n$ *is an orthonormal basis of* Y*,* $Px \in Y$ *is equivalently characterized via*

$$Px = \sum_{i=1}^{n} \langle x, \varphi_i \rangle_X \, \varphi_i, \quad \forall x \in X.$$

Proof: See Exercise 1.2. □

Theorem 1.8 (Riesz Representation Theorem)
Let X *be a Hilbert space. Then,*

$$J : X \to X' : v \mapsto J(v)(\cdot) := \langle v, \cdot \rangle_X$$

is a linear, continuous, and bijective Isometry. In particular, for every $l \in X'$ *there exists a unique* Riesz-representative $v_l := J^{-1}(l) \in X$ *satisfying* $l(v) = \langle v_l, v \rangle_X$ *for all* $v \in X$*.*

Proof: The linearity of J follows from the linearity of the scalar product as for $v_1, v_2 \in X$ and $\lambda_1, \lambda_2 \in \mathbb{R}$ we have

$$\begin{aligned} J(\lambda_1 v_1 + \lambda_2 v_2)(\cdot) &= \langle \lambda_1 v_1 + \lambda_2 v_2, \cdot \rangle_X = \lambda_1 \langle v_1, \cdot \rangle_X + \lambda_2 \langle v_2, \cdot \rangle_X \\ &= \lambda_1 J(v_1)(\cdot) + \lambda_2 J(v_2)(\cdot). \end{aligned}$$

For $v \in X \setminus \{0\}$ we calculate

$$\begin{aligned} \|J(v)(\cdot)\|_{X'} &= \sup_{w \in X \setminus \{0\}} \frac{|J(v)(w)|}{\|w\|_X} = \sup_{w \in X \setminus \{0\}} \frac{|\langle v, w \rangle_X|}{\|w\|_X} \\ &\overset{CSU}{\leq} \sup_{w \in X \setminus \{0\}} \frac{\|v\|_X \|w\|_X}{\|w\|_X} = \|v\|_X, \end{aligned}$$

so that J is continuous. Furthermore,

$$\begin{aligned}\|v\|_X = \frac{\|v\|_X \cdot \|v\|_X}{\|v\|_X} &\overset{CSU}{=} \frac{|\langle v, v\rangle_X|}{\|v\|_X} = \frac{|J(v)(v)|}{\|v\|_X} \\ &\leq \sup_{w\in X\setminus\{0\}} \frac{|J(v)(w)|}{\|w\|_X} = \|J(v)(\cdot)\|_{X'},\end{aligned}$$

so that $\|J(v)(\cdot)\|_{X'} = \|v\|_X$ follows for all $v \in X$ and J is an Isometry. Thus, J has to be injective as well.

Regarding the surjectivity: Let $l \in X'$ with $l \neq 0$ and we note that $\mathcal{N}(l)$ is a closed subspace of X such that from Theorem 1.7 there exists an orthogonal projection $P : X \to \mathcal{N}(l)$. As $l \neq 0$, there exists a $v_0 \in X$ satisfying $c := l(v_0) \neq 0$. Define $v_1 := v_0 - Pv_0 \in X$. Then,

$$l(v_1) = l(v_0) + l(Pv_0) = l(v_0) = c.$$

For $v \in X$ we can write

$$v = v - \frac{l(v)}{c}v_1 + \frac{l(v)}{c}v_1$$

and it is $v - \frac{l(v)}{c}v_1 \in \mathcal{N}(l)$ since

$$l\left(v - \frac{l(v)}{c}v_1\right) = l(v) - \frac{l(v)}{c}l(v_1) = l(v) - l(v) = 0.$$

As $v_1 \perp \mathcal{N}(l)$ from Theorem 1.7, we have

$$\begin{aligned}\left\langle c\frac{v_1}{\|v_1\|_X^2}, v\right\rangle_X &= \underbrace{\left\langle c\frac{v_1}{\|v_1\|_X^2}, v - \frac{l(v)}{c}v_1\right\rangle_X}_{=0} + \left\langle c\frac{v_1}{\|v_1\|_X^2}, \frac{l(v)}{c}v_1\right\rangle_X \\ &= l(v)\left\langle \frac{v_1}{\|v_1\|_X^2}, v_1\right\rangle_X = l(v)\frac{\|v_1\|_X^2}{\|v_1\|_X^2} = l(v).\end{aligned}$$

Therefore, $l \in \mathcal{R}(J)$ such that J is bijective and $v_l := c\frac{v_1}{\|v_1\|_X^2}$ is a Riesz-representative of l. The Riesz-representative is also unique: if there was a $\tilde{v}_l \in X$ satisfying

$$\langle \tilde{v}_l, v\rangle_X = l(v), \quad \forall v \in X$$

then

$$\langle v_l - \tilde{v}_l, v\rangle_X = l(v) - l(v) = 0, \quad \forall v \in X$$

implying $v_l = \tilde{v}_l$ which concludes the proof. □

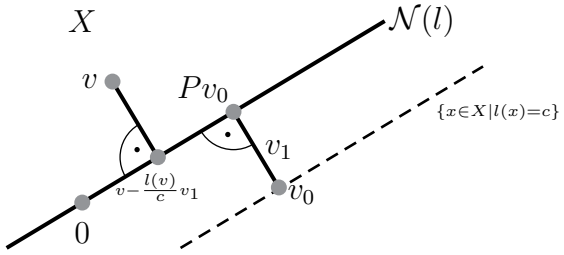


Figure 1.6: Illustration of the Riesz Representation Theorem.

Theorem 1.9 (Adjoint Operator)
Let X, Y be Hilbert spaces. For $A \in \mathcal{L}(X, Y)$ there exists a unique operator $A^* \in \mathcal{L}(Y, X)$ satisfying

$$\langle Ax, y \rangle_Y = \langle x, A^* y \rangle_X , \quad \forall x \in X,\ y \in Y,$$

the so-called adjoint Operator. *If $X = Y$ and $A = A^*$, we call A* self-adjoint.

Proof: For any $y \in Y$, it is $\langle A\cdot, y \rangle_X \in X'$ such that a unique Riesz-representative $v_{A,y} \in X$ exists. The map $A^* : Y \to X : y \mapsto v_{A,y}$ thus satisfies

$$\langle Ax, y \rangle_Y = \langle x, A^* y \rangle_X , \quad \forall x \in X,\ y \in Y$$

and is uniquely defined.
For $\lambda_1, \lambda_2 \in \mathbb{R}$, $y_1, y_2 \in Y$ and $x \in X$ we obtain the linearity of A^* via

$$\begin{aligned} \langle x, A^*(\lambda_1 y_1 + \lambda_2 y_2) \rangle_X &= \langle Ax, \lambda_1 y_1 + \lambda_2 y_2 \rangle_Y = \lambda_1 \langle Ax, y_1 \rangle_Y + \lambda_2 \langle Ax, y_2 \rangle_Y \\ &= \lambda_1 \langle x, A^* y_1 \rangle_X + \lambda_2 \langle x, A^* y_2 \rangle_X \\ &= \langle x, \lambda_1 A^* y_1 + \lambda_2 A^* y_2 \rangle_X . \end{aligned}$$

For the continuity of A^* we calculate for $y \neq 0$ with $A^* y \neq 0$

$$\begin{aligned} \|A^* y\|_X = \frac{\|A^* y\|_X^2}{\|A^* y\|_X} &\overset{x = A^* y}{=} \frac{\langle x, A^* y \rangle_X}{\|x\|_X} = \frac{\langle Ax, y \rangle_Y}{\|x\|_X} \leq \sup_{x \in X \setminus \{0\}} \frac{\langle Ax, y \rangle_Y}{\|x\|_X} \\ &\overset{CSU}{\leq} \sup_{x \in X \setminus \{0\}} \frac{\|Ax\|_Y}{\|x\|_X} \|y\|_Y = \|A\| \|y\|_Y \\ &\Rightarrow \|A^*\| \leq \|A\| . \end{aligned}$$

Since $A \in \mathcal{L}(X, Y)$, we obtain $A^* \in \mathcal{L}(Y, X)$. □

Theorem 1.10 (Spectral-Theorem)
Let X be a Hilbert space and $A \in \mathcal{K}(X, X)$ self-adjoint. Then, there exists a finite or countably infinite Orthonormalsystem of eigenvectors $\{\varphi_i\}_{i\in I}$, $I \subset \mathbb{N}$, for eigenvalues $\{\lambda_i\}_{i\in I} \subset \mathbb{R} \setminus \{0\}$ satisfying

$$Ax = \sum_{i\in I} \lambda_i \langle x, \varphi_i \rangle_X \varphi_i, \quad \forall x \in X.$$

If I is inifinite, we have $\lim_{i\to\infty} \lambda_i = 0$.

Proof: [Alt, Theorem 10.12]. □

Definition 1.11 (Bilinear Forms)
Let X be a Hilbert space and $a : X \times X \to \mathbb{R}$ a bilinear form.

(a) If

$$\gamma := \sup_{u,v\in X\setminus\{0\}} \frac{|a(u, v)|}{\|u\|_X \|v\|_X} < \infty,$$

we call a continuous *with* continuity constant γ.

(b) If

$$\alpha := \inf_{u\in X\setminus\{0\}} \frac{a(u, u)}{\|u\|_X^2} > 0,$$

we call a coercive *with* coercivity constant α.

(c) We define for $u,\ v \in X$ via

$$a_s(u, v) := \frac{1}{2}(a(u, v) + a(v, u)) \quad \text{and} \quad a_a(u, v) := \frac{1}{2}(a(u, v) - a(v, u))$$

the symmetric *and* antisymmetric *part of $a = a_s + a_a$.*

Theorem 1.12 (Operators and Bilinear Forms)
Let X be a Hilbert space.

(a) For every $A \in \mathcal{L}(X)$ there exists a continuous bilinear form $a : X\times X \to \mathbb{R}$ uniquely defined by

$$a(u, v) = \langle Au, v \rangle_X, \quad \forall u, v \in X. \tag{1.1}$$

(b) For every continuous bilinear form $a : X \times X \to \mathbb{R}$ there exists a unique $A \in \mathcal{L}(X)$ which satisfies (1.1).

Proof: (a) a defined via (1.1) is bilinear via the bilinearity of the inner product $\langle \cdot, \cdot \rangle_X$ and the linearity of $A \in \mathcal{L}(X)$. The continuity of a follows from the continuity of A as for $u, v \in X$ it is

$$\begin{aligned} |a(u,v)| = |\langle Au, v\rangle_X| &\overset{CSU}{\leq} \|A\| \|u\|_X \|v\|_X \\ \Rightarrow \frac{a(u,v}{\|u\|_X \|v\|_X} &\leq \|A\| < \infty \quad \forall u, v \in X. \end{aligned}$$

(b) Let $u \in X$ be fixed. Then, $a(u, \cdot) : X \to \mathbb{R}$ is linear and continuous since with the continuity of $a(\cdot, \cdot)$ it is $|a(u,v)| \leq \gamma \|u\|_X \|v\|_X$, for all $u, v \in X$ and thus

$$\sup_{v \in X \setminus \{0\}} \frac{|a(u,v)|}{\|v\|_X} \leq \sup_{v \in X \setminus \{0\}} \gamma \frac{\|u\|_X \cancel{\|v\|_X}}{\cancel{\|v\|_X}} = \gamma \|u\|_X < \infty.$$

Therefore, $a(u, \cdot) \in X'$ and with Theorem 1.8 there exists a unique riesz-representative $v_u \in X$ with $a(u,v) = \langle v_u, v\rangle_X$ for all $v \in X$. Define

$$A : X \to X : u \mapsto Au := v_u$$

such that A is unique and (1.1) is fulfilled. For the linearity of A, we know for $u_1, u_2 \in X$ that $u_1 + u_2 \in X$ such that $A(u_1 + u_2) = v_{u_1+u_2}$ satisfying $\langle v_{u_1+u_2}, v\rangle_X = a(u_1 + u_2, v)$ for all $v \in X$. Together with

$$\begin{aligned} \langle v_{u_1+u_2}, v\rangle_X &= a(u_1 + u_2, v) = a(u_1, v) + a(u_2, v) \\ &= \langle v_{u_1}, v\rangle_X + \langle v_{u_2}, v\rangle_X = \langle v_{u_1} + v_{u_2}, v\rangle_X, \quad \forall v \in X, \end{aligned}$$

we have $v_{u_1+u_2} = v_{u_1} + v_{u_2}$ and thus $A(u_1 + u_2) = Au_1 + Au_2$. Similarly, we obtain $A(\lambda u) = \lambda Au$ for $\lambda \in \mathbb{R}$ and $u \in X$ and the linearity of A follows. For the continuity of A, we obtain with the continuity of a for all $u \in X$ with $Au \neq 0$

$$\begin{aligned} \|Au\|_X^2 = \langle Au, Au\rangle_X = \langle v_u, Au\rangle_X = a(u, Au) &\leq \gamma \|u\|_X \|Au\|_X \\ \Rightarrow \|Au\|_X \leq \gamma \|u\|_X, \quad \forall u \in X \quad \Rightarrow \sup_{u \in X \setminus \{0\}} \frac{\|Au\|_X}{\|u\|_X} &\leq \gamma. \end{aligned}$$

This concludes the proof. □

Theorem 1.13 (Lax-Milgram)
Let X be a Hilbert space and $a : X \times X \to \mathbb{R}$ a coercive and continuous bilinear form with coercivity constant α. Then, there exists a unique operator $A \in \mathcal{L}(X)$ fulfilling

$$a(u,v) = \langle Au, v\rangle_X, \quad \forall u, v \in X.$$

Furthermore, A is bijective with $A^{-1} \in \mathcal{L}(X)$ and

$$\left\|A^{-1}\right\| \leq \frac{1}{\alpha}.$$

Proof: The existence and uniqueness of $A \in \mathcal{L}(X)$ follow from Theorem 1.12. With the coercivity of a we obtain

$$\begin{aligned} & \alpha \left\|u\right\|_X^2 \leq a(u,u) = \left\langle Au, u\right\rangle_X \overset{CSU}{\leq} \left\|Au\right\|_X \left\|u\right\|_X \\ \Rightarrow\ & \alpha \left\|u\right\|_X \leq \left\|Au\right\|_X, \quad \forall u \in X. \end{aligned} \tag{1.2}$$

For the injectivity of A, let $u \in X$ so that $Au = 0$. This implies

$$\left\|Au\right\|_X = 0 \overset{(1.2)}{\Rightarrow} \left\|u\right\|_X = 0 \Rightarrow u = 0.$$

To obtain the surjectivity, we first show that $\mathcal{R}(A)$ is closed.
Consider a Cauchy sequence $(y_i)_{i\in\mathbb{N}}$ in $\mathcal{R}(A)$ and we have a corresponding sequence $(u_i)_{i\in\mathbb{N}}$ in the inverse image of A satisfying $y_i = Au_i$ for $i \in \mathbb{N}$. Using (1.2) for $u_m - u_n \in X$, for $m, n \in \mathbb{N}$, yields

$$\alpha \left\|u_m - u_n\right\|_X \leq \left\|Au_m - Au_n\right\|_X = \left\|y_m - y_n\right\|_X \to 0, \quad m,\ n \to \infty.$$

Therefore, $(u_i)_{i\in\mathbb{N}}$ is a Cauchy sequence as well and with the completeness of X we have $\bar{u} := \lim_{i\to\infty} u_i \in X$. With A being continuous, we obtain

$$\lim_{i\to\infty} Au_i = Au \in \mathcal{R}(A)$$

and $\mathcal{R}(A)$ is closed. Since A is linear, $\mathcal{R}A$ is then a closed subspace of X so that Theorem 1.7 yields the existence of an orthogonal Projection P onto $\mathcal{R}(A)$.
Surjectivity: assume that there exists a $v \in X$ with $v \notin \mathcal{R}(A)$ such that $\bar{v} := v - Pv \neq 0$. With Theorem 1.7 we obtain $\left\langle Au, \bar{v}\right\rangle_X = 0$ for all $u \in X$. Remembering that $\bar{v} \in X$ and using the coercivity of a and get

$$0 < \alpha \left\|\bar{v}\right\|_X^2 \leq a(\bar{v}, \bar{v}) = \left\langle A\bar{v}, \bar{v}\right\rangle_X = 0,$$

which is a contradiction. Therefore, it has to be $v \in \mathcal{R}A$ such that $X \equiv \mathcal{R}A$ and A is surjective and thus bijective and A^{-1} exists. A^{-1} is linear, as for $\lambda_1,\ \lambda_2 \in \mathbb{R}$ and $u_1,\ u_2 \in X$ we know from the bijectivity of A that there exist $\tilde{u}_1,\ \tilde{u_2} \in X$ so that

$$u_1 = A\tilde{u}_1,\ u_2 = A\tilde{u}_2 \quad \Leftrightarrow \quad \tilde{u}_1 = A^{-1}u_1,\ \tilde{u}_2 = A^{-1}u_2$$

and we obtain

$$\begin{aligned} A^{-1}(\lambda_1 u_1 + \lambda_2 u_2) &= A^{-1}(\lambda_1 A\tilde{u}_1 + \lambda_2 A\tilde{u}_2) = A^{-1}(A(\lambda_1\tilde{u}_1 + \lambda_2\tilde{u}_2)) \\ &= \lambda_1\tilde{u}_1 + \lambda_2\tilde{u}_2 = \lambda_1 A^{-1}u_1 + \lambda_2 A^{-1}u_2. \end{aligned}$$

As for every $w \in X \setminus \{0\}$ there exists a $u \in X$ with $w := Au$, we obtain the continuity of A^{-1} via

$$\begin{aligned} &\alpha \left\|A^{-1}w\right\|_X = \alpha \|u\|_X \overset{(1.2)}{\leq} \|w\|_X \\ \Rightarrow &\left\|A^{-1}w\right\|_X \leq \frac{1}{\alpha}\|w\|_X \quad \Rightarrow \quad \frac{\|A^{-1}w\|_X}{\|w\|_X} \leq \frac{1}{\alpha} \end{aligned}$$

for all $w \in X \setminus \{0\}$ such that $\|A^{-1}\| = \sup_{w\in X\setminus\{0\}} \frac{\left\|A^{-1}w\right\|_X}{\|w\|_X} \leq \frac{1}{\alpha}$ which concludes the proof. □

1.2.2 Parameter Dependence

If not specified otherwise, X always denotes a Hilbert space and $\mathcal{P} \subset \mathbb{R}^p$ denotes a bounded parameter set.

Definition 1.14 (Parametric Linear and Bilinear Forms)
We call

(a) $l : X \times \mathcal{P} \to \mathbb{R}$ *a* parametric continuous linear form*, if*

$$l(\cdot;\mu) \in X' \quad \textit{for all } \mu \in \mathcal{P}.$$

(b) $a : X \times X \times \mathcal{P} \to \mathbb{R}$ *a* parametric (symmetric) bilinear form*, if*

$$a(\cdot,\cdot;\mu) : X \times X \to \mathbb{R}$$

is a (symmetric) bilinear form for all $\mu \in \mathcal{P}$*. Furthermore, if*

$$\gamma(\mu) := \sup_{u,v\in X\setminus\{0\}} \frac{|a(u,v;\mu)|}{\|u\|_X \|v\|_X} < \infty, \quad \forall \mu \in \mathcal{P},$$

we call $a(\cdot,\cdot;\mu)$ parametrically continuous *with* continuity constant $\gamma(\mu)$ *and if*

$$\alpha(\mu) := \inf_{u\in X\setminus\{0\}} \frac{a(u,u;\mu)}{\|u\|_X^2} > 0, \quad \forall \mu \in \mathcal{P},$$

we call $a(\cdot,\cdot;\mu)$ parametrically coercive *with* coercivity constant $\alpha(\mu)$.

A parametrically continuous linear or bilinear form does not have to be continuous with respect to μ. A counter example for a linear form is

$$X = \mathbb{R}, \quad \mathcal{P} = [0,1], \quad l : X \times \mathcal{P} \to \mathbb{R} \text{ with } l(x;\mu) = \begin{cases} x, & \mu \leq \frac{1}{2}, \\ \frac{1}{2}x, & \mu > \frac{1}{2}. \end{cases}$$

We formalize properties of linear and bilinear forms with respect to μ.

Definition 1.15 (Parametric Properties of Forms)
We call

(a) a parametrically continuous linear form l or bilinear form a uniformly continuous w.r.t *μ if there exist $\bar{\gamma}_l,\ \bar{\gamma} < \infty$ satisfying*

$$\sup_{\mu \in \mathcal{P}} \|l(\cdot;\mu)\|_{X'} \leq \bar{\gamma}_l \quad \text{or} \quad \sup_{\mu \in \mathcal{P}} \gamma(\mu) \leq \bar{\gamma}.$$

(b) a parametrically coercive bilinear form a uniformly coercive w.r.t. *μ if there exists an $\bar{\alpha} > 0$ satisfying*

$$\inf_{\mu \in \mathcal{P}} \alpha(\mu) \geq \bar{\alpha}.$$

(c) a linear form l or a bilinear form a Lipschitz-continuous w.r.t. *μ if there exist $L_l,\ L_a \geq 0$ satisfying*

$$|l(u;\mu_1) - l(u;\mu_2)| \leq L_l \|u\|_X \|\mu_1 - \mu_2\|_2, \quad \forall u \in X,\ \forall \mu_1, \mu_2 \in \mathcal{P},$$

or

$$|a(u,v;\mu_1) - a(u,v;\mu_2)| \leq L_a \|u\|_X \|v\|_X \|\mu_1 - \mu_2\|_2$$

for all $u, v \in X$ and all $\mu_1, \mu_2 \in \mathcal{P}$.

In the following we define an important property required for the computational efficiency of RB-methods that will be investigated in Section 1.3.3.

Definition 1.16 (Parameter-Separability)
We call

(a) a function $v : \mathcal{P} \to X$ parameter separable*, if there exist* components *$v_q \in X$ and* coefficient functions *$\theta_q^v : \mathcal{P} \to \mathbb{R}$ for $q = 1, \dots Q_v$ satisfying*

$$v(\mu) = \sum_{q=1}^{Q_v} \theta_q^v(\mu) v_q, \quad \forall \mu \in \mathcal{P}.$$

(b) *a parametric continuous linear form* $l : X \times \mathcal{P} \to \mathbb{R}$ parameter separable, *if there exist* $l_q \in X'$ *and* $\theta_q^l : \mathcal{P} \to \mathbb{R}$ *for* $q = 1, \dots, Q_l$ *satisfying*

$$l(v; \mu) = \sum_{q=1}^{Q_l} \theta_q^l(\mu) l_q(v), \quad \forall \mu \in \mathcal{P}, v \in X.$$

(c) *a parametrically continuous bilinear form* $a : X \times X \times \mathcal{P} \to \mathbb{R}$ parameter separable, *if there exist bilinear and continuous* $a_q : X \times X \to \mathbb{R}$ *and* $\theta_q^a : \mathcal{P} \to \mathbb{R}$ *for* $q = 1, \dots, Q_a$ *satisfying*

$$a(u, v; \mu) = \sum_{q=1}^{Q_a} \theta_q^a(\mu) a_q(u, v), \quad \forall \mu \in \mathcal{P}, \ u, v \in X.$$

We will see in Section 1.3.3, that Q_a and Q_l should preferably be small.

Lemma 1.17 (Transfer of Coefficient Properties)
Let l *be a parametric continuous, parameter separable linear form and* a *a parametrically continuous, parameter separable bilinear form.*

(a) *If* $\theta_q^l(\mu)$ *or* $\theta_q^a(\mu)$ *are bounded for all* $q = 1, \dots, Q_l$ *or* $q = 1, \dots Q_a$, *then* l *or* a *are uniformly continuous w.r.t.* μ.

(b) *If there exists a constant* $c > 0$ *such that* $\theta_q^a(\mu) \geq c$ *for all* $\mu \in \mathcal{P}$, $q = 1, \dots, Q_a$, *and if* $a_q(v, v) \geq 0$ *for all* $v \in X$, $q = 1, \dots Q_a$, *and if* $a(\cdot, \cdot; \mu)$ *is coercive for at least one* $\bar{\mu} \in \mathcal{P}$, *then* a *is uniformly coercive w.r.t.* μ.

(c) *If* $\theta_q^l(\mu)$ *or* $\theta_q^a(\mu)$ *are Lipschitz-continuous for all* $q = 1, \dots, Q_l$ *or* $q = 1, \dots, Q_a$, *then* l *or* a *are Lipschitz-continuous w.r.t.* μ.

Proof: See Exercise 2.1. □

We close this section with a comprehensive example.

Example 1.18 (Thermal Block)
For $B_1, B_2 \in \mathbb{N}$, *let* $\Omega := (0, 1)^2$ *be decomposed into* $p := B_1 \cdot B_2$ *congruent rectangles* Ω_i, $i = 1, \dots, p$, *see Figure 1.7. Let* $\mathcal{P} := [\mu_{\min}, \mu_{\max}]^p \subset \mathbb{R}^p$ *be the parameter set with* $0 < \mu_{\min} < \mu_{max} < \infty$.

Then, $\mu := (\mu_1, \dots, \mu_p)^\intercal \in \mathcal{P}$ *is a vector of heat conductivities on the subdomains and* $\sigma(x; \mu)$ *defines a piecewise constant heat conductivity field via*

$$\sigma(x; \mu) = \sum_{q=1}^{p} \mu_q \chi_{\Omega_q}(x).$$

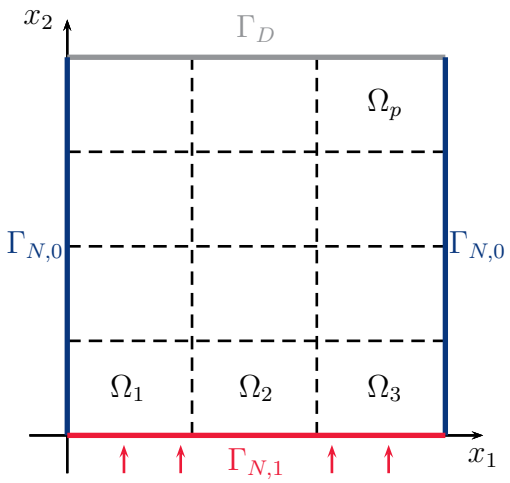


Figure 1.7: Illustration of Thermal Block domain for $B_1 = 3$, $B_2 = 4$.

Now, consider the following elliptic partial differential align with Dirichlet and Neumann boundary conditions.

$$\begin{aligned} -\nabla \cdot (\sigma(x;\mu)\nabla u(x;\mu)) &= 0, \quad \text{for } x \in \Omega, \\ u(x;\mu) &= 0, \quad \text{for } x \in \Gamma_D, \\ (\sigma(x;\mu)\nabla u(x;\mu)) \cdot n(x) &= i, \quad \text{for } x \in \Gamma_{N,i},\ i = 0, 1, \end{aligned}$$

where the boundary parts Γ_D, $\Gamma_{N,0}$, and $\Gamma_{N,1}$ are as in Figure 1.7 and $n(x)$ denotes the outward unit normal.

Interpretation: there is a unit influx on $\Gamma_{N,1}$, no flow on $\Gamma_{N,0}$, no heat sources inside Ω and a prescribed cooling of the material to $u = 0$ on Γ_D, while μ prescribes the heat conductivity of the material on the subdomains $\Omega_1, \ldots, \Omega_p$.

The weak form of this problem is: for $\mu \in \mathcal{P}$ find $u(\mu) \in H^1_{\Gamma_D}(\Omega)$ with $H^1_{\Gamma_D}(\Omega) := \{u \in H^1(\Omega) \mid u\,|_{\Gamma_D} = 0\}$ satisfying

$$\underbrace{\int_\Omega \sigma(x;\mu)\nabla u(x;\mu) \cdot \nabla v(x)\,\mathrm{d}x}_{=:a(u,v;\mu)} = \underbrace{\int_{\Gamma_{N,1}} v(x)\,\mathrm{d}x}_{=:f(v;\mu)}, \quad \forall v \in H^1_{\Gamma_D}(\Omega). \qquad (\diamond)$$

One might also desire the average temperature over $\Gamma_{N,1}$ as a possible output

$$s(\mu) := \int_{\Gamma_{N,1}} u(x;\mu)dx.$$

We will show in Exercise 2.2 that

- *$a(\cdot,\cdot;\mu)$ defined in ($\diamond$) is a parameter separable continuous bilinear form, which is uniformly coercive and continuous and Lipschitz-continuous w.r.t. μ.*
- *$f(\cdot;\mu)$ defined in ($\diamond$) is a parameter separable continuous linear form, which is uniformly continuous and Lipschitz-continuous w.r.t. μ.*

1.3 Reduced Basis Methods for linear coercive Problems

If not specified otherwise, X always denotes a Hilbert space and $\mathcal{P} \subset \mathbb{R}^p$ denotes a bounded parameter set.

1.3.1 Problem Formulation and Properties

Definition 1.19 (Detailed Problem)
Let the parametric bilinear form a be uniformly coercive and continuous w.r.t. μ and let the parametric linear forms f and l be uniformly continuous w.r.t. μ. For a given $\mu \in \mathcal{P}$ find the detailed solution $u(\mu) \in X$ *and* output $s(\mu) \in \mathbb{R}$ *satisfying*

$$\begin{aligned} a(u(\mu), v; \mu) &= f(v;\mu), \quad \forall v \in X, \\ s(\mu) &= l(u(\mu);\mu). \end{aligned} \tag{$P(\mu)$}$$

Before we introduce the reduced problem and prove well-posedness of both problems, we formally define and discuss the *solution manifold*.

Definition 1.20 (Solution Manifold)
We define the solution manifold $\mathcal{M}$ *of the detailed problem ($P(\mu)$) as*

$$\mathcal{M} := \{u(\mu) \mid u(\mu) \text{ solves } (P(\mu)) \text{ for } \mu \in \mathcal{P}\}.$$

The term "manifold" is not used in the strict differential geometric sense here, since we do not assume continuity or differentiability of $\mathcal{M}$. We want to better understand the solution manifold via the following examples.

Example 1.21 ($\mathcal{M}$ of Thermal Block)
As shown in Exercise 2.2, the problem proposed in Example 1.18 is an instance of problem ($P(\mu)$). Regarding its solution manifold, we make the following observations.

1. *Simple solution structure: if* $B_1 = 1$ *(or* $B_1 > 1$ *but the* μ_q *are identical in each row) the solution* $u(\mu)$ *is piecewise linear and constant in* x_1*-direction. Thus the solution manifold is contained in a* B_2*-dimensional and thus finite dimensional subspace of* $H^1_{\Gamma_D}(\Omega)$*. This will be shown in Exercise 2.3.*

2. *Complex solution structure: if* $B_1 > 1$ *and* $\mu \in \mathcal{P}$ *is arbitrary, no finite dimensional space contains all solutions* $u(\mu)$*.*

3. *Parameter redundancy: if* $u(\mu)$ *is a solution for* $\mu \in \mathcal{P}$ *and* $c \in \mathbb{R} \setminus \{0\}$*, then* $\frac{1}{c}u(\mu)$*, is a solution for* $c\mu \in \mathcal{P}$*. Thus, the solution manifold is invariant under scaling of the parameter* μ*.*

Example 1.22 ($\mathcal{M}$ in the case of $Q_a = 1$)
If a *and* f *in* $(P(\mu))$ *are parameter separable with* $Q_a = 1$ *and* $Q_f \in \mathbb{N}$ *arbitrary,* $\mathcal{M}$ *is contained in an at most* Q_f*-dimensional linear subspace of* X*. In this case* $(P(\mu))$ *reads*

$$\theta_1^a(\mu)a_1(u,v) = \sum_{q=1}^{Q_f} \theta_q^f(\mu) f_q(v), \quad \forall v \in X.$$

Due to the uniform coercivity of a*, it is* $\theta_1^a(\mu) \neq 0$ *and thus*

$$a_1(u,v) = \sum_{q=1}^{Q_f} \frac{\theta_q^f(\mu)}{\theta_1^a(\mu)} f_q(v), \quad \forall v \in X. \tag{$\diamond$}$$

Furthermore, w.l.o.g. $a_1(\cdot,\cdot)$ *is coercive (otherwise* $-a_1(\cdot,\cdot)$ *is coercive) such that the Lax-Milgram Theorem 1.13 yields for each* $q = 1, \dots, Q_f$*,*

$$\begin{aligned} a_1(u,v) = f_q(v) \quad &\Leftrightarrow \quad \langle Au, v\rangle_X = \langle v_{f,q}, v\rangle_X, \quad \forall v \in X \\ \Leftrightarrow Au = v_{f,q} \quad &\Leftrightarrow \quad u = A^{-1}v_{f,q}, \end{aligned}$$

where $v_{f,q} \in X$ *is the unique riesz-representative of* f_q*. Now,* $u_q := A^{-1}v_{f,q}$ *is the unique solution of* $a_1(u,v) = f_q(v)$ *for all* $v \in X$ *and* $q = 1, \dots Q_f$*. In total,* $u(\mu) := \sum_{q=1}^{Q_f} \frac{\theta_q^f(\mu)}{\theta_1^a(\mu)} u_q \in \operatorname{span}\left(\{u_q\}_{q=1}^{Q_f}\right)$ *uniquely solves* $(\diamond)$ *for* $\mu \in \mathcal{P}$*, such that* $\mathcal{M}$ *is contained in an at most* Q_f*-dimensional linear subspace of* X

Example 1.23 (Arbitrary complex $\mathcal{M}$)
Let $u : \mathcal{P} \to X$ *be an arbitrarily complex mapping. Defining a detailed problem* $(P(\mu))$ *via*

$$a(u,v;\mu) := \langle u, v\rangle_X \quad \text{and} \quad f(v;\mu) := \langle u(\mu), v\rangle_X$$

results in $u(\mu)$ being the solution to this detailed problem for $\mu \in \mathcal{P}$. Thus, the corresponding solution manifold $\mathcal{M}$ can be arbitrarily complex, nonsmooth or even discontinuous.

Remark 1.24 (Solution Manifold)
From these examples we can conclude that the dimension $p \in \mathbb{N}$ of $\mathcal{P}$ does not necessarily correlate to the complexity of the solution manifold $\mathcal{M}$. In the Thermal Block Example 1.21, we have seen that many parameters do not have to induce a complex solution manifold as there might be a redundancy in the parameters. Furthermore, one can show: for $p \in \mathbb{N}$ arbitrary and suitably chosen $a(\cdot,\cdot;\mu)$ and $f(\cdot;\mu)$, problem ($P(\mu)$) may have a solution manifold $\mathcal{M}$ which is contained in a 1-dimensional subspace.

Example 1.23 demonstrates the other extreme case: even for $p = 1$, a suitably chosen ($P(\mu)$) can have a highly complex solution manifold $\mathcal{M}$, e.g., when $u(\mu)$ is a "space filling curve" (e.g. a Hilbert curve).

We now want to introduce the *reduced problem*. In order to do so, we assume to have a low-dimensional *Reduced Basis space*

$$X_N := \operatorname{span}(\Phi_N) = \operatorname{span}(\{u(\mu^{(1)}), \dots, u(\mu^{(N)})\}) \subset X \tag{1.3}$$

with *Reduced Basis* Φ_N available. The functions $u(\mu^{(i)}) \in X$ are suitably chosen *snapshots* of the detailed problem at parameter samples $\mu^{(i)} \in \mathcal{P}$. We will present sophisticated techniques to construct X_N in section 1.4 and for now only assume that the snapshots $\{u(\mu^{(i)})\}_{i=1}^N$ are linearly independent.

Definition 1.25 (Reduced Problem)
Let a detailed problem ($P(\mu)$) be given. For a given $\mu \in \mathcal{P}$ find the reduced solution $u_N(\mu) \in X_N$ *and* reduced output $s_N(\mu) \in \mathbb{R}$ *satisfying*

$$\begin{aligned} a(u_N(\mu), v; \mu) &= f(v;\mu), \quad \forall v \in X_N, \\ s_N(\mu) &= l(u_N(\mu);\mu). \end{aligned} \tag{$P_N(\mu)$}$$

Both problems ($P(\mu)$) and ($P_N(\mu)$) are well-posed.

Proposition 1.26 (Well-posedness and Stability)
For $\mu \in \mathcal{P}$ there exist a unique detailed solution $u(\mu) \in X$ and output $s(\mu)$ of ($P(\mu)$) as well as a unique reduced solution $u_N(\mu) \in X_N$ and reduced output $s_N(\mu)$ of ($P_N(\mu)$). Furthermore, they are bounded by

$$\|u_N(\mu)\|_X \leq \frac{1}{\alpha(\mu)} \|f(\cdot;\mu)\|_{X'} \leq \frac{\bar{\gamma}_f}{\bar{\alpha}}, \qquad \|u(\mu)\|_X \leq \frac{\bar{\gamma}_f}{\bar{\alpha}},$$

$$|s_N(\mu)| \leq \frac{1}{\alpha(\mu)} \|f(\cdot;\mu)\|_{X'} \|l(\cdot;\mu)\|_{X'} \leq \frac{\bar{\gamma}_f \bar{\gamma}_l}{\bar{\alpha}}, \qquad |s(\mu)| \leq \frac{\bar{\gamma}_f \bar{\gamma}_l}{\bar{\alpha}},$$

where $\bar{\alpha}$ is the uniform coercivity constant of a and $\bar{\gamma}_f$ and $\bar{\gamma}_l$ are the uniform continuity constants of f and l.

Proof: We only prove the statements for the reduced problem ($P_N(\mu)$), the statements for ($P(\mu)$) follow analogously.

With $X_N \subset X$ the stability constants of the full problem carry over to the reduced problem:

$$\sup_{u,v\in X_N\setminus\{0\}} \frac{|a(u,v;\mu)|}{\|u\|_X \|v\|_X} \leq \sup_{u,v\in X\setminus\{0\}} \frac{|a(u,v;\mu)|}{\|u\|_X \|v\|_X} = \gamma(\mu) \leq \bar{\gamma},$$
$$\inf_{u\in X_N\setminus\{0\}} \frac{a(u,u;\mu)}{\|u\|_X^2} \geq \inf_{u\in X\setminus\{0\}} \frac{a(u,u;\mu)}{\|u\|_X^2} = \alpha(\mu) \geq \bar{\alpha}.$$

Analogously, the linear forms f and l are continuous on X_N and existence and uniqueness of $u_N(\mu)$ follow from Theorem 1.13 via

$$a(u_N(\mu),v;\mu) = f(v;\mu) \quad \Rightarrow \quad \langle A(\mu)u_N(\mu),v\rangle_X = \langle v_f(\mu),v\rangle_X \quad \forall v\in X_N$$
$$\Leftrightarrow A(\mu)u_N(\mu) = v_f(\mu) \quad \Leftrightarrow \quad u_N(\mu) = A^{-1}(\mu)v_f(\mu),$$

where $v_f \in X$ is the unique riesz-representative of $f(\cdot;\mu)$. As a consequence, $s_N(\mu) = l(u_N(\mu);\mu)$ is unique and the boundedneess follows via

$$\begin{aligned}\|u_N(\mu)\|_X &= \left\|A^{-1}(\mu)v_f(\mu)\right\|_X \leq \left\|A^{-1}(\mu)\right\| \|v_f(\mu)\|_X \\ &\leq \frac{1}{\alpha(\mu)} \|f(\cdot;\mu)\|_{X'} \leq \frac{\bar{\gamma}_f}{\bar{\alpha}},\\ |s_N(\mu)| &= |l(u_N(\mu))| = |\langle v_l(\mu),u_N(\mu)\rangle_X| \leq \|v_l(\mu)\|_X \|u_N(\mu)\|_X \\ &\leq \frac{1}{\alpha(\mu)} \|f(\cdot;\mu)\|_{X'} \|l(\cdot;\mu)\|_{X'} \leq \frac{\bar{\gamma}_f\bar{\gamma}_l}{\bar{\alpha}},\end{aligned}$$

where $v_l \in X$ is the unique riesz-representative of $l(\cdot;\mu)$. □

Furthermore, if the bilinear form and the linear forms are Lipschitz-continuous w.r.t. μ, the solutions of ($P(\mu)$) and ($P_N(\mu)$) are Lipschitz-continuous w.r.t. μ as well.

Proposition 1.27 (Lipschitz-Continuity)
In addition to the assumptions of ($P(\mu)$), let the bilinear form a and the linear forms f and l be Lipschitz-continuous w.r.t. μ with Lipschitz-constants L_a, L_f, and L_l. Then,

$$\|u(\mu_1)-u(\mu_2)\|_X \leq C_1\|\mu_1-\mu_2\|_2, \quad \|u_N(\mu_1)-u_N(\mu_2)\|_X \leq C_1\|\mu_1-\mu_2\|_2,$$
$$|s(\mu_1)-s(\mu_2)| \leq C_2\|\mu_1-\mu_2\|_2, \quad |s_N(\mu_1)-s_N(\mu_2)| \leq C_2\|\mu_1-\mu_2\|_2,$$

for all $\mu_1,\mu_2 \in \mathcal{P}$ with Lipschitz-constants $C_1 := \frac{L_f}{\bar{\alpha}} + \frac{L_a\bar{\gamma}_f}{\bar{\alpha}^2}$ as well as $C_2 := L_l\frac{\bar{\gamma}_f}{\bar{\alpha}} + \bar{\gamma}_l C_1$.

Proof: See Exercise 3.1. □

From a computational point of view, solving problem ($P_N(\mu)$) amounts to solving a simple linear align system.

Proposition 1.28 (Discrete Reduced Problem & Numerical Stability) *For $\mu \in \mathcal{P}$ and a given reduced basis $\Phi_N = \{\varphi_1, \dots, \varphi_N\}$ we define the following matrix, right hand side and output vector*

$$\begin{aligned}\mathbf{A}_N(\mu) &:= (a(\varphi_j, \varphi_i; \mu))_{i,j=1}^N \in \mathbb{R}^{N\times N},\\ \mathbf{f}_N(\mu) &:= (f(\varphi_i; \mu))_{i=1}^N \in \mathbb{R}^N, \quad \mathbf{l}_N(\mu) := (l(\varphi_i; \mu))_{i=1}^N \in \mathbb{R}^N.\end{aligned}$$

(a) By solving the following linear system for $\mathbf{u}_N(\mu) := (u_{N,i})_{i=1}^N \in \mathbb{R}^N$

$$\mathbf{A}_N(\mu)\mathbf{u}_N(\mu) = \mathbf{f}_N(\mu)$$

we obtain the solution of ($P_N(\mu)$) via

$$u_N(\mu) = \sum_{j=1}^N u_{N,j}\varphi_j \quad \text{and} \quad s_N(\mu) = \mathbf{l}_N^\mathsf{T}(\mu)\mathbf{u}_N(\mu).$$

(b) If $a(\cdot,\cdot;\mu)$ is symmetric and Φ_N is orthonormal, then the condition number of $\mathbf{A}_N(\mu)$ is bounded (independently of N) by

$$\mathrm{cond}_2(\mathbf{A}_N(\mu)) := \|\mathbf{A}_N(\mu)\|_2 \left\|\mathbf{A}_N(\mu)^{-1}\right\|_2 \le \frac{\gamma(\mu)}{\alpha(\mu)}.$$

Proof: The statement of (a) follows by inserting the expressions for $u_N(\mu)$ and $s_N(\mu)$ into ($P_N(\mu)$).

Regarding (b): since $a(\cdot,\cdot;\mu)$ is symmetric and coercive, $\mathbf{A}_N(\mu)$ is symmetric and positive definite and we have $\mathrm{cond}_2(\mathbf{A}_N(\mu)) = \frac{\lambda_{\max}}{\lambda_{\min}}$ with $\lambda_{\max}$, $\lambda_{\min}$ being the largest/smallest magnitude eigenvalue of $\mathbf{A}_N(\mu)$.

Let $\mathbf{u}_{\max} = (u_i)_{i=1}^N \in \mathbb{R}^N$ be an eigenvector of $\mathbf{A}_N(\mu)$ for the eigenvalue $\lambda_{\max}$ and set $u_{\max} := \sum_{i=1}^N u_i\varphi_i \in X$. Then, we have on the one hand

$$\|u_{\max}\|_X^2 = \left\langle \sum_{i=1}^N u_i\varphi_i, \sum_{j=1}^N u_j\varphi_j \right\rangle_X = \sum_{i,j=1}^N u_i u_j \langle\varphi_i, \varphi_j\rangle_X = \sum_{i=1}^N u_i^2 = \|\mathbf{u}_{\max}\|_2^2,$$

and on the other hand by definition of $\mathbf{A}_N(\mu)$ and the continuity of $a(\cdot,\cdot;\mu)$

$$\lambda_{\max}\|\mathbf{u}_{\max}\|_2^2 = \mathbf{u}_{\max}^\intercal \lambda_{\max}\mathbf{u}_{\max} = \mathbf{u}_{\max}^\intercal \mathbf{A}_N(\mu)\mathbf{u}_{\max} = \sum_{i,j=1}^{N} u_i u_j a(\varphi_i,\varphi_j;\mu)$$
$$= a\left(\sum_{i=1}^{N} u_i\varphi_i, \sum_{j=1}^{N} u_j\varphi_j\right) \leq \gamma(\mu)\,\|u_{\max}\|_X^2 .$$

In total, we conclude $\lambda_{\max} \leq \gamma(\mu)$ and as $\lambda_{\min} \geq \alpha(\mu)$ can be shown analogously we obtain the desired result. □

At this point we can note some differences between the reduced problem and a discretized full problem.

Remark 1.29 (FEM vs. RB)
Let $\mathbf{A}_H \in \mathbb{R}^{H\times H}$ for some large $H \in \mathbb{N}$ denote the Finite Element (or Finite Difference, Finite Volume, etc.) matrix of the linear system of ($P(\mu)$). Then we note that

- *the RB-matrix $\mathbf{A}_N \in \mathbb{R}^{N\times N}$, with $N \ll H$, is small but typically dense (as the basis functions φ_i usually do not have a disjoint support) in contrast to $\mathbf{A}_H$ which is large but typically sparse,*
- *the condition number of $\mathbf{A}_N$ does not deteriorate with growing N if an orthonormal basis is used, in contrast to the large $\mathbf{A}_H$ whose condition number usually grows polynomially in H.*

A basis property of the RB methodology is the reproduction of solutions. The property trivially follows if one has error estimators available (as we will see later), but it can also be proven without. It states, that if a solution of ($P(\mu)$) is in the reduced space, the reduced basis approximation is exact.

Proposition 1.30 (Reproduction of Solutions)
If $u(\mu) \in X_N$ for some $\mu \in \mathcal{P}$, then $u_N(\mu) = u(\mu)$.

Proof: If $u(\mu) \in X_N$ the error $e(\mu) = u(\mu) - u_N(\mu)$ is in X_N as well. Since $u(\mu)$ and $u_N(\mu)$ are solutions of ($P(\mu)$) and ($P_N(\mu)$) respectively, we obtain with the coercivity of a

$$\alpha(\mu)\,\|e(\mu)\|_X^2 \leq a(e(\mu),e(\mu);\mu) = a(u(\mu),e(\mu);\mu) - a(u_N(\mu),e(\mu);\mu)$$
$$= f(e(\mu);\mu) - f(e(\mu);\mu) = 0$$

such that $e(\mu) = 0$. □

Remark 1.31 (Validation of implementation)
The reproduction property from Proposition 1.30 is useful to validate your implementation of a RB-scheme. By setting the reduced basis Φ_N as a snapshot basis, i.e.,

$$\Phi_N := \{\varphi_1, \dots, \varphi_N\} = \{u(\mu^{(1)}), \dots, u(\mu^{(N)})\},$$

the reduced solution for $\mu^{(i)}$ then must be $\mathbf{u}_N(\mu^{(i)}) = \mathbf{e}_i$, the i-th unit vector, since $u_N(\mu^{(i)}) = \sum_{j=1}^{N} \delta_{ji}\varphi_j$ (with δ_{ji} denoting the Kronecker δ) is a solution of ($P_N(\mu)$) and the solution has to be unique.

1.3.2 Error analysis & Error estimators

As first approximation property, the RB-approximation will always be as good as the best-approximation, up to a constant.

Lemma 1.32 (Céa's Lemma)
For all $\mu \in \mathcal{P}$ holds

$$\|u(\mu) - u_N(\mu)\|_X \leq \frac{\gamma(\mu)}{\alpha(\mu)} \inf_{v \in X_N} \|u(\mu) - v\|_X .$$

Proof: For any $v \in X_N$ it is $v - u_N(\mu) \in X_N$ and we obtain

$$\begin{aligned} a(u(\mu) - u_N(\mu), & v - u_N(\mu); \mu) \\ &= a(u(\mu), v - u_N(\mu); \mu) - a(u_N(\mu), v - u_N(\mu); \mu) \\ &= f(v - u_N(\mu); \mu) - f(v - u_N(\mu); \mu) = 0. \end{aligned}$$

Coercivity and continuity of a yield

$$\begin{aligned} \alpha(\mu) \|u(\mu) - u_N(\mu)\|_X^2 &\leq a(u(\mu) - u_N(\mu), u(\mu) - u_N(\mu); \mu) \\ &= a(u(\mu) - u_N(\mu), v - u_N(\mu); \mu) \\ &\quad + a(u(\mu) - u_N(\mu), u(\mu) - v; \mu) \\ &= a(u(\mu) - u_N(\mu), u(\mu) - v; \mu) \\ &\leq \gamma(\mu) \|u(\mu) - u_N(\mu)\|_X \|u(\mu) - v\|_X \\ &\Rightarrow \|u(\mu) - u_N(\mu)\|_X \leq \frac{\gamma(\mu)}{\alpha(\mu)} \|u(\mu) - v\|_X , \ \forall v \in X_N, \end{aligned}$$

which concludes the proof. □

The property of Céa's lemma is also called "quasi-optimality" of the RB-approximation and it is important to note that the constant $\frac{\gamma(\mu)}{\alpha(\mu)}$ does not grow with increasing dimension N of the RB-space. Furthermore, we can say that the approximation error will be small if the space X_N is "good" in the sense that $\inf_{v \in X_N} \|u(\mu) - v\|_X$ is small.

For further error analysis of the reduced basis error, we provide the following error-residual relation, where the error satisfies a variational problem with the same bilinear form but the residual as the right hand side.

Proposition 1.33 (Error-Residual Relation)
For $\mu \in \mathcal{P}$ we define the residual $r(\cdot;\mu) \in X'$ as

$$r(v;\mu) = f(v;\mu) - a(u_N(\mu), v;\mu), \quad v \in X. \tag{1.4}$$

The reduced basis error $e(\mu) := u(\mu) - u_N(\mu) \in X$ *then satisfies*

$$a(e(\mu), v;\mu) = r(v;\mu), \quad v \in X.$$

Proof: A simple calculation yields

$$\begin{aligned} a(e(\mu), v;\mu) &= a(u(\mu), v;\mu) - a(u_N(\mu), v;\mu) \\ &= f(v;\mu) - a(u_N(\mu), v;\mu) = r(v;\mu), \end{aligned}$$

which concludes the proof. □

In particular, the residual vanishes on X_N since for $v \in X_N$ it is $a(u_N(\mu), v;\mu) = f(v;\mu)$ and thus $r(v;\mu) = 0$.

In the following, we discuss an important topic for RB-methods, the *certification* of the method via a-posteriori error bounds, which will be based on the residual. Here, we assume to have a rapidly computable lower bound $\alpha_{LB}(\mu)$ of the coercivity constant available.

Proposition 1.34 (A-posteriori Error Bounds)
Let for $\mu \in \mathcal{P}$ denote $v_r(\mu) \in X$ be the riesz-representative of the residual defined in (1.4) and $v_l(\mu) \in X$ denote the riesz-representative of the output functional $l(\cdot;\mu)$ from ($P_N(\mu)$). Then, for all $\mu \in \mathcal{P}$, the reduced basis error $e(\mu) = u(\mu) - u_N(\mu)$ is bounded by

$$\|u(\mu) - u_N(\mu)\|_X \leq \Delta_N(\mu) := \frac{\|v_r(\mu)\|_X}{\alpha_{LB}(\mu)}$$

and the output error $|s(\mu) - s_N(\mu)|$ is bounded by

$$|s(\mu) - s_N(\mu)| \leq \Delta_s(\mu) := \|v_l(\mu)\|_X \, \Delta_N(\mu).$$

Proof: With coercivity and the error-residual relation with $e(\mu) \in X$ we obtain

$$\begin{aligned}\alpha_{LB}(\mu) \left\|e(\mu)\right\|_X^2 \leq \alpha(\mu) \left\|e(\mu)\right\|_X^2 \leq a(e(\mu), e(\mu); \mu) = r(e(\mu); \mu) \\ = (v_r(\mu), e(\mu)) \leq \left\|v_r(\mu)\right\|_X \left\|e(\mu)\right\|_X .\end{aligned}$$

Division by $e(\mu)$ and $\alpha_{LB}(\mu)$ yield the first result. For the output error we obtain

$$\begin{aligned}|s(\mu) - s_N(\mu)| = |l(u(\mu); \mu) - l(u_N(\mu; \mu)| = |\left\langle v_l(\mu), e(\mu)\right\rangle_X| \\ \leq \left\|v_l(\mu)\right\|_X \left\|e(\mu)\right\|_X \leq \left\|v_l(\mu)\right\|_X \Delta_N(\mu),\end{aligned}$$

which concludes the proof. □

Remark 1.35 (RB-error bound)

(a) Bounding the error by the residual is a technique well known from finite element methods (FEM) where the FEM-solution is compared to the analytical solution. In that case, X is infinite-dimensional and $\left\|v_r\right\|_X$ is not available analytically. In our case, by using a fine discrete FEM-space as X, the residual norm $\left\|v_r\right\|_X$ then becomes a computable quantity, which is available after *the reduced solution $u_N(\mu)$ is computed. Therefore, our error bound derived in Proposition 1.34 is an* a-posteriori *bound.*

(b) Since the error bounds in Proposition 1.34 are provable upper bounds, we call them rigorous *error bounds. Thus, we not only obtain an RB-approximation but also a* certification *via a rigorous error bound, which motivates to call the approach made so far a* certified *RB-method.*

Having an error bound available, we want to investigate how tight the bound is. As a first desirable property, we can show that the bound is zero when the error is zero.

Proposition 1.36 (Vanishing Error Bound)
If $u(\mu) = u_N(\mu)$ for some $\mu \in \mathcal{P}$, then $\Delta_N(\mu) = \Delta_s(\mu) = 0$.

Proof: With $e(\mu) = u(\mu) - u_N(\mu) = 0$, we have

$$0 = a(0, v) = a(e(\mu), v) = r(v; \mu) = \left\langle v_r(\mu), v\right\rangle_X , \quad \forall v \in X$$

such that $v_r(\mu) = 0$. Therefore, $\Delta_N(\mu) = 0$ and thus $\Delta_s(\mu) = 0$. □

This statement gives hope that the a-posteriori error bounds are *effective*, meaning that the amount of overestimation is bounded. In particular we want to investigate the quotient of the error bound and the true error. For now, we only consider the error bound for the reduced basis error and tackle the error bound for the output error later.

Proposition 1.37 (Effectivity Bound)
For $\mu \in \mathcal{P}$ let $\gamma_{UB}(\mu) \geq \gamma(\mu)$ denote a (computable) upper bound of the continuity constant. The effectivity *is then defined as and bounded by*

$$\eta_N(\mu) := \frac{\Delta_N(\mu)}{\|u(\mu) - u_N(\mu)\|_X} \leq \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}.$$

Proof: The definition of the error bound yields

$$\eta_N(\mu) = \frac{\Delta_N(\mu)}{\|e(\mu)\|_X} = \frac{\|v_r(\mu)\|_X}{\alpha_{LB}(\mu)\,\|e(\mu)\|_X}.$$

With the error-residual relation and the continuity, we obtain

$$\begin{aligned} \|v_r(\mu)\|_X^2 = \langle v_r(\mu), v_r(\mu)\rangle_X = r(v_r(\mu);\mu) = a(e(\mu), v_r(\mu);\mu) \\ \leq \gamma(\mu)\,\|e(\mu)\|_X\,\|v_r(\mu)\|_X\,. \end{aligned} \tag{1.5}$$

This implies

$$\eta_N(\mu) \leq \frac{\gamma(\mu)}{\alpha_{LB}(\mu)} \leq \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)},$$

which concludes the proof. □

As $\Delta_N(\mu)$ is now proven to be reliable and effective, we wan call it an *error estimator*.

We can also derive an error estimator for the relative error.

Proposition 1.38 (Relative Error Estimator and Effectivity)
If for all $\mu \in \mathcal{P}$

$$\Delta_N^{rel}(\mu) := \frac{2}{\|u_N(\mu)\|_X} \cdot \frac{\|v_r(\mu)\|_X}{\alpha_{LB}(\mu)} \leq 1,$$

then the relative reduced basis error is bounded by

$$\frac{\|u(\mu) - u_N(\mu)\|_X}{\|u(\mu)\|_X} \leq \Delta_N^{rel}(\mu).$$

The bound is then also effective via

$$\eta_N^{rel}(\mu) := \frac{\Delta_N^{rel}(\mu)}{\|e(\mu)\|_X \,/\, \|u(\mu)\|_X} \leq 3 \cdot \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}.$$

Proof: Using the error estimator $\Delta_N(\mu)$ we get

$$\frac{\|u(\mu) - u_N(\mu)\|_X}{\|u(\mu)\|_X} \leq \frac{1}{\|u(\mu)\|_X} \cdot \frac{\|v_r(\mu)\|_X}{\alpha_{LB}(\mu)}$$

and it remains to show that

$$\frac{1}{\|u(\mu)\|_X} \leq \frac{2}{\|u_N(\mu)\|_X} \Leftrightarrow \frac{1}{2} \|u_N(\mu)\|_X \leq \|u(\mu)\|_X .$$

Using the reverse triangle inequality, we obtain

$$\begin{aligned} \left| \frac{\|u(\mu)\|_X - \|u_N(\mu)\|_X}{\|u_N(\mu)\|_X} \right| &\leq \frac{\|u(\mu) - u_N(\mu)\|_X}{\|u_N(\mu)\|_X} = \frac{\|e(\mu)\|_X}{\|u_N(\mu)\|_X} \\ &\leq \frac{\|v_r(\mu)\|_X}{\alpha_{LB}(\mu) \|u_N(\mu)\|_X} = \frac{1}{2} \Delta_N^{rel}(\mu) \leq \frac{1}{2}. \end{aligned}$$

If $\|u_N(\mu)\|_X > \|u(\mu)\|_X$, we get $\|u_N(\mu)\|_X - \|u(\mu)\|_X \leq \frac{1}{2} \|u_N(\mu)\|_X$ from the above statement and it follows

$$\frac{1}{2} \|u_N(\mu)\|_X \leq \|u(\mu)\|_X .$$

If $\|u_N(\mu)\|_X \leq \|u(\mu)\|_X$, we directly obtain $\frac{1}{2} \|u_N(\mu)\|_X \leq \|u(\mu)\|_X$ and the error bound for the relative error follows.

For the effectivity, we note that as in (1.5) we have $\|v_r(\mu)\|_X \leq \gamma_{UB}(\mu) \|e(\mu)\|_X$ and obtain

$$\begin{aligned} \eta_N^{rel}(\mu) &= \frac{2 \cdot \|v_r(\mu)\|_X}{\alpha_{LB}(\mu) \|u_N(\mu)\|_X} \cdot \frac{1}{\|e(\mu)\|_X / \|u(\mu)\|_X} \\ &\leq 2 \frac{\gamma_{UB}(\mu) \cancel{\|e(\mu)\|_X}}{\alpha_{LB}(\mu) \cancel{\|e(\mu)\|_X}} \cdot \frac{\|u(\mu)\|_X}{\|u_N(\mu)\|_X}. \end{aligned}$$

It remains to show that $\|u(\mu)\|_X \leq \frac{3}{2} \|u_N(\mu)\|_X$. If $\|u(\mu)\|_X < \|u_N(\mu)\|_X$ this is obvious, if $\|u(\mu)\|_X \geq \|u_N(\mu)\|_X$, we obtain from

$$\left| \frac{\|u(\mu)\|_X - \|u_N(\mu)\|_X}{\|u_N(\mu)\|_X} \right| \leq \frac{1}{2}$$

that

$$\|u(\mu)\|_X - \|u_N(\mu)\|_X \leq \frac{1}{2} \|u_N(\mu)\|_X$$

and thus $\|u(\mu)\|_X \leq \frac{3}{2} \|u_N(\mu)\|_X$. □

Effectivity bounds for the output estimator $\Delta_s(\mu)$ can not be proven without further assumptions. This is due to $\frac{\Delta_s(\mu)}{|s(\mu)-s_N(\mu)|}$ not being well-defined as soon as $\Delta_s(\mu) \neq 0$ but $s(\mu) = s_N(\mu)$. This can be obtained as follows: choose X_N and $\mu \in \mathcal{P}$ such that $u_N(\mu) \neq u(\mu)$ which is achieved by $u(\mu) \notin X_N$. Then

$$e(\mu) \neq 0 \quad \Rightarrow \quad v_r(\mu) \neq 0 \quad \Rightarrow \quad \Delta_N(\mu) \neq 0 \text{ and } \Delta_s(\mu) \neq 0.$$

Now, choosing $l(v;\mu) := \langle v_l, v \rangle_X \in X'$ with $v_l \in X$ such that $v_l \perp u(\mu) - u_N(\mu)$ yields

$$s(\mu) - s_N(\mu) = l(u(\mu) - u_N(\mu);\mu) = \langle v_l, u(\mu) - u_N(\mu) \rangle_X = 0.$$

1.3.2.1 Estimators for Symmetric Bilinear Forms

In the following we assume that the bilinear form a is, in addition to the assumptions made in Definition 1.19, symmetric. Based on that we can define the μ-dependent *energy norm*.

Lemma 1.39 (Energy Norm)
For $\mu \in \mathcal{P}$ and all $u, v \in X$ we define the form

$$\langle u, v \rangle_\mu := a(u, v; \mu).$$

This is a positive definite form and thus a scalar product. We call the norm induced by this the energy norm

$$\|u\|_\mu := \sqrt{\langle u, u \rangle_\mu}.$$

The energy norm is equivalent to the X-norm

$$\sqrt{\alpha(\mu)}\, \|u\|_X \leq \|u\|_\mu \leq \sqrt{\gamma(\mu)}\, \|u\|_X\,, \quad \forall u \in X.$$

Proof: For $u \in X$ and $\mu \in \mathcal{P}$ we have by coercivity

$$0 \leq \alpha(\mu)\, \|u\|_X^2 \leq a(u, u; \mu) = \langle u, u \rangle_\mu$$

so that $\langle \cdot, \cdot \rangle_\mu$ is positive definite. Symmetry and bilinearity are inherited from the bilinear form such that $\langle \cdot, \cdot \rangle_\mu$ is a scalar product. We obtain the equivalency to the X-norm via coercivity and continuity

$$\begin{aligned} &\alpha(\mu)\, \|u\|_X^2 \leq a(u, u; \mu) = \|u\|_\mu^2 \leq \gamma(\mu)\, \|u\|_X^2 \\ \Leftrightarrow &\sqrt{\alpha(\mu)}\, \|u\|_X \leq \|u\|_\mu \leq \sqrt{\gamma(\mu)}\, \|u\|_X \end{aligned}$$

which concludes the proof. □

We can show that the reduced solution is the orthogonal projection with respect to the energy scalar product.

Proposition 1.40 (Galerkin Projection & Galerkin-Orthogonality) *Let $X_N \subset X$. Let for $\mu \in \mathcal{P}$ denote $\mathcal{P}_\mu : X \to X_N$ denote the orthogonal projection with respect to $\langle \cdot, \cdot \rangle_\mu$ and let $u(\mu), u_N(\mu)$ be solutions of ($P(\mu)$), ($P_N(\mu)$) respectively. We then have*

$$u_N(\mu) = \mathcal{P}_\mu u(\mu)$$

which implies $\langle u(\mu) - u_N(\mu), v \rangle_\mu = 0$ for all $v \in X_N$, the so-called Galerkin-Orthogonality.

Proof: Due to the norm equivalency in Lemma 1.39, $(X, \langle \cdot, \cdot \rangle_\mu)$ is a Hilbert space and since $(X_N, \langle \cdot, \cdot \rangle_\mu)$ is finite dimensional it is a closed subspace. Therefore, the orthogonal projection $\mathcal{P}_\mu$ is well-defined and the orthogonality of the projection error from Theorem 1.7 yields for all $v \in X_N$

$$\begin{aligned} & \langle \mathcal{P}_\mu u(\mu) - u(\mu), v \rangle_\mu = 0 \\ \Leftrightarrow \; & a(\mathcal{P}_\mu u(\mu) - u(\mu), v; \mu) = 0 \\ \Leftrightarrow \; & a(\mathcal{P}_\mu u(\mu), v; \mu) = b(u(\mu), v; \mu) \\ \Leftrightarrow \; & a(\mathcal{P}_\mu u(\mu), v; \mu) = f(v; \mu). \end{aligned}$$

Therefore, $\mathcal{P}_\mu u(\mu)$ is a solution of ($P_N(\mu)$) and as the solution has to be unique due to Proposition 1.26 we get $u_N(\mu) = \mathcal{P}_\mu u(\mu)$. □

For non-symmetric bilinear forms, we previously already used the simpler "Galerkin-Orthogonality"

$$a(u(\mu) - u_N(\mu), v; \mu) = 0, \quad \forall v \in X_N, \mu \in \mathcal{P}.$$

One consequence of the previous result is that $u_N(\mu)$ is the best approximation with respect to the energy norm. We further obtain an improvement of Lemma 1.32 in this symmetric case.

Corollary 1.41 (Energy Norm: Error Statements) *Let $\mu \in \mathcal{P}$ and $u(\mu), u_N(\mu)$ be solutions of ($P(\mu)$), ($P_N(\mu)$) respectively.*

(a) The error in the (μ-dependent) energy norm satisfies

$$\|u(\mu) - u_N(\mu)\|_\mu = \inf_{v \in X_N} \|u(\mu) - v\|_\mu .$$

(b) The error in the (μ-independent) X-norm satisfies

$$\|u(\mu) - u_N(\mu)\|_X \leq \sqrt{\frac{\gamma(\mu)}{\alpha(\mu)}} \inf_{v \in X_N} \|u(\mu) - v\|_X .$$

Proof: The first statement follows from Proposition 1.40 and Theorem 1.7 as

$$\|u(\mu) - u_N(\mu)\|_\mu = \|u(\mu) - \mathcal{P}_\mu u(\mu)\|_\mu = \inf_{v \in X_N} \|u(\mu) - v\|_\mu .$$

The second statement follows from the equivalency of the X-norm and the energy norm as well as the first statement

$$\begin{aligned}\sqrt{\alpha(\mu)}\, \|u(\mu) - u_N(\mu)\|_X \leq \|u(\mu) - u_N(\mu)\|_\mu &= \inf_{v \in X_N} \|u(\mu) - v\|_\mu \\ &\leq \sqrt{\gamma(\mu)} \inf_{v \in X_N} \|u(\mu) - v\|_X ,\end{aligned}$$

which concludes the proof. □

An immediate consequence is a monotone decrease of the reduced basis error in the energy norm for so-called *hierarchical* RB-spaces.

Corollary 1.42 (Monotone Error Decrease in Energy Norm)
Let $\{X_N\}_{N=1}^{N_{\max}}$ be a sequence of hierarchical *RB-spaces, i.e., $X_N \subseteq X_{N'}$ for $N \leq N'$. For $\mu \in \mathcal{P}$ we define $e_N(\mu) := u(\mu) - u_N(\mu)$ and the sequence $\{\|e_N(\mu)\|_\mu\}_{N=1}^{N_{\max}}$ is then monotone decreasing.*

Proof: Since $X_N \subseteq X_{N'}$ for $N \leq N'$, we obtain

$$\begin{aligned}\|u(\mu) - u_N(\mu)\|_\mu = \inf_{v \in X_N} \|u(\mu) - v\|_\mu &\geq \inf_{v \in X_{N'}} \|u(\mu) - v\|_\mu \\ &= \|u(\mu) - u_{N'}(\mu)\|_\mu ,\end{aligned}$$

which was the desired result. □

Remark 1.43 (RB-error Decrease in X-norm)
We do not obtain a strictly monotone decrease of $\|e_N(\mu)\|_\mu$. Nevertheless, the "worst case" scenario of a stagnating error is unrealistic as every new basis vector would have to be orthogonal to the projection error. In fact, with clever basis construction one can obtain exponential convergence as we will see in section 1.4.

The error in a different norm $\|e_N(\mu)\|_\star$ does not have to decrease monotonously. But, as the RB-spaces are finite dimensional, we have

$$c\, \|e_N(\mu)\|_\mu \leq \|e_N(\mu)\|_\star \leq C\, \|e_N(\mu)\|_\mu$$

with some constants $c, C > 0$ independent of N. Therefore, the error $\|e_N(\mu)\|_\star$ does remain inside a "corridor" given by the energy norm.

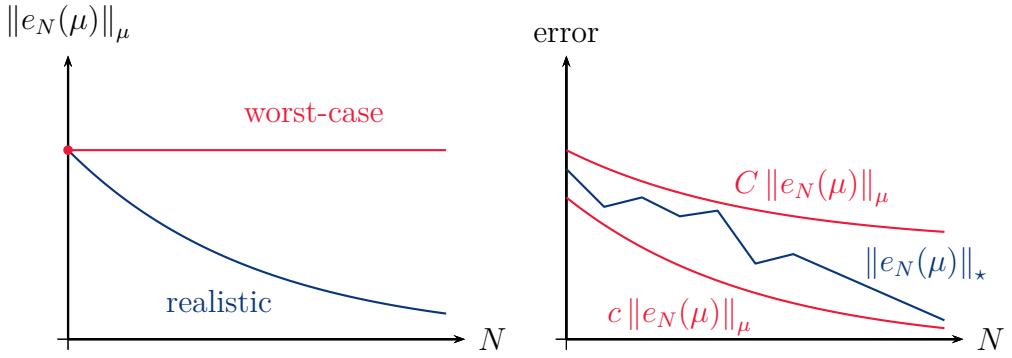


Figure 1.8: Error decrease for increasing dimension N of X_N.

Remark 1.44 (Uniform convergence of Lagrange-RB-Approximation) *Let $\mathcal{P} \subset \mathbb{R}^p$ be compact. We define for $S_N := \{\mu^{(1)}, \dots, \mu^{(N)}\} \subset \mathcal{P}$, for $N \in \mathbb{N}$, the so-called* fill-distance *and the* Lagrange-RB-space

$$h_N := \sup_{\mu\in\mathcal{P}} \inf_{\mu'\in S_N} \|\mu - \mu'\|_2 \quad \text{and} \quad X_N = \operatorname{span}(\{u(\mu^{(i)}) \mid \mu^{(i)} \in S_N\}).$$

Furthermore, let $u(\mu)$, $u_N(\mu)$ be Lipschitz-continuous w.r.t. μ with Lipschitz-constant L_u that is independent of N (this is for example the case in the setting of Proposition 1.27) and let for $\mu \in \mathcal{P}$ denote $\mu^* := \arg\min_{\mu'\in S_N} \|\mu - \mu'\|_2$ the "closest" parameter. Then, we obtain for all $\mu \in \mathcal{P}$ and $N \in \mathbb{N}$

$$\begin{aligned}\|u(\mu) - u_N(\mu)\|_X &\le \|u(\mu) - u(\mu^*)\|_X + \|u(\mu^*) - u_N(\mu^*)\|_X + \|u_N(\mu^*) - u_N(\mu)\|_X \\ &\le L_u \|\mu - \mu^*\|_2 + 0 + L_u \|\mu - \mu^*\|_2 \le 2h_N L_u,\end{aligned}$$

where $\|u(\mu^*) - u_N(\mu^*)\|_X = 0$ due to the reproductions of solutions from Proposition 1.30. Now, choosing S_N such that the fill distance converges to zero, i.e., $\lim_{N\to\infty} h_N = 0$, we can conclude uniform convergence of the RB-error as

$$\lim_{N\to\infty} \sup_{\mu\in\mathcal{P}} \|u(\mu) - u_N(\mu)\|_X = 0.$$

But, the linear convergence rate in h_N is practically not relevant, since h_N decays very slowly and thus N would have to be very large to guarantee a small error. This further motivates to investigate sophisticated basis construction techniques in Section 1.4.

Proposition 1.45 (Energy Norm: Error Estimators & Effectivities)
For all $\mu \in \mathcal{P}$ we have the error estimator for the error in the energy norm

$$\|u(\mu) - u_N(\mu)\|_\mu \leq \Delta_N^{en}(\mu) := \frac{\|v_r(\mu)\|_X}{\sqrt{\alpha_{LB}(\mu)}}$$

and the estimator is effective with

$$\eta_N^{en}(\mu) := \frac{\Delta_N^{en}(\mu)}{\|u(\mu) - u_N(\mu)\|_\mu} \leq \sqrt{\frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}}.$$

Furthermore, if for all $\mu \in \mathcal{P}$

$$\Delta_N^{en,rel}(\mu) := \frac{2}{\|u_N(\mu)\|_\mu} \cdot \frac{\|v_r(\mu)\|_X}{\sqrt{\alpha_{LB}(\mu)}} \leq 1,$$

then the relative reduced basis error in the energy norm is bounded by

$$\frac{\|u(\mu) - u_N(\mu)\|_\mu}{\|u(\mu)\|_\mu} \leq \Delta_N^{en,rel}(\mu)$$

and the estimator is effective with

$$\eta_N^{en,rel}(\mu) := \frac{\Delta_N^{en,rel}(\mu)}{\|e(\mu)\|_\mu / \|u(\mu)\|_\mu} \leq 3 \cdot \sqrt{\frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}}.$$

Proof: See Exercise 3.2. □

We obtain an effective output error estimator in the so-called *compliant* case that occurs when the output functional and the right hand side functional are equal.

Proposition 1.46 (Output Error Bound in Compliant Case)
Let $\mu \in \mathcal{P}$ and $u(\mu), u_N(\mu)$ be solutions of ($P(\mu)$), ($P_N(\mu)$) respectively. If for all $\mu \in \mathcal{P}$ we have $f(v;\mu) = l(v;\mu)$ for all $v \in X$, then the output error $s(\mu) - s_N(\mu)$ satisfies

$$0 \leq s(\mu) - s_N(\mu) = \|u(\mu) - u_N(\mu)\|_\mu^2 .$$

We thus obtain the error estimator

$$s(\mu) - s_N(\mu) \leq \Delta_s(\mu) := \frac{\|v_r(\mu)\|_X^2}{\alpha_{LB}(\mu)}$$

and the estimator is effective with

$$\eta_s(\mu) := \frac{\Delta_s(\mu)}{s(\mu) - s_N(\mu)} \leq \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}.$$

Proof: From Proposition 1.40 we know that

$$a(u_N(\mu), u(\mu) - u_N(\mu); \mu) = \langle u_N(\mu), u(\mu) - u_N(\mu) \rangle_\mu = 0$$

and since $u(\mu)$ solves $(P(\mu))$ we obtain

$$\begin{aligned} s(\mu) - s_N(\mu) &= l(u(\mu); \mu) - l(u_N(\mu); \mu) = f(u(\mu); \mu) - f(u_N(\mu); \mu) \\ &= a(u(\mu), u(\mu) - u_N(\mu); \mu) - \underbrace{a(u_N(\mu), u(\mu) - u_N(\mu); \mu)}_{=0} \\ &= a(u(\mu) - u_N(\mu), u(\mu) - u_N(\mu); \mu) = \|u(\mu) - u_N(\mu)\|_\mu^2 \end{aligned}$$

which implies $s(\mu) - s_N(\mu) \geq 0$. From Proposition 1.45 we obtain

$$s(\mu) - s_N(\mu) = \|u(\mu) - u_N(\mu)\|_\mu^2 \leq \frac{\|v_r(\mu)\|_X^2}{\alpha_{LB}(\mu)} = \Delta_s(\mu)$$

as well as

$$\eta_s(\mu) = \frac{\Delta_s(\mu)}{s(\mu) - s_N(\mu)} = \frac{(\Delta_N^{en}(\mu))^2}{\|u(\mu) - u_N(\mu)\|_\mu^2} = (\eta_N^{en}(\mu))^2 \leq \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}$$

which concludes the proof. □

Assuming $\Delta_N(\mu) \approx h \ll 1$, we observe a quadratic dependence of $\Delta_s(\mu)$ on h in this symmetric/compliant case compared to the linear dependence of $\Delta_s(\mu)$ in Proposition 1.34 such that this improved output error bound is expected to be much better.

1.3.2.2 Output Error Estimators: an additional Dual Problem

In the following we define a dual problem as well as a reduced dual problem in order to get an improved reduced output and thus obtain the "quadratic effect" observed in Proposition 1.46 also for non-symmetric/non-compliant problems. We assume that the bilinear form a and the linear form l are the forms from the detailed problem $(P(\mu))$.

Definition 1.47 (Dual Problem)
For a given $\mu \in \mathcal{P}$ find the dual solution $u^{du}(\mu) \in X$ *of*

$$a(v, u^{du}(\mu); \mu) = -l(v; \mu), \quad \forall v \in X. \qquad (P^{du}(\mu))$$

Again, we assume to have a Reduced Basis space $X_N^{du} \subset X$ with dimension $N^{du} \in \mathbb{N}$ given, where N and N^{du} can be different.

Definition 1.48 (Primal-Dual Reduced Problem)
For a given $\mu \in \mathcal{P}$ let $u_N(\mu)$ be the corresponding solution of ($P_N(\mu)$). Then, find the dual reduced solution $u_N^{du}(\mu) \in X_N^{du}$ *and* improved reduced output $s'_N(\mu) \in \mathbb{R}$ *satisfying*

$$\begin{aligned} a(v, u_N^{du}(\mu); \mu) &= -l(v; \mu), \quad \forall v \in X_N^{du}, \\ s'_N(\mu) &= l(u_N(\mu); \mu) - r(u_N^{du}(\mu); \mu). \end{aligned} \tag{$P_N^{du}(\mu)$}$$

Well-posedness of both problems ($P^{du}(\mu)$) and ($P_N^{du}(\mu)$) follow from the coercivity and continuity as in Proposition 1.26.

We observe that compared to the reduced output in ($P_N(\mu)$), the improved reduced output $s'_N(\mu)$ incorporates a "correction term" given by the residual evaluated with the dual reduced solution. This "correction" yields sharper output error bounds.

Proposition 1.49 (Dual Error Bound & Improved Output Bound)
For a given $\mu \in \mathcal{P}$ we introduce the dual residual

$$r^{du}(v; \mu) := -l(v; \mu) - a(v, u_N^{du}(\mu); \mu), \quad \forall v \in X$$

and obtain the error estimator

$$\left\| u^{du}(\mu) - u_N^{du}(\mu) \right\|_X \leq \Delta_N^{du}(\mu) := \frac{\left\| v_r^{du}(\mu) \right\|_X}{\alpha_{LB}(\mu)},$$

where $v_r^{du}(\mu) \in X$ is the Riesz-representative of the dual residual. The estimator is effective with

$$\eta_N^{du}(\mu) := \frac{\Delta_N^{du}(\mu)}{\left\| u^{du}(\mu) - u_N^{du}(\mu) \right\|_X} \leq \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)},$$

and we further obtain the improved output error bound

$$|s(\mu) - s'_N(\mu)| \leq \Delta'_s(\mu) := \frac{\left\| v_r^{du}(\mu) \right\|_X \left\| v_r(\mu) \right\|_X}{\alpha_{LB}(\mu)}.$$

Proof: See Exercise 3.3. □

Comparing this primal/dual scenario to the symmetric/compliant case in Proposition 1.46, it is easy to see that the problems ($P_N(\mu)$) and ($P_N^{du}(\mu)$) are equivalent in the symmetric/compliant case, when $X_N \equiv X_N^{du}$: as $f = l$ and a is symmetric, $u_N^{du}(\mu) \equiv -u_N(\mu)$ solves the dual problem and $r(v; \mu) = r^{du}(v; \mu)$

for all $v \in X$ and $\mu \in \mathcal{P}$ such that $\Delta_N(\mu) = \Delta_N^{du}(\mu)$ for all $\mu \in \mathcal{P}$. Furthermore, as $u_N^{du}(\mu) \in X_N$, $r(u_N^{du}(\mu);\mu) = 0$ and $s_N(\mu) = s'_N(\mu)$ for all $\mu \in \mathcal{P}$. Similarly, both problems $(P(\mu))$ and $(P^{du}(\mu))$ are equivalent in the symmetric/compliant case. Therefore, this primal/dual approach is an extension enabling an improved output and sharper output error bounds in the non-symmetric/non-compliant case.

Remark 1.50 (Summary: Relevance Error Estimators)
The error estimators

- *are rigorous upper bounds for the approximation error, not just "error indicators" as in FEM.*
- *are effective as the degree of overestimation of the error is bounded with a known bound. In particular: $e(\mu) = 0 \Leftrightarrow \Delta_N(\mu) = 0$ for all $\mu \in \mathcal{P}$ and thus "a-posteriori" verification of an exact approximation.*
- *together with the reduced solution can be efficiently computed via an Offline/Online decomposition (⇝ Section 1.3.3).*
- *can be utilized offline for the basis generation (⇝ Section 1.4), online for the dimension choice, or even in a problem specific adaptive basis generation strategy (for example in optimization or inverse problems).*

1.3.3 Offline/Online Decomposition

We now focus on the efficient implementation of the RB-methodology and start by revisiting the full problem $(P(\mu))$. In an implementation this corresponds to a high-dimensional discrete problem. Assuming that

$$X = \operatorname{span}(\{\psi_1, \dots, \psi_H\})$$

is spanned by a large number of basis functions ψ_i, we introduce the following notation

$$\mathbf{A}(\mu) := (a(\psi_i, \psi_j; \mu))_{i,j=1}^H \in \mathbb{R}^{H\times H},$$
$$\mathbf{f}(\mu) := (f(\psi_i;\mu))_{i=1}^H \in \mathbb{R}^H, \quad \mathbf{l}(\mu) := (l(\psi_i;\mu))_{i=1}^H \in \mathbb{R}^H.$$

Then, for $\mu \in \mathcal{P}$, the full problem $(P(\mu))$ can be solve by determining the coefficient vector $\mathbf{u}(\mu) = (u_i)_{i=1}^H \in \mathbb{R}^H$ of the detailed solution $u(\mu) = \sum_{i=1}^H u_i\psi_i$ by solving the linear system

$$\mathbf{A}(\mu)\mathbf{u}(\mu) = \mathbf{f}(\mu), \quad s(\mu) = \mathbf{l}(\mu)^\mathsf{T}\mathbf{u}(\mu).$$

In the following, we want to give a rough complexity analysis for computing a detailed and a reduced solution. For this, we assume that $\mathbf{A}(\mu)$ is a sparse matrix such that a solution of $(P(\mu))$ requires $\mathcal{O}(H^2)$ operations. In contrast, the solution of the reduced problem $(P_N(\mu))$ from Proposition 1.28 requires $\mathcal{O}(N^3)$ operations as the matrix $\mathbf{A}_N(\mu)$ is usually dense as mentioned in Remark 1.29. Clearly, the RB-approach only pays off computationally if $N \ll H$. Let us further investigate the relevant steps for the computation of a RB-solution:

1. construction of X_N via N snapshots of $(P(\mu))$: $\mathcal{O}(NH^2)$,
2. obtaining $\mathbf{A}_N(\mu)$ via N^2 evaluations of $a(\varphi_i, \varphi_j; \mu)$: $\mathcal{O}(N^2H)$,
3. obtaining $\mathbf{f}_N(\mu)$ via N evaluations of $f(\varphi_i; \mu)$: $\mathcal{O}(NH)$,
4. solution of the $N \times N$ linear system in $(P_N(\mu))$: $\mathcal{O}(N^3)$.

As we can see, the approach does not pay off if only the solution for a single parameter μ is required. If one needs the solutions for many different parameters, the RB-approach will pay off due to an efficient implementation via the Offline/Online decomposition. Ideally, we want to achieve the following split:

- *Offline-phase:* μ-independent, high-dimensional quantities are computed, operation count depends on H. Expensive but only done *once*.
- *Online-phase:* performed for every new $\mu \in \mathcal{P}$. The offline data is combined to give the small μ-dependent discretized reduced system and thus the reduced solution $u_N(\mu)$ and $s_N(\mu)$ can be computed rapidly. Operation count is ideally independent of H and scales polynomially in N.

Comparing this desired splitting, we can already assign step 1 above (construction of X_N) to the offline phase and step 4 (solve reduced system) to the online phase. Steps 2 and 3 can not be clearly assigned as they involve both low- and highdimensional operations. Dividing these steps into an offline and an online part, is made possible by assuming that the bilinear form a and the linear forms f, l in problem $(P(\mu))$ are parameter separable as defined in

Definition 1.16 such that

$$a(u,v;\mu) = \sum_{q=1}^{Q_a} \theta_q^a(\mu) a_q(u,v), \quad \text{and}$$
$$f(v;\mu) = \sum_{q=1}^{Q_f} \theta_q^f(\mu) f_q(v), \qquad l(v;\mu) = \sum_{q=1}^{Q_l} \theta_q^l(\mu) l_q(v),$$

for all $\mu \in \mathcal{P}$ and all $u, v \in X$.

Due to the linearity of the problem $(P_N(\mu))$ the parameter-separability of a, f, l transfers over to $\mathbf{A}_N, \mathbf{f}_N, \mathbf{l}_N$ and we obtain the Offline/Online decomposition of $(P_N(\mu))$.

Corollary 1.51 (Offline/Online decomposition of $(P_N(\mu))$)
Offline Phase

- *Compute a reduced basis* $\Phi_N = \{\varphi_1, \dots, \varphi_N\}$.
- *Construct the parameter-independent component matrices and vectors*

$$\begin{aligned} \mathbf{A}_{N,q} &:= (a_q(\varphi_i, \varphi_j))_{i,j=1}^N \mathbb{R}^{N\times N}, \quad q = 1, \dots, Q_a, \\ \mathbf{f}_{N,q} &:= (f_q(\varphi_i))_{i=1}^N \in \mathbb{R}^N, \quad q = 1, \dots, Q_f, \\ \mathbf{l}_{N,q} &:= (l_q(\varphi_i))_{i=1}^N \in \mathbb{R}^N, \quad q = 1, \dots, Q_l. \end{aligned}$$

Online Phase

- *For a given* $\mu \in \mathcal{P}$ *evaluate the coefficient functions* $\theta_q^a(\mu)$, $\theta_q^f(\mu)$, $\theta_q^l(\mu)$ *and assemble the matrix and vectors*

$$\mathbf{A}_N(\mu) = \sum_{q=1}^{Q_a} \theta_q^a(\mu)\mathbf{A}_{N,q}, \ \mathbf{f}_N(\mu) = \sum_{q=1}^{Q_f} \theta_q^f(\mu)\mathbf{f}_{N,q}, \ \mathbf{l}_N(\mu) = \sum_{q=1}^{Q_l} \theta_q^l(\mu)\mathbf{l}_{N,q}.$$

- *The resulting linear system coincides with the system in Proposition 1.28 and can thus be solved for* $u_N(\mu)$ *and* $s_N(\mu)$.

The second step in the offline phase can be realized in a simple way: let the reduced basis vectors φ_j be expanded in the basis $\{\psi_i\}_{i=1}^H$ of the discrete full problem by $\varphi_j = \sum_{i=1}^H \varphi_{i,j}\psi_i$ with the resulting coefficient matrix

$$\mathbf{\Phi}_N := (\varphi_{i,j})_{i,j=1}^{H,N} \in \mathbb{R}^{H\times N}. \tag{1.6}$$

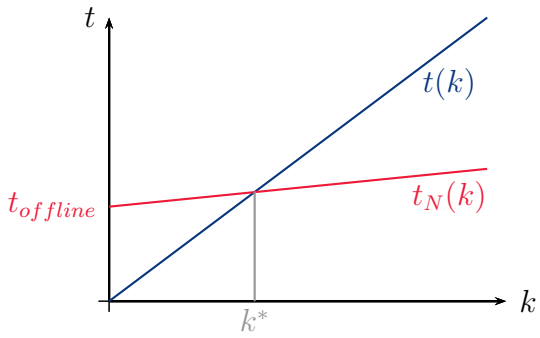


Figure 1.9: Runtime comparison of full and reduced model for increasing number k of simulations.

Then, using the component matrices and vectors of the full problem

$$\begin{aligned}\mathbf{A}_q &:= (a_q(\psi_i, \psi_j))_{i,j=1}^{H} \in \mathbb{R}^{H\times H}, \quad q = 1, \dots, Q_a,\\ \mathbf{f}_q &:= (f_q(\psi_i))_{i=1}^{H} \in \mathbb{R}^{H}, \quad q = 1, \dots, Q_f,\\ \mathbf{l}_q &:= (l_q(\psi_i))_{i=1}^{H} \in \mathbb{R}^{H}, \quad q = 1, \dots Q_l,\end{aligned}$$

we obtain the reduced matrices and vectors via

$$\mathbf{A}_{N,q} = \mathbf{\Phi}_N^{\mathsf{T}} \mathbf{A}_q \mathbf{\Phi}_N, \quad \mathbf{f}_{N,q} = \mathbf{\Phi}_N^{\mathsf{T}} \mathbf{f}_q, \quad \mathbf{l}_{N,q} = \mathbf{\Phi}_N^{\mathsf{T}} \mathbf{l}_q.$$

Regarding the complexities, we observe that the offline phase scales in the order of $\mathcal{O}(NH^2 + NH(Q_f + Q_l) + N^2HQ_a)$ and the online phase scales in the order of $\mathcal{O}(N^2Q_a + N(Q_f + Q_l) + N^3)$ and thus completely independent of H.

Let t_{full}, $t_{offline}$, t_{online} denote the computational time required for a solution of $(P(\mu))$, the offline phase of $(P_N(\mu))$ and the online phase of $(P_N(\mu))$. Then, assuming that these times do not vary for different $\mu \in \mathcal{P}$, we need $t(k) := k \cdot t_{full}$ for k full solutions and $t_N(k) := t_{offline} + k \cdot t_{online}$ for k reduced solutions, see Figure 1.9.

It can be seen that we need $k > k^* := \frac{t_{offline}}{t_{full} - t_{online}}$ simulation requests before the reduced model pays off.

Remark 1.52 (No Discrimination between $u(\mu)$ and $u_h(\mu)$)
Throughout this chapter, we usually have no separate notation for the discrete detailed solution $u_h(\mu)$ (FEM) and the detailed solution $u(\mu)$ (weak solution). This can now be motivated as follows:

- *As the online phase of ($P_N(\mu)$) is independent of H and the offline phase is only done once, H can be chosen arbitrarily large. Therefore, $u_h(\mu)$ will be arbitrarily accurate (due to suitably refined grids) such that $u(\mu)$ and $u_h(\mu)$ are practically the same ($\varepsilon(\mu) := \|u(\mu) - u_h(\mu)\|_X$ is arbitrarily small) whilst ($P_N(\mu)$) is still rapidly solvable.*

- *Therefore, due to*

$$\|u(\mu) - u_N(\mu)\|_X \leq \varepsilon(\mu) + \|u_h(\mu) - u_N(\mu)\|_X ,$$

the true approximation error $\|u(\mu) - u_N(\mu)\|_X$ will in practice be dominated by the reduction error $\|u_h(\mu) - u_N(\mu)\|_X$. As a consequence, by controlling the reduction error (which can be controlled via our error estimators!), we also control the true approximation error up to $\varepsilon(\mu)$.

The parmeter-separability can also be exploited in the full problem.

Remark 1.53 (Decomposition of ($P(\mu)$) & Parameter-Separability)
(a) Using the parameter-separability, problem ($P(\mu)$) can also be decomposed. The component matrices $\mathbf{A}_q \in \mathbb{R}^{H\times H}$, $q = 1, \dots, Q_a$, and vectors $\mathbf{f}_q \in \mathbb{R}^H$, $q = 1, \dots, Q_f$, $\mathbf{l}_q \in \mathbb{R}^H$, $q = 1, \dots Q_l$, of the full problem can be precomputed once and then, for a new $\mu \in \mathcal{P}$, we can assemble the full system

$$\mathbf{A}(\mu) = \sum_{q=1}^{Q_a} \theta_q^a(\mu)\mathbf{A}_q, \quad \mathbf{f}(\mu) = \sum_{q=1}^{Q_f} \theta_q^f(\mu)\mathbf{f}_q, \quad \mathbf{l}(\mu) = \sum_{q=1}^{Q_l} \theta_q^l(\mu)\mathbf{l}_q$$

and solving $\mathbf{A}(\mu)\mathbf{u}(\mu) = \mathbf{f}(\mu)$ for $\mathbf{u}(\mu) = (u_i)_{i=1}^H \in \mathbb{R}^H$ yields the solution of ($P(\mu)$) via

$$u(\mu) = \sum_{i=1}^{H} u_i\psi_i \quad \text{and} \quad s(\mu) = \mathbf{l}^\intercal(\mu)\mathbf{u}(\mu).$$

Obviously, both parts of this decomposition depend on H and the only gain is, that solving ($P(\mu)$) in the presence of parameter-separability boils down to solving a large linear system.

(b) We have seen that the parameter-separability of a problem plays a major role in the computational efficiency of the RB-approach. Thus, we mention that for non-parameter-separable problems the Empirical Interpolation Method [EIM] is available.

(c) *Finally, the availability of the components* $\mathbf{A}_q$, $\mathbf{f}_q$, $\mathbf{l}_q$ *is non-trivial when third-party discretization packages are used and this is a major challenge in a broad practical application of the RB-approach.*

We want to derive an Offline/Online decomposition of the a-posteriori error estimators and thus begin with the parameter separability of the residual.

Proposition 1.54 (Parameter-Separability of the Residual)
Set $Q_r := Q_f + NQ_a$ *and let* $\Phi_N = \{\varphi_1, \dots, \varphi_N\}$ *be a reduced basis. From the Riesz-representation Theorem 1.8 we obtain the unique representatives*

$$\left\langle v_f^q, v \right\rangle_X = f_q(v), \quad \forall v \in X,\ 1 \leq q \leq Q_f$$

and

$$\left\langle v_a^{q,n}, v \right\rangle_X = a_q(\varphi_n, v), \quad \forall v \in X,\ 1 \leq q \leq Q_a,\ 1 \leq n \leq N.$$

We now define $r_q \in X'$ *and* $v_r^q \in X$ *for* $1 \leq q \leq Q_r$ *via*

$$\begin{aligned}(r_1(\cdot), \dots, r_{Q_r}(\cdot)) := (f_1(\cdot), \dots, f_{Q_f}(\cdot), &a_1(\varphi_1, \cdot), \dots, a_{Q_a}(\varphi_1, \cdot), \\ &\dots, a_q(\varphi_N, \cdot), \dots, a_{Q_a}(\varphi_N, \cdot))\end{aligned}$$

and

$$\left(v_r^1, \dots, v_r^{Q_r}\right) := \left(v_f^1, \dots, v_f^{Q_f}, v_a^{1,1}, \dots, v_a^{Q_a,1}, \dots, v_a^{1,N}, \dots, v_a^{Q_a,N}\right).$$

Letting for $\mu \in \mathcal{P}$ *be* $u_N(\mu) = \sum_{i=1}^N u_{N,i}(\mu)\varphi_i$ *the solution of* ($P_N(\mu)$), *we define* $\theta_q^r(\mu) : \mathcal{P} \to \mathbb{R},\ q = 1, \dots, Q_r$ *via*

$$\begin{aligned}\left(\theta_1^r(\mu), \dots, \theta_{Q_r}^r(\mu)\right) := (\theta_1^f(\mu), \dots, \theta_{Q_f}^f(\mu), &\\ -\theta_1^a(\mu) \cdot u_{N,1}(\mu), \dots, -\theta_{Q_a}^a(\mu) \cdot u_{N,1}(\mu), &\\ \dots, -\theta_1^a(\mu) \cdot u_{N,N}(\mu), \dots, -\theta_{Q_a}^a(\mu) \cdot u_{N,N}(\mu)).&\end{aligned}$$

Then, for all $\mu \in \mathcal{P}$ *and* $v \in X$, *the residual and its riesz-representatives are parameter-separable via*

$$r(v; \mu) = \sum_{q=1}^{Q_r} \theta_q^r(\mu) r_q(v), \quad v_r(\mu) = \sum_{q=1}^{Q_r} \theta_q^r(\mu) v_r^q.$$

Proof: Using the parameter-separability of a, f and the definition of the residual, we obtain for all $\mu \in \mathcal{P}$ and $v \in X$

$$\begin{aligned}r(v; \mu) &= f(v; \mu) - a(u_N(\mu), v; \mu) \\ &= \underbrace{\sum_{q=1}^{Q_f} \theta_q^f(\mu) f_q(v) - \sum_{q=1}^{Q_a} \sum_{n=1}^{N} \theta_q^a(\mu) u_{N,n}(\mu) a_q(\varphi_n, v)}_{= \sum_{q=1}^{Q_r} \theta_q^r(\mu) r_q(v)}\end{aligned}$$

and thus

$$\begin{aligned}\langle v_r(\mu), v\rangle_X &= \sum_{q=1}^{Q_f} \theta_q^f(\mu)\left\langle v_f^q, v\right\rangle_X - \sum_{q=1}^{Q_a}\sum_{n=1}^{N} \theta_q^a(\mu) u_{N,n}(\mu)\left\langle v_a^{q,n}, v\right\rangle_X \quad \forall v \in X \\ &= \Bigg\langle \underbrace{\sum_{q=1}^{Q_f} \theta_q^f(\mu) v_f^q - \sum_{q=1}^{Q_a}\sum_{n=1}^{N} \theta_q^a(\mu) u_{N,n}(\mu) v_a^{q,n}}_{\sum_{q=1}^{Q_r} \theta_q^r(\mu) v_r^q}, v \Bigg\rangle_X \quad \forall v \in X,\end{aligned}$$

which concludes the proof. □

Remembering $X = \operatorname{span}(\{\psi_1, \dots, \psi_H\})$, we introduce

$$\mathbf{K} := (\langle \psi_i, \psi_j\rangle)_{i,j=1}^H \in \mathbb{R}^{H\times H}$$

the inner product matrix of X which is typically sparse. This allows for an easy computation of riesz-representatives.

Lemma 1.55 (Computation of Riesz-representatives)
Let $g \in X'$ and $\mathbf{v} = (v_i)_{i=1}^H \in \mathbb{R}^H$ be the coefficient vector of its riesz-representative $v_g = \sum_{i=1}^H v_i\psi_i \in X$. Introducing $\mathbf{g} := (g(\psi_i))_{i=1}^H \in \mathbb{R}^H$, we can obtain $\mathbf{v}$ by solving the linear system

$$\mathbf{K}\mathbf{v} = \mathbf{g}.$$

Proof: For any test function $w = \sum_{i=1}^H w_i\psi_i \in X$ with coefficient vector $\mathbf{w} = (w_i)_{i=1}^H \in \mathbb{R}^H$ we obtain

$$g(w) = \sum_{i=1}^H w_i g(\psi_i) = \mathbf{w}^\mathsf{T}\mathbf{g}$$

and

$$\langle v_g, w\rangle_X = \left\langle \sum_{i=1}^H v_i\psi_i, \sum_{j=1}^H w_j\psi_j \right\rangle_X = \mathbf{w}^\mathsf{T}\mathbf{K}\mathbf{v}.$$

Since $g(w) = \langle v_g, w\rangle$ has to hold for all $w \in X$ this is equivalent to $\mathbf{g} = \mathbf{K}\mathbf{v}$. □

We continue with the Offline/Online decomposition of the residual norm and the solution norms of the relative error estimators.

Proposition 1.56 (Offline/Online decomposition of error estimators)
Offline Phase

After the offline phase of Corollary 1.51 and after the computation of v_r^q, $1 \leq q \leq Q_r$, *according to Proposition 1.54, we define the matrices*

$$\mathbf{K}_N := \left(\langle \varphi_i, \varphi_j \rangle_X\right)_{i,j=1}^N \in \mathbb{R}^{N\times N} = \mathbf{\Phi}_N^\mathsf{T} \mathbf{K} \mathbf{\Phi}_N,$$

where $\mathbf{\Phi}_N$ *was the coefficient matrix of the reduced basis defined in (1.6), and*

$$\mathbf{G}_r := \left(\left\langle v_r^q, v_r^{q'} \right\rangle_X\right)_{q,q'=1}^{Q_r} \in \mathbb{R}^{Q_r \times Q_r}.$$

Online Phase

For a given $\mu \in \mathcal{P}$ *and corresponding reduced solution* $u_N(\mu) = \sum_{i=1}^N u_{N,i}\varphi_i$ *with coefficient vector* $\mathbf{u}_N(\mu) := (u_{N,i})_{i=1}^N \in \mathbb{R}^N$ *we compute the residual coefficient vector* $\theta_r(\mu) := \left(\theta_1^r(\mu), \dots, \theta_{Q_r}^r(\mu)\right)^\mathsf{T} \in \mathbb{R}^{Q_r}$. *Remembering the reduced system matrix* $\mathbf{A}_N(\mu)$ *from Corollary 1.51, we obtain*

$$\begin{aligned}
\|v_r(\mu)\|_X &= \sqrt{\theta_r(\mu)^\mathsf{T} \mathbf{G}_r \theta(\mu)},\\
\|u_N(\mu)\|_X &= \sqrt{\mathbf{u}_N(\mu)^\mathsf{T} \mathbf{K}_N \mathbf{u}_N(\mu)},\\
\|u_N(\mu)\|_\mu &= \sqrt{\mathbf{u}_N(\mu)^\mathsf{T} \mathbf{A}_N \mathbf{u}_N(\mu)}.
\end{aligned}$$

Proof: Straight forward calculations reveal

$$\begin{aligned}
\|v_r(\mu)\|_X^2 &= \left\langle \sum_{q=1}^{Q_r} \theta_q^r(\mu) v_r^q, \sum_{q'=1}^{Q_r} \theta_{q'}^r(\mu) v_r^{q'} \right\rangle_X = \sum_{q,q'=1}^{Q_r} \theta_q^r(\mu)\theta_{q'}^r(\mu) \left\langle v_r^q, v_r^{q'} \right\rangle_X\\
&= \theta_r(\mu)^\mathsf{T} \mathbf{G}_r \theta(\mu),\\
\|u_N(\mu)\|_X^2 &= \left\langle \sum_{i=1}^{N} u_{N,i}\varphi_i, \sum_{j=1}^{N} u_{N,j}\varphi_j \right\rangle_X = \sum_{i,j=1}^{N} u_{N,i}u_{N,j} \left\langle \varphi_i, \varphi_j \right\rangle_X\\
&= \mathbf{u}_N(\mu)^\mathsf{T} \mathbf{K}_N \mathbf{u}_N(\mu),\\
\|u_N(\mu)\|_\mu^2 &= a(u_N(\mu), u_N(\mu); \mu) = \sum_{i,j=1}^{N} u_{N,i}u_{N,j} a(\varphi_i, \varphi_j; \mu)\\
&= \mathbf{u}_N(\mu)^\mathsf{T} \mathbf{A}_N \mathbf{u}_N(\mu),
\end{aligned}$$

which concludes the proof. □

The last ingredient required for the error estimators is $\alpha_{LB}(\mu)$, a lower bound for the coercivity constant. Surely, for some model problems the true coercivity constant $\alpha(\mu)$ is available and rapidly computable (Exercise 2.2) and one can choose $\alpha_{LB}(\mu) \equiv \alpha(\mu)$. For general problems, one can rapidly compute a lower bound via the following *min-theta-approach*.

Lemma 1.57 (Min-Theta-Approach for $\alpha_{LB}(\mu)$)
Let the components and coefficient functions of $a(\cdot,\cdot;\mu)$ satisfy $a_q(u,u) \geq 0$ and $\theta_q^a(\mu) > 0$, for $q = 1,\ldots,Q_a$ and all $u \in X$, $\mu \in \mathcal{P}$. Furthermore, let $\bar{\mu} \in \mathcal{P}$ such that $\alpha(\bar{\mu})$ is available. Then, we have

$$0 < \alpha_{LB}(\mu) \leq \alpha(\mu), \quad \forall \mu \in \mathcal{P},$$

with the lower bound

$$\alpha_{LB}(\mu) := \alpha(\bar{\mu}) \cdot \min_{q=1,\ldots,Q_a} \frac{\theta_q^a(\mu)}{\theta_q^a(\bar{\mu})}.$$

Proof: As $\alpha(\mu) > 0$ and $\theta_q^a(\mu) > 0$ for all $\mu \in \mathcal{P}$ we have

$$\alpha_{LB}(\mu) = \alpha(\bar{\mu}) \cdot \min_{q=1,\ldots,Q_a} \frac{\theta_q^a(\mu)}{\theta_q^a(\bar{\mu})} > 0.$$

By definition, we have

$$\alpha(\mu) = \inf_{u \in X \setminus \{0\}} \frac{a(u,u;\mu)}{\|u\|_X^2}$$

and it is sufficient to show that $a(u,u;\mu) \geq \alpha_{LB}(\mu)\,\|u\|_X^2$ for all $u \in X$. We calculate

$$\begin{aligned}
a(u,u,;\mu) &= \sum_{q=1}^{Q_a} \theta_q^a(\mu) a_q(u,u) = \sum_{q=1}^{Q_a} \frac{\theta_q^a(\mu)}{\theta_q^a(\bar{\mu})} \theta_q^a(\bar{\mu}) a_q(u,u) \\
&\geq \sum_{q=1}^{Q_a} \left(\min_{q'=1,\ldots,Q_a} \frac{\theta_{q'}^a(\mu)}{\theta_{q'}^a(\bar{\mu})} \right) \theta_q^a(\bar{\mu}) a_q(u,u) \\
&= \left(\min_{q'=1,\ldots,Q_a} \frac{\theta_{q'}^a(\mu)}{\theta_{q'}^a(\bar{\mu})} \right) a(u,u;\bar{\mu}) \geq \left(\min_{q'=1,\ldots,Q_a} \frac{\theta_{q'}^a(\mu)}{\theta_{q'}^a(\bar{\mu})} \right) \alpha(\bar{\mu})\,\|u\|_X^2 \\
&= \alpha_{LB}(\mu)\,\|u\|_X^2\,,
\end{aligned}$$

which is valid for all $u \in X$ such that the statement follows. □

For the min-theta-approach we require $\alpha(\bar{\mu})$ for one $\bar{\mu} \in \mathcal{P}$ and we can compute this via a high-dimensional eigenvalue problem.

Proposition 1.58 (Computation of $\alpha(\mu)$ for ($P(\mu)$))
We remember $\mathbf{A}(\mu)$, $\mathbf{K} \in \mathbb{R}^{H \times H}$ *the system matrix of* $(P(\mu))$ *and the inner product matrix of* X. *Defining* $\mathbf{A}_s(\mu) := \frac{1}{2}(\mathbf{A}(\mu) + \mathbf{A}(\mu))^\mathsf{T}$ *as the* symmetric part *of* $\mathbf{A}$, *we obtain*

$$\alpha(\mu) = \lambda_{\min}(\mathbf{K}^{-1}\mathbf{A}_s(\mu))$$

where $\lambda_{\min}$ *denotes the smallest eigenvalue.*

Proof: See Exercise 4.1. □

In order to prevent the inversion of $\mathbf{K}$, one can either use an eigenvalue solver that requires only matrix-vector products so that as soon as a product $\mathbf{w} = \mathbf{K}^{-1}\mathbf{A}_s\mathbf{v}$, for $\mathbf{w}$, $\mathbf{v} \in \mathbb{R}^H$, has to be evaluated, one can solve the system $\mathbf{K}\mathbf{w} = \mathbf{A}_s\mathbf{v}$ instead. Alternatively, $\lambda_{\min}$ can be computed as the smallest eigenvalue of the *generalized eigenvalue problem* $\mathbf{A}_s\mathbf{v} = \lambda\mathbf{K}\mathbf{w}$.

For problems, where the min-theta approach cannot be applied, the *Successive Constaint Method* [SCM] is available.

Similarly to $\alpha_{LB}(\mu)$, we can obtain $\gamma_{UB}(\mu)$, the upper bound for the continuity constant, if a is symmetric.

Lemma 1.59 (Max-Theta-Approach for $\gamma_{UB}(\mu)$)
Let $a(\cdot,\cdot;\mu)$ *be symmetric and let the components and coefficient functions satisfy* $a_q(u,u) \geq 0$ *and* $\theta_q^a(\mu) > 0$, *for* $q = 1,\dots,Q_a$ *and all* $u \in X$, $\mu \in \mathcal{P}$. *Furthermore, let* $\bar{\mu} \in \mathcal{P}$ *such that* $\gamma(\bar{\mu})$ *is available. Then, we have*

$$\gamma(\mu) \leq \gamma_{UB}(\mu) < \infty, \quad \forall \mu \in \mathcal{P},$$

with the upper bound

$$\gamma_{UB}(\mu) := \gamma(\bar{\mu}) \cdot \max_{q=1,\dots,Q_a} \frac{\theta_q^a(\mu)}{\theta_q^a(\bar{\mu})}.$$

Proof: See Exercise 4.2. □

1.4 Basis Construction

Again, if not specified otherwise, X always denotes a Hilbert space and $\mathcal{P} \subset \mathbb{R}^p$ denotes a bounded parameter set.

We begin with a formal definition of the type of reduced basis that we have been using so far.

Definition 1.60 (Lagrange Reduced Basis Space)
Let $S_N := \{\mu^{(1)}, \dots, \mu^{(N)}\} \subset \mathcal{P}$ be a sample set of parameters such that the snapshots $\{u(\mu^{(i)}\}_{i=1}^N \subset X$ *are linearly independent. We then call*

$$\Phi_N := \{u(\mu^{(1)}, \dots, u^{(N)}\}$$

a Lagrangian reduced basis. *The resulting space $X_N := \operatorname{span}(\Phi_N)$ is then called a* Lagrangian reduced basis space.

Aim of this section

- Good global approximation of the solution manifold

  $$\mathcal{M} := \{u(\mu) \mid \mu \in \mathcal{P}\}.$$

  Can be achieved by optimizing with respect to different error measures, e.g., minimizing the maximum error in the X-norm (energy norm also possible)

  $$\inf_{\substack{Y \subset X \\ \dim(Y)=N}} \sup_{\mu \in \mathcal{P}} \|u(\mu) - u_N(\mu)\|_X ,$$

  or minimizing over the mean squared projection error

  $$\inf_{\substack{Y \subset X \\ \dim(Y)=N}} \int_{\mathcal{P}} \|u(\mu) - P_Y u(\mu)\|_X^2 \, \mathrm{d}\mu.$$

- Desirable properties of the basis:
  - orthonormality to ensure numerical stability,
  - "hierarchical basis", i.e., the basis vectors posses a meaningful sorting with respect to accuracy of the reduced basis space. If $X_{N'} := \operatorname{span}(\{\varphi_1, \dots, \varphi_{N'}\})$ for $1 \leq N' \leq N$ is a sequence of spaces, then increasing N' should increase the accuracy of $X_{N'}$.
- Generate numerical basis construction procedures from the above optimization problems via several simplifications:

- discretization of $\mathcal{P}$: instead of mean/sup over $\mathcal{P}$ choose a *finite* set $S_{train} \subset \mathcal{P}$ of training parameters somehow *resembling* $\mathcal{P}$ (equidistant grid, randomly or adaptively chosen).
- instead of $Y \subset X$, search for for $Y \subset \operatorname{span}(\{u(\mu^i)\}_{i=1}^n)$ with some set of snapshots, e.g., $\mathcal{M}_{train} = \{u(\mu^i) \mid \mu^i \in S_{train}\}$.
- instead of error measure involving the full solution $u(\mu)$, use error estimators that can be rapidly evaluated.

We met various inner product matrices in the previous section. Now we formalize the construct.

Definition and Theorem 1.61 (Gramian Matrix)
Let $\{u_i\}_{i=1}^n \subset X$. We define the Gramian matrix *as*

$$\mathbf{K}_u := \left(\langle u_i, u_j \rangle_X\right)_{i,j=1}^n \in \mathbb{R}^{n \times n}.$$

$\mathbf{K}_u$ *has the following properties.*

(a) $\mathbf{K}_u$ *is symmetric and positive semidefinite.*

(b) $\operatorname{rank}(\mathbf{K}_u) = \dim\left(\operatorname{span}(\{u_i\}_{i=1}^n)\right)$.

(c) $\mathbf{K}_u$ *is positive definite* $\Leftrightarrow$ $\{u_i\}_{i=1}^n$ *are linearly independent.*

Proof: See Exercise 4.3. □

We present a first trivial reduced basis generation technique.

Definition and Theorem 1.62 (Gram-Schmidt basis & properties)
Let $\{u_i\}_{i=1}^n \subset X$ be linearly independent. We define for $1 \leq m \leq n$ the Gram-Schmidt basis $\Phi_{GR,m} := \{\varphi_1, \ldots, \varphi_m\}$ *via*

$$\bar{\varphi}_m := u_m - \sum_{i=1}^{m-1} \langle u_m, \varphi_i \rangle_X \, \varphi_i$$
$$\varphi_m := \frac{\bar{\varphi}_m}{\|\bar{\varphi}_m\|_X}$$

and by $X_{GR,m} = \operatorname{span}(\Phi_{GR,m})$ the corresponding Gram-Schmidt-RB-space.
We have the following properties for $1 \leq m \leq n$:

(a) $\Phi_{GR,m}$ *is an orthonormal basis,*

(b) $X_{GR,m} = \operatorname{span}(\{u_i\}_{i=1}^m)$,

(c) $\max_{j=1,\dots,m} \inf_{v \in X_{GR,m}} \|u_j - v\|_X = 0$.

Proof: We proof the properties.

(a) We have

$$\langle \varphi_i, \varphi_i \rangle_X = \left\langle \frac{\bar{\varphi}}{\|\bar{\varphi}_i\|_X}, \frac{\bar{\varphi}}{\|\bar{\varphi}_i\|_X} \right\rangle_X = \frac{\langle \bar{\varphi}_i, \bar{\varphi}_i \rangle_X}{\|\bar{\varphi}_i\|_X^2} = 1$$

for $i = 1, \dots, n$. The orthogonality follows via Induction: for $m = 2$, we have

$$\begin{aligned}\langle \varphi_1, \varphi_2 \rangle_X =& \frac{1}{\|\bar{\varphi}_2\|_X} \left\langle \frac{u_1}{\|u_1\|_X}, u_2 \right\rangle_X \\ & - \frac{1}{\|\bar{\varphi}_2\|_X} \underbrace{\left\langle \frac{u_1}{\|u_1\|_X}, \frac{u_1}{\|u_1\|_X} \right\rangle_X}_{=1} \cdot \left\langle u_2, \frac{u_1}{\|u_1\|_X} \right\rangle_X = 0.\end{aligned}$$

Then, assuming $\langle \varphi_i, \varphi_j \rangle_X = \delta_{ij}$ for all $1 \leq i, j \leq m' < n$, we have

$$\begin{aligned}\langle \varphi_i, \varphi_{m'+1} \rangle_X &= \frac{1}{\|\bar{\varphi}_{m'+1}\|_X} \left(\langle \varphi_i, u_{m'+1} \rangle_X - \sum_{k=1}^{(m'+1)-1} \langle u_{m'+1}, \varphi_k \rangle_X \underbrace{\langle \varphi_i, \varphi_k \rangle_X}_{\delta_{ik}} \right) \\ &= \frac{1}{\|\bar{\varphi}_{m'+1}\|_X} \left(\langle \varphi_i, u_{m'+1} \rangle_X - \langle \varphi_i, u_{m'+1} \rangle_X \right) = 0.\end{aligned}$$

(b) By construction, we have $X_{GR,m} \subset \operatorname{span}(\{u_i\}_{i=1}^m)$ and equality follows as $\dim(X_{GR,m}) = m = \dim(\operatorname{span}(\{u_i\}_{i=1}^m))$ since the $\{u_i\}_{i=1}^n \subset X$ were linearly independent.

(c) Trivially holds, as $u_j \in X_{GR,m}$ for $j \leq m$ follows from (b). □

Remark 1.63 (Gram-Schmidt)

(a) We obtain an orthonormal basis, which guarantees stability of the RB-scheme via Proposition 1.28.

(b) Resulting RB-space $X_{GR,m}$ only allows for trivial approximation statements: the snapshots $\{u_i\}_{i=1}^m$ that were used for the construction of $X_{GR,m}$ are perfectly approximated, but the remaining snapshots $\{u_i\}_{i=m+1}^n$ could be approximated arbitrarily bad.

(c) *The basis $\Phi_{GR,m}$ depends on the order of the snapshots, which is only reasonable if the snapshots $\{u_i\}_{i=1}^m$ already had a meaningful order.*

(d) *The Gram-Schmidt procedure is usually not used to create a reduced basis space, but is rather used as post-processing in order to orthonormalize a reduced basis, see, e.g., section 1.4.2.*

1.4.1 Proper Orthogonal Decomposition

Theorem 1.64 (Proper Orthogonal Decomposition)
Let $\{u_i\}_{i=1}^n \subset X$ be a given set of snapshots. We define the empirical correlation operator

$$R : X \to X : u \mapsto Ru := \frac{1}{n} \sum_{i=1}^{n} \langle u_i, u \rangle_X \, u_i.$$

R is a compact self-adjoint linear operator and there exists an orthonormal set $\{\varphi_i\}_{i=1}^{n'}$ of $0 < n' \leq n$ eigenvectors with real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{n'} > 0$ satisfying

$$Ru = \sum_{i=1}^{n'} \lambda_i \langle \varphi_i, u \rangle_X \, \varphi_i.$$

For $1 \leq m \leq n'$ we define $\Phi_{POD,m} := \{\varphi_i\}_{i=1}^m$ as the POD-basis *and via $X_{POD,m} = \operatorname{span}(\Phi_{POD,m}) \subset X$ the corresponding* POD-RB-space.

Proof: R is surely linear and also continuous via

$$\|R\| = \sup_{u \in X \setminus \{0\}} \frac{\|Ru\|_X}{\|u\|_X} = \sup_{u \in X \setminus \{0\}} \frac{\left\| \frac{1}{n} \sum_{i=1}^n \langle u_i, u \rangle_X \, u_i \right\|_X}{\|u\|_X} \overset{CSU}{\leq} \frac{1}{n} \sum_{i=1}^{n} \|u_i\|_X^2 .$$

As $\mathcal{R}(R)$ is finite dimensional, R is then also a compact operator. For $u, v \in X$ we obtain

$$\langle Ru, v \rangle_X = \frac{1}{n} \sum_{i=1}^{n} \langle u_i, u \rangle_X \langle u_i, v \rangle_X = \langle u, Rv \rangle_X$$

so that R is also self-adjoint and from the Spectral-Theorem 1.10 we obtain the desired spectral decomposition of the operator where the decomposition must be finite ($n' < \infty$) since $\mathcal{R}(R)$ is finite dimensional. Furthermore, as

$$\langle Ru, u \rangle_X = \frac{1}{n} \sum_{i=1}^{n} \langle u_i, u \rangle_X^2 , \quad \forall u \in X,$$

the operator is positive semidefinite such that it follows from [Alt, Remark 10.13(2)] that all eigenvalues of the operator are positive and there exists at least one strictly positive eigenvalue such that $n' > 0$. □

We provide a short geometrical interpretation of the POD.

Illustration

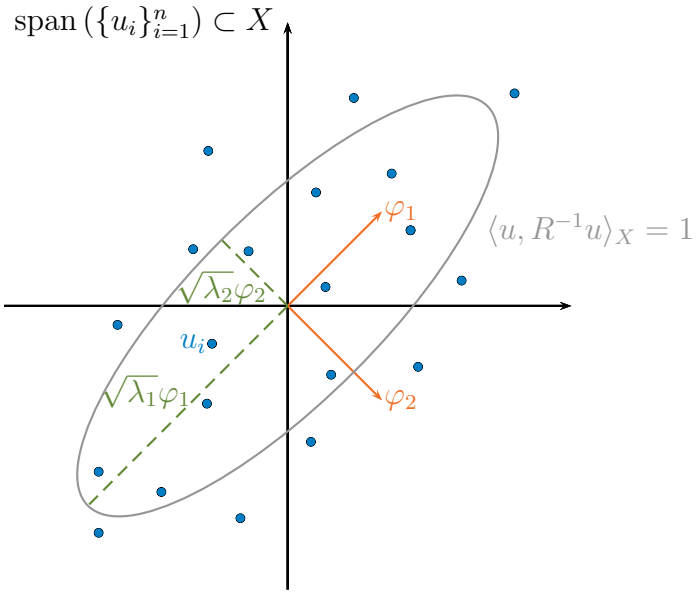


Figure 1.10: Illustration of POD.

- $\{\varphi_i\}_{i=1}^{n'}$ is an orthonormal basis for $\text{span}(\{u_i\}_{i=1}^n)$ but it is not unique (reflections, multiple eigenvalues of same magnitude).
- φ_1 is the direction of 'highest variance' of $\{u_i\}_{i=1}^n$, φ_2 is the direction of 'highest variance' of $\{P_{X_{POD,1}^\perp} u_i\}_{i=1}^n$, etc.
- The coordinates of the data w.r.t. the POD-basis, i.e., $\langle u_k, \varphi_j\rangle_X \in \mathbb{R}$, are uncorrelated, see Exercise 5.1.
- $R : \text{span}(\{u_i\}_{i=1}^n) \to \text{span}(\{u_i\}_{i=1}^n)$ is bijective. Then, $\{\varphi_i\}_{i=1}^{n'}$ and $\{\sqrt{\lambda_i}\}_{i=1}^{n'}$ are the the principal axes and principal axes sections of the ellipsoid $\{u \in X \mid \langle u, R^{-1}u\rangle_X = 1\}$, see Exercise 5.1.
- The Terminology POD, especially 'proper' relates to the french expression 'valeur propre' which directly translates to 'own value'.

- In the literature, the POD is also known as *Principal Component Analysis*, *Karhunen-Loeve-Transformation* or *Hotelling-Transformation* and the basis vectors of $\Phi_{POD,m}$ are often referred to as *POD-modes*.

In order to compute a POD-basis, one has to solve the eigenvalue problem for the correlation operator which is either high-dimensional ($\dim X = H$) or infinite-dimensional ($\dim X = \infty$). If $n < H$, one can alternatively solve an n-dimensional eigenvalue problem for the Gramian matrix of the snapshots.

Proposition 1.65 (Computation of $X_{POD,m}$ via Gramian Matrix) *Let $\{u_i\}_{i=1}^n \subset X$ be a set of snapshots with corresponding Gramian matrix $\mathbf{K}_u = \left(\langle u_i, u_j\rangle_X\right)_{i,j=1}^n \in \mathbb{R}^{n\times n}$. Then, the following statements are equivalent*

(a) $\varphi \in X$ is an eigenvector of R for an eigenvalue $\lambda > 0$ with $\|\varphi\|_X = 1$ and a representation $\varphi = \sum_{i=1}^n a_i u_i$ with $a = (a_i)_{i=1}^n \in \mathcal{N}(\mathbf{K}_u)^\perp$,

(b) $a = (a_i)_{i=1}^n \in \mathbb{R}^n$ is an eigenvector of $\frac{1}{n}\mathbf{K}_u$ for an eigenvalue $\lambda > 0$ with $\|a\|_2 = \frac{1}{\sqrt{n\lambda}}$.

Proof: Since $\mathbf{K}_u$ is a positive semidefinite real symmetric matrix, there exists an orthonormal system of eigenvectors where the eigenvectors for the zero-eigenvalue span $\mathcal{N}(\mathbf{K}_u)$ and the remaining eigenvectors span $\mathcal{R}(\mathbf{K}_u) = \mathcal{N}(\mathbf{K}_u)^\perp$. The same holds true for $\frac{1}{n}\mathbf{K}_u$.

(b) $\Rightarrow$ (a): let $a = (a_i)_{i=1}^n \in \mathbb{R}^n$ be an eigenvector of $\frac{1}{n}\mathbf{K}_u$ for an eigenvalue $\lambda > 0$ with $\|a\|_2 = \frac{1}{\sqrt{n\lambda}}$ so that

$$\lambda a = \frac{1}{n}\mathbf{K}_u a.$$

Multiplying the i-th component of this align with u_i yields

$$\lambda a_i u_i = \frac{1}{n}\sum_{j=1}^n \langle u_i, u_j\rangle_X a_j u_i$$

and summation over i yields

$$\lambda \sum_{i=1}^n a_i u_i = \sum_{i=1}^n \frac{1}{n}\sum_{j=1}^n \langle u_i, u_j\rangle_X a_j u_i.$$

Defining $\varphi := \sum_{j=1}^n u_j a_j$ yields

$$\lambda\varphi = \lambda\sum_{i=1}^n a_i u_i = \frac{1}{n}\sum_{i=1}^n \left\langle u_i, \sum_{j=1}^n a_j u_j \right\rangle_X u_i = \frac{1}{n}\sum_{i=1}^n \langle u_i, \varphi\rangle_X u_i = R\varphi$$

so that φ is an eigenvector of R for the eigenvalue λ. Regarding the norm, we obtain

$$\begin{aligned}\|\varphi\|_X^2 &= \left\langle \sum_{i=1}^n a_i u_i, \sum_{j=1}^n a_j u_j \right\rangle_X = \sum_{i,j=1}^n a_i a_j \langle u_i, u_j \rangle_X = a^\mathsf{T} \mathbf{K} a = a^\mathsf{T} n\lambda a \\ &= n\lambda \|a\|_2^2 = n\lambda \frac{1}{(\sqrt{n\lambda})^2} = 1.\end{aligned}$$

(a) $\Rightarrow$ (b): we first show that the representation

$$\varphi = \sum_{i=1}^n a_i u_i, \text{ with } a = (a_i)_{i=1}^n \in \mathcal{N}(\mathbf{K}_u)^\perp$$

is not an assumption but can always be obtained.

Let φ be an eigenvector of R for an eigenvalue $\lambda > 0$ with $\|\varphi\|_X = 1$. Therefore, $\varphi \in \mathcal{R}(R)$ and we can find $\bar{a} \in \mathbb{R}^n$ so that $\varphi = \sum_{i=1}^n \bar{a}_i u_i$. Since $\mathcal{N}(\mathbf{K}_u)$ is a closed subspace of $\mathbb{R}^n$, we have the orthogonal projection $P : \mathbb{R}^n \to \mathcal{N}(\mathbf{K}_u)$ and define the projection error $a := \bar{a} - P\bar{a} \in \mathbb{R}^n$ so that $P\bar{a} \in \mathcal{N}(\mathbf{K}_u)$ and $a \in \mathcal{N}(\mathbf{K}_u)^\perp$. Defining $\varphi' := \sum_{i=1}^n a_i u_i$, we obtain

$$\begin{aligned}\langle \varphi', u_k \rangle_X &= \left\langle \sum_{i=1}^n \bar{a}_i u_i, u_k \right\rangle_X - \left\langle \sum_{i=1}^n (P\bar{a})_i u_i, u_k \right\rangle_X \\ &= \langle \varphi, u_k \rangle_X - \underbrace{\sum_{i=1}^n (P\bar{a})_i \langle u_i, u_k \rangle_X}_{=(\mathbf{K}_u \cdot P\bar{a})_k = 0} = \langle \varphi, u_k \rangle_X, \quad k = 1, \ldots, n,\end{aligned}$$

so that $\varphi = \varphi' = \sum_{i=1}^n a_i u_i$ with $a \in \mathcal{N}(\mathbf{K}_u)^\perp$ and we always obtain the desired representation.

Since φ is an eigenvector of R for an eigenvalue $\lambda > 0$, we have

$$R\varphi = \frac{1}{n} \sum_{i=1}^n \langle u_i, \varphi \rangle_X u_i = \frac{1}{n} \sum_{i=1}^n \left\langle u_i, \sum_{j=1}^n a_j u_j \right\rangle_X u_i = \lambda\varphi = \lambda \sum_{j=1}^n a_j u_j.$$

Testing this align with u_k yields

$$\frac{1}{n} \underbrace{\sum_{i,j=1}^n \langle u_i, u_j \rangle_X a_j \langle u_i, u_k \rangle_X}_{=(\mathbf{K}_u^2 a)_k} = \lambda \underbrace{\sum_{j=1}^n a_j \langle u_j, u_k \rangle_X}_{=(\mathbf{K}_u a)_k}$$

so that $\frac{1}{n}\mathbf{K}_u^2 a = \lambda \mathbf{K}_u a$ and thus $\mathbf{K}_u a$ being an eigenvector of $\frac{1}{n}\mathbf{K}_u$ for the eigenvalue $\lambda > 0$. Since $a \in \mathcal{N}(\mathbf{K}_u)^\perp = \mathcal{R}(\mathbf{K}_u)$, we have a representation $a = \sum_{i=1}^{n'} \nu_i a^i$ where the a^i are the eigenvectors of $\frac{1}{n}\mathbf{K}_u$ for the positive eigenvalues $\lambda_i > 0$. We calculate

$$\mathbf{K}_u a = n \sum_{i=1}^{n'} \nu_i \lambda_i a^i \quad \Rightarrow \quad \frac{1}{n}\mathbf{K}_u^2 a = n \sum_{i=1}^{n'} \nu_i \lambda_i \frac{1}{n}\mathbf{K}_u a^i = n \sum_{i=1}^{n'} \nu_i \lambda_i^2 a^i.$$

But since $\mathbf{K}_u a$ is an eigenvector of $\frac{1}{n}\mathbf{K}_u$ for the eigenvalue $\lambda > 0$ it has to be

$$\frac{1}{n}\mathbf{K}_u^2 a = \lambda \mathbf{K}_u a \Rightarrow n \sum_{i=1}^{n'} \nu_i \lambda_i^2 a^i \stackrel{!}{=} n \sum_{i=1}^{n'} \nu_i \lambda \lambda_i a^i$$

so that $\nu_i = 0$ for all $i \in I := \{i \in \{1, \dots, n'\} \mid \lambda_i \neq \lambda\}$. Therefore,

$$a = \sum_{i \in I} \nu_i a^i \quad \Rightarrow \quad \frac{1}{n}\mathbf{K}_u a = \sum_{i \in I} \nu_i \lambda_i a^i = \lambda a$$

so that a is an eigenvector of $\frac{1}{n}\mathbf{K}_u$ for an eigenvalue $\lambda > 0$. For the norm, we obtain

$$1 = \|\varphi\|_X^2 = a^\mathsf{T} \mathbf{K}_u a = n\lambda a^\mathsf{T} a = n\lambda \|a\|_2^2 \quad \Leftrightarrow \quad \|a\|_2 = \frac{1}{\sqrt{n\lambda}},$$

which concludes the proof. □

Sometimes, the POD-basis can also be generated via a singular value decomposition of the snapshot-matrix.

Proposition 1.66 (Computation of $X_{POD,n'}$ via SVD for $X = \mathbb{R}^H$) *Let $X = \mathbb{R}^H$, $U = [u_1, \dots, u_n] \in \mathbb{R}^{H \times n}$ be the so-called* snapshot-matrix *with* $\operatorname{rank}(U) = n'$. *Let $U = \Phi S V^\mathsf{T}$ be a truncated singular value decomposition, i.e.,*

- $\Phi \in \mathbb{R}^{H \times n'}$ *with orthonormal columns,*
- $S \in \mathbb{R}^{n' \times n'}$ *diagonal,*
- $V \in \mathbb{R}^{H \times n'}$ *with orthonormal columns.*

Further let $S = \operatorname{diag}(\sigma_1, \dots, \sigma_{n'})$ with $\sigma_1 > \sigma_2 > \cdots > \sigma_{n'} > 0$. Then, $\Phi = \Phi_{POD,n'}$.

Proof: Let $\Phi = (\tilde{\varphi}_1, \dots, \tilde{\varphi}_{n'})$. Since $X = \mathbb{R}^H$, we obtain from the definition of R

$$Ru = \frac{1}{n}\sum_{i=1}^{n}(u_i^\intercal u)u_i = \frac{1}{n}UU^\intercal u, \quad \forall u \in X.$$

Therefore,

$$R\tilde{\varphi}_i = \frac{1}{n}UU^\intercal\tilde{\varphi}_i = \frac{1}{n}\Phi S\underbrace{V^\intercal V}_{=I_{n'}}\underbrace{S^\intercal}_{=S}\underbrace{\Phi^\intercal\tilde{\varphi}_i}_{e_i} = \frac{1}{n}\Phi e_i\sigma_i^2 = \frac{1}{n}\sigma_i^2\tilde{\varphi}_i$$

such that $\tilde{\varphi}_i$ is an eigenvector of R for the eigenvalue $\frac{1}{n}\sigma_i^2$. Since these eigenvalues are strictly monotone decreasing, the sorting coincides with the spectral decomposition of the POD such that $\frac{1}{n}\sigma_i^2 = \lambda_i$ and $\tilde{\varphi}_i = \varphi_i$ or $\tilde{\varphi}_i = -\varphi_i$. □

Finally, we provide two non-trivial statements about the approximation quality of the POD-RB-space.

Theorem 1.67 (Approximation Properties of $X_{POD,m}$)
Let $\{u_i\}_{i=1}^n \subset X$. Define for a closed subspace $Y \subset X$ with $\dim(Y) < \infty$ *the* mean square projection error

$$J(Y) := \frac{1}{n}\sum_{i=1}^{n}\|u_i - P_Y u_i\|_X^2$$

where $P_Y \in \mathcal{L}(X, Y)$ denotes the orthogonal projection onto Y. Then, we obtain for $Y = X_{POD,m}$ defined in Theorem 1.64

(a) $J(X_{POD,m}) = \sum_{i=m+1}^{n'}\lambda_i,$

(b) $J(X_{POD,m}) = \inf_{\substack{Y\subset X\\ \dim(Y)=m}} J(Y).$

Proof: (a) Let $\{\psi_1, \dots, \psi_m\}$ be an orthonormal basis of a generic $Y \subset X$

with $\dim(Y) < \infty$. We obtain via Theorem 1.7

$$\begin{aligned} J(Y) &= \frac{1}{n}\sum_{i=1}^{n} \|u_i - P_Y u_i\|_X^2 = \frac{1}{n}\sum_{i=1}^{n} \left\| u_i - \sum_{j=1}^{m} \langle u_i, \psi_j\rangle \psi_j \right\|_X^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \|u_i\|_X^2 - \frac{2}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \langle u_i, \psi_j\rangle_X^2 \\ &+ \frac{1}{n}\sum_{i=1}^{n}\sum_{j,k=1}^{m} \langle u_i, \psi_j\rangle_X \langle u_i, \psi_k\rangle_X \underbrace{\langle \psi_j, \psi_k\rangle_X}_{=\delta_{j,k}} \\ &= \frac{1}{n}\sum_{i=1}^{n} \|u_i\|_X^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \langle u_i, \psi_j\rangle_X^2 . \end{aligned}$$

Now, let $Y = X_{POD,m}$ so that $\{\varphi_1, \dots, \varphi_m\}$ are an onthonormal basis of Y, where φ_i, $i = 1, \dots, n'$ were the eigenvectors of the correlation operator R. Since $u_i \in \mathcal{R}(R)$ and since $\{\varphi_1, \dots, \varphi_{n'}\}$ are an orthonormal basis of $\mathcal{R}(R)$ we also obtain

$$u_i = \sum_{j=1}^{n'} \langle u_i, \varphi_j\rangle_X \varphi_j \quad \text{and} \quad \|u_i\|_X^2 = \sum_{j=1}^{n'} \langle u_i, \varphi_j\rangle_X^2 .$$

In total, we obtain for the mean projection error

$$\begin{aligned} J(X_{POD,m}) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n'} \langle u_i, \varphi_j\rangle_X^2 - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} \langle u_i, \varphi_j\rangle_X^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=m+1}^{n'} \langle u_i, \varphi_j\rangle_X^2 = \sum_{j=m+1}^{n'} \left\langle \underbrace{\frac{1}{n}\sum_{i=1}^{n} \langle u_i, \varphi_j\rangle_X u_i}_{=R\varphi_j = \lambda_j \varphi_j}, \varphi_j \right\rangle \\ &= \sum_{j=m+1}^{n'} \lambda_j \langle \varphi_j, \varphi_j\rangle_X = \sum_{j=m+1}^{n'} \lambda_j . \end{aligned}$$

(b) Exercise 5.2. □

Remark 1.68 (Summary of POD)

(a) The POD yields an orthonormal basis, which guarantees numerical stability of the resulting RB-Scheme.

(b) *There exist approximation statements via the mean squared error and even optimality can be proven. The POD bases approximate* all *snapshots* $\{u_i\}_{i=1}^n$ *and allow for error control via the truncated eigenvalues.*

(c) *The bases are furthermore hierarchical, i.e.,* $\Phi_{POD,m} \subset \Phi_{POD,m'}$ *for* $m' \leq m$ *and do not depend on the ordering of the snapshots.*

(d) *The POD can furthermore be used for the extension of an existing ONB, denoted* Φ *here, by first orthogonalizing* $\{u_i\}_{i=1}^n$ *w.r.t.* Φ *via*

$$\tilde{u}_i := u_i - P_{\mathrm{span}(\Phi)} u_i$$

and then computing a POD basis of $\{\tilde{u}_i\}_{i=1}^n$.

(e) *Finally, the POD can be interpreted as an incremental procedure of 1-dimensional optimization problems of the mean squared projection error: for given snapshots* $\{u_i\}_{i=1}^n \subset X$ *define*

$$\bar{\varphi}_1 := POD_1(\{u_i\}_{i=1}^n) \in \arg \min_{\substack{\varphi \in X \\ \|\varphi\|_X = 1}} \frac{1}{n} \sum_{i=1}^n \|u_i - \langle u_i, \varphi \rangle_X \varphi\|_X^2$$

as well as $\bar{X}_1 := \mathrm{span}(\bar{\varphi}_1)$. *For* $i = 2, \ldots, n'$ *define*

$$\bar{\varphi}_i := POD_1\left(\{u_i - P_{\bar{X}_{i-1}} u_i\}_{i=1}^n\right), \quad \text{and} \quad \bar{X}_i := \mathrm{span}(\{\bar{\varphi}_1, \ldots, \bar{\varphi}_i\}).$$

Then $\{\bar{\varphi}_1, \ldots, \bar{\varphi}_{n'}\}$ *is a POD-basis for* $\{u_i\}_{i=1}^n$. *In this sense, it can be seen as an approximate solution of the optimization problem*

$$\inf_{\substack{Y \subset X \\ \dim(Y) = N}} \int_{\mathcal{P}} \|u(\mu) - P_Y u(\mu)\|_X^2 \, \mathrm{d}\mu$$

from the beginning of this section. Instead of the mean over $\mu \in \mathcal{P}$, *we average over the set of snapshots and instead of the infimum over* $Y \subset X$ *with* $\dim(Y) = N$ *we have an iterative sequence of spaces* $Y = X_{POD,m}$.

1.4.2 Greedy Search

The central idea of the Greedy Search is to incrementally construct both the sample set of parameters S_N and the reduced basis Φ_N by repeatedly selecting the "currently worst approximated" parameter and then computing the corresponding snapshot.

We formulate the abstract algorithm utilizing

- a general error indicator $\Delta(Y, \mu) \in \mathbb{R}^+$, that predicts the expected approximation error for the parameter $\mu \in \mathcal{P}$ when using the subspace $X_N = Y$,
- $S_{train} \subset \mathcal{P}$, a given training set of parameters,
- $\varepsilon_{tol} > 0$, a prescribed error tolerance.

Algorithm 1 Greedy-Algorithm(S_{train}, ε_{tol}, $\Delta(Y, \mu)$)

$m = 0$, $S_m := \emptyset$, $\Phi_{GRE,m} := \emptyset$, $X_{GRE,m} := \{0\}$
while $\varepsilon_m := \max_{\mu \in S_{train}} \Delta(X_{GRE,m}, \mu) > \varepsilon_{tol}$ **do**
  $\mu_{m+1} := \arg\max_{\mu \in S_{train}} \Delta(X_{GRE,m}, \mu)$
  $S_{m+1} = S_m \cup \{\mu_{m+1}\}$
  $\varphi_{m+1} := u(\mu_{m+1})$, i.e., the solution of $(P(\mu_{m+1}))$
  $\Phi_{GRE,m+1} := \Phi_{GRE,m} \cup \{\varphi_{m+1}\}$
  $X_{GRE,m+1} := X_{GRE,m} + \operatorname{span}(\varphi_{m+1})$
  $m = m + 1$
end while
$N := m$
return $\Phi_{GRE,N}$, $X_{GRE,N}$

Remark 1.69 (Greedy-Procedure)

(a) $\Phi_{GRE,m}$ *is a Lagrange reduced basis for the sample set* S_m *and it is hierarchical, i.e.,* $\Phi_{GRE,m} \subset \Phi_{GRE,m'}$ *for* $m \leq m'$.

(b) In general, the basis is not orthonormal but it can be orthonormalized via for example the Gram-Schmidt procedure after each step of Algorithm 1 to ensure the numerical stability of the resulting RB-scheme.

(c) The search for μ_1*, i.e., the first iteration of the Greedy-Procedure, is often skipped and one simply starts with a random element of* S_{train}.

(d) The Greedy-Algorithm is an accumulative basis generation procedure, which in each iteration selects $\mu_{m+1} \in S_{train}$*, the parameter that is currently worst approximated, computes* $\varphi_{m+1} = u(\mu_{m+1})$*, the corresponding snapshot, and chooses this snapshot as the new basis vector.*

It can thus be seen as an approximate solution of the optimization problem

$$\inf_{\substack{Y \subset X \\ \dim(Y) = N}} \sup_{\mu \in \mathcal{P}} \|u(\mu) - u_N(\mu)\|_X$$

from the beginning of this section. Instead of the supremum over $\mu \in \mathcal{P}$, we maximize over $\mu \in S_{train}$ and instead of the infimum over $Y \subset X$ with $\dim(Y) = N$ we have an iterative sequence of spaces $Y = X_{GRE,m}$.

(e) *For $X_{GRE,N}$ to be a reasonable space, S_{train} has to somehow represent $\mathcal{P}$. Often, S_{train} is chosen as a random or structured subset of $\mathcal{P}$ with finitely many parameters, but there are also very sophisticated approaches in the literature ("randomized greedy", "partitioning approaches" and "adaptive training set extensions").*

We list various choices of the error indicator $\Delta(Y, \mu)$ for which the Greedy-Algorithm terminates after a finite number of steps and further comment on these choices.

Lemma and Remark 1.70 (Error Indicators and Termination)
If the error indicator $\Delta(Y, \mu)$ satisfies for all $\mu \in \mathcal{P}$ and all subspaces $Y \subset X$

$$u(\mu) \in Y \Rightarrow \Delta(Y, \mu) = 0,$$

Algorithm 1 terminates after $N \leq |S_{train}|$ steps with

$$\max_{\mu \in S_{train}} \Delta(X_{GRE,N}, \mu) \leq \varepsilon_{tol}.$$

This is satisfied by the following error indicators.

(a) ***Projection error:*** $\Delta(Y, \mu) := \inf_{v \in Y} \|u(\mu) - v\|_X = \|u(\mu) - P_Y u(\mu)\|_X$ *with $P_Y : X \to Y$ the orthogonal projection.*
Motivation: *Lemma 1.32 means small projection error $\Rightarrow$ small RB-error.*
Disadvantage: *expensive, as all snapshots $u(\mu)$, $\mu \in S_{train}$ must be available and high-dimensional operations for the projection are required. Thus, S_{train} has to be of moderate size.*
Advantage: *generation of X_N decoupled from RB-model. Greedy can be applied without RB-model/error estimator available.*

(b) ***RB-error:*** $\Delta(Y, \mu) = \|u(\mu) - u_N(\mu)\|_X$.
Motivation: *the ultimate quantity to be controlled.*
Disadvantage: *expensive, as all snapshots $u(\mu)$, $\mu \in S_{train}$ must be available such that S_{train} has to be of moderate size.*
Advantage: *is the error measure one wants to be small.*

(c) ***Error estimator:*** $\Delta(Y,\mu) = \Delta_N(\mu)$ *from Proposition 1.34 (or for symmetric a* $\Delta(Y,\mu) = \Delta_N^{en}(\mu)$ *from Proposition 1.45).*
Motivation: *for rigorous error estimators, RB-error will also be small.*
Disadvantage: *if "overestimation" of the true error is too large, resulting space can be unnecessarily large.*
Advantage: *cheap to evaluate (Offline/Online decomposition), hence* S_{train} *can be much larger and thus represent* $\mathcal{P}$ *much better; Algorithm 1 only requires* N *snapshots computations and is thus fast.*

Proof: We verify that all proposed indicators fulfill

$$u(\mu) \in Y \Rightarrow \Delta(Y,\mu) = 0,$$

for all $\mu \in \mathcal{P}$ and all subspaces $Y \subset X$.

(a) For $\Delta(Y,\mu) = \|u(\mu) - P_Y u(\mu)\|_X$ this is fulfilled as $P_Y u(\mu) = u(\mu)$ if $u(\mu) \in Y$.

(b) For $\Delta(Y,\mu) = \|u(\mu) - u_N(\mu)\|_X$, Proposition 1.30 (reproduction of solutions) yields $u_N(\mu) = u(\mu)$ such that $\Delta(Y,\mu) = 0$.

(c) For $\Delta(Y,\mu)$ being a residual-based error estimator, Proposition 1.30 yields $e(\mu) = 0$ and Proposition 1.33 then yields $v_r(\mu) = 0$ such that the corresponding error estimator is also zero.

The rest of the statement follows since with

$$u(\mu) \in Y \Rightarrow \Delta(Y,\mu) = 0,$$

no element in S_{train} can be selected twice during Algorithm 1. □

Besides the error indicators described so far, one could also use a *goal-oriented* error indicator, e.g., having the output as goal $\Delta(Y,\mu)$ could either be the output error or the output error estimator. This would result in a possibly very small RB-space that approximates the output very well, but u is potentially not well approximated. This is in contrast to the various error indicator choices discussed above, where the approximation for both u and thus the output is good but the RB space is potentially larger.

We list a simple quality statement of the RB-spaces constructed via Algorithm 1 and comment on the issue of *overfitting*.

Lemma and Remark 1.71 (Quality Statement, Overfitting)
With $\Delta(X_{GRE,N'}, \mu)$ being either $\|u(\mu) - u_{N'}(\mu)\|_X$ or $\Delta_{N'}(\mu)$ for $1 \leq N' \leq N$, we obtain

$$\max_{\mu \in S_{train}} \|u(\mu) - u_{N'}(\mu)\|_X \leq \varepsilon_{N'}.$$

Furthermore, $\varepsilon_N = \max_{\mu \in S_{train}} \Delta(X_{GRE,N}, \mu)$ is called the training error *of the Greedy-Algorithm. If S_{train} is too small/does not represent $\mathcal{P}$ very well, so-called "overfitting" can occur, i.e.,*

$$\sup_{\mu \in \mathcal{P}} \|u(\mu) - u_N(\mu)\|_X \gg \varepsilon_N.$$

Therefore, a small training error is not sufficient and one should also aim at a small test error

$$\varepsilon_{test} := \max_{\mu \in S_{test}} \Delta(X_{GRE,N}, \mu),$$

where $S_{test} \subset \mathcal{P}$ is a test set independent of S_{train}.

Proof: As

$$\|u(\mu) - u_{N'}(\mu)\|_X \leq \Delta(X_{GRE,N'}, \mu)$$

for both choices $\Delta(X_{GRE,N'}, \mu) = \Delta_{N'}(\mu)$ or $\Delta(X_{GRE,N'}, \mu) = \|u(\mu) - u_{N'}(\mu)\|_X$, it follows from Algorithm 1 that

$$\max_{\mu \in S_{train}} \|u(\mu) - u_{N'}(\mu)\|_X \leq \max_{\mu \in S_{train}} \Delta(X_{GRE,N'}, \mu) = \varepsilon_{N'},$$

which concludes the proof. □

In some special cases one can prove monotonic decrease of the training error.

Lemma and Remark 1.72 (Monotonicity of Training Error)
For a general $\Delta(Y, \mu)$ the training error $\varepsilon_n = \max_{\mu \in S_{train}} \Delta(X_{GRE,n}, \mu)$ does not have to decrease monotonically, such that $\varepsilon_{n+1} \geq \varepsilon_n$ is possible. Nevertheless, if

(a) $\Delta(Y, \mu) = \|u(\mu) - P_Y u(\mu)\|_X$ or

(b) problem ($P(\mu)$) is compliant ($a(\cdot, \cdot; \mu)$ symmetric and $f = l$) and $\Delta(Y, \mu) = \|u(\mu) - u_N(\mu)\|_\mu$,

the sequence $(\varepsilon_n)_{n \geq 1}$ generated by Algorithm 1 decays monotonically.

Proof: As Algorithm 1 produces a hierarchical sequence of RB-spaces and both error indicators posses the best-approximation property

$$\Delta(Y,\mu) = \inf_{v\in Y} \|u(\mu) - v\|_\star, \quad \mu \in \mathcal{P}$$

in their respective setting, the statement follows as in the proof of Corollary 1.42. □

We can see that the Greedy-Algorithm is a perfect application of our error estimators as error indicators $\Delta(Y,\mu)$. They can be rapidly evaluated for all $\mu \in S_{train}$ (Offline/Online decomposition) without having to compute $u(\mu)$ for all $\mu \in S_{train}$ such that S_{train} can be chosen very large. Therefore, S_{train} can represent $\mathcal{P}$ much better such that the RB-approximation for $\mu \in \mathcal{P} \setminus S_{train}$ should be of high quality.

Furthermore, we can compare the Greedy-Algorithm using the projection error as error indicator with the POD.

- While both methods require the set of snapshots $\{u(\mu) \mid \mu \in S_{train}\}$ to be available and are thus expensive, they are guided by different error measures. The POD wants to minimize the *mean squared projection error* while the Greedy-Algorithm wants to minimize the *maximum projection error*. As a result, single "outliers" with a large projection error are allowed in the POD while they are prevented by the Greedy.
- Furthermore, while the Greedy-Algorithm produces a Lagrangian RB-space, the POD produces a space that is a subset of a span of snapshots but not a Lagrangian RB-space.

As mentioned in Remark 1.69, we can orthonormalize the Greedy-basis after each iteration of Algorithm 1. This can conveniently be done via the respective Gramian matrix.

Proposition 1.73 (Orthonormalization of Reduced Basis)
Let $\Phi_N = \{\varphi_1, \dots, \varphi_N\}$ be a reduced basis with Gramian matrix

$$\mathbf{K}_N := \left(\langle \varphi_i, \varphi_j \rangle_X\right)_{i,j=1}^N$$

which has the Cholesky-factorization $\mathbf{K}_N = \mathbf{L}\mathbf{L}^\mathsf{T}$. Letting $c_{ij} := \left(\mathbf{L}^{-\mathsf{T}}\right)_{ij}$, we obtain the Gram-Schmidt orthonormalized basis $\tilde{\Phi}_N := \{\tilde{\varphi}_1, \dots, \tilde{\varphi}_N\}$ via $\tilde{\varphi}_j := \sum_{i=1}^j c_{ij}\varphi_i$.

Proof: Exercise 5.3. □

We close this section with a discussion on the theoretical convergence of the Greedy-Algorithm, i.e., when $\varepsilon_m \to 0$ for $m \to \infty$ can be expected. Therefore, we introduce the *Kolmogorov n-width*, which is defined as the maximum projection error of the n-dimensional linear subspace that is best-approximating the solution manifold $\mathcal{M} = \{u(\mu) \mid \mu \in \mathcal{P}\}$

$$d_n(\mathcal{M}) := \inf_{\substack{Y \subset X \\ \dim(Y)=n}} \sup_{\mu \in \mathcal{P}} \|u(\mu) - P_Y u(\mu)\|_X .$$

Choosing $\Delta(Y, \mu)$ so that $\Delta(Y, \mu) \geq \|u(\mu) - P_Y u(\mu)\|_X$ for all $\mu \in \mathcal{P}$, we get

$$\sup_{\mu \in \mathcal{P}} \Delta(X_{GRE,n}, \mu) \geq d_n(\mathcal{M}).$$

Therefore, the decay of the Kolmogorov-n-width $d_n(\mathcal{M})$ is a necessary condition for the convergence of the Greedy-Algorithm: if $\sup_{\mu \in \mathcal{P}} \Delta(X_{GRE,n}, \mu)$ becomes small, $d_n(\mathcal{M})$ has to become small as well, and if $d_n(\mathcal{M})$ does not decay, the left-hand-side cannot decay either and $X_{GRE,m}$ cannot approximate $\mathcal{M}$ well.

Now, the converse statement is of interest: does a decay of $d_n(\mathcal{M})$ also imply a decay of the greedy error (estimator)? A positive answer has been given in recent literature: we list the results adapted to our notation and refer to the article for the proofs.

Theorem 1.74 (Greedy Convergence Rates)
Let $S_{train} = \mathcal{P}$ be compact and let $\Delta(Y, \mu)$ be chosen such that for every $\mu_{n+1} = \arg\sup_{\mu \in S_{train}} \Delta(X_{GRE,n}, \mu)$, $n > 0$, there exists a $\xi \in (0, 1]$ such that

$$\left\|u(\mu_{n+1}) - P_{X_{GRE,n}} u(\mu_{n+1})\right\|_X \geq \xi \sup_{u \in \mathcal{M}} \left\|u - P_{X_{GRE,n}} u\right\|_X . \tag{1.7}$$

We then obtain

(a) algebraic convergence: *if $d_n(\mathcal{M}) \leq M n^{-\alpha}$ for some $\alpha, M > 0$ and all $n > 0$ and $d_0(\mathcal{M}) \leq M$, then*

$$\varepsilon_n \leq C M n^{-\alpha}, \quad \text{for all } n > 0,$$

where the constant C can be explicitly computed.

(b) exponential convergence: *if $d_n(\mathcal{M}) \leq M e^{-a n^\alpha}$ for some $a, \alpha, M > 0$ and all $n \geq 0$, then*

$$\varepsilon_n \leq C M e^{-c n^\beta}, \quad \text{for all } n \geq 0,$$

with $\beta := \frac{\alpha}{\alpha+1}$ and constants c, C that can be explicitly computed.

Proof: [BCDPWD, Theorem 3.1 & Theorem 3.2] □

The Greedy-Procedure is called

- *strong* ("strong greedy") if $\xi = 1$, which is e.g. obtained by the projection error $\Delta(Y, \mu) = \|u(\mu) - P_Y u(\mu)\|_X$.
- *weak* ("weak greedy") if $\xi < 1$, since instead of the currently worst approximated element, a sufficiently bad approximated element is chosen for the basis extension.

Thus, Theorem 1.74 is also called *quasi-optimality* of the Greedy-Procedure in the literature. The question remains, if condition (1.7) is fulfilled by our error estimator. The answer is yes and we can show that the error estimator $\Delta_N(\mu)$ yields a weak Greedy-Procedure.

Proposition 1.75 (Weak Greedy via $\Delta_N(\mu)$)
Let $S_{train} = \mathcal{P}$ be compact. The Greedy-Procedure with $\Delta(Y, \mu) = \Delta_N(\mu)$ is a weak greedy scheme with weakness parameter

$$\xi = \frac{\bar{\alpha}^2}{\bar{\gamma}^2} \in (0, 1],$$

where $\bar{\alpha}$ and $\bar{\gamma}$ were the uniform lower/upper bounds for the coercivity/continuity constant.

Proof: We obtain for all $n \geq 1$

$$\begin{aligned}
\left\|u(\mu_{n+1}) - P_{X_{GRE,n}} u(\mu_{n+1})\right\|_X &\overset{1.7}{=} \inf_{v \in X_{GRE,n}} \|u(\mu_{n+1}) - v\|_X \\
&\overset{1.32}{\geq} \frac{\alpha(\mu_{n+1})}{\gamma(\mu_{n+1})} \|u(\mu_{n+1}) - u_N(\mu_{n+1})\|_X \\
&\overset{1.37}{\geq} \frac{\alpha(\mu_{n+1})}{\gamma(\mu_{n+1})\eta_N(\mu_{n+1})} \Delta_N(\mu_{n+1}) \\
&= \frac{\alpha(\mu_{n+1})}{\gamma(\mu_{n+1})\eta_N(\mu_{n+1})} \sup_{\mu \in \mathcal{P}} \Delta_N(\mu) \\
&\overset{1.34}{\geq} \frac{\bar{\alpha}^2}{\bar{\gamma}^2} \sup_{\mu \in \mathcal{P}} \|u(\mu_{n+1}) - u_N(\mu_{n+1})\|_X \\
&\geq \xi \sup_{\mu \in \mathcal{P}} \left\|u(\mu_{n+1}) - P_{X_{GRE,n}} u(\mu_{n+1})\right\|_X
\end{aligned}$$

and $\xi \in (0, 1]$ follows from the definitions of $\bar{\alpha}$, $\bar{\gamma}$ as well as the coercivity and continuity. □

Chapter 2

Balanced Truncation for Linear Time Invariant Control Systems

2.1 Introduction

We begin with the definition of a *linear time invariant (LTI) control systems.*

Definition 2.1 (LTI Control System)
Given time-invariant system matrices $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$, *a time horizon* $[t_0, t_f]$, *an initial state* $x_0 \in \mathbb{R}^n$ *and the* control $u : [t_0, t_f] \to \mathbb{R}^m$, *we seek the* state $x : [t_0, t_f] \to \mathbb{R}^n$ *and the observable* output $y : [t_0, t_f] \to \mathbb{R}^p$ *satisfying*

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

We denote with $\Sigma_{n,m,p}$ *the set of all LTI systems with state space dimension* n, m *inputs and* p *outputs and shortly write* $[A, B, C, D] \in \Sigma_{n,m,p}$. *Furthermore, the set of admissible controls is denoted* $\mathcal{U}_{ad}$.

We illustrate the concept with an example.

Example 2.2 (Controlled Parabolic Heat align)
Similar to Example 1.18, let $\Omega := (0, 1)^2$ *be decomposed into* m *congruent rectangles* Ω_i, $i = 1, \ldots, m$ *and let the time interval* $[0, t_f]$, *with some*

final time $t_f > 0$, be given. We investigate the parabolic PDE: for a time-dependent control vector $u(t) := (u_1(t), \dots, u_m(t))^\mathsf{T} \in \mathbb{R}^m$, find the 'temperature' $T \in L^2(0, t_f, H_0^1(\Omega))$ solving

$$\dot{T}(t,x) - \Delta T(t,x) = \sum_{i=1}^{m} u_i(t)\chi_i(x), \quad (t,x) \in (0,t_f) \times \Omega,$$
$$T(0,x) = 0, \quad x \in \overline{\Omega},$$

where $\dot{T}(t,x) \equiv \frac{\partial}{\partial t} T(t,x)$. One can show that an equivalent variational form of this problem is given by: find $T \in L^2(0, t_f, H_0^1(\Omega))$ solving

$$\int_\Omega \dot{T}(t,x)v(x)\,\mathrm{d}x + \int_\Omega \nabla T(t,x) \cdot \nabla v(x)\,\mathrm{d}x = \int_\Omega \sum_{i=1}^{m} u_i(t)\chi_i(x)v(x)\,\mathrm{d}x$$

for all $v \in H_0^1(\Omega)$ and $t \in (0, t_f)$ almost everywhere. Discretizing the problem in space with piecewise linear finite elements yields the Ansatz-space $X_H := \operatorname{span}(\{\varphi_i, \dots, \varphi_n\})$. We then seek $T_H : [0, t_f] \to X_H$ satisfying

$$\int_\Omega \dot{T}_H(t,x)v(x)\,\mathrm{d}x + \int_\Omega \nabla T_H(t,x) \cdot \nabla v(x)\,\mathrm{d}x \tag{$\diamond$}$$
$$= \int_\Omega \sum_{i=1}^{m} u_i(t)\chi_i(x)v(x)\,\mathrm{d}x, \quad \forall v \in X_H, \quad T_H(0) = 0.$$

As $T_H \in X_H$ for all $t \in [0, t_f]$, there exist coefficient functions $x(t) : [0, t_f] \to \mathbb{R}$ such that $T_H(t,x) = \sum_{i=1}^n x_i(t)\varphi_i(x)$ and introducing the notation $x(t) := (x_i(t))_{i=1}^n$, the mass- *and* stiffness matrices

$$M := \left(\int_\Omega \varphi_i\varphi_j\,\mathrm{d}x\right)_{i,j=1}^n \in \mathbb{R}^{n\times n} \text{ and } K := \left(\int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j\,\mathrm{d}x\right)_{i,j=1}^n \in \mathbb{R}^{n\times n},$$

as well as the matrix $B \in \mathbb{R}^{n\times m}$ which contains in the i-th column the evaluation of $\chi_i(x)$ at the n grid points, solving ($\diamond$) is equal to solving the system of ordinary differential aligns

$$M\dot{x}(t) + Kx(t) = MBu(t) \quad \Leftrightarrow \quad \dot{x}(t) = -M^{-1}Kx(t) + Bu(t).$$

Introducing as output the average temperature over Ω

$$y(t) := \int_\Omega T_H(t,x)\,\mathrm{d}x \quad \Leftrightarrow \quad y(t) = Cx(t)$$

with the row vector $C := \left(\int_\Omega \varphi_i(x)\,\mathrm{d}x\right)_{i=1}^n \in \mathbb{R}^n$, we obtain a system of the form (LTI) by setting $A := -M^{-1}K$ as well as $D = 0$. The system has state space dimension n, m inputs and 1 output.

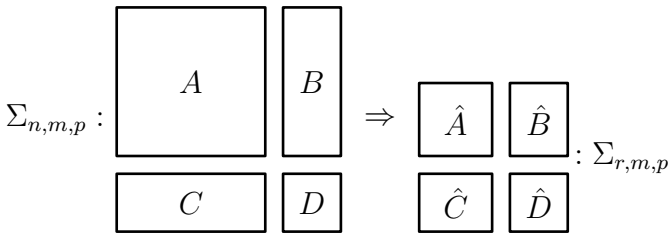


Figure 2.1: Schematic model order reduction approach for LTI systems.

Aside from the previous example, time-dependent linear partial differential aligns occur in various branches such as mechanics, biological systems, and weather prediction. Furthermore, (large) ODE systems naturally arise when modeling electrical circuits. Example 2.2 shows that LTI systems can get very large in practice (Ω could also be a 3-dimensional domain) such that model order reduction techniques are required.

The general idea for model order reduction in LTI systems is to find a coordinate transformation for the state such that *insignificant* parts of the transformed state can be truncated. In formulas, we want to approximate $[A, B, C, D] \in \Sigma_{n,m,p}$ by some $[\hat{A}, \hat{B}, \hat{C}, \hat{D}] \in \Sigma_{r,m,p}$ with $r \ll n$ and this is visually sketched in Figure 2.1.

Especially, the dimension of the control and the output remain unchanged. Ideally, the reduced system $[\hat{A}, \hat{B}, \hat{C}, \hat{D}] \in \Sigma_{r,m,p}$ should

- have a small approximation error (with possibly a global error bound),
- preserve stability of the original system (what concept of stability is present here?).

Finally, the procedure to generate the reduced system should be computationally stable and efficient. We will investigate the following aspects:

- How can a reduced system be obtained? How can we identify insignificant parts of the state and how to trucnate them?
- How can the approximation error be measured/quantified? An intuitive measure would be that outputs $y(t)$ and $\hat{y}(t)$ obtained by the full system $\Sigma_{n,m,p}$ and the reduced system $\Sigma_{r,m,p}$ using the same control $u(t)$ are close (the norm difference being small).

2.2 Theoretical Background

For simplicity, we work with continuous controls in this section, i.e., $\mathcal{U}_{ad} = C([t_0, t_f], \mathbb{R}^m)$.

In order to discuss the solution of (LTI), we introduce the *matrix exponential*.

Definition and Theorem 2.3 (Matrix Exponential)
For any $A \in \mathbb{R}^{n\times n}$ we remember that $A^0 = I_{n\times n}$. Then, for a given $t \in \mathbb{R}$, we define the matrix exponential

$$e^{At} := \lim_{N\to\infty} \sum_{k=0}^{N} \frac{1}{k!} (At)^k = \sum_{k=0}^{\infty} \frac{1}{k!} (At)^k .$$

As the series converges uniformly for all $t \in \mathbb{R}$, the map $t \mapsto e^{At}$ is a well-defined analytic function from $\mathbb{R}$ to $\mathbb{R}^{n\times n}$. It has the properties

- $e^{A0} = I_{n\times n}$ *and* $e^{A+B} = e^A e^B$ *for any* $B \in \mathbb{R}^{n\times n}$ *with* $AB = BA$,
- $e^{A(t+\tau)} = e^{At} e^{A\tau}$, *for any* $\tau \in \mathbb{R}$, *and hence* $e^{-At} = \left(e^{At}\right)^{-1}$,
- $\frac{\partial}{\partial t} e^{At} = Ae^{At} = e^{At} A$.

For $x_0 \in \mathbb{R}^n$, $x(t) = e^{At} x_0$ is the unique solution of

$$\dot{x}(t) = Ax(t), \quad x(0) = x_0.$$

Proof: $e^{A0} = I_{n\times n}$ follows from the definition of $e^A t$ as a power series and that $A^0 = I_{n\times n}$ for $A \in \mathbb{R}^{n\times n}$. Now, given any $B \in \mathbb{R}^{n\times n}$ with $AB = BA$, we obtain with the cauchy product

$$e^A e^B = \left(\sum_{k=0}^{\infty} \frac{1}{k!} A^k\right) \left(\sum_{j=0}^{\infty} \frac{1}{j!} B^j\right) = \sum_{i=0}^{\infty} c_i,$$

with

$$c_i = \sum_{l=0}^{i} \frac{1}{l!} A^l \frac{1}{(i-l)!} B^{i-l} = \frac{1}{i!} \sum_{l=0}^{i} \binom{i}{l} A^l B^{i-l} = \frac{1}{i!} (A+B)^i,$$

where the binomial theorem was used in the last equality which is applicable here since A and B commute. Thus,

$$e^A e^B = \sum_{i=0}^{\infty} \frac{1}{i!} (A+B)^i = e^{A+B}.$$

Using this result and setting $X := At$, $Y := A\tau$ for some $t,\ \tau \in \mathbb{R}$, it is clear that $XY = AtA\tau = A\tau At = YX$ and thus

$$e^{At}e^{A\tau} = e^X e^Y = e^{X+Y} = e^{A(t+\tau)}.$$

Thus, $e^{At}e^{-At} = I_{n\times n}$ such that e^{At} is invertible with $e^{-At} = \left(e^{At}\right)^{-1}$. For the third property the differentiation and sum can be switched due to the uniform convergence of the series and thus

$$\frac{\partial}{\partial t}\sum_{k=0}^{\infty}\frac{1}{k!}(At)^k = \sum_{k=1}^{\infty}\frac{1}{k!}\frac{\partial}{\partial t}(At)^k = \sum_{k=1}^{\infty}\frac{1}{(k-1)!}(At)^{k-1}A = \sum_{k=0}^{\infty}\frac{1}{k!}(At)^k A$$

and it is clear that A can also be dragged out of the series to the left. Regarding the second part, we first verify that $x(t) = e^{At}x_0$ is a solution since

$$x(0) = e^{A0}x_0 = x_0, \quad \dot{x}(t) = \frac{\partial}{\partial t}e^{At}x_0 = Ae^{At}x_0 = Ax(t).$$

For the uniqueness, assume that $y(t)$ is also a solution for $y(0) = x_0$. Using the product rule, we obtain

$$\begin{aligned}\frac{\partial}{\partial t}\left(e^{-At}y(t)\right) &= \left(\frac{\partial}{\partial t}e^{-At}\right)y(t) + e^{-At}\frac{\partial}{\partial t}y(t) = e^{-At}(-A)y(t) + e^{-At}Ay(t)\\ &= e^{-At}(-A+A)y(t) = 0.\end{aligned}$$

Thus, $e^{-At}y(t) = c$ for some constant $c \in \mathbb{R}^n$ and plugging in $t = 0$ yields $c = x_0$ such that

$$e^{-At}y(t) = x_0 \quad \Leftrightarrow \quad y(t) = e^{At}x_0 = x(t)$$

for all $t \in \mathbb{R}$. □

The matrix exponential can be calculated using for example the *Jordan normal form*. This leads to an explicit formula for the matrix exponential.

Definition and Theorem 2.4 (Jordan normal form & e^{At})
For every $A \in \mathbb{C}^{n\times n}$ there exists an invertible $S \in \mathbb{C}^{n\times n}$ such that

$$S^{-1}AS = J = \operatorname{diag}(J_1, \dots J_g), \quad \textit{with} \quad J_l := \begin{pmatrix}\lambda_l & 1 & \cdots & 0\\ \vdots & \ddots & \ddots & \vdots\\ 0 & \cdots & \lambda_l & 1\\ 0 & \cdots & 0 & \lambda_l\end{pmatrix},$$

where J_l are the Jordan blocks. *This* Jordan normal form $\mathrm{diag}(J_1, \dots J_g)$ *is (up to permutation) uniquely determined by A, $\lambda_1, \dots, \lambda_g$, $g \leq n$, are the (not necessarily different) eigenvalues of A, and S collects in its columns the (generalized) eigenvectors corresponding to the Jordan blocks. Regarding the matrix exponential, we obtain*

$$e^{At} = Se^{Jt}S^{-1} = S\mathrm{diag}(e^{J_1 t}, \dots, e^{J_g t})S^{-1}$$

and letting d denote the dimension of the Jordan block J_l, we have

$$e^{J_l t} = \begin{pmatrix} 1 & t & \frac{t^2}{2!} & \cdots & \frac{t^{d-2}}{(d-2)!} & \frac{t^{d-1}}{(d-1)!} \\ 0 & 1 & t & \cdots & \frac{t^{d-3}}{(d-3)!} & \frac{t^{d-2}}{(d-2)!} \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & t \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{pmatrix} e^{\lambda_l t} \in \mathbb{R}^{d\times d} \quad \textit{for} \quad l = 1, \dots, g.$$

Proof: Regarding the existence of the Jordan normal form and its properties, we refer to a basic linear algebra lecture. The remaining statements are proven in Exercise 6.2. □

We introduce the concept of stability for ODEs.

Definition and Theorem 2.5 (Stability of (LTI))
We call the LTI system $[A, B, C, D] \in \Sigma_{n,m,p}$ 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 $x(t_0) = x_0 \in \mathbb{R}^n$.

The LTI system $[A, B, C, D] \in \Sigma_{n,m,p}$ is asymptotically stable *if and only if all eigenvalues $\{\lambda_i\}_{i=1}^n$ of A satisfy $\lambda_i \in \mathbb{C}^- := \{\lambda \in \mathbb{C} : \mathrm{Re}(\lambda) < 0\}$.*

Proof: "⇐": As a consequence of Theorem 2.4, it is clear that e^{At} can be explicitly written as a linear combination of terms of the form $t^k e^{\lambda t}$, where λ is an eigenvalue of A and $k \in \mathbb{N}_0$. Now, letting $\lambda = \sigma + i\omega$ be such an eigenvalue with $\sigma < 0$, we obtain (as w.l.o.g. $t \geq 0$)

$$|t^k e^{\lambda t}| = t^k |e^{\sigma t} e^{i\omega t}| = t^k e^{\sigma t} |e^{i\omega t}| = e^{\sigma t} t^k \underbrace{\sqrt{\cos^2(\omega t) + \sin^2(\omega t)}}_{=1} = e^{\sigma t} t^k$$

and $e^{\sigma t}t^k$ goes to 0 when t goes to infinity since $\sigma < 0$. As e^{At} is a linear combination of such terms, we obtain $\lim_{t\to\infty} e^{At} = 0$ such that

$$\lim_{t\to\infty} x(t) = \lim_{t\to\infty} e^{At}x_0 = 0$$

for all $x_0 \in \mathbb{R}^n$.

"⇒": we prove that if there is an eigenvalue of A with non-negative real part, then the system $[A, B, C, D] \in \Sigma_{n,m,p}$ is not asymptotically stable. Therefore, let w.l.o.g. the first eigenvalue $\lambda_1 = \sigma + i\omega$ fulfill $\sigma \geq 0$. According to Theorem 2.4, $S = (v_1, \ldots, v_n)$ collects the eigenvectors of A in the sorting corresponding to the Jordan normal form and since $S^{-1}S = I_{n\times n}$ it is clear that $S^{-1}v_1 = e_1$, where $v_1 \neq 0$. We thus obtain using Theorem 2.4

$$\begin{aligned} e^{At}v_1 &= S\mathrm{diag}(e^{J_1 t}, \ldots, e^{J_g t})S^{-1}v_1 = S\mathrm{diag}(e^{J_1 t}, \ldots, e^{J_g t})e_1 = e^{\lambda_1 t}v_1 \\ &= e^{(\sigma+i\omega)t}v_1 = e^{\sigma t}(\cos(\omega t) + i\sin(\omega t)) \cdot (\mathrm{Re}(v_1) + i\mathrm{Im}(v_1)). \end{aligned}$$

As $e^{At} \in \mathbb{R}^{n\times n}$, we have

$$e^{At}\mathrm{Re}(v_1) = \mathrm{Re}(e^{At}v_1) = \mathrm{Re}(e^{\lambda_1 t}v_1) = e^{\sigma t}(\mathrm{Re}(v_1)\cos(\omega t) - \mathrm{Im}(v_1)\sin(\omega t)).$$

Thus, choosing $x_0 = \mathrm{Re}(v_1) \in \mathbb{R}^n$ as the initial value of the homogeneous ODE $\dot{x}(t) = Ax(t)$, it is clear that $x(t) = e^{At}\mathrm{Re}(v_1) \not\to 0$ for $t \to \infty$ since $\sigma \geq 0$ and since either $\mathrm{Re}(v_1) \neq 0$ or $\mathrm{Im}(v_1) \neq 0$ (it was $v_1 \neq 0$). Therefore, the system $[A, B, C, D] \in \Sigma_{n,m,p}$ is not asymptotically stable. □

We are now ready to formulate the analytical solution of (LTI).

Proposition 2.6 (Solution of (*LTI*))
For any input $u \in \mathcal{U}_{ad} = C([t_0, t_f], \mathbb{R}^m)$ and initial state $x_0 \in \mathbb{R}^n$ the unique state solution *of (LTI) is given by*

$$x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau,$$

where the vector-valued integral has to be understood component-wise. The corresponding output-response *of (LTI) is then given by*

$$y(t) = Ce^{A(t-t_0)}x_0 + \int_{t_0}^{t} Ce^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau + Du(t).$$

Proof: See Exercise 6.3. □

We note that $u \in C([t_0, t_f], \mathbb{R}^m)$ is necessary for the uniqueness of the solution, but the existence of a solution is already guaranteed for controls $u \in L^2([t_0, t_f], \mathbb{R}^m)$.

In case the control is constant in time, i.e., $u(t) \equiv u_c \in \mathbb{R}^m$ for all $t \in [t_0, t_f]$, and the system is asymptotically stable, the solution formula from Proposition 2.6 further simplifies. For some $x_0 \in \mathbb{R}^n$, we obtain

$$x(t) = e^{A(t-t_0)}x_0 + \left(\int_{t_0}^{t} e^{A(t-\tau)}\,\mathrm{d}\tau\right) Bu_c$$

and since A is invertible, we get

$$\begin{aligned}\int_{t_0}^{t} e^{A(t-\tau)}\,\mathrm{d}\tau &= \int_{t_0}^{t} e^{At} \underbrace{e^{-A\tau}}_{=\frac{\partial}{\partial\tau}e^{-A\tau}(-A^{-1})}\,\mathrm{d}\tau = e^{At} \underbrace{\int_{t_0}^{t} \frac{\partial}{\partial\tau}e^{-A\tau}\,\mathrm{d}\tau}_{=[e^{-A\tau}]_{t_0}^{t}=e^{-At}-e^{-At_0}} \quad (-A^{-1}) \\ &= e^{A(t-t_0)}A^{-1} - A^{-1}\end{aligned}$$

such that in total

$$x(t) = e^{A(t-t_0)}(x_0 + A^{-1}Bu_c) - A^{-1}Bu_c.$$

We continue with two more important system characterizations of (LTI).

Definition 2.7 (Further Properties of (*LTI*))
We call the LTI system $[A, B, C, D] \in \Sigma_{n,m,p}$

(a) controllable, *if for every initial condition* $x(t_0) = x_0 \in \mathbb{R}^n$ *and every state* $\bar{x} \in \mathbb{R}^n$, *there exists a time* $t_1 > t_0$ *and a control function* $u \in \mathcal{U}_{ad}$ *such that* $x(t_1) = \bar{x}$. *We also say that every state* $\bar{x} \in \mathbb{R}^n$ *can be* reached.

(b) observable, *if for two solution trajectories* $x(t)$ *and* $\tilde{x}(t)$ *(obtained with the same control function* $u \in \mathcal{U}_{ad}$*) it holds:*

$$Cx(t) = C\tilde{x}(t) \quad \text{for all } t \geq t_0 \quad \Rightarrow \quad x(t) = \tilde{x}(t) \quad \text{for all } t \geq t_0.$$

Thus, the state solution can fully be recovered from the output-response of the system.

As for the asymptotic stability, controllability and observability have equivalent algebraic characterizations as shown in the following lemma.

Lemma 2.8 (Algebraic Characterizations)
The LTI system $[A, B, C, D] \in \Sigma_{n,m,p}$ *is*

(a) controllable $\Leftrightarrow \operatorname{rank}\left(\begin{bmatrix}\lambda I_{n\times n} - A & B\end{bmatrix}\right) = n \quad \forall \lambda \in \mathbb{C}$
$\Leftrightarrow \operatorname{rank}\left(\begin{bmatrix}B & AB & \dots & A^{n-1}B\end{bmatrix}\right) = n.$

(b) observable $\Leftrightarrow \operatorname{rank}\left(\begin{bmatrix}\lambda I_{n\times n} - A \\ C\end{bmatrix}\right) = n \quad \forall \lambda \in \mathbb{C}$
$\Leftrightarrow \operatorname{rank}\left(\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1}\end{bmatrix}\right) = n.$

Proof: We refer to [ANT, Theorem 4.15 & Theorem 4.26]. □

The matrix

$$\begin{bmatrix}B & AB & \dots & A^{n-1}B\end{bmatrix} \in \mathbb{R}^{n\times nm}$$

is also called the *controllability matrix* of the system, while

$$\begin{bmatrix}C & CA & \dots & CA^{n-1}\end{bmatrix}^{\mathsf{T}} \mathbb{R}^{np\times n}$$

is called the *observability matrix*. One can show (see [ANT, Theorem 4.23]) that both concepts are *dual* in the sense that an LTI system is observable if and only if the dual system

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

is controllable, where $z(t) \in \mathbb{R}^n$ is the dual state and $v(t) \in \mathbb{R}^p$ is the dual control.

With the following considerations, we want to motivate the two *Gramians* that shall be defined next. We first define the *input-to-state map*

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

which represents the effect of an impulsive control input on the solution of the state align, when $t_0 = 0$ and $x(0) = x_0 = 0$. An impulsive control $u(t)$ only takes a value $u_0 \in \mathbb{R}^m$ at $t = 0$ and is zero everywhere else and we write this as $u(t) = u_0\delta(t)$ with δ denoting the Dirac delta distribution. Then,

$$\begin{aligned} x(t) &= e^{At}x_0 + \int_0^t e^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau = \int_0^t e^{A(t-\tau)}Bu_0\delta(\tau)\,\mathrm{d}\tau \\ &= e^{At}Bu_0 = \zeta(t)u_0. \end{aligned}$$

The above calculation is mathematically *not correct*, since δ is not a function mapping from $\mathbb{R}$ to $\mathbb{R}$, but a *distribution* (also called generalized function) that is defined as

$$\delta : C_0^\infty(\mathbb{R}^k, \mathbb{R}) \to \mathbb{R}, \quad f \mapsto f(0).$$

Thus, we should write

$$x(t) := \delta\left(e^{A(t-\cdot)}Bu_0\right) = e^{At}Bu_0,$$

where the application of the delta distribution has to be understood componentwise.

Next, we define the *state-to-output map*

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

which represents the effect of the (initial) state on the output and is motivated by the fact that for $x(0) = x_0 \in \mathbb{R}^n$ and $u(t) \equiv 0$, we obtain

$$y(t) = Ce^{At}x_0 + C\int_0^t e^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau = Ce^{At}x_0 = \eta(t)x_0.$$

For the analysis of LTI control systems we now define the following *Gramians*, which are based on the input-to-state and state-to-output map.

Definition 2.9 (Finite Gramians)
The matrix

$$P(T) = \int_0^T e^{At}BB^\mathsf{T}e^{A^\mathsf{T}t}\,\mathrm{d}t \quad \in \mathbb{R}^{n\times n}$$

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

$$Q(T) = \int_0^T e^{A^\mathsf{T}t}C^\mathsf{T}Ce^{At}\,\mathrm{d}t \quad \in \mathbb{R}^{n\times n}$$

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

Do note that these matrices are indeed Gramian matrices: introducing for functions $u, v \in L^2([0,T],\mathbb{R}^n)$ the inner product

$$\langle u, v\rangle_{L^2([0,T],\mathbb{R}^n)} := \int_0^T u(t)^\mathsf{T}v(t)\,\mathrm{d}t,$$

and given a matrix-valued function $W : \mathbb{R} \to \mathbb{R}^{n\times n} : t \mapsto [w_1(t), \dots, w_n(t)]$, we indeed observe

$$\langle w_i, w_j \rangle_{L^2([0,T],\mathbb{R}^n)} = \int_0^T w_i(t)^\mathsf{T} w_j(t)\,\mathrm{d}t = \left(\int_0^T W(t)^\mathsf{T} W(t)\,\mathrm{d}t \right)_{ij}.$$

Thus, $P(T)$ is the Gramian for the columns of $B^\mathsf{T} e^{A^\mathsf{T} t}$, $Q(T)$ is the Gramian for the columns of Ce^{At} and according to Def/Theorem 1.61 they are both symmetric and positive semidefinite. We conclude this section by considering these Gramians for $T \to \infty$.

Lemma 2.10 (Infinite Gramians)
Let the system (LTI) be asymptotically stable. Then,

(a) the infinite controllability Gramian

$$P := \lim_{T\to\infty} P(T) = \int_0^\infty e^{At} B B^\mathsf{T} e^{A^\mathsf{T} t}\,\mathrm{d}t \quad \in \mathbb{R}^{n\times n}$$

as well as the infinite observability Gramian

$$Q := \lim_{T\to\infty} Q(T) = \int_0^\infty e^{A^\mathsf{T} t} C^\mathsf{T} C e^{At}\,\mathrm{d}t \quad \in \mathbb{R}^{n\times n}$$

exist. Furthermore, the they are the unique solutions of the two Lyapunov aligns

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

(b) If the system (LTI) is also controllable and observable, both Gramians are positive definite, i.e., $x^\mathsf{T} P x > 0$ and $x^\mathsf{T} Q x > 0$ for all $0 \neq x \in \mathbb{R}^n$.

Proof: (a) See Exercise 7.1.

(b) Assume P is not positive definite, then there exists $0 \neq x \in \mathbb{R}^n$ with

$$0 = x^\mathsf{T} P x = \int_0^\infty x^\mathsf{T} e^{At} B B^\mathsf{T} e^{A^\mathsf{T} t} x\,\mathrm{d}t = \int_0^\infty \left\| B^\mathsf{T} e^{A^\mathsf{T} t} x \right\|_2^2 \mathrm{d}t.$$

Since $\left\| B^\mathsf{T} e^{A^\mathsf{T} t} x \right\|_2 \geq 0$ for all $t \geq 0$, we obtain $B^\mathsf{T} e^{A^\mathsf{T} t} x = 0$ for all $t \geq 0$. As e^{At} is an analytic function, this implies that also all derivatives are zero at $t = 0$, i.e., $B^\mathsf{T} (A^\mathsf{T})^i x = 0$ for all $i \geq 0$. But the system is

controllable, such that we get from Lemma 2.8 (note that the rank of a matrix and its transpose are the same)

$$\operatorname{rank}\left(\begin{bmatrix} B & AB & \dots & A^{n-1}B \end{bmatrix}\right) = \operatorname{rank}\left(\underbrace{\begin{bmatrix} B^\mathsf{T} \\ B^\mathsf{T}A^\mathsf{T} \\ \vdots \\ B^\mathsf{T}(A^{n-1})^\mathsf{T} \end{bmatrix}}_{=:M}\right) = n$$

and M has full columnrank. But $Mx = 0$, a contradiction.

The proof for Q follows similar lines: assume it is not positive definite, then there exists $0 \neq x \in \mathbb{R}^n$ with

$$0 = x^\mathsf{T}Qx = \int_0^\infty x^\mathsf{T}e^{A^\mathsf{T}t}C^\mathsf{T}Ce^{At}x \,\mathrm{d}t = \int_0^\infty \left\|Ce^{At}x\right\|_2^2 \,\mathrm{d}t.$$

Therefore, it has to be $Ce^{At}x = 0$ for all $t \geq 0$, such that again all derivatives have to be zero at $t = 0$ and $CA^ix = 0$ for all $i \geq 0$. But as the system is observable, we have from Lemma 2.8

$$\operatorname{rank}\left(\underbrace{\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}}_{=:N}\right) = n,$$

but $Nx = 0$, which is again a contradiction.

This concludes the proof. □

2.3 Balanced Truncation

As described in the Introduction, the aim of model reduction for LTI systems is to reduce the state dimension n, that is to *truncate* state variables. But what variables shall be truncated? The main idea of Balanced Truncation is to truncate states that are difficult to observe and at the same time difficult to reach. We will introduce both concepts in the following section and then present a state space transformation that results in a *balanced system*.

2.3.1 Input and Output Energy

Let us consider the state align $\dot{x}(t) = Ax(t) + Bu(t)$ with initial state $x(-\infty) = 0$ and an input $u \in L^2((-\infty, 0], \mathbb{R}^m)$ that acts on the negative time-horizon leading to $x(0) = x_0$. By switching off the control input at $t = 0$, the output align $y(t) = Cx(t) + Du(t)$ then gives an output signal $y \in L^2([0, \infty), \mathbb{R}^p)$ on the positive time-horizon. Based on this setup, we define via

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

the corresponding *input energy* and via

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

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

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

$$E_u(x_0) := \inf_{\substack{u \in L^2((-\infty,0],\mathbb{R}^m) \\ x(-\infty)=0,\, x(0)=x_0}} \|u\|_{L^2((-\infty,0],\mathbb{R}^m)}, \tag{2.1}$$

as the minimal input energy required to steer the system from the 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* (as the input energy required to reach it is small), otherwise it is *hard to reach*. Do note that $E_u(x_0) = \infty$ is possible, such that the state x_0 is said to be *unreachable* and the system is *uncontrollable*.

Regarding the output, the control is switched off at $t = 0$, meaning that it is always $u|_{[0,\infty)} = 0$ here. Thus, we have $y(t) = Ce^{At}x_0$ and

$$E_y(x_0) := \|y\|_{L^2([0,\infty),\mathbb{R}^p)} = \left\| Ce^{A\cdot}x_0 \right\|_{L^2([0,\infty),\mathbb{R}^p)},$$

as 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 state x_0 is *unobservable*, and therefore the system is *unobservable*.

We now know what it means if a state is easy/hard to reach or easy/hard to observe. But how can we easily compute these quantities? It turns out that the energies $E_u(x_0)$ and $E_y(x_0)$ can be conveniently expressed using the Gramians from the previous section.

Theorem 2.11 (Input & Output Energy)
Let the system $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable and controllable. The minimal input energy and corresponding output energy can then be obtained via

$$E_u(x_0)^2 = x_0^\mathsf{T} P^{-1} x_0 \quad \text{and} \quad E_y(x_0)^2 = x_0^\mathsf{T} Q x_0,$$

where P and Q are the controllability and observability Gramians defined in Lemma 2.10. Moreover, $u_*(t) := B^\mathsf{T} e^{-A^\mathsf{T} t} P^{-1} x_0$ is a control for which the infimum in (2.1) is attained.

Proof: We begin with the statements regarding the controllability Gramian and note that due to the controllability of the system and Lemma 2.10, P is positive definite and thus invertible (and the inverse is also symmetric). Furthermore, by setting $t = -\tau$, we obtain

$$P = \int_0^\infty e^{At} B B^\mathsf{T} e^{A^\mathsf{T} t} \,\mathrm{d}t = \int_{-\infty}^0 e^{-A\tau} B B^\mathsf{T} e^{-A^\mathsf{T} \tau} \,\mathrm{d}\tau.$$

Now, let $x(t)$ be a solution trajectory obtained with initial state $x(-\infty) = 0$, final state $x(0) = x_0$, and control $u \in L^2((-\infty, 0], \mathbb{R}^m)$ having finite energy $E_u < \infty$. Then, due to Proposition 2.6, we have

$$x_0 = x(0) = \int_{-\infty}^0 e^{-A\tau} B u(\tau) \,\mathrm{d}\tau.$$

We show that $E_u \geq E_{u_*}$ for $u_*(t) := B^\mathsf{T} e^{-A^\mathsf{T} t} P^{-1} x_0$. Define $v(t) := u(t) - u_*(t)$, such that

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

Hence, we obtain

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

such that the infimum is attained at u_* and $E_u(x_0) = E_{u_*}$. Moreover, we have

$$E_{u_*}^2 = x_0^\mathsf{T} P^{-1} \underbrace{\int_{-\infty}^{0} e^{-A\tau} B B^\mathsf{T} e^{-A^\mathsf{T}\tau} \,\mathrm{d}\tau}_{=P} P^{-1} x_0 = x_0^\mathsf{T} P^{-1} P P^{-1} x_0$$
$$= x_0^\mathsf{T} P^{-1} x_0.$$

Regarding the corresponding output energy, as the control is switched off at $t = 0$ (such that $u|_{[0,\infty)} = 0$), we have $y(t) = Ce^{At}x_0$ and thus

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

which concludes the proof. □

Since P is real and symmetric, it has an eigendecomposition $P = U\Sigma U^\mathsf{T}$ with orthogonal $U = [u_1, \ldots, u_n]$ and $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$, where $\sigma_1, \ldots, \sigma_n > 0$ (the system is assumed to be controllable). Then, according to Theorem 2.11, the energy needed to reach the state $x_0 = u_i$ from $x(-\infty) = 0$ is given by

$$E_u(u_i)^2 = u_i^\mathsf{T} P^{-1} u_i = \frac{1}{\sigma_i} u_i^\mathsf{T} u_i = \frac{1}{\sigma_i}.$$

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, Q has an eigendecomposition $Q = V\Lambda V^\mathsf{T}$ with orthogonal $V = [v_1, \ldots, v_n]$ and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, where $\lambda_1, \ldots, \lambda_n \geq 0$ and the output energy gained from the state $x_0 = v_i$ is given by

$$E_y(v_i)^2 = x_i^\mathsf{T} Q v_i = \lambda_i v_i^\mathsf{T} v_i = \lambda_i.$$

Thus, 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.

2.3.2 Model Reduction by Balanced Truncation

We begin with an example indicating the need for a so-called *balancing transformation* (which we will define later).

Example 2.12 (Balancing Energies)
Consider the following asymptotically stable, controllable and observable system

$$A = \begin{pmatrix} 1 & 3 \\ -1 & -2 \end{pmatrix}, \quad B = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad C = \begin{pmatrix} 0 & 1 \end{pmatrix}, \quad D = 0.$$

According to Lemma 2.10, we can compute P and Q by solving the corresponding Lyapunov aligns and obtain

$$P = \begin{pmatrix} \frac{5}{2} & -1 \\ -1 & \frac{1}{2} \end{pmatrix}, \quad \text{and} \quad Q = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}.$$

Computing the eigenvalues and eigenvectors, we obtain (rounding to 5 digits)

$$\Sigma_P = \begin{pmatrix} 2.91421 & 0 \\ 0 & 0.08578 \end{pmatrix}, \quad U_P = \begin{pmatrix} 0.92388 & 0.38268 \\ -0.38268 & 0.92388 \end{pmatrix},$$
$$\Lambda_Q = \begin{pmatrix} 1.30901 & 0 \\ 0 & 0.19098 \end{pmatrix}, \quad V_Q = \begin{pmatrix} 0.52573 & -0.85865 \\ 0.85865 & 0.52573 \end{pmatrix}.$$

Thus, the eigenvector $u_1 = \begin{pmatrix} 0.92388 & -0.38268 \end{pmatrix}^\intercal$ is easy to reach and the eigenvector $u_2 = \begin{pmatrix} 0.38268 & 0.92388 \end{pmatrix}^\intercal$ is hard to reach. Calculating the corresponding output energies, we obtain

$$u_1^\intercal Q u_1 = 0.21966 \quad \text{and} \quad u_2^\intercal Q u_2 = 1.28033.$$

This means, that u_1 is at the same time easy to reach and hard to observe and conversely u_2 is hard to reach but easy to observe.

The example shows that if we want to do model order reduction via truncation of states that are difficult to reach and difficult to observe, we have to find a coordinate transformation, in which states that are difficult to reach are *at the same time* difficult to observe (and vice versa). This motivates the following definition.

Definition 2.13 (Balanced System)
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*

$$P = Q = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n),$$

with some $\sigma_1, \ldots, \sigma_n \in \mathbb{R}$.

Now, given a state-space transformation, i.e., an invertible matrix $T \in \mathbb{R}^{n\times n}$, and introducing the transformed state $\hat{x} = Tx \;\Leftrightarrow\; x = T^{-1}\hat{x}$, we obtain from the original system $[A, B, C, D] \in \Sigma_{n,m,p}$

$$\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t), & y(t) &= Cx(t) + Du(t) \\ \Leftrightarrow T^{-1}\dot{\hat{x}}(t) &= AT^{-1}\hat{x}(t) + Bu(t), & y(t) &= CT^{-1}\hat{x}(t) + Du(t), \\ \Leftrightarrow \dot{\hat{x}}(t) &= TAT^{-1}\hat{x}(t) + TBu(t), & y(t) &= CT^{-1}\hat{x}(t) + Du(t), \end{aligned}$$

the transformed system $\left[\hat{A}, \hat{B}, \hat{C}, \hat{D}\right] \in \Sigma_{n,m,p}$ with

$$\hat{A} = TAT^{-1}, \quad \hat{B} = TB, \quad \hat{C} = CT^{-1}, \quad \hat{D} = D, \quad \text{and} \quad \hat{x}(t_0) = Tx_0.$$

We will show that we can find a state-space transformation such that the transformed system is balanced. In order to do so, we have to investigate how such transformations affect the Gramians.

Lemma 2.14 (Transformed Gramians)
Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable. Given $T \in \mathbb{R}^{n\times n}$ invertible with associated transformed system $\left[\hat{A}, \hat{B}, \hat{C}, \hat{D}\right] := [TAT^{-1}, TB, CT^{-1}, D]$,

(a) P is the controllability Gramian of $[A, B, C, D]$ $\Leftrightarrow$ $\hat{P} := TPT^{\mathsf{T}}$ is the controllability Gramian of $\left[\hat{A}, \hat{B}, \hat{C}, \hat{D}\right]$,

(b) Q is the observability Gramian of $[A, B, C, D]$ $\Leftrightarrow$ $\hat{Q} := T^{-\mathsf{T}}QT^{-1}$ is the observability Gramian of $\left[\hat{A}, \hat{B}, \hat{C}, \hat{D}\right]$.

Proof: See Exercise 7.2. □

We are ready to show how to balance a system using a so-called *balancing transformation*.

Theorem 2.15 (Balanced Transformation)
Let the system $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable, controllable, and observable. Then, there exists an invertible matrix $T \in \mathbb{R}^{n\times n}$ such that the transformed system $[TAT^{-1}, TB, CT^{-1}, D] \in \Sigma_{n,m,p}$ is balanced.

Proof: From Lemma 2.10, the Gramians P and Q are positive definite, so that there exist Cholesky decompositions $P = RR^{\mathsf{T}}$ and $Q = LL^{\mathsf{T}}$, where R and L are lower triangular with positive diagonal and thus invertible. We further introduce the singular value decomposition $L^{\mathsf{T}}R = U\Sigma V^{\mathsf{T}}$ with orthogonal $U, V \in \mathbb{R}^{n\times n}$ and $\Sigma = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$.

Since L and R are invertible, so is $L^\intercal R$ and therefore, we even have $\sigma_n > 0$ such that Σ is invertible as well.

Introducing the transformation $T := \Sigma^{-\frac{1}{2}} U^\intercal L^\intercal$ we see that

$$\Sigma^{-\frac{1}{2}} U^\intercal L^\intercal R V \Sigma^{-\frac{1}{2}} = \Sigma^{-\frac{1}{2}} U^\intercal U \Sigma V^\intercal V \Sigma^{-\frac{1}{2}} = I_{n\times n},$$

such that $T^{-1} = RV\Sigma^{-\frac{1}{2}}$. For the controllability Gramian $\hat{P}$ of the transformed system we thus have

$$\hat{P} = TPT^\intercal = \Sigma^{-\frac{1}{2}} U^\intercal L^\intercal R R^\intercal L U \Sigma^{-\frac{1}{2}} = \Sigma^{-\frac{1}{2}} U^\intercal U \Sigma V^\intercal V \Sigma U^\intercal U \Sigma^{-\frac{1}{2}} = \Sigma.$$

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

$$\begin{aligned}\hat{Q} &= T^{-\intercal} Q T^{-1} = \Sigma^{-\frac{1}{2}} V^\intercal R^\intercal L L^\intercal R V \Sigma^{-\frac{1}{2}} = \Sigma^{-\frac{1}{2}} V^\intercal V \Sigma U^\intercal U \Sigma V^\intercal V \Sigma^{-\frac{1}{2}} \\ &= \Sigma = \hat{P}.\end{aligned}$$

and the transformed system $[TAT^{-1}, TB, CT^{-1}, D] \in \Sigma_{n,m,p}$ is balanced. □

We continue with our example.

Example 2.16 (Example 2.12 continued)
Following the steps of the proof of Theorem 2.15, we obtain the singular values $\sigma_1 = 0.80902$ *and* $\sigma_2 = 0.30902$ *(again rounding to* 5 *digits) and, defining the quantity* $\gamma = 0.66874$*, we obtain the transformation matrix*

$$T = \gamma \begin{pmatrix} -1 & -\frac{1}{4\sigma_1^2} \\ 1 & \frac{1}{4\sigma_2^2} \end{pmatrix}.$$

Thus, the transformed system $\left[\hat{A}, \hat{B}, \hat{C}, \hat{D}\right] \in \Sigma_{2,1,1}$ *is given by*

$$\begin{aligned}\hat{A} &= TAT^{-1} = \begin{pmatrix} -\frac{\gamma^2}{2\sigma_1} & -\frac{\gamma^2}{\sigma_1-\sigma_2} \\ -\frac{\gamma^2}{\sigma_2-\sigma_1} & -\frac{\gamma^2}{2\sigma_2} \end{pmatrix} = \begin{pmatrix} -0.27639 & -0.89443 \\ 089443 & -0.72361 \end{pmatrix}, \\ \hat{B} &= TB = \begin{pmatrix} -\gamma \\ \gamma \end{pmatrix} = \begin{pmatrix} -0.66874 \\ 0.66874 \end{pmatrix}, \\ \hat{C} &= CT^{-1} = \begin{pmatrix} \gamma & \gamma \end{pmatrix} = \begin{pmatrix} 0.66874 & 0.66874 \end{pmatrix}, \\ \hat{D} &= D = 0.\end{aligned}$$

Indeed, the system is balanced as

$$\hat{P} = TPT^\intercal = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} = T^{-\intercal} Q T^{-1} \hat{Q}.$$

such that the eigenvector e_2 *of* $\hat{P} = \hat{Q}$ *is both harder to to reach and harder to observe than* e_1*.*

Truncating states of a balanced system that are at the same time difficult to reach and difficult to observe results in truncating states that correspond to small eigenvalues of the transformed gramians. Thus, we can formulate (given a desired reduced state dimension r) our model order reduction algorithm for asympotically stable, controllable and observable LTI systems $[A, B, C, D] \in \Sigma_{n,m,p}$ that is called *Balanced Truncation*, see Algorithm 2.

Algorithm 2 Balanced Truncation($[A, B, C, D] \in \Sigma_{n,m,p}$, $r \leq n$)

1: Solve the Lyapunov equations

$$AP + PA^\intercal = -BB^\intercal, \quad A^\intercal Q + QA = -C^\intercal C$$

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

2: Compute Cholesky factorizations $P = RR^\intercal$ and $Q = LL^\intercal$.

3: Compute the singular value decomposition $L^\intercal R = U\Sigma V^\intercal$.

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

5: Do the balancing transformation

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

with $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$ denoting the *reduced system*.

6: **return** $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$

2.4 Properties of Balanced Truncation

It is now clear how to generate a reduced LTI system, but what properties does the reduced system have? Is it asymptotically stable? And is there a meaningful way of choosing r the reduced state dimension?

We first revisit the setup introduced in section 2.3.1 and discover the theoretical background of the approach made in Algorithm 2. Thus, consider the state equation on the time horizon $(-\infty, 0]$ with $x(-\infty) = 0$ and a control $u \in L^2((-\infty, 0], \mathbb{R}^m)$ steering the system towards

$$x(0) = x_0 = \int_{-\infty}^{0} e^{-A\tau} B u(\tau)\, \mathrm{d}\tau.$$

For the output equation on the time horizon $[0, \infty)$ the control is then switched off so that one obtains an output signal $y \in L^2([0, \infty), \mathbb{R}^p)$ with

$y(t) = Ce^{At}x_0$. This motivates the definition of the following operator mapping "past inputs to future outputs".

Definition and Theorem 2.17 (Hankel Operator & Singular Values) *Let the system $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable. The operator $\mathcal{H} : L^2((-\infty, 0], \mathbb{R}^m) \to L^2([0, \infty), \mathbb{R}^p)$ defined via*

$$u \mapsto (\mathcal{H}u)(t) = y(t) = \int_{-\infty}^{0} Ce^{A(t-\tau)}Bu(\tau)\,\mathrm{d}\tau$$

is called the Hankel Operator *of the system and it is a linear and bounded operator. Thus, its adjoint operator $\mathcal{H}^* : L^2([0, \infty), \mathbb{R}^p) \to L^2((-\infty, 0], \mathbb{R}^m)$ exists and it is defined via*

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

Furthermore, $\sigma \geq 0$ is called a singular value *of $\mathcal{H}$, if σ^2 is an eigenvalue of $\mathcal{H}^*\mathcal{H}$, i.e., there exists an eigenfunction $v \in L^2((-\infty, 0], \mathbb{R}^m) \setminus \{0\}$ such that $\mathcal{H}^*\mathcal{H}v = \sigma^2 v$. In particular, the positive singular values of $\mathcal{H}$ are called* Hankel singular values.

Proof: With the system being asymptotically stable the operator is well-defined and the linearity is obvious due to the linearity of the integral. Letting $x_0 := \int_{-\infty}^{0} e^{-A\tau}Bu(\tau)\,\mathrm{d}\tau$, we have $(\mathcal{H}u)(t) = Ce^{At}x_0$ and thus

$$\begin{aligned}\|\mathcal{H}u\|^2_{L^2([0,\infty),\mathbb{R}^p)} &= x_0^\mathsf{T} \int_0^\infty e^{A^\mathsf{T}\tau} C^\mathsf{T} C e^{A\tau}\,\mathrm{d}\tau x_0 = x_0^\mathsf{T} Q x_0 = \left\| Q^{\frac{1}{2}} x_0 \right\|_2^2 \\ &\leq \left\| Q^{\frac{1}{2}} \right\|_2^2 \|x_0\|_2^2 .\end{aligned}$$

Regarding the second term, we note that for any $v \in \mathbb{R}^n$, one can prove that $\|v\|_2^2 = \max\limits_{\substack{u\in\mathbb{R}^n \\ \|u\|_2=1}} |\langle v, u\rangle_2|^2$. Using Cauchy-Schwartz, we thus obtain

$$\begin{aligned}\|x_0\|_2^2 &= \max_{\substack{x\in\mathbb{R}^n \\ \|x\|_2=1}} |\langle x_0, x\rangle_2|^2 = \max_{\substack{x\in\mathbb{R}^n \\ \|x\|_2=1}} \left| \int_{-\infty}^{0} u(\tau)^\mathsf{T} B^\mathsf{T} e^{-A^\mathsf{T}\tau} x\,\mathrm{d}\tau \right|^2 \\ &= \max_{\substack{x\in\mathbb{R}^n \\ \|x\|_2=1}} \left| \left\langle u, B^\mathsf{T} e^{-A^\mathsf{T}\cdot} x \right\rangle_{L^2((-\infty,0],\mathbb{R}^m)} \right|^2 \\ &\overset{CSU}{\leq} \max_{\substack{x\in\mathbb{R}^n \\ \|x\|_2=1}} \|u\|^2_{L^2((-\infty,0],\mathbb{R}^m)} \left\| B^\mathsf{T} e^{-A^\mathsf{T}\cdot} x \right\|^2_{L^2((-\infty,0],\mathbb{R}^m)} .\end{aligned}$$

But

$$\left\|B^\mathsf{T} e^{-A^\mathsf{T}\cdot} x\right\|^2_{L^2((-\infty,0],\mathbb{R}^m)} = x^\mathsf{T} \underbrace{\int_{-\infty}^0 e^{-A\tau} B B^\mathsf{T} e^{-A^\mathsf{T}\tau} \,\mathrm{d}\tau}_{=P} x = x^\mathsf{T} P x = \left\|P^{\frac{1}{2}} x\right\|_2^2$$
$$\leq \left\|P^{\frac{1}{2}}\right\|_2^2 \|x\|_2^2,$$

so that in total

$$\|\mathcal{H}u\|_{L^2([0,\infty),\mathbb{R}^p)} \leq \left\|Q^{\frac{1}{2}}\right\|_2 \left\|P^{\frac{1}{2}}\right\|_2 \|u\|_{L^2((-\infty,0],\mathbb{R}^m)}$$

and the operator is also continuous. Regarding the adjoint, we calculate

$$\begin{aligned}
\langle \mathcal{H}u, y\rangle_{L^2([0,\infty),\mathbb{R}^p)} &= \int_0^\infty ((\mathcal{H}u)(t))^\mathsf{T} y(t)\,\mathrm{d}t \\
&= \int_0^\infty \int_{-\infty}^0 u(\tau)^\mathsf{T} B^\mathsf{T} e^{A^\mathsf{T}(t-\tau)} C^\mathsf{T} y(t)\,\mathrm{d}\tau\,\mathrm{d}t \\
&= \int_{-\infty}^0 u(\tau)^\mathsf{T} \int_0^\infty B^\mathsf{T} e^{A^\mathsf{T}(t-\tau)} C^\mathsf{T} y(t)\,\mathrm{d}t\,\mathrm{d}\tau \\
&= \langle u, \mathcal{H}^* y\rangle_{L^2((-\infty,0],\mathbb{R}^m)}
\end{aligned}$$

and since the adjoint operator satisfying this property is unique due to Theorem 1.9, the operator $\mathcal{H}^*$ is indeed the adjoint operator. □

It turns out that we already encountered the Hankel singular values during Algorithm 2.

Theorem 2.18 (Hankel Singular Values & Balanced truncation) *Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable, P and Q its controllability and observability Gramians, and $\mathcal{H}$ its Hankel operator. Then, $\sigma > 0$ is a Hankel singular value if and only if σ^2 is an eigenvalues of PQ.*

Proof: We first want to obtain a representation of $(\mathcal{H}^*\mathcal{H}u)(t)$ for some function $u \in L^2((-\infty, 0], \mathbb{R}^m)$. Again, letting $x_0 := \int_{-\infty}^0 e^{-A\tau} B u(\tau)\,\mathrm{d}\tau$, we have $(\mathcal{H}u)(t) = Ce^{At}x_0$. On the other hand, we obtain for some $y \in L^2([0,\infty), \mathbb{R}^p)$

$$(\mathcal{H}^* y)(t) = \int_0^\infty B^\mathsf{T} e^{A^\mathsf{T}(\tau-t)} C^\mathsf{T} y(\tau)\,\mathrm{d}\tau = B^\mathsf{T} e^{-A^\mathsf{T} t} \int_0^\infty e^{A^\mathsf{T}\tau} C^\mathsf{T} y(\tau)\,\mathrm{d}\tau.$$

This leads to

$$(\mathcal{H}^*\mathcal{H}u)(t) = B^\mathsf{T} e^{-A^\mathsf{T} t} \int_0^\infty e^{A^\mathsf{T}\tau} C^\mathsf{T} C e^{A\tau} x_0\,\mathrm{d}\tau = B^\mathsf{T} e^{-A^\mathsf{T} t} Q x_0.$$

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

$$(\mathcal{H}^*\mathcal{H}u)(t) = B^\mathsf{T} e^{-A^\mathsf{T} t} Q x_0 = \sigma^2 u(t) \quad \Leftrightarrow \quad u(t) = \frac{1}{\sigma^2} B^\mathsf{T} e^{-A^\mathsf{T} t} Q x_0.$$

Inserting this u into the definition of x_0 yields

$$x_0 = \int_{-\infty}^{0} e^{-A\tau} B \frac{1}{\sigma^2} B^\mathsf{T} e^{-A^\mathsf{T}\tau} Q x_0 \,\mathrm{d}\tau = \frac{1}{\sigma^2} \underbrace{\int_{-\infty}^{0} e^{-A\tau} B B^\mathsf{T} e^{-A^\mathsf{T}\tau} \,\mathrm{d}\tau}_{=P} Q x_0 = \frac{1}{\sigma^2} P Q x_0$$

such that σ^2 is an eigenvalue of PQ.

"⇐": Now assume that $\sigma^2 > 0$ is an eigenvalue of PQ with an eigenvector $v \in \mathbb{R}^n \setminus \{0\}$ and define $u(t) = \frac{1}{\sigma^2} B^\mathsf{T} e^{-A^\mathsf{T} t} Q v \in L^2((-\infty, 0], \mathbb{R}^m)$. We obtain

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

and σ is a singular value of $\mathcal{H}$. □

Coming back to Algorithm 2, the system is then also controllable and observable, such that P and Q are positive definite and the Cholesky factorizations $P = RR^\mathsf{T}$ and $Q = LL^\mathsf{T}$ exist. Thus, if σ^2 is an eigenvalue of PQ for an eigenvector $v \in \mathbb{R}^n$, then we have

$$PQv = \left(RR^\mathsf{T}\right)\left(LL^\mathsf{T}\right)v = \sigma^2 v.$$

Multiplying both sides with R^{-1} from the left, this is equivalent to

$$\underbrace{\left(R^\mathsf{T} L\right)}_{=(L^\mathsf{T} R)^\mathsf{T}} \left(L^\mathsf{T} R\right) R^{-1} v = \sigma^2 R^{-1} v,$$

which implies that σ is a singular value of $L^\mathsf{T} R$. Therefore, the positive entries of Σ computed in Algorithm 2 are exactly the Hankel singular values of the system. Thus, the model order reduction is based on truncating the states of the transformed system that correspond to the neglectable Hankel singular values.

We now turn our attention to the question: is the reduced system obtained with Algorithm 2 asymptotically stable? Is it also controllable and observable? The answers are given in the upcoming Theorem.

Theorem 2.19 (Stability of the Reduced System)
Let $[A, B, C, D] \in \Sigma_{n,m,p}$ be asymptotically stable, controllable and observable and let $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$ be the reduced system obtained with Algorithm 2. Assume that $\sigma_r > \sigma_{r+1}$ holds for the Hankel singular values σ_i, $i = 1, \ldots, n$ of the full system. Then, the reduced system $[A_{11}, B_1, C_1, D]$ is asymptotically stable, observable, controllable, and balanced with the Gramians $P_{11} = Q_{11} = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) =: \Sigma_1$.

Proof: Since the system $[A, B, C, D] \in \Sigma_{n,m,p}$ is controllable and observable, the balancing transformation introduced in Theorem 2.15 leads to the transformed Gramians

$$\hat{P} = \hat{Q} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) =: \operatorname{diag}(\Sigma_1, \Sigma_2) > 0$$

and the Lyapunov equations in balanced coordinates read

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} A_{11}^\mathsf{T} & A_{21}^\mathsf{T} \\ A_{12}^\mathsf{T} & A_{22}^\mathsf{T} \end{bmatrix} = - \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \begin{bmatrix} B_1^\mathsf{T} & B_2^\mathsf{T} \end{bmatrix}, \quad (2.2)$$

$$\begin{bmatrix} A_{11}^\mathsf{T} & A_{21}^\mathsf{T} \\ A_{12}^\mathsf{T} & A_{22}^\mathsf{T} \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} + \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = - \begin{bmatrix} C_1^\mathsf{T} \\ C_2^\mathsf{T} \end{bmatrix} \begin{bmatrix} C_1 & C_2 \end{bmatrix}. \quad (2.3)$$

Thus, if the reduced system is asymptotically stable, Lemma 2.10 yields that $\Sigma_1 > 0$ is the controllability and observability Gramian of the reduced system which is thus balanced. Furthermore, the reduced system is then also controllable and observable (one can show, see e.g., [ANT, Theorem 4.15 & Theorem 4.26], that the converse of Lemma 2.10 holds as well, i.e., that the positive definiteness of a Gramian implies controllability/observability).

Thus, it remains to show that all eigenvalues of A_{11} have negative real part. Let $\lambda \in \mathbb{C}$ be an eigenvalue of A_{11}^T with eigenvector $0 \neq v \in \mathbb{C}^r$ (thus, $\bar{\lambda}$ is an eigenvalue of A_{11}^T for the eigenvector $\bar{v}$). Multiplying the reduced Lyapunov equation

$$A_{11}\Sigma_1 + \Sigma_1 A_{11}^\mathsf{T} = -B_1 B_1^\mathsf{T}$$

with v^* from the left and with v from the right, yields

$$-\left\|B_1^\mathsf{T} v\right\|_2^2 = v^\star A_{11}\Sigma_1 v + v^\star \Sigma_1 A_{11}^\mathsf{T} v = v^\star A_{11}\Sigma_1 v + \lambda v^\star \Sigma_1 v$$

and since

$$v^\star A_{11}\Sigma_1 v = \left(\left(v^\star A_{11} v\right)^\mathsf{T}\right)^\mathsf{T} \Sigma_1 = \left(v^\mathsf{T} A_{11}^\mathsf{T} \bar{v}\right)^\mathsf{T}\Sigma_1 = \left(v^\mathsf{T}\bar{\lambda}\bar{v}\right)^\mathsf{T}\Sigma_1 = \bar{\lambda} v^\star \Sigma_1 v$$

we have

$$\underbrace{-\left\|B_1^\mathsf{T} v\right\|_2^2}_{\leq 0} = 2\mathrm{Re}(\lambda)\,\underbrace{v^\star \Sigma_1 v}_{>0},$$

such that $\mathrm{Re}(\lambda) \leq 0$ and it remains to show that A_{11} has no eigenvalues on the imaginary axis. Therefore, assume that there exist imaginary eigenvalues. Let $i\omega \in i\mathbb{R}$ be an imaginary eigenvalue and $\{v_1, \ldots, v_q\} \subset \mathbb{C}^r$ be an orthonormal basis of $\mathcal{N}(A_{11} - i\omega I_r)$, i.e., $V := \begin{bmatrix} v_1 & \ldots & v_q \end{bmatrix} \in \mathbb{C}^{r\times q}$ spans the eigenspace for the eigenvalue $i\omega$. Then, we have

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

and from the reduced Lyapunov equations, we have

$$(A_{11} - i\omega I_r)\Sigma_1 + \Sigma_1\left(A_{11}^\mathsf{T} + i\omega I_r\right) = -B_1 B_1^\mathsf{T}, \tag{2.4}$$
$$\left(A_{11}^\mathsf{T} + i\omega I_r\right)\Sigma_1 + \Sigma_1(A_{11} - i\omega I_r) = -C_1^\mathsf{T} C_1. \tag{2.5}$$

Multiplying (2.5) with $V^\star$ from the left and with V from the right gives

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

resulting in $C_1 V = 0$. Multiplying (2.5) with V from the right yields

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

and thus $\left(A_{11}^\mathsf{T} + i\omega I_r\right)\Sigma_1 V = 0$. Now, multiplying (2.4) with $V^\star\Sigma_1$ from the left and with $\Sigma_1 V$ from the right results in

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

giving $B_1^\mathsf{T}\Sigma_1 V = 0$. By multiplying (2.4) with $\Sigma_1 V$ from the right, we obtain

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

such that $(A_{11} - i\omega I_r)\Sigma_1^2 V = 0$ and each column of $\Sigma_1^2 V$ is an element of $\mathcal{N}(A_{11} - i\omega I_r)$. Since V is a basis of this kernel, there exists a matrix $\Xi \in \mathbb{C}^{q\times q}$ such that

$$\Sigma_1^2 V = V\Xi \quad \text{with} \quad \Lambda(\Xi) \subseteq \Lambda\big(\Sigma_1^2\big), \tag{2.6}$$

where the second statement follows from the first one: let $\lambda \in \mathbb{C}$ be an eigenvalue of Ξ with eigenvector $y \in \mathbb{C}^q$, then

$$\Sigma_1^2 V y = V\Xi y = \lambda V y,$$

such that λ is also an eigenvalue of Σ_1^2 with eigenvector Vy. Multiplying the (2,1) block of (2.2) 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 (2.3) 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 (2.6) and both of the last two equations we get

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

This is a *Sylvester matrix equation*

$$N_1 X + X N_2 = M$$

with $N_1 \equiv -\Sigma_2^2$, $N_2 \equiv \Xi$, unknown $X \equiv A_{21}V$, and $M \equiv 0$. One can show that this matrix equation has a unique solution, if N_1 and $-N_2$ have disjoint spectra. Since we have $\Lambda(\Xi) \cap \Lambda(\Sigma_2^2) = \emptyset$ from (2.6), the solution is unique here and since the zero matrix is a solution to this equation, we obtain $A_{21}V = 0$. In total, we have

$$\hat{A}\begin{bmatrix} V \\ 0 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} V \\ 0 \end{bmatrix} = \begin{bmatrix} A_{11}V \\ A_{21}V \end{bmatrix} = i\omega \begin{bmatrix} V \\ 0 \end{bmatrix},$$

and since $\hat{A}$ and A are similar (T is a similarity transform), $i\omega$ is also an imaginary eigenvalue of A, contradicting the asymptotic stability of the original system. □

To conclude this chapter,we want to present an error bound for the approximation error of Balanced Truncation. We first define the necessary concepts and begin with the Laplace transformation.

Definition 2.20 (Laplace Transformation)
Let $f : [0, \infty) \to \mathbb{R}^n$ be exponentially bounded, i.e., there exist $M \geq 0$ and $\alpha \geq 0$ such that $\|f(t)\|_2 \leq Me^{\alpha t}$ for all $t \geq 0$. Then,

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

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

We want to apply the Laplace transform to (LTI): assuming that each of the Laplace transforms $X(s) := \mathcal{L}\{x\}(s)$, $U(s) := \mathcal{L}\{u\}(s)$, and $Y(s) := \mathcal{L}\{y\}(s)$ exists, one can show that the Laplace transformed system reads

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

and 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 2.21 (Transfer Function)
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.*

We can see, that the transfer function maps the input of the system to the corresponding output (in the frequency domain). Thus, the error between the transfer functions of the full and reduced system $G(s) - \hat{G}(s)$ is an interesting quantity. In order to measure this error in the correct norm, we introduce the following *Hardy Space*.

Definition 2.22 (The Space $\mathcal{H}_\infty^{p\times m}$)
Introducing the $\mathcal{H}_\infty$-norm $\|G\|_{\mathcal{H}_\infty} := \sup_{\omega\in\mathbb{R}} \|G(i\omega)\|_2$, the space $\mathcal{H}_\infty^{p\times m}$

$$\mathcal{H}_\infty^{p\times m} := \left\{G : \mathbb{C}^+ \to \mathbb{C}^{p\times m} : G \text{ is analytic in } \mathbb{C}^+ \text{ and } \|G\|_{\mathcal{H}_\infty} < \infty\right\}$$

equipped with this norm is a Banach space.

Based on these concepts, one can show the following error bound.

Theorem 2.23 (Balanced Truncation Error Bound)
Let $[A, B, C, D] \in \Sigma_{n,m,p}$ with transfer function $G \in \mathcal{H}_\infty^{p\times m} \cap \mathbb{R}(s)^{p\times m}$ be asymptotically stable and balanced with Gramians

$$P = Q = \operatorname{diag}(\sigma_1 I_{s_1}, \sigma_2 I_{s_2}, \ldots, \sigma_k I_{s_k}) \text{ where } \sigma_1 > \sigma_2 > \ldots > \sigma_k \geq 0.$$

Let $[A_{11}, B_1, C_1, D] \in \Sigma_{r,m,p}$ be the reduced of order r obtained with Algorithm 2 with $r = s_1 + s_2 + \ldots + s_l$ for some $l \leq k$ and with transfer function $\hat{G} \in \mathcal{H}_\infty^{p\times m} \cap \mathbb{R}(s)^{p\times m}$. Then, it holds

$$\left\|G - \hat{G}\right\|_{\mathcal{H}_\infty} \leq \sum_{j=l+1}^{k} 2\sigma_j.$$

Proof: See, e.g., [ANT, Theorem 7.9]. □

Finally, one can show

$$\|G\|_{\mathcal{H}_\infty} = \sup_{\substack{u\in L^2([0,\infty),\mathbb{R}^m)\\ u\neq 0}} \frac{\|y\|_{L^2([0,\infty),\mathbb{R}^p)}}{\|u\|_{L^2([0,\infty),\mathbb{R}^p)}},$$

such that this bound also relates back to the time domain of the system.

Bibliography

[Alt] Alt, H. W. (1992). Linear functional analysis. *An application oriented introduction*.

[EIM] Barrault, M., Maday, Y., Nguyen, N. C., & Patera, A. T. (2004). An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *Comptes Rendus Mathematique*, 339(9), 667-672.

[SCM] Huynh, D. B. P., Rozza, G., Sen, S., & Patera, A. T. (2007). A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. *Comptes Rendus Mathematique*, 345(8), 473-478.

[BCDPWD] Binev, P., Cohen, A., Dahmen, W., DeVore, R., Petrova, G., & Wojtaszczyk, P. (2011). Convergence rates for greedy algorithms in reduced basis methods. *SIAM journal on mathematical analysis*, 43(3), 1457-1472.

[ANT] Antoulas, A. C. (2005). Approximation of large-scale dynamical systems (Vol. 6). *Siam*.