

Vorlesungsmitschrift

# REDUZIERTE BASIS METHODEN

UNIVERSITÄT STUTTGART, SS15  
Prof. Dr. Bernard Haasdonk

AUTOREN:  
Stefan Simeonov  
Frank Schneider

STAND:  
19. Juli 2015

## Inhaltsverzeichnis

|          |                                                  |            |
|----------|--------------------------------------------------|------------|
| <b>1</b> | <b>Einleitung</b>                                | <b>2</b>   |
| 1.1      | Modellreduktion . . . . .                        | 3          |
| <b>2</b> | <b>Grundlagen</b>                                | <b>8</b>   |
| <b>3</b> | <b>RB-Methoden für lineare koerzive Probleme</b> | <b>16</b>  |
| 3.1      | Primales RB-Problem . . . . .                    | 16         |
| 3.2      | Fehleranalyse . . . . .                          | 19         |
| 3.3      | Offline/Online-Zerlegung . . . . .               | 30         |
| 3.4      | Basisgenerierung . . . . .                       | 41         |
| 3.5      | Primal-Duale RB-Verfahren . . . . .              | 69         |
| 3.6      | Geometrieparametrisierung . . . . .              | 73         |
| <b>4</b> | <b>Allgemeinere Lineare Probleme</b>             | <b>81</b>  |
| 4.1      | Allgemeine Parameterabhängigkeit . . . . .       | 81         |
| 4.2      | Inf-sup stabile Probleme . . . . .               | 89         |
| <b>5</b> | <b>Nichtlineare Probleme</b>                     | <b>97</b>  |
| <b>6</b> | <b>Zeitabhängige Probleme</b>                    | <b>101</b> |

# 1 Einleitung

## Parameterabhängige Probleme

### Beispiel 1.1 (Parameterabhängige PDE)

Sie  $\Omega \subseteq \mathbb{R}^d$  polygonales Gebiet. Zu Parametervektor  $\mu \in \mathcal{P} \subset \mathbb{R}^p$  aus einer Menge  $\mathcal{P}$  von „erlaubten“ Parametern ist eine Funktion (z. B. „Temperatur“)  $u(\mu) : \Omega \rightarrow \mathbb{R}$ , s. d.:

$$\begin{aligned} \nabla \cdot (\kappa(\mu) \nabla u) &= q(\mu) && \text{in } \Omega \\ u(\mu) &= 0 && \text{auf } \delta\Omega \end{aligned}$$

mit  $\kappa(\mu) : \Omega \rightarrow \mathbb{R}$  (z. B. „Wärmeleitungskoeffizient“)

und  $q(\mu) : \Omega \rightarrow \mathbb{R}$  (z. B. „Wärmequelle/-senke“)

$$\text{z. B. } q(x; \mu) := \begin{cases} 1 & \text{für } x \in \Omega_q \\ 0 & \text{sonst} \end{cases}$$

Weiter kann Ausgabe erwünscht, z. B. mittlere Temperatur

$$s(\mu) = \frac{1}{|\Omega_s|} \int u(x; \mu) dx$$

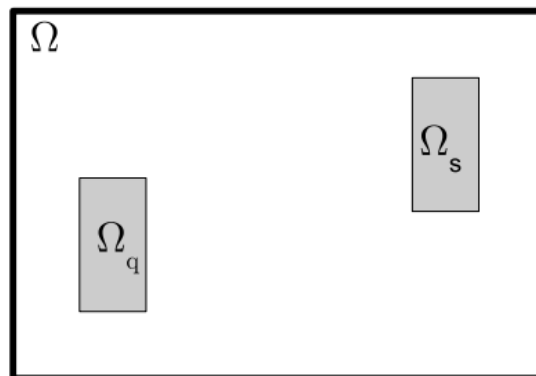


Abbildung 1: Beispiel Wärmeleitung mit Quelle  $\Omega_q$  und Messbereich  $\Omega_s$   
(aus B. Haasdonk, Reduzierte-Basis-Methoden, Skript zur Vorlesung SS 2011, Universität Stuttgart, IANS-Report 4/11, 2011.)

### Beispiel 1.2 (Parametrisches stationäres System)

Zu Parameter  $\mu \in \mathcal{P} \subseteq \mathbb{R}^p$  ist Zustandsvektor  $u(\mu) \in \mathbb{R}^n$  und Ausgabe  $s(\mu) \in \mathbb{R}^k$  gesucht, s. d.:

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

mit parameterabhängigen Matrizen  $A(\mu) \in \mathbb{R}^{n \times n}$ ,  $B(\mu) \in \mathbb{R}^{n \times m}$ ,  $C(\mu) \in \mathbb{R}^{k \times n}$  mit Eingabevektor  $w \in \mathbb{R}^m$ .

## Schwache Formulierung in Hilberträumen

Sie  $X$  reeller Hilbertraum (reel, seperabel). Zu  $\mu \in \mathcal{P}$  ist gesucht ein  $u(\mu) \in X$  und  $s(\mu) \in \mathbb{R}$

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

Mit Bilinearform  $a(\cdot, \cdot; \mu)$  und Linearform  $f(\cdot; \mu)$ ,  $l(\cdot; \mu)$ . Beide Beispiele lassen sich so formulieren.

z. B. 1.1:

$$\begin{aligned} X &= H_0^1(\Omega) := \{f \in L^2(\Omega) \mid \frac{\partial}{\partial x_i} f \in L^2(\Omega), f|_{\partial\Omega} = 0\} \\ \underbrace{\int_{\Omega} \kappa(x; \mu) \nabla u(x; \mu) \cdot \nabla v(x) dx}_{a(u(\mu), v; \mu)} &= \underbrace{\int_{\Omega} q(x; \mu) \cdot v(x) dx}_{f(v; \mu)} \quad \forall v \in X \\ s(\mu) &= \frac{1}{|\Omega_s|} \int_{\Omega_s} u(x; \mu) =: l(u(\mu); \mu) \end{aligned}$$

Zu Bsp. 1.2 ( $k = 1$ , „single output“)  $X = \mathbb{R}^n$

$$\begin{aligned} \underbrace{v^T A(\mu) u(\mu)}_{a(u(\mu), v; \mu)} &= \underbrace{-v^T B w}_{f(v; \mu)} \\ s(\mu) &:= \underbrace{C(\mu) u(\mu)}_{l(u(\mu); \mu)} \end{aligned}$$

In der Vorlesung werden weitere Verallgemeinerungen zu  $a : X_1 \times X_2 \rightarrow \mathbb{R}$  mit  $X_1 \neq X_2$ , nichtlinear und instationäre Probleme behandelt.

## 1.1 Modellreduktion

### Grundidee/Motivation

- $\mathcal{M} := \{u(\mu) \mid \mu \in \mathcal{P}\} \subset X$  für  $\mathcal{P} \subseteq \mathbb{R}^p$  ist die durch  $\mu$  parametrisierte Lösungsmanigfaltigkeit.
- $X$  ist im allgemeinen  $\infty$ -dimensional Sobolev-Raum) oder endlich- aber sehr hochdimensional (FEM, FV, FD-Raum).  $\mathcal{M}$  ist aber höchstens  $p$ -dimensional.  
 $\Rightarrow$  Motivation für Suche nach einem niedrigdimensionalen Teilraum  $X_n \subseteq X$  zur Approximation von  $\mathcal{M}$  und einer Approximation  $u_N(\mu) \approx u(\mu)$ ,  $u_N(\mu) \in X_N$
- Insbesondere bei Reduzierten-Basis-Methoden (RB-Methoden):  
 $X_N$  durch Beispiellösungen erzeugt, sog. „Snapshots“  
 $X_N \subseteq \text{span}\{u(\mu_1), \dots, u(\mu_n)\}$  für geeignete Parameterwerte  $\mu_i \in \mathcal{P}$ .  
Ziel ist außerdem Fehlerkontrolle durch Schranken  $\Delta_N(\mu)$ :

$$\|u(\mu) - u_N(\mu)\| \leq \Delta_N(\mu)$$

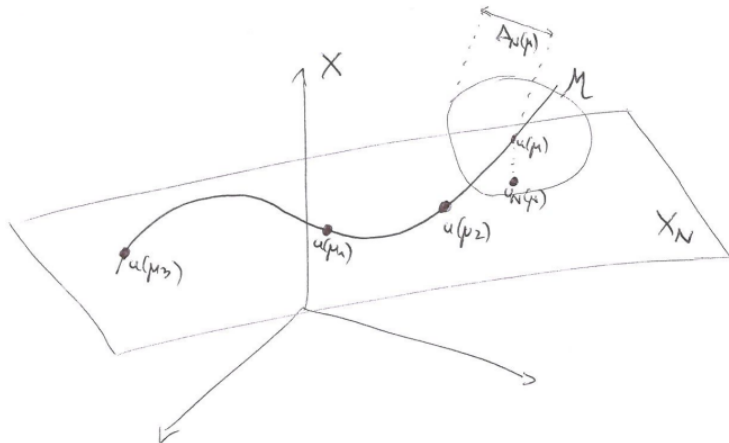


Abbildung 2: Parametrisierte niedrigdimensionale Lösungsmenge  
(aus dem Online-Skript von Prof. Dr. Haasdonk zu Reduzierte Basen 2015)

### Illustration

#### Beispiel 1.3

Gesucht ist  $u(\mu) \in C^2([0,1])$  mit

$$\begin{aligned} (1 + \mu)u'' &= 1 & \text{auf } (0,1) \\ u(0) &= u(1) = 1 \end{aligned}$$

Für  $\mu \in [0,1] =: \mathcal{P} \subseteq \mathbb{R}$ . Spezielle Lösungen („Snapshots“)

$$\begin{aligned} \mu = 0 &\Rightarrow u_0(x) = u(x; \mu = 0) = \frac{1}{2}x - \frac{1}{2}x + 1 \\ \mu = 1 &\Rightarrow u_1(x) = u(x; \mu = 1) = \frac{1}{4}x - \frac{1}{4}x + 1 \end{aligned}$$

RB-Raum:  $X_N := \text{span}(u_0, u_1)$  Reduzierte Lösung gegeben durch

$$\begin{aligned} u_N(\mu) &:= \alpha_0(\mu)u_0 + \alpha_1(\mu)u_1 \\ \alpha_0(\mu) &= \frac{2}{\mu + 1} - 1; \quad \alpha_1(\mu) = 2 - \frac{2}{\mu + 1} \end{aligned}$$

Diese erfüllt

$$\|u_N(\mu) - u(\mu)\|_\infty = \sup_{\mu} \|u(\mu) - u_N(\mu)\| = 0$$

ist somit exakt.  $\mathcal{M}$  ist enthalten in 2-dimensionalem Unterraum  $X_N$ : Genauer  $\alpha_0 + \alpha_1 = 1, 0 \leq \alpha_0, \alpha_1 \leq 1$ , also ist  $\mathcal{M}$  Menge der Konvexkombinationen von  $u_0, u_1$ .

## Begriffe

- Eine PDE ist ein *analytisches* Modell, welches die *exakte Lösung*  $u(\mu) \in X$  in einem typischerweise  $\infty$ -dimensionalen Funktionenraum  $X$  charakterisiert.
- Ein *detailliertes Modell* (auch *hochdimensionales Modell*) ist ein Berechnungsverfahren oder charakterisiert eine Approximation  $u(\mu) \in X$  in hochdimensionalen Raum mit sehr allgemeinen Approximationseigenschaften. (z.B. FEM/FV/FD,  $\dim X = 10^3 - 10^8$ ). In dieser Vorlesung kann  $u(\mu)$  sowohl eine analytische als auch eine detaillierte Lösung darstellen.
- Ein *reduziertes* Modell ist ein Berechnungsverfahren bzw. eine Charakterisierung einer reduzierten Lösung  $u_N(u)$  in einem sehr problemangepassten Raum  $X_N$  ( $\dim X_N = 1 - 10^3$ ).
- *Modellreduktion* beschäftigt sich mit Methoden der Erzeugung reduzierter Modelle und Untersuchung ihrer Eigenschaften
- Modellreduktion ist ein modernes Gebiet der angewandten Mathematik und Ingenieurwissenschaften (Schwerpunkt in SimTech PN3, MOR-Seminar)

## Anwendungen für parametrische reduzierte Modelle

„Kleinere“ Modelle stellen geringere Anforderungen an Rechenzeit und Speicher, daher Einsatz in:

- „multi-query“-Kontext, d. h. Vielfachanfragen unter Parametervariation: Parameterstudien, Design, Parameteridentifikation, Inverse Probleme, Optimierung, statistische Analyse
- Multi-skalen-Modelle (reduzierte Mikrolöser)
- „real-time“-Kontext, d. h. Anwendungen mit schneller Simulationsantwort: Interaktive Benutzeroberfläche, Web-Formulare, Echtzeitsteuerung von Prozessen
- „cool-computing“-Kontext, d. h. Simulation auf „einfacher“ Hardware: elektronische Regler, Smartphones, Ubiquitous Computing

## Demonstration

demo\_thermalblock.m aus RBmatlab, Smartphone App JaRMoS

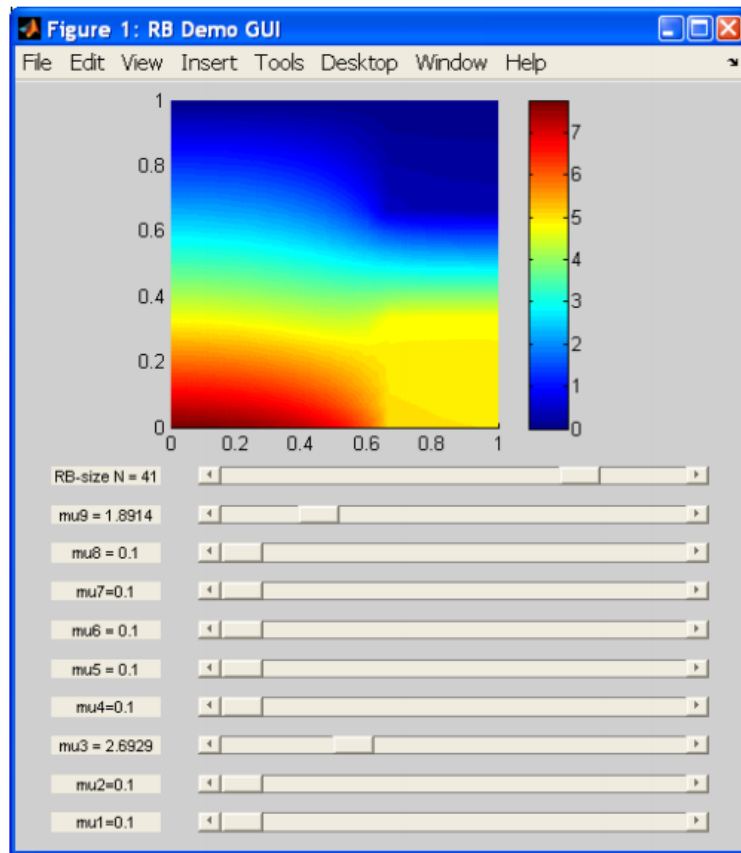


Abbildung 3: Beispiel des Thermischen Blocks aus `demo_thermalblock.m`  
 (aus B. Haasdonk, Reduzierte-Basis-Methoden, Skript zur Vorlesung SS 2011, Universität  
 Stuttgart, IANS-Report 4/11, 2011.)

### Offline/Online Zerlegung

Typischerweise wird eine Verechnungsintensive Generierung des reduzierten Modells akzeptiert, sog. *Offline-Phase*. Dies ermöglicht schnelle Anwendbarkeit des reduzierten Modells in der *Online-Phase*. Offline-Kosten werden gerechtfertigt durch Amortisierung im multi-query-Kontext, d. h. Laufzeitgewinn bei genügend großer Anzahl an Online-Simulationen

multi-query mit detailliertem Modell:



multi-query mit reduziertem Modell:

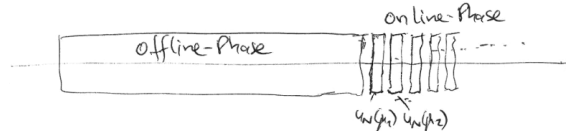


Abbildung 4: Laufzeitvergleich eines detaillierten mit einem reduzierten Modell  
(aus dem Online-Skript von Prof. Dr. Haasdonk zu Reduzierte Basen 2015)

### Zentrale Fragen

- Reduzierte Basis: Wie kann ein möglichst kompakter Teilraum konstruiert werden? Können solche Verfahren *beweisbar* gut sein?
- Reduziertes Modell: Wie kann eine Lösung  $u_N(\mu) \in X_N$  bestimmt werden
- Berechnungs-Effizienz: Wie kann  $u_N(\mu)$  *schnell* berechnet werden?
- Stabilität: Wie kann Stabilität des reduzierten Modells garantiert werden bei wachsendem  $N := \dim X_N$ ?
- Fehlerschätzer: Kann der Fehler des reduzierten zum detaillierten oder analytischen Modells beschränkt werden? Sind die Fehlerschätzer schnell berechenbar?
- Effektivität der Fehlerschätzer: Kann garantiert werden, dass der Schätzer den Fehler nicht zu pessimistisch überschätzt?
- Für welche Problemklassen kann ein RB-Ansatz funktionieren, für welche nicht?

### Vorläufige Gliederung

- 1 Einleitung
- 2 Grundlagen
- 3 RB Verfahren für lineare koerzive Probleme
- 4 Allgemeinere lineare Probleme
- 5 Nichtlineare Probleme
- 6 Instationäre Probleme
- 7 Weiterführende Aspekte

## 2 Grundlagen

Im Folgenden sei  $X$  (oder  $X_1, X_2$ ) stets reeller Hilbertraum mit Skalarprodukt  $\langle \cdot, \cdot \rangle_X$ , Norm  $\| \cdot \|_X$  und Dualraum  $X'$ . Subskript wird weggelassen falls keine Verwechslungsgefahr besteht.

**Definition 2.1** (Parametrische Formen)

Sei  $\mathcal{P} \subset \mathbb{R}^p$  beschränkte Parametermenge. Dann nennen wir

i)  $l : X \times \mathcal{P} \rightarrow \mathbb{R}$  *parametrische stetige Linearform* falls  $\forall \mu \in \mathcal{P}$ :

$$l(\cdot; \mu) \in X'$$

ii)  $a : X_1 \times X_2 \times \mathcal{P} \rightarrow \mathbb{R}$  eine *parametrische stetige* (symmetrische) *Bilinearform*, falls für alle  $\mu \in \mathcal{P}$

$$a(\cdot, \cdot; \mu) : X_1 \times X_2 \rightarrow \mathbb{R} \quad \text{ist bilinear und stetig (symmetrisch)}$$

Wir bezeichnen die Stetigkeitskonstante mit

$$\gamma(\mu) := \sup_{u \in X_1} \sup_{v \in X_2} \frac{a(u, v; \mu)}{\|u\|_{X_1} \|v\|_{X_2}}$$

Falls  $X_1 = X_2 =: X$  und  $a(\cdot, \cdot; \mu)$  ist koerziv für alle  $\mu \in \mathcal{P}$ , so ist  $a(\cdot, \cdot; \cdot)$  *parametrisch koerziv* und wir bezeichnen die Koerzitivitätskonstante mit

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

**Bemerkung.** Eine parametrische stetige Bi-/Linearform ist nicht unbedingt stetig bzgl.  $\mu$ . Beispiel:  $X = \mathbb{R}$ ,  $\mathcal{P} = [0, 1]$ ,  $l : X \times \mathcal{P} \rightarrow \mathbb{R}$  definiert durch

$$l(x; \mu) := \begin{cases} x & \text{falls } \mu < \frac{1}{2} \\ \frac{1}{2}x & \text{sonst} \end{cases}$$

**Definition 2.2** (Parametrische Beschränktheit / Lipschitz-Stetigkeit / Koerzitivität)

Wir nennen

i) eine parametrische stetige Linearform  $l$  bzw. Bilinearform  $a$  *gleichmäßig beschränkt* bzgl.  $\mu$  falls ex.  $\bar{\gamma}_l, \bar{\gamma} < \infty$  mit

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

ii)  $a$  *gleichmäßig koerziv* bzgl.  $\mu$  falls ex.  $\bar{\alpha} > 0$  mit

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



iii)  $l$  bzw.  $a$  *Lipschitz-stetig* bzgl.  $\mu$  falls ex.  $L_l$  bzw.  $L_a \in \mathbb{R}^+$ , sodass  $\forall \mu_1, \mu_2 \in \mathcal{P}$  gilt

$$|l(u; \mu_1) - l(u; \mu_2)| \leq L_l \|u\| \|\mu_1 - \mu_2\| \quad \forall u \in X$$

bzw.

$$|a(u, v; \mu_1) - a(u, v; \mu_2)| \leq L_a \|u\| \|v\| \|\mu_1 - \mu_2\| \quad \forall u \in X_1, v \in X_2$$

**Definition 2.3** (Sensitivitätsableitung)

Sei  $\mu_0 \in \mathcal{U} \subset \mathcal{P}$  in Umgebung  $\mathcal{U}$  von  $\mu_0$ . Wir nennen  $f : \mathcal{U} \rightarrow X$  (Frechet)-differenzierbar in  $\mu_0$ , falls ex. ein  $Df(\mu_0) \in L(\mathbb{R}^p, X)$  mit

$$\lim_{h \rightarrow 0} \frac{\|f(\mu_0 + h) - f(\mu_0) - Df(\mu_0)h\|}{\|h\|} = 0$$

Falls  $f$  in jedem  $\mu \in \mathcal{U}$  diffbar, dann existieren insbesondere partielle Ableitungen

$$\frac{\partial}{\partial \mu_i} f(\cdot) := Df(\cdot) e_i : \mathcal{U} \rightarrow X$$

für  $e_i \in \mathbb{R}^p$  Einheitsvektor  $i = 1, \dots, p$ . Falls diese wiederum diffbar in  $\mathcal{U}$  bezeichnet allgemein

$$\partial_\sigma f(\cdot) := \frac{\partial^{|\sigma|}}{\partial \mu_1^{\sigma_1} \dots \partial \mu_p^{\sigma_p}} f(\cdot) : \mathcal{U} \rightarrow X$$

die Sensitivitätsableitung der Ordnung  $|\sigma| := \sum_{i=1}^p \sigma_i$  für Multiindex  $\sigma = (\sigma_i)_{i=1}^p \in \mathbb{N}_0^p$ .

**Bemerkung.** Diese Ableitungen werden später insbesondere bei parameterabhängigen Lösungen  $u(x; \mu)$  verwendet:

$u : \Omega \times \mathcal{P} \rightarrow \mathbb{R}$  mit  $u(\cdot; \mu) \in X$  kann auch als

$u : \mathcal{P} \rightarrow X$  aufgefasst werden mit Sensitivitätsableitungen

$\partial_\sigma u : \mathcal{P} \rightarrow X$ , d.h.  $\partial_\sigma u(\cdot; \mu) \in X \quad \forall \mu \in \mathcal{P}$  und insbesondere

$\partial_\sigma u : \Omega \times \mathcal{P} \rightarrow \mathbb{R}$ , d.h.  $\partial_\sigma$  sind wieder Funktionen auf  $\Omega$

**Definition 2.4** (Separierbare Parameterabhängigkeit)

- i) Eine Funktion  $v : \mathcal{P} \rightarrow X$  nennen wir *separierbar parametrisch*, falls existieren Komponenten  $v^q \in X$  und Koeffizientenfunktionen  $\Theta_v^q : \mathcal{P} \rightarrow \mathbb{R}$  für  $q = 1, \dots, Q_v$  mit

$$v(\mu) = \sum_{q=1}^{Q_v} \Theta_v^q(\mu) v^q$$

- ii) Eine parametrische stetige Linearform  $l : X \times \mathcal{P} \rightarrow \mathbb{R}$  bzw. Bilinearform  $a : X_1 \times X_2 \times \mathcal{P} \rightarrow \mathbb{R}$  ist separierbar parametrisch, falls existieren  $l^q \in X'$  und  $\Theta_l^q : \mathcal{P} \rightarrow \mathbb{R}$  für  $q = 1, \dots, Q_l$  bzw.  $a^q : X_1 \times X_2 \rightarrow \mathbb{R}$  stetig, bilinear und  $\Theta_a^q : \mathcal{P} \rightarrow \mathbb{R}$  für  $q = 1, \dots, Q_a$  mit

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

$$a(u, v; \mu) = \sum_{q=1}^{Q_a} \Theta_a^q(\mu) a^q(u, v) \quad \forall u \in X_1, v \in X_2, \mu \in \mathcal{P}$$

**Bemerkung.**

- i) In Literatur auch “affine Annahme” oder “affin parametrisch” verwendet. Wir verwenden jedoch “separierbar”, da  $\Theta_l^q$  auch nichtlinear sein können.
- ii)  $Q_a, Q_l$  sollten möglichst klein sein, weil diese in die Online-Komplexität eingehen, siehe *Abschnitt 3*.

**Satz 2.5** (Energienorm)

Sei  $a : X \times X \times \mathcal{P} \rightarrow \mathbb{R}$  parametrische stetige, koerzive Bilinearform, und  $a_s(u, v; \mu) = \frac{1}{2}(a(u, v; \mu) + a(v, u; \mu))$  der symmetrische Anteil. Dann ist für  $\mu \in \mathcal{P}$

$$\langle u, v \rangle_\mu := a_s(u, v; \mu) \quad \text{bzw.} \quad \|u\|_\mu := \sqrt{\langle u, u \rangle_\mu}$$

das *Energie-Skalarprodukt* bzw. die *Energienorm* bzgl.  $\mu$ . Diese ist äquivalent zu  $\|\cdot\|_X$ :

$$\sqrt{\alpha(\mu)}\|u\| \leq \|u\|_\mu \leq \sqrt{\gamma(\mu)}\|u\|$$

*Beweis.* Skalarprodukt: klar wegen Bilinearität, Stetigkeit und Koerzivität. Normäquivalenz folgt aus Stetigkeit und Koerzivität von  $a_s$ .

$$\alpha(\mu)\|u\|^2 \leq \underbrace{a(u, u; \mu)}_{\leq \|u\|^2 \gamma(\mu)} = a_s(u, u; \mu) = \|u\|_\mu^2$$

□

**Satz 2.6** (Übertragung von Koeffizienten-Eigenschaften)

Seien  $f$  bzw.  $a$  separierbar parametrische stetige Linear- bzw. Bilinearform.

- i) Falls  $\Theta_f^q(\mu)$  bzw.  $\Theta_a^q(\mu)$  beschränkt sind, dann sind  $f$  bzw.  $a$  gleichmäßig beschränkt bzgl.  $\mu$ .
- ii) Falls  $\Theta_a^q(\mu)$  strikt positiv, d.h. ex.  $\bar{\Theta}$  mit  $\Theta_a^q(\mu) \geq \bar{\Theta} > 0 \forall \mu \in \mathcal{P}$  alle Komponenten positiv semidefinit, d.h.  $a^q(v, v) \geq 0 \forall v, q$  und  $a(\cdot, \cdot; \bar{\mu})$  ist koerziv für mindestens ein  $\bar{\mu} \in \mathcal{P}$ , dann ist  $a$  gleichmäßig koerziv bzgl.  $\mu$ .

iii) Falls  $\Theta_f^q, \Theta_a^q$  Lipschitz-stetig, so ist  $f, a$  Lipschitz-stetig bzgl.  $\mu$ .

*Beweis.*

i) Sei  $\bar{\Theta}_f^q \in \mathbb{R}^+$  mit  $|\Theta_f^q(\mu)| \leq \bar{\Theta}_f^q \forall \mu$ . Dann gilt

$$\|f(\cdot; \mu)\| = \left\| \sum_q \Theta_f^q(\mu) f^q \right\| \leq \sum_q |\Theta_f^q(\mu)| \|f^q\| \leq \sum_q \bar{\Theta}_f^q \|f^q\| =: \bar{\gamma}_f < \infty$$

analog für  $a(\cdot, \cdot; \mu)$ .

ii) Für  $u \in X, \mu \in \mathcal{P}$  gilt

$$a(u, u; \mu) = \sum_q \Theta_a^q(\mu) a^q(u, u) = \sum_q \underbrace{\frac{\Theta_a^q(\mu)}{\Theta_a^q(\bar{\mu})}}_{>0} \underbrace{\Theta_a^q(\bar{\mu}) a^q(u, u)}_{\sum(\cdot) = a(u, u; \bar{\mu})} \geq \underbrace{\sum_q \frac{\bar{\Theta}}{\max_{q'} \Theta_a^{q'}(\bar{\mu})} \alpha(\bar{\mu})}_{=: \bar{\alpha} > 0} \|u\|^2$$

iii) Sei  $|\Theta_f^q(\mu_1) - \Theta_f^q(\mu_2)| \leq L_f^q |\mu_1 - \mu_2| \forall \mu_1, \mu_2 \in \mathcal{P}$  mit geeignetem  $L_f^q \in \mathbb{R}$ . Dann gilt

$$\begin{aligned} |f(v; \mu_1) - f(v; \mu_2)| &= \left| \sum_q \Theta_f^q(\mu_1) f^q(v) - \sum_q \Theta_f^q(\mu_2) f^q(v) \right| \\ &\leq \sum_q |\Theta_f^q(\mu_1) - \Theta_f^q(\mu_2)| \|f^q\| \|v\| \\ &\leq \underbrace{\sum_q L_f^q \|f^q\|}_{=: L_f} \|\mu_1 - \mu_2\| \|v\| \end{aligned}$$

analog für  $a(\cdot, \cdot; \mu)$ .

□

**Definition 2.7** (Volles Problem  $(P(\mu))$ )

Seien  $a$  bzw.  $f, l$  parametrische Bilinearform bzw. Linearform und gleichmäßig stetig bzgl.  $\mu$ , sei  $a$  gleichmäßig koerziv bzgl.  $\mu$ . Dann ist für  $\mu \in \mathcal{P}$  gesucht  $u(\mu) \in X$  und  $s(\mu) \in \mathbb{R}$  als Lösung von

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

**Bemerkung.**

- Das volle Problem kann also ein analytisches Modell (PDE) oder ein detailliertes Modell (PDE-Diskretisierung) darstellen.
- Symmetrie von  $a$  wird nicht vorausgesetzt.

- In Abschnitt 4, Abschnitt 5 werden Verallgemeinerungen von  $(P(\mu))$  betrachtet.

**Satz 2.8** (Wohlgestelltheit und Stabilität)

Das Problem  $(P(\mu))$  besitzt eine eindeutige Lösung mit

$$\|u(\mu)\| \leq \frac{\|f(\mu)\|_{X'}}{\alpha(\mu)} \leq \frac{\bar{\gamma}_f}{\bar{\alpha}}, \quad |s(\mu)| \leq \|l(\mu)\|_{X'} \|u(\mu)\| \leq \frac{\bar{\gamma}_l \bar{\gamma}_f}{\bar{\alpha}}$$

*Beweis.* Existenz, Eindeutigkeit und Schranke für  $u(\mu)$  folgen mit Lax-Milgram (siehe z.B. Satz 2.5 in Braess'03). Gleichmäßige Stetigkeit und Koerzivität ergeben  $\mu$ -unabhängige Schranke für  $u(\mu)$ . Definition von  $s(\mu)$  ergibt Eindeutigkeit und entsprechende Schranken.  $\square$

**Definition 2.9** (Lösungsmannigfaltigkeit)

Wir definieren

$$\mathcal{M} := \{u(\mu) \in X : \mu \in \mathcal{P} \text{ und } u(\mu) \text{ löst } (P(\mu))\}$$

**Bemerkung.** Wir verwenden den Begriff “Mannigfaltigkeit” nicht im strengen differentialgeometrischen Sinn, weil keine Stetigkeit / Diffbarkeit von  $\mathcal{M}$  gefordert wird.

**Beispiel 2.10** (Thermischer Block)

TODO

**Beispiel 2.11** (Matrixgleichung)

- Zu  $\mu \in \mathcal{P}$  suche  $u(\mu) \in \mathbb{R}^H$  als Lösung von

$$A(\mu)u(\mu) = b(\mu)$$

für  $A(\mu) \in \mathbb{R}^{H \times H}$  und  $b(\mu) \in \mathbb{R}^H$ .

- Dies ist Beispiel für  $(P(\mu))$  via

$$X := \mathbb{R}^H, \quad a(u, v; \mu) := v^\top A(\mu)u, \quad f(v; \mu) := v^\top b(\mu)$$

und beliebiger linearer Ausgabe  $l(v; \mu) := \underline{l}^\top v$  für  $\underline{l} \in \mathbb{R}^H$ .

**Beispiel 2.12** ( $Q_a = 1$ )

Falls  $a(\cdot, \cdot; \mu)$ ,  $f(\cdot; \mu)$  separierbar parametrisch mit  $Q_a = 1$  und  $Q_f$  beliebig, so ist  $\mathcal{M}$  enthalten in einem  $Q_f$ -dimensionalen linearen Teilraum von  $X$ :

$$(P(\mu)) \Rightarrow \Theta_a^1(\mu)a^1(u, v) = \sum_q \Theta_f^q(\mu)f^q(v) \quad \forall v \in X$$

$\Theta_a^1(\mu) \neq 0$  wegen  $a$  gleichmäßig koerziv

$$a^1(u, v) = \sum_q \frac{\Theta_f^q(\mu)}{\Theta_a^1(\mu)} f^q(v) \quad \forall v \in X \quad (*)$$

$a^1(\cdot, \cdot)$  ist koerziv,  $f^q$  linear und stetig

$$\begin{aligned} & \xRightarrow{\text{Lax-Milgram}} \text{ex. } u^q, q = 1, \dots, Q_f \text{ mit } a^1(u^q, v) = f^q(v), \quad v \in X \\ & \Rightarrow u := \sum_q \frac{\Theta_f^q(\mu)}{\Theta_a^1(\mu)} u^q \text{ löst } (*) \text{ wegen Linearität} \\ & \Rightarrow u \in \text{span}\{u^q\}_{q=1}^{Q_f} \end{aligned}$$

**Beispiel 2.13** ( $(P(\mu))$  mit vorgegebener Lösung)

Sei  $u : \mathcal{P} \rightarrow X$  beliebig komplizierte Abbildung. Dann existiert ein  $(P(\mu))$  mit  $u(\mu)$  als Lösung via Skalarprodukten:

$$a(v, w; \mu) := \langle w, v \rangle_X, \quad f(v; \mu) := \langle u(\mu), v \rangle_X$$

d.h. Klasse der Probleme  $(P(\mu))$  können beliebig komplizierte, nichtglatte oder sogar unstetige Lösungsmannigfaltigkeit  $\mathcal{M}$  besitzen.

**Bemerkung** (Parameter-Anzahl und Lösungskomplexität). Es gibt (sogar in der Literatur) ein Missverständnis zwischen Parameteranzahl  $p \in \mathbb{N}$  und Komplexität der Lösungsmannigfaltigkeit  $\mathcal{M}$ , denn es kann Redundanz in Parametern vorliegen (siehe Thermischer Block). Extremfall:  $p \in \mathbb{N}$  beliebig, für geeignetes  $a(\cdot, \cdot; \mu)$ ,  $f(\mu)$  hat  $(P(\mu))$  ein  $\mathcal{M}$ , welches in einem 1D-Raum enthalten ist. (Übung) Beispiel 2.13 zeigt andererseits einen anderen Extremfall: Sogar für  $p = 1$  kann bei geeignetem  $(P(\mu))$  das  $\mathcal{M}$  beliebig kompliziert sein (z.B. "Raumfüllende Kurve"). Unter geeigneter Annahmen an  $a(\cdot, \cdot; \mu)$  und  $f(\cdot; \mu)$  können einfache Regularitätseigenschaften von  $u(\mu)$  bzw.  $\mathcal{M}$  geschlossen werden.

**Korollar 2.14** (Beschränktheit von  $\mathcal{M}$ )

Weil  $a(\cdot, \cdot; \mu)$  gleichmäßig koerziv und  $f(\cdot; \mu)$  gleichmäßig beschränkt, so ist  $\mathcal{M}$  beschränkt

$$\mathcal{M} \subseteq B_{\frac{\bar{\gamma}_f}{\bar{\alpha}}}(0)$$

*Beweis.* Klar weil  $\|u(\mu)\| \leq \frac{\bar{\gamma}_f}{\bar{\alpha}}$  nach Satz 2.8. □

**Satz 2.15** (Lipschitz-Stetigkeit)

Falls  $a(\cdot, \cdot; \mu)$ ,  $f(\cdot; \mu)$ ,  $l(\cdot; \mu)$  Lipschitz-stetig bzgl.  $\mu$ , so sind  $u(\mu)$  und  $s(\mu)$  Lipschitz-stetig bzgl.  $\mu$  mit Lipschitz-Konstanten

$$L_u = \frac{L_f}{\bar{\alpha}} + \bar{\gamma}_f \frac{L_a}{\bar{\alpha}^2} \quad \text{und} \quad L_s = L_l \frac{\bar{\gamma}_f}{\bar{\alpha}} + \bar{\gamma}_l L_u$$

*Beweis.* Übung. □

**Satz 2.16** (Diffbarkeit)

Sei  $a(u, \cdot; \mu) \in X'$  Frechet-diffbar in Umgebung von  $(u_0, \mu_0) \subset X \times \mathcal{P}$  und  $f(\cdot; \mu) \in X'$

Frechet-diffbar in Umgebung von  $\mu_0 \in \mathcal{P}$ . Dann ist Lösung  $u(\mu)$  von  $(P(\mu))$  Frechet-diffbar in Umgebung von  $\mu_0 \in \mathcal{P}$  mit

$$D_\mu u(\mu) := - \left( \frac{\partial}{\partial u} F(u, \mu) \right)^{-1} \frac{\partial}{\partial \mu} F(u, \mu)$$

wobei  $F(u, \mu) := a(u, \cdot; \mu) - f(\cdot; \mu) \in X'$ .

*Beweis.* Aus Frechet-Diffbarkeit von  $a(\cdot, \cdot; \cdot)$  und  $f(\cdot; \cdot)$  folgt Frechet-Diffbarkeit von  $F : X \times \mathcal{P} \rightarrow X'$  in Umgebung von  $(u_0, \mu_0)$  mit partiellen Ableitungen

$$\frac{\partial}{\partial \mu} F(u_0, \mu_0) := \frac{\partial}{\partial \mu} a(u_0, \cdot; \mu_0) - \frac{\partial}{\partial \mu} f(\cdot; \mu_0) \in L(\mathbb{R}^p, X')$$

und  $\frac{\partial}{\partial u} F(u_0, \mu_0) \in L(X, X')$  durch

$$\frac{\partial}{\partial u} F(u_0, \mu_0) h_u := a(h_u, \cdot; \mu_0) \in X' \quad \forall h_u \in X$$

Dann erfüllt  $u(\mu)$  als Lösung von  $(P(\mu))$  gerade

$$F(u(\mu), \mu) = 0$$

in Umgebung von  $\mu_0$ . Dann ist (z.B. mit Folgerung 2.15 in Ruzicka: Nichtlineare Funktionalanalysis, Springer 2004) auch  $u(\mu)$  Frechet-diffbar in Umgebung von  $\mu_0$  mit Ableitung

$$D_\mu u(\mu) := - \left( \frac{\partial}{\partial u} F(u, \mu) \right)^{-1} \frac{\partial}{\partial \mu} F(u, \mu)$$

□

### Bemerkung.

- Plausibilität der Ableitungsformel folgt aus formellem Ableiten:

$$\begin{aligned} & D_\mu (F(u(\mu), \mu)) = 0 \\ \Rightarrow & \frac{\partial}{\partial u} F(u(\mu), \mu) D_\mu u(\mu) + \frac{\partial}{\partial \mu} F(u, \mu) = 0 \\ \Rightarrow & \frac{\partial}{\partial u} F(u(\mu), \mu) D_\mu u(\mu) = - \frac{\partial}{\partial \mu} F(u, \mu) \\ \Rightarrow & D_\mu u(\mu) = - \left( \frac{\partial}{\partial u} F(u(\mu), \mu) \right)^{-1} \frac{\partial}{\partial \mu} F(u, \mu) \end{aligned}$$

- Man kann zeigen, dass die Sensitivitäts-Ableitungen  $\partial_{\mu_i} u(\mu) \in X$  für  $i = 1, \dots, p$  erfüllen das sogenannte *Sensitivitätsproblem*

$$a(\partial_{\mu_i} u(\mu), v; \mu) = \tilde{f}_i(v; u(\mu), \mu)$$

mit rechter Seite  $\tilde{f}_i(\cdot; u(\mu), \mu) \in X'$  gegeben durch

$$\tilde{f}_i(\cdot; w, \mu) := \partial_{\mu_i} f(\cdot; \mu) - \partial_{\mu_i} a(w, \cdot; \mu)$$

d.h. das Problem  $(P(\mu))$  mit modifizierter rechter Seite, in welcher insbesondere  $u(\mu)$  eingeht. (Übung)

- Hinreichend für die Diffbarkeit von  $a, f$  in Satz 2.16 sind z.B. im Fall von separierbarer Parameterabhängigkeit die Diffbarkeit der Koeffizienten  $\Theta_a^q(\mu), \Theta_f^q(\mu)$ ,  $q = 1, \dots$  (Übung)
- Ähnliche Aussagen / Sensitivitätsprobleme gelten für Ableitungen höherer Ordnung. Also überträgt sich Glattheit der Koeffizientenfunktionen auf Glattheit der Lösung / Mannigfaltigkeit.

### 3 RB-Methoden für lineare koerzive Probleme

#### 3.1 Primales RB-Problem

**Definition 3.1** (Reduzierte Basis, RB-Räume)

Sei  $S_N = \{\mu_1, \dots, \mu_N\} \subset \mathcal{P}$  Menge von Parametern mit (o.B.d.A.) linear unabhängigen Lösungen  $\{u(\mu_i)\}_{i=1}^N$  von  $(P(\mu_i))$ . Dann ist  $X_N := \text{span} \{u(\mu_i)\}_{i=1}^N$  ein sog. *Lagrange-RB-Raum*.

Sei  $\mu^0 \in \mathcal{P}$  und  $u(\mu)$  Lösung von  $(P(\mu^0))$   $k$ -mal diffbar in Umgebung von  $\mu^0$ . Dann ist

$$X_{k,\mu^0} := \text{span} \{ \partial_\sigma u(\mu^0) : \sigma \in \mathbb{N}_0^p, |\sigma| \leq k \}$$

ein *Taylor-RB-Raum*. Eine Basis  $\Phi_N = \{\varphi_1, \dots, \varphi_N\} \subseteq X$  eines RB-Raums ist eine *reduzierte Basis*.

**Bemerkung.**

- $\Phi_N$  kann direkt aus Snapshots  $u(\mu^i)$  oder, für numerische Stabilität (siehe ??), auch orthonormiert sein.
- Wahl der Parameter  $\{\mu^i\}$  ist entscheidend für Güte des RB-Modells:  
Hier: zufällige oder äquidistante Menge ausreichend  
Später: intelligente Wahl durch a-priori Analysis oder Greedy-Verfahren
- Es ex. auch andere Arten von RB-Räumen (Hermite, POD). Gemeinsam ist diesen die Konstruktion aus Snapshots von  $u$  bzw.  $\partial_\sigma u$ .
- Andere MOR-Techniken:  $\Phi_N$  kann auch komplett unabhängig von Snapshots auf andere Weise konstruiert werden: Balanced Truncation, Krylov-Räume, etc. (siehe z.B. Antoulas: Approximation of large scale dynamical systems, SIAM 2004)

**Definition 3.2** (Reduziertes Problem  $(P_N(\mu))$ )

Sei eine Instanz von  $(P(\mu))$  gegeben und  $X_N \subseteq X$  ein RB-Raum. Zu  $\mu \in \mathcal{P}$  ist die RB-Lösung  $u_N(\mu) \in X_N$  und Ausgabe  $s_N(\mu) \in \mathbb{R}$  gesucht mit

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

**Bemerkung.**

- Wir nennen obiges “primal” weil im Fall  $f \neq l$  oder  $a$  asymmetrisch, kann mit Hilfe eines geeigneten dualen Problems bessere Schätzung für  $s$  erreicht werden.
- Obiges ist “Ritz-Galerkin”-Projektion im Gegensatz zu “Petrov-Galerkin”-Projektion, welches für nicht-koerzive Probleme notwendig ist.  $\rightsquigarrow$  4

**Satz 3.3** (Galerkin-Projektion, Galerkin-Orthogonalität)

Sei  $P_\mu : X \rightarrow X_N$  die orthogonale Projektion bzgl. Energieskalarprodukt  $\langle \cdot, \cdot \rangle_\mu$ , sei  $a$  symmetrisch und  $u(\mu)$ ,  $u_N(\mu)$  Lösung von  $(P(\mu))$  bzw.  $(P_N(\mu))$ . Dann:



- i)  $u_N(\mu) = P_\mu u(\mu)$  “Galerkin-Projektion”  
 ii)  $\langle e(\mu), v \rangle_\mu = 0 \quad \forall v \in X_N$ , wobei  $e(\mu) := u(\mu) - u_N(\mu)$

*Beweis.* Nach Aufgabe 1/Blatt 1 ist  $P_\mu$  wohldefiniert, denn  $(X, \langle \cdot, \cdot \rangle_\mu)$  ist Hilbertraum und  $X_N \subseteq X$  abgeschlossen weil endlichdimensional. Orthogonale Projektion des Fehlers ergibt

$$\begin{aligned} & \langle P_\mu u(\mu) - u(\mu), \varphi_i \rangle_\mu = 0 & \forall i = 1, \dots, N \\ \Leftrightarrow & a(P_\mu u(\mu) - u(\mu), \varphi_i; \mu) = 0 & \forall i = 1, \dots, N \\ \Leftrightarrow & a(P_\mu u(\mu), \varphi_i; \mu) = a(u(\mu), \varphi_i; \mu) = f(\varphi_i; \mu) & \forall i = 1, \dots, N \end{aligned}$$

- i) also ist  $P_\mu u(\mu)$  Lösung von  $(P_N(\mu))$   
 ii)  $e(\mu)$  ist also Projektions-Fehler, orthogonal nach Aufgabe 1/Blatt 1

□

**Bemerkung.** Für  $a$  nichtsymmetrisch gilt immer noch folgende “Galerkin-Orthogonalität”

$$a(u - u_N, v; \mu) = 0 \quad \forall v \in X_N$$

(auch wenn  $a$  kein Skalarprodukt)

**Satz 3.4** (Existenz und Eideutigkeit für  $(P_N(\mu))$ )

Zu  $\mu \in \mathcal{P}$  ex. eindeutige Lösung  $u_N(\mu) \in X_N$  und RB-Ausgabe  $s_n(\mu) \in \mathbb{R}$  von  $(P_N(\mu))$ . Diese sind beschränkt

$$\begin{aligned} \|u_N(\mu)\| &\leq \frac{\|f(\cdot; \mu)\|_{X'}}{\alpha(\mu)} \leq \frac{\bar{\gamma}_f}{\bar{\alpha}} \\ \|s_N(\mu)\| &\leq \|l(\cdot; \mu)\| \|u_N(\mu)\| \leq \frac{\bar{\gamma}_l \bar{\gamma}_f}{\bar{\alpha}} \end{aligned}$$

*Beweis.* Weil  $X_N \subset X$  ist  $a(\cdot, \cdot; \mu)$  stetig und koerziv auf  $X_N$ .

$$\begin{aligned} \alpha_N(\mu) &:= \inf_{v \in X_N} \frac{a(v, v; \mu)}{\|v\|^2} \geq \inf_{v \in X} \frac{a(v, v; \mu)}{\|v\|^2} = \alpha(\mu) > 0 \\ \gamma_N(\mu) &:= \sup_{u, v \in X_N} \frac{a(u, v; \mu)}{\|u\| \|v\|} \leq \sup_{u, v \in X} \frac{a(u, v; \mu)}{\|u\| \|v\|} = \gamma(\mu) < \infty \end{aligned}$$

analog  $f, l$  stetig auf  $X_N$ . Existenz, Eindeutigkeit und Schranken folgen also mit Lax-Milgram analog zu 2.8. □

**Korollar 3.5** (Lipschitz-Stetigkeit)

Seien  $f, l$  gleichmäßig beschränkt und  $a, f, l$  Lipschitz-stetig bzgl.  $\mu$ , dann sind auch  $u_N(\mu), s_N(\mu)$  Lipschitz-stetig bzgl.  $\mu$  mit  $L_u, L_s$  wie in 2.15.

*Beweis.* Analog zu 2.15 / Übung. □

**Satz 3.6** (Diskrete RB-Probleme)

Sei  $\Phi_N = \{\varphi_1, \dots, \varphi_N\}$  eine reduzierte Basis für  $X_N$ . Für  $\mu \in \mathcal{P}$ ,

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

und  $\underline{u}_N = (u_{N,i})_{i=1}^N \in \mathbb{R}^N$  als Lösung von

$$A_N(\mu) \underline{u}_N = \underline{f}_N(\mu) \quad (3.1)$$

Dann ist  $u_N(\mu) := \sum_{i=1}^N u_{N,i} \varphi_i$  und  $s_N(\mu) := \underline{l}_N^\top(\mu) \underline{u}_N$ .

*Beweis.* Einsetzen und Linearität zeigt, dass

$$a\left(\sum u_{N,j} \varphi_j, \varphi_i; \mu\right) = (A_N(\mu) \underline{u}_N)_i = (\underline{f}_N)_i = f(\varphi_i; \mu)$$

□

**Satz 3.7** (Kondition bei ONB und Symmetrie)

Falls  $a(\cdot, \cdot; \mu)$  symmetrisch und  $\Phi_N$  ist ONB, so ist Kondition von (3.1) unabhängig von  $N$  beschränkt

$$\text{cond}_2(A_N) := \|A_N\|_2 \|A_N^{-1}\|_2 \leq \frac{\gamma(\mu)}{\alpha(\mu)}$$

*Beweis.* Wegen Symmetrie gilt

$$\text{cond}_2(A_N) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|} \quad (3.2)$$

mit betragsmäßig größtem/kleinstem Eigenwert  $\lambda_{\max}/\lambda_{\min}$  von  $A_N(\mu)$ . Sei  $\underline{u}_{\max} = (u_i)_{i=1}^N \in \mathbb{R}^N$  Eigenvektor zu  $\lambda_{\max}$  und

$$\underline{u}_{\max} := \sum_{i=1}^N u_i \varphi_i \in X_N$$

Dann gilt

$$\begin{aligned} \lambda_{\max} \|\underline{u}_{\max}\|^2 &= \lambda_{\max} \underline{u}_{\max}^\top \underline{u}_{\max} = \underline{u}_{\max}^\top A_N \underline{u}_{\max} \\ &= \sum_{i,j=1}^N u_i u_j a(\varphi_j, \varphi_i; \mu) = a\left(\sum_j u_j \varphi_j, \sum_i u_i \varphi_i; \mu\right) \\ &= a(\underline{u}_{\max}, \underline{u}_{\max}; \mu) \leq \gamma(\mu) \|\underline{u}_{\max}\|^2 \end{aligned}$$

Wegen

$$\|\underline{u}_{\max}\|^2 = \left\langle \sum u_i \varphi_i, \sum u_j \varphi_j \right\rangle = \sum u_i u_j \langle \varphi_i, \varphi_j \rangle = \sum u_i^2 = \|\underline{u}_{\max}\|^2$$

folgt  $|\lambda_{\max}| \leq \gamma(\mu)$ . Analog zeigt man  $|\lambda_{\min}| \geq \alpha(\mu)$  also folgt mit (3.2) die Behauptung. □

**Bemerkung** (Unterschied FEM zu RB). Es bezeichne  $A_h(\mu) \in \mathbb{R}^{H \times H}$  die FEM Matrix (oder FV/FD).

- i) Die RB-Matrix  $A_N(\mu) \in \mathbb{R}^{H \times H}$  ist klein aber typischerweise vollbesetzt im Gegensatz zur großen aber dünnbesetzten Matrix  $A_h$ .
- ii) Die Kondition von  $A_N$  verschlechtert sich nicht mit wachsendem  $N$  (falls eine ONB verwendet wird), während die Konditionszahl von  $A_h$  typischerweise polynomiell in  $H$  wächst, also schlechter wird.

**Satz 3.8** (Reproduktion von Lösungen)

Seien  $u(\mu)$ ,  $u_N(\mu)$  Lösungen von  $(P(\mu))$  bzw.  $(P_N(\mu))$ ,  $e_i \in \mathbb{R}^n$   $i$ -ter Einheitsvektor

- i) Falls  $u(\mu) \in X_N \Rightarrow u_N(\mu) = u(\mu)$
- ii) Falls  $u(\mu) = \varphi_i \in \Phi_N \Rightarrow u_N(\mu) = e_i \in \mathbb{R}^N$

*Beweis.*

- i) Mit  $u(\mu), u_N(\mu) \in X_N \Rightarrow e := u(\mu) - u_N(\mu) \in X_N$ . Wegen Galerkin-Orthogonalität ( $a(e, v; \mu) = 0 \forall v \in X_N$ ) und Koerzivität folgt:

$$0 = a(e, e; \mu) \geq \underbrace{\alpha(\mu)}_{>0} \underbrace{\|e\|^2}_{\geq 0} \Rightarrow \|e\| = 0 \Rightarrow e = 0 \Rightarrow u = u_N$$

- ii)  $u_N(\mu) = \varphi_i$ , nach i). Mit Eindeutigkeit der Basisexpansion folgt die Behauptung. □

**Bemerkung.**

- Reproduktion von Lösungen ist grundlegende Konsistenzeigenschaft. Es gilt trivialerweise falls/sobald Fehlerschranken vorliegen, aber für komplexe RB-Probleme ohne Fehlerschranken ist obiges ein guter Test.
- Validierung für Programmcode: Wähle Basis aus Snapshots  $\varphi_i = u(\mu^i)$ ,  $i = 1, \dots, N$ , ohne Orthonormierung, dann muss  $u_N(\mu^i) = e_i \in \mathbb{R}^N$  ein Einheitsvektor sein.

## 3.2 Fehleranalyse

**Satz 3.9** (Céa, Beziehung zur Bestapproximation)

Für alle  $\mu \in \mathcal{P}$  gilt

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

*Beweis.*  $\forall v \in X_N$  mit Stetigkeit und Koerzivitat

$$\begin{aligned} \alpha \|u - u_N\|^2 &\leq a(u - u_N, u - u_N) = a(u - u_N, u - v) + \underbrace{a(u - u_N, v - u_N)}_{=0 \text{ (Galerkin-Orth.)}} \\ &\leq \gamma(\mu) \|u - u_N\| \|u - v\| \end{aligned}$$

Division durch  $\alpha$ ,  $\|u - u_N\|$  liefert

$$\|u - u_N\| \leq \frac{\gamma}{\alpha} \|u - v\|$$

also Behauptung durch Infimum-Bildung.  $\square$

**Bemerkung.**

- i) hnliche Bestapproximationsaussagen gelten auch fur andere Interpolationstechniken, aber die zugehorige Lebesgue-Konstante divergiert meist mit wachsender Dimension  $N$ . Obiges ist konzeptioneller Vorteil von Galerkin-Projektion uber anderen Interpolationstechniken, da  $\frac{\gamma}{\alpha}$  unabhangig von  $N$  beschrankt bleibt. “Quasi-Optimalitat” der Galerkin-Projektion/des RB-Ansatzes.
- ii) Falls  $a(\cdot, \cdot; \mu)$  zusatzlich symmetrisch ist, kann um eine “Wurzel” verbessert werden mittels Normaquivalenz 2.5 und Bestapproximation der orthogonalen Projektion (Aufg. 1/Blatt 1)

$$\begin{aligned} \sqrt{\alpha} \|u - u_N\| &\stackrel{2.5}{\leq} \|u - u_N\|_\mu = \|u - P_\mu u\|_\mu = \inf_{v \in X_N} \|u - v\|_\mu \stackrel{2.5}{\leq} \sqrt{\gamma} \inf_{v \in X_N} \|u - v\| \\ \Rightarrow \|u - u_N\| &\leq \sqrt{\frac{\gamma}{\alpha}} \inf_{v \in X_N} \|u - v\| \end{aligned}$$

- iii) Implikation von 3.9: Wahle guten Approximationsraum  $X_N$ , so wird Galerkin-Projektion/RB-Approximation auch garantiert gut sein.

**Satz 3.10** (Ausgabe und Bestapproximation)

- i) Fur alle  $\mu \in \mathcal{P}$  gilt

$$|s(\mu) - s_N(\mu)| \leq \|l(\cdot; \mu)\|_{X'} \frac{\gamma(\mu)}{\alpha(\mu)} \inf_{v \in X_N} \|u - v\|$$

- ii) Fur den sog. “compliant” Fall (d.h.  $a(\cdot, \cdot; \mu)$  symmetrisch und  $l = f$ ) gilt sogar

$$\begin{aligned} 0 \leq s(\mu) - s_N(\mu) &= \|u - u_N\|_\mu^2 \\ &= \inf_{v \in X_N} \|u - v\|_\mu^2 \\ &\leq \gamma(\mu) \inf_{v \in X_N} \|u - v\|^2 \end{aligned}$$

*Beweis.*

i) Klar mit Céa, Bestapproximation und Linearität

$$|s(\mu) - s_N(\mu)| = |l(u) - l(u_N)| = |l(u - u_N)| \leq \|l\| \|u - u_N\| \leq \|l\| \frac{\gamma}{\alpha} \inf_{v \in X_N} \|u - v\|$$

ii) Wegen  $a(\cdot, \cdot; \mu)$  symmetrisch gilt wie in voriger Bemerkung

$$\|u - u_N\|_\mu = \|u - P_\mu u\|_\mu = \inf_{v \in X_N} \|u - v\| \quad (3.3)$$

Damit

$$\begin{aligned} s(\mu) - s_N(\mu) &= l(u) - l(u_N) \stackrel{f=l}{=} f(u) - f(u_N) = f(u - u_N) \\ &= a(u, u - u_N) - \underbrace{a(u_N, u - u_N)}_{=0 \text{ (Gal.-Orth./Symm.)}} = \|u - u_N\|_\mu^2 \\ &\stackrel{3.3}{=} \inf_{v \in X_N} \|u - v\|_\mu^2 \\ &\stackrel{2.5}{\leq} \gamma \inf_{v \in X_N} \|u - v\|^2 \end{aligned}$$

Also insbesondere  $s - s_N = \|u - u_N\|_\mu^2 \geq 0$ .

□

**Bemerkung.**

- Im “compliant” Fall ist der Ausgabefehler i.A. sehr klein, da das Quadrat des RB-Fehlers eingeht.
- Im “nicht-compliant” Fall geht der RB-Fehler nur linear in die Schranke ein, das wird später durch primal-duale Technik verbessert.
- Aus ii) folgt nicht nur Fehlerschranke, sondern sogar Vorzeichen-Information,  $s_N(\mu)$  ist untere Schranke für  $s$ .

**Korollar 3.11** (Monotoner Fehlerabfall in Energienorm)

Falls  $a(\cdot, \cdot; \mu)$  symmetrisch,  $(X_N)_{N=1}^{N_{\max}}$  Folge von RB-Räumen, mit  $X_N \subseteq X_{N'}, \forall N \leq N'$  (“hierarchische Räume”) und für  $\mu \in \mathcal{P}$  setze  $e_{u,N} := u(\mu) - u_N(\mu)$ ,  $e_{s,N} := s(\mu) - s_N(\mu)$ .

i) Dann ist  $(\|e_{u,N}\|_\mu)_{N=1}^{N_{\max}}$  monoton fallend.

ii) Falls  $l = f$  (also “compliant” Fall) ist  $e_{s,N}$  monoton fallend.

*Beweis.*

i) Mit (3.3) gilt für  $N \leq N'$

$$\|e_{u,N}\|_\mu = \|u - u_N\|_\mu \stackrel{(3.3)}{=} \inf_{v \in X_N} \|u - v\|_\mu \geq \inf_{v \in X_{N'}} \|u - v\|_\mu \stackrel{(3.3)}{=} \|e_{u,N'}\|_\mu$$

ii) Mit Satz 3.10 ii) gilt

$$e_{s,N} = \|e_{u,N}\|_\mu^2, \text{ also Behauptung folgt mit i)}$$

□

**Bemerkung.**

- “Worst-case” ist Stagnation des Fehlers (unrealistisch, jeder neue Basisvektor müsste orthogonal zum Fehler  $e_N(\mu)$  sein). In Praxis ist bei geschickter Basiswahl und “glatten” Problemen exponentielle Konvergenz zu erwarten, siehe Basisgenerierung, §3.4.
- Monotonie gilt nicht notwendigerweise bezüglich anderen Normen trotz Normäquivalenz

$$c\|e_{u,N}\|_\mu \leq \|e_{u,N}\| \leq C\|e_{u,N}\|_\mu, \text{ mit } c, C \text{ unabhängig von } N$$

Fehlernorm  $\|e_{u,N}\|$  kann gelegentlich anwachsen, bleibt aber in einem “Korridor”, welcher monoton fällt.

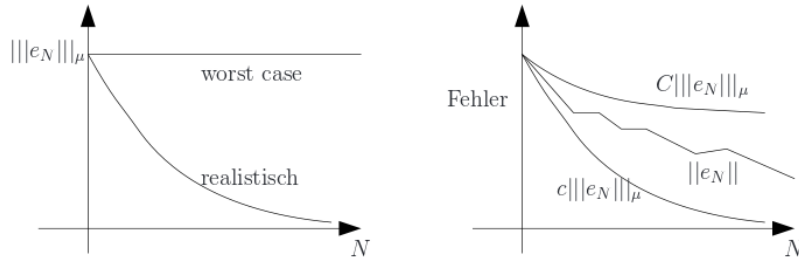


Abbildung 5: Fehlerabfall mit wachsender reduzierter Dimension.  
(aus B. Haasdonk, Reduzierte-Basis-Methoden, Skript zur Vorlesung SS 2011, Universität Stuttgart, IANS-Report 4/11, 2011.)

**Bemerkung** (Gleichmäßige Konvergenz von Lagrange RB-Ansatz).

- Sei  $\mathcal{P}$  kompakt und  $S_N := \{\mu^1, \dots, \mu^N\} \subset \mathcal{P}$ ,  $N \in \mathbb{N}$ , sodass die sog. Füll-Distanz (fill-distance)  $h_N$  gegen 0 geht:

$$h_N := \sup_{\mu \in \mathcal{P}} \text{dist}(\mu, S_N), \quad \text{dist}(\mu, S_N) := \min_{\mu' \in S_N} \|\mu - \mu'\|$$

$$\lim_{N \rightarrow \infty} h_N = 0$$

- Falls  $u(\mu)$ ,  $u_N(\mu)$  Lipschitz-stetig mit Lipschitz-Konstante  $L_u$  unabhängig von  $N$ , so folgt für alle  $N$ ,  $\mu$  und “nächstes”  $\mu^* := \arg \min_{\mu' \in S_N} \|\mu - \mu'\|$ :

$$\begin{aligned} \|u(\mu) - u_N(\mu)\| &\leq \|u(\mu) - u(\mu^*)\| + \|u(\mu^*) - u_N(\mu^*)\| + \|u_N(\mu^*) - u_N(\mu)\| \\ &\leq L_u \underbrace{\|\mu - \mu^*\|}_{\leq h_N} + 0 + L_u \underbrace{\|\mu - \mu^*\|}_{\leq h_N} \leq 2L_u h_N \end{aligned}$$

- Also folgt uniforme Konvergenz

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

- Jedoch Konvergenzrate linear in  $h_N$  ist nicht praktisch bedeutsam, weil  $h_N$  sehr langsam mit  $N$  abfällt, also muss  $N$  sehr groß sein, um kleinen Fehler zu garantieren.
- Wir werden sehen, dass bei gleichmäßig koerziven Problemen und geschickter Wahl der  $\mu^i$  sogar exponentielle Konvergenz erreicht wird.

**Lemma 3.12** (Fehler-Residuums-Beziehung)

Für  $\mu \in \mathcal{P}$  definieren wir mittels der RB-Lösung  $u_N$  das Residuum  $r(\cdot; \mu) \in X'$  bzw. seinen Riesz-Repräsentanten  $v_r(\mu) \in X$

$$\langle v_r(\mu), v \rangle_X := r(v; \mu) := f(v; \mu) - a(u_N(\mu), v; \mu) \quad \forall v \in X$$

Dann erfüllt der Fehler  $e(\mu) := u(\mu) - u_N(\mu)$

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

*Beweis.*  $a(e(\mu), v; \mu) = \underbrace{a(u, v)}_{f(v)} - a(u_N, v) = r(v)$

□

**Bemerkung.**

- Fehler erfüllt “ $(P(\mu))$  mit Residuum als rechte Seite”
- Insbesondere ist  $r(v; \mu) = 0 \quad \forall v \in X_N$  (wegen Galerkin-Orthogonalität)
- $r(\cdot; \mu) = 0 \quad \Rightarrow \quad e = 0$

**Satz 3.13** (A-posteriori Fehlerschätzer, absoluter Fehler)

Sei  $\mu \in \mathcal{P}$ ,  $u(\mu)$  bzw.  $u_N(\mu)$  Lösung von  $(P(\mu))$ ,  $(P_N(\mu))$  und  $e = u - u_N$ . Sei  $\alpha_{LB}(\mu)$  eine untere Schranke für  $\alpha(\mu)$  und  $v_r \in X$  Riesz-Repräsentant von  $r(\cdot; \mu)$  aus Lemma 3.12. Dann gelten folgende Schranken

i) Fehler in Energienorm

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

ii) Fehler in  $X$ -Norm  $\|\cdot\|$

$$\|e(\mu)\| \leq \Delta_N(\mu) := \frac{\|v_r\|}{\alpha_{LB}(\mu)}$$

iii) Ausgabefehler

$$|s(\mu) - s_N(\mu)| \leq \Delta_{N,s}(\mu) := \|l(\cdot; \mu)\| \Delta_N(\mu)$$

*Beweis.*

i) Normäquivalenz 2.5 impliziert

$$\|e\| \leq \frac{\|e\|_\mu}{\sqrt{\alpha_{LB}(\mu)}}$$

Damit folgt

$$\|e\|_\mu^2 = a_s(e, e) = a(e, e) = r(e) = \langle v_r, e \rangle \leq \|v_r\| \|e\| \leq \frac{\|v_r\|}{\sqrt{\alpha_{LB}(\mu)}} \|e\|_\mu$$

Division durch  $\|e\|_\mu$  liefert Behauptung i).

ii) Koerzivität liefert

$$\alpha_{LB}(\mu) \|e\|^2 \leq a(e, e) = r(e) = \langle v_r, e \rangle \leq \|v_r\| \|e\|$$

Division durch  $\alpha_{LB}$  und  $\|e\|$  liefert ii).

iii) Stetigkeit von  $l$  liefert

$$|s(\mu) - s_N(\mu)| = |l(u - u_N; \mu)| \leq \|l\| \|u - u_N\| \stackrel{ii)}{\leq} \|l\| \Delta_N$$

□

### Bemerkung.

- $\alpha_{LB}(\mu)$  soll eine *schnell berechenbare* untere Schranke an  $\alpha(\mu)$  sein, z.B.  $\alpha_{LB}(\mu) := \bar{\alpha}$  falls  $\bar{\alpha}$  bekannt, andere Möglichkeiten folgen später (“min  $\Theta$ ”, “SCM”).
- $\Delta_N$  ist also immer um Faktor  $\sqrt{\alpha_{LB}(\mu)}$  schlechter.
- Beschränkung des Fehlers durch Residuums-Norm ist bekannte Technik aus FEM, um FEM-Lösung  $u_h$  gegen Sobolev-Raum Lösung  $u$  abzuschätzen. In diesem Fall ist  $X$   $\infty$ -dimensional und Residuums-Norm algorithmisch nicht berechenbar. In RB-Methoden wird  $\|v_r\|$  eine *berechenbare* Größe sobald  $X$  endlich-dimensional, z.B. FEM-Raum, ist. Für Residuum ist  $u_N(\mu)$  erforderlich, daher sind Schranken “*a posteriori*”.
- Allgemeines Vorgehen (und alternative Begründung für ii)) zur Herleitung von Fehlerschranken: Zeige, dass Fehler  $e$  erfüllt  $(P(\mu))$  mit rechter Seite, genannt  $r$  (Residuum), wende a-priori Stabilitätsaussage an:

$$\|e\| \leq \frac{\|r\|}{\alpha(\mu)} \quad \text{z.B. Lax-Milgram}$$

und erhalte berechenbare Größe durch Wahl  $X = X_{FEM}$  und untere Schranke  $\alpha_{LB}(\mu) \leq \alpha(\mu)$ .



- Weil die Schranken beweisbare obere Schranken an Fehler darstellen, nennt man sie “rigorose” Fehlerschranken (vgl. “zuverlässige” Schätzer in FEM, bei denen jedoch die Konstante unbekannt ist).
- Fehlerschranken liefern eine Absicherung für RB-Methoden, “certified” RB-Methode, im Gegensatz zu vielen anderen Reduktionsmethoden (z.B. Krylov-Raum-Methoden).
- Ausgabefehler ist grob, indem  $\Delta_N$  nur linear eingeht. Verbesserungen können für den “compliant” Fall oder mit primal-dual Techniken erreicht werden. ( $\rightsquigarrow$  §3.5)

**Korollar 3.14** (Verschwindende Fehlerschranke)

Falls  $u(\mu) = u_N(\mu)$  dann ist  $\Delta_N(\mu) = \Delta_N^{en}(\mu) = \Delta_{N,s}(\mu) = 0$

*Beweis.*

$$\begin{aligned} 0 &= a(0, v; \mu) = a(e, v; \mu) = r(v; \mu) \\ \Rightarrow r &\equiv 0 \Rightarrow \|v_r\| = 0 \Rightarrow \Delta_N = \Delta_N^{en} = \Delta_{N,s} = 0 \end{aligned}$$

□

**Bemerkung.**

- Dies ist initialer Wunsch an eine Fehlerschranke: diese soll verschwinden falls exakte Approximation vorliegt. Dies ist Grundlage dafür, dass der Faktor der Überschätzung endlich ist.
- Aussage ist trivial für *effektive* Fehlerschätzer (sehen wir bald), aber in komplexen Problemen kann 3.14 schon das maximal erreichbare sein.
- 3.14 ist wieder sinnvoll um Programmcode zu validieren.

**Satz 3.15** (A-posteriori Fehlerschranken, relative Fehler)

Mit Bezeichnungen/Voraussetzungen aus 3.13 und unter Annahme, dass alle Brüche im Folgenden wohldefiniert sind, gilt:

i) Für den relativen Fehler gilt in Energienorm:

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

ii) Für den relativen Fehler gilt in  $X$ -Norm:

$$\frac{\|e(\mu)\|}{\|u(\mu)\|} \leq \Delta_N^{rel}(\mu) := 2 \frac{\|v_r\|}{\alpha_{LB}(\mu)} \cdot \frac{1}{\|u_N(\mu)\|} \quad \text{falls} \quad \Delta_N^{rel} \leq 1$$

*Beweis.*

i) Falls  $\Delta_N^{en,rel}(\mu) \leq 1$ , so ist

$$\begin{aligned} \left| \frac{\|u\|_\mu - \|u_N\|_\mu}{\|u_N\|_\mu} \right| &\stackrel{\Delta\text{-Ungl.}}{\leq} \frac{\|u - u_N\|_\mu}{\|u_N\|_\mu} = \frac{\|e\|_\mu}{\|u_N\|_\mu} \stackrel{3.13 \text{ i)}}{\leq} \frac{\|v_r\|}{\sqrt{\alpha_{LB}(\mu)} \|u_N\|_\mu} \\ &= \frac{1}{2} \Delta_N^{en,rel}(\mu) \leq \frac{1}{2} \end{aligned}$$

Falls  $\|u_N\|_\mu > \|u\|_\mu$  gilt  $\|u_N\|_\mu - \|u\|_\mu \leq \frac{1}{2} \|u_N\|_\mu$  also

$$\frac{1}{2} \|u_N\|_\mu \leq \|u\|_\mu \quad (*)$$

Falls  $\|u\|_\mu \geq \|u_N\|_\mu$ , so ist  $(*)$  klar. Damit folgt

$$\frac{\|e\|_\mu}{\|u\|_\mu} \stackrel{3.13 \text{ i)}}{\leq} \frac{\|v_r\|}{\sqrt{\alpha_{LB}}} \cdot \frac{1}{\|u\|_\mu} \stackrel{(*)}{\leq} \frac{\|v_r\|}{\sqrt{\alpha_{LB}}} \cdot \frac{1}{\|u_N\|_\mu} \cdot 2 = \Delta_N^{en,rel}(\mu)$$

ii) analog zu i).

□

### Bemerkung.

- Analog folgt auch relativer Ausgabefehlerschätzer

$$\frac{|s(\mu) - s_N(\mu)|}{|s(\mu)|} \leq \Delta_{N,s}^{rel}(\mu) := \frac{\|l(\cdot; \mu)\| \cdot \Delta_N}{|s_N(\mu)|} \cdot 2 \quad \text{falls} \quad \Delta_{N,s}^{rel}(\mu) \leq 1$$

- Relative Fehlerschranken sind nur mit Zusatzbedingung  $(\Delta_*^{rel} \leq 1)$  gültig. Diese Bedingung ist jedoch konkret überprüfbar. Falls  $\Delta_N^{rel}(\mu) > 1$ , sollte der RB-Raum verbessert werden.

### Satz 3.16 (Effektivität der Fehlerschranken)

Mit Bezeichnungen aus 3.13 sei  $u(\mu) \neq u_N(\mu)$  und  $\gamma_{UB}(\mu) < \infty$  eine obere Schranke an  $\gamma(\mu)$ . Dann sind die *Effektivitäten*  $\eta_N^{en}(\mu)$  und  $\eta_N(\mu)$  definiert und beschränkt durch

i)

$$\eta_N^{en}(\mu) := \frac{\Delta_N^{en}(\mu)}{\|e\|_\mu} \leq \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}$$

Falls  $a(\cdot, \cdot; \mu)$  symmetrisch, gilt sogar  $\eta_N^{en}(\mu) \leq \sqrt{\frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}}$

ii)

$$\eta_N(\mu) := \frac{\Delta_N(\mu)}{\|e\|_\mu} \leq \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}$$

*Beweis.*

$$\begin{aligned} \text{ii) } \|v_r\|^2 &= \langle v_r, v_r \rangle = r(v_r) = a(e, v_r) \leq \gamma_{UB}(\mu) \|e\| \|v_r\| \\ \|v_r\| &\leq \gamma_{UB}(\mu) \|e\| \end{aligned} \quad (3.4)$$

Damit

$$\frac{\Delta_N(\mu)}{\|e\|} = \frac{\|v_r\|}{\alpha_{LB}} \cdot \frac{1}{\|e\|} \stackrel{(3.4)}{\leq} \frac{\gamma_{UB}}{\alpha_{LB}} \cdot \frac{\|e\|}{\|e\|}$$

i)

$$\frac{\Delta_N^{en}(\mu)}{\|e\|_\mu} = \frac{\|v_r\|}{\sqrt{\alpha_{LB}}} \cdot \frac{1}{\underbrace{\|e\|_\mu}_{\geq \sqrt{\alpha_{LB}} \cdot \|e\|}} \leq \frac{\|v_r\|}{\alpha_{LB}} \cdot \frac{1}{\|e\|} \stackrel{\text{ii)}}{\leq} \frac{\gamma_{UB}}{\alpha_{LB}}$$

Falls  $a(\cdot, \cdot)$  symmetrisch, gilt wegen Normäquivalenz

$$\|v_r\|_\mu \leq \sqrt{\gamma_{UB}} \|v_r\|$$

und

$$\|v_r\|^2 = a(e, v_r) \stackrel{\text{CS}}{\leq} \|e\|_\mu \|v_r\|_\mu \Rightarrow \|v_r\| \leq \|e\|_\mu \cdot \sqrt{\gamma_{UB}}$$

Damit

$$\frac{\Delta_N^{en}(\mu)}{\|e\|_\mu} = \frac{\|v_r\|}{\sqrt{\alpha_{LB}}} \cdot \frac{1}{\|e\|_\mu} \leq \frac{\|e\|_\mu \cdot \sqrt{\gamma_{UB}}}{\sqrt{\alpha_{LB}} \cdot \|e\|_\mu}$$

□

### Bemerkung.

- Wir nennen  $\Delta_N, \Delta_N^{en}$  daher “effektive” Fehlerschranken weil Faktor der Überschätzung höchstens  $\frac{\gamma_{UB}}{\alpha_{LB}}$  beträgt.
- “Rigorousität” also äquivalent mit  $\eta_N(\mu) \geq 1$ .
- Für den Ausgabefehler  $\Delta_{N,s}(\mu)$  ohne weitere Annahmen keine Effektivität beweisbar. Tatsächlich kann  $\frac{\Delta_{N,s}}{|s-s_N|}$  beliebig groß oder nicht definiert sein, falls  $\Delta_{N,s} \neq 0$ , aber  $s(\mu) = s_N(\mu)$ :

Wähle  $X_N$  und  $\mu$  so dass  $u(\mu) \neq u_N(\mu)$ , wird erreicht durch  $u(\mu) \notin X_N$

$$\Rightarrow e(\mu) \neq 0 \Rightarrow \Delta_N \neq 0, \Delta_{N,s} \neq 0 \quad \text{falls } l \neq 0$$

Wähle  $l(\cdot; \mu) \neq 0$ , so dass  $l(u - u_N; \mu) = 0$

$$\Rightarrow s(\mu) - s_N(\mu) = l(u - u_N; \mu) = 0$$

- Wir nennen die Fehlerschranken auch *Fehlerschätzer* weil sie äquivalent zum Fehler sind.

$$\|e\| \leq \Delta_N \leq \eta_N \|e\|$$

**Satz 3.17** (Effektivität, relative Fehlerschätzer)

Für  $\Delta_N^{rel}(\mu)$  aus 3.15 ist Effektivität definiert und beschränkt durch

$$\eta_N^{rel}(\mu) := \frac{\Delta_N^{rel}(\mu)}{\frac{\|e\|}{\|u\|}} \leq 3 \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)} \quad \text{falls} \quad \Delta_N^{rel}(\mu) \leq 1$$

*Beweis.* Wie in Beweis zu 3.15 impliziert  $\Delta_N^{rel} \leq 1$ :

$$\left| \frac{\|u\| - \|u_N\|}{\|u\|} \right| \leq \frac{1}{2}$$

Falls  $\|u_N\| \leq \|u\|$  so gilt  $\|u\| - \|u_N\| \leq \frac{1}{2}\|u_N\|$  also

$$\|u\| \leq \frac{3}{2}\|u_N\|$$

Falls  $\|u_N\| > \|u\|$ , so ist (\*) klar. Dann gilt

$$\eta_N^{rel}(\mu) = \underbrace{\frac{2\|v_r\|}{\alpha_{LB}(\mu)\|u_N\|}}_{\Delta_N^{rel}} \cdot \frac{1}{\frac{\|e\|}{\|u\|}} \stackrel{(3.4)}{\leq} 2 \frac{\gamma_{UB}\|e\|}{\alpha_{LB}\|e\|} \cdot \frac{\|u\|}{\|u_N\|} \stackrel{(*)}{\leq} 3 \frac{\gamma_{UB}}{\alpha_{LB}}$$

□

**Bemerkung.**

- Ähnlich für  $\Delta_N^{en,rel}$
- Verbesserung von Schranken und Effektivität durch Normwechsel.

Wähle  $\bar{\mu} \in \mathcal{P}$  und  $\|u\| := \|u\|_{\bar{\mu}}$  als neue Norm auf  $X$ . Dann gilt für symmetrisches  $a$ :  $\alpha(\bar{\mu}) = 1 = \gamma(\bar{\mu})$  also Effektivitäten  $\eta_N, \eta_N^{en} = 1$ , Schätzer sind genau der echte Fehler. Dies lässt  $u_N$  unberührt, liefert aber bessere Fehlerschätzung. Im Fall von Stetigkeit bzgl.  $\mu$  kann auch in Umgebung von  $\bar{\mu}$  gute Effektivität erwartet werden.

**Satz 3.18** (Ausgabefehlerschranke und Effektivität, compliant Fall)

Sei  $a(\cdot, \cdot; \mu)$  symmetrisch,  $l = f$ . Dann erhalte verbesserte Ausgabeschranke

$$0 \leq s(\mu) - s_N(\mu) \leq \bar{\Delta}_{N,s}(\mu) := \frac{\|v_r\|^2}{\alpha_{LB}}$$

und Effektivität

$$\bar{\eta}_{N,s}(\mu) := \frac{\bar{\Delta}_{N,s}(\mu)}{s(\mu) - s_N(\mu)} \leq \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}$$

*Beweis.* Nach Satz 3.10 ii) und 3.13 gilt

$$0 \stackrel{3.10}{\leq} s(\mu) - s_N(\mu) = \|u - u_N\|_{\mu}^2 = \|e\|_{\mu}^2 \stackrel{3.13}{\leq} \Delta_N^{en}(\mu)^2 = \bar{\Delta}_{N,s}(\mu)$$

Für Effektivität gilt entsprechend mit 3.16 i)

$$\bar{\eta}_{N,s}(\mu) = \frac{\bar{\Delta}_{N,s}}{s(\mu) - s_N(\mu)} \stackrel{3.10}{=} \frac{\Delta_N^{en}(\mu)^2}{\|u - u_N\|_\mu^2} = \eta_N^{en}(\mu)^2 \stackrel{3.16}{=} \sqrt{\frac{\gamma_{UB}}{\alpha_{LB}}}^2 = \frac{\gamma_{UB}}{\alpha_{LB}}$$

□

**Bemerkung.** Analog kann man im compliant Fall eine relative Ausgabefehlerschranke und Effektivität beweisen.

$$\frac{s(\mu) - s_N(\mu)}{s(\mu)} \leq \bar{\Delta}_{N,s}^{rel}(\mu) := \frac{\|v_r\|^2}{\alpha_{LB} s_N(\mu)}$$

und

$$\bar{\eta}_{N,s}^{rel}(\mu) := \frac{\bar{\Delta}_{N,s}^{rel}}{\frac{s(\mu) - s_N(\mu)}{s(\mu)}} \leq 2 \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)}$$

falls  $\bar{\Delta}_{N,s}^{rel}(\mu) \leq 1$ .

**Bemerkung** (Zusammenfassende Relevanz der Fehlerschätzer).

- Rigorose obere Schranke für tatsächlichen Fehler nicht nur “Indikatoren” wie bei FEM.
- Effektivität Faktor der Überschätzung des Fehlers ist klein und bleibt beschränkt. Insbesondere:

$$e(\mu) = 0 \Rightarrow \Delta_N(\mu) = 0$$

also “a-posteriori” exakte Approximation verifizierbar.

- Theoretische Untermauerung der i.A. empirischen Basiswahl.
- Unabhängig von Basiswahl sind Fehlerschätzer anwendbar, auch für nicht-Snapshot-Basen (z.B. Krylov-Unterräume, etc.).
- Effiziente Berechnung: Durch Offline-Online-Zerlegung ( $\rightsquigarrow$  §3.3) ist neben reduzierter Simulation auch Fehlerschranken & Effektivitätsschranken schnell berechenbar.
- Weitere Einsatzmöglichkeiten: Offline zur Basisgenerierung ( $\rightsquigarrow$  §3.4) und Online zur adaptiven Dimensionswahl.

## Numerische Beispiele

**demos\_chapter3(1)** Thermischer Block aus Beispiel 2.10,  $B_1 = B_2 = 2$ ;  $N = 5$ ,  $\langle \cdot, \cdot \rangle_X := \langle \cdot, \cdot \rangle_{H_0^1}$ ,

$$S_N = \{0.1, 0.5, 0.9, 1.4, 1.7\} \times \{0.1\}^3 \subseteq \mathbb{R}^4$$

Erkenntnisse:

- Fehlerschätzer kann günstig für sehr feines Parametergitter berechnet werden, Fehler ist teuer zu berechnen, daher nur in wenigen Punkten.
- Fehler und Schätzer sind 0 für Basisparameter (bestätigt 3.8, 3.14).
- Fehlerschätzer ist obere Schranke für Fehler gemäß 3.13.
- Für kleine Werte von  $\mu_1$  größere Fehler  $\Rightarrow$  gute Wahl von  $S_N$  wird vermutlich (und später bewiesen) hier mehr Samples benötigen.

**demos\_\_chapter3(2)** Effektivitäten  $\eta_N(\mu)$  und obere Schranke  $\frac{\gamma}{\alpha} \leq \frac{\mu_{max}}{\mu_{min}}$ .  
Erkenntnisse:

- Effektivitäten sind gut, nur etwa Faktor 10 über Fehler.
- Obere Schranke für Effektivität gemäß 3.16.
- Effektivitäten sind undefiniert für Parametersamples  $\mu \in S_N$  (Division durch Null).

**demos\_\_chapter3(3)** Fehlerkonvergenz bezüglich  $N$ .

$$B_1 = B_2 = 3, \quad \mu_1 \in [0.5, 2], \quad \mu = (\mu_1, 1, \dots, 1) \in \mathbb{R}^9$$

Lagrange-Basis mit Gram-Schmidt-Orthonormierung,  $\{\mu_i\}_{i=1}^N$  äquidistant. Erkenntnisse für Testfehler: (Maximierung über 100 zufällige Parameter)

$$S_{test} \subset \mathcal{P}, \quad |S_{test}| = 100$$

- Exponentielle Konvergenz für Fehler und Schätzer.
- Obere Schranke sehr gut.
- Numerische Ungenauigkeiten für Schätzer.

### 3.3 Offline/Online-Zerlegung

Bisher:

- $(P_N(\mu))$  niedrigdimensional, aber noch keine schnelle Berechnungsvorschrift.
- Um “berechenbares” Verfahren zu erhalten: Forderung  $\dim X < \infty$  in diesem Kapitel.
- Für effiziente Berechnung ist separierbare Parameterabhängigkeit von  $(P(\mu))$  essenziell.

Offline-Phase:

- Typischerweise berechnungsintensiv, Komplexität polynomiell in  $H := \dim X$

- Einmal durchgeführt.
- Berechnung *hochdimensionaler* Daten: Snapshots, reduzierte Basis, Riesz-Repräsentanten. (“detailed\_data” in RBmatlab)
- Projektion der hochdimensionalen Daten in *parameterunabhängigen niedrigdimensionalen* Daten. (“reduced\_data”)

Online-Phase:

- Schnelle Berechnung, Komplexität polynomiell in  $N$ ,  $Q_a$ ,  $Q_f$ ,  $Q_l$ , *unabhängig von  $H$* .
- Typischerweise häufig ausgeführt für variierendes  $\mu$ .
- Assemblierung des reduzierten parametrischen Systems für  $(P_N(\mu))$ .
- Lösen von  $(P_N(\mu))$ .
- Berechnung von Fehlerschranken und Effektivität.

### Komplexitätsbetrachtung der bisherigen Formulierung

- Mit  $\dim X = H$  und dünnbesetzter Matrix für  $(P(\mu))$  ist Lösung z.B. in  $\mathcal{O}(H^2)$  erreichbar (z.B.  $H$  Schritte eines iterativen Löser mit  $\mathcal{O}(H)$  Komplexität für Matrix-Vektor-Multiplikation dank Dünnbesetztheit).
- $N \times N$  System für  $(P_N(\mu))$  ist vollbesetzt, also in  $\mathcal{O}(N^3)$  lösbar, also  $N \ll H$  erforderlich, um Gewinn zu bewirken.
- Genaue Betrachtung der Berechnung von  $u_N(\mu)$ :
  1.  $N$  Snapshots berechnen mittels  $(P(\mu))$ :  $\mathcal{O}(N \cdot H^2)$
  2.  $N^2$  Auswertungen von  $a(\varphi_i, \varphi_j; \mu)$ :  $\mathcal{O}(N^2 \cdot H)$
  3.  $N$  Auswertungen von  $f(\varphi_i; \mu)$ :  $\mathcal{O}(N \cdot H)$
  4. Lösen des  $N \times N$  Systems für  $(P_N(\mu))$ :  $\mathcal{O}(N^3)$
- Wir haben noch keine Offline/Online-Zerlegung: 1. gehört zur Offline-Phase, 4. gehört zur Online-Phase, aber 2. und 3. können nicht in Offline-Phase berechnet werden (wegen Parameterabhängigkeit) und nicht in Online-Phase (wegen  $H$ -Abhängigkeit).  
→ Zerlegung von 2. und 3. mittels separierbarer Parameterabhängigkeit

**Definition 3.19** (Notation für Zerlegung von  $(P(\mu))$ )

Unter Annahme  $H = \dim X < \infty$ ,  $X = \text{span} \{\psi_i\}_{i=1}^H$ , definiere Matrix

$$K := (\langle \psi_i, \psi_j \rangle)_{i,j=1}^H \in \mathbb{R}^{H \times H} \quad \text{“Gram’sche Matrix” / “Skalarprodukt-Matrix”}$$

Mit separierbare Parameterabhängigkeit definiere Matrizen und Vektoren

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

**Korollar 3.20** (Lösung von  $(P(\mu))$ )

Lösung von  $(P(\mu))$  wird erhalten durch Assemblieren des vollen Systems

$$A(\mu) = \sum_{q=1}^{Q_a} \Theta_a^q(\mu) \cdot A^q, \quad \underline{f}(\mu) = \sum_{q=1}^{Q_f} \Theta_f^q(\mu) \underline{f}^q, \quad \underline{l}(\mu) = \sum_{q=1}^{Q_l} \Theta_l^q(\mu) \underline{l}^q$$

und Lösen von  $A(\mu) \underline{u}(\mu) = \underline{f}(\mu)$  nach  $\underline{u}(\mu) = (u_i)_{i=1}^H \in \mathbb{R}^H$  und

$$u(\mu) = \sum_{i=1}^H u_i \varphi_i \in X, \quad s(\mu) = \underline{l}^T(\mu) \cdot \underline{u}(\mu)$$

*Beweis.* Klar mit Definitionen. □

**Bemerkung.**

- Das Vorliegen der  $A^q, \underline{f}^q, \underline{l}^q$  ist nicht trivial im Fall von “fremden” Diskretisierungspaketen und stellt wesentliche Schwierigkeit in breiter praktischer Anwendung dar. Motivation für Eigenentwicklung von Diskretisierungscode.
- Sinn von Matrix  $K$  ist Berechnung von Skalarprodukten und Normen, z.B. für

$$\begin{aligned} u &= \sum u_i \psi_i, \quad v = \sum v_i \psi_i \in X \quad \text{für} \quad \underline{u} = (u_i), \underline{v} = (v_i)_{i=1}^H \in \mathbb{R}^H \\ &\Rightarrow \langle u, v \rangle_X = \sum_{i,j} u_i v_j \langle \psi_i, \psi_j \rangle = \underline{u}^T K \underline{v} \end{aligned}$$

**Korollar 3.21** (Offline-/Online- Zerlegung für  $(P_N(\mu))$ )

(Offline:) Nach Konstruktion einer Basis  $\Phi_N = \{\varphi_1, \dots, \varphi_N\}$  berechne parameter-unabhängige Komponenten-Matrizen & Vektoren

$$A_N^q := (a^q(\varphi_j, \varphi_i))_{i,j=1}^N \in \mathbb{R}^{N \times N}, \quad q = 1, \dots, Q_n$$

$$\underline{f}_N^q := (f^q(\varphi_i))_{i=1}^N, \quad \underline{l}_N^q := (l^q(\varphi_i))_{i=1}^N \in \mathbb{R}^N, \quad q = 1, \dots, Q_f/Q_l$$

(Online:) Zu  $\mu \in \mathcal{P}$  berechne Koeffizienten  $\Theta_a^q(\mu), \Theta_f^q(\mu), \Theta_l^q(\mu)$  und

$$A_N(\mu) := \sum_q \Theta_a^q(\mu) A_N^q$$

$$\underline{f}_N(\mu) := \sum_q \Theta_f^q(\mu) \underline{f}_N^q, \quad \underline{l}_N(\mu) := \sum_q \Theta_l^q(\mu) \underline{l}_N^q$$

Dies liefert genau das diskrete System  $A_N(\mu) \underline{u}_N = \underline{f}_N(\mu)$  aus 3.6 welches nach  $\underline{u}_N$  gelöst wird und  $u_N(\mu), s_N(\mu)$  ergibt



*Beweis.* klar wg. Separierbarkeit □

**Bemerkung** (Einfache Berechnung von  $A_N^q, f_N^q, l_N^q$ ). Die reduzierten Komponenten benötigen keinerlei Integration über  $\Omega$  oder Gitterdurchlauf, falls hochdim.  $A^q$  vorliegen. Sei Basis  $\Phi_N$  gegeben durch Koeffizientenmatrix

$$\Phi_N := (\varphi_{ji})_{i=1, j=1}^{H, N} \in \mathbb{R}^{N \times N} \quad \text{mit} \quad \varphi_j = \sum_{i=1}^H \varphi_{ji} \psi_i$$

Dann erhalte reduzierten Komponenten durch Matrix-Multi

$$A_N^q := \Phi_N^T A^q; f_N^q := \Phi_N^T f^q; l_N^q := \Phi_N^T l^q$$

**Bemerkung.**

- Offline-Phase benötigt  $\mathcal{O}(NH^2 + NH(Q_f + Q_l) + N^2 H Q_a)$  für die Berechnung von  $\Phi_N, f_N^q, l_N^q, A_N^q$  dominiert von der Basisgenerierung.
- Online-Phase skaliert mit  $\mathcal{O}(N^2 Q_a + N(Q_f + Q_l) + N^3)$  für Berechnung von  $A_N(\mu), f_N(\mu), l_N(\mu)$  und  $\underline{u}_N(\mu)$  dominiert durch LGS lösen falls  $Q_a, Q_f, Q_l$  klein sind. Insbesondere komplett unabhängig von  $H$ , wie gewünscht.
- Laufzeitdiagramm Seien  $t_{detail}, t_{offline}, t_{online}$ , die Laufzeiten für einzelne Lösungen von  $(P(\mu))$ , Offline-Phase bzw. Online-Phase von  $(P_N(\mu))$ . Unter Annahme, dass diese konstant unter Parametervariation, erhalte affin-lineare Beziehung der Gesamtlaufzeit für  $k$  parameterische Lösungen

$$t(k) := k \cdot t_{detail}, \quad t_N(k) = t_{offline} + k \cdot t_{online}$$

Das reduzierte Modell zahlt sich aus, sobald mehr als  $k^* := \frac{t_{offline}}{t_{detail} - t_{online}}$  Lösungen berechnet werden sollen.

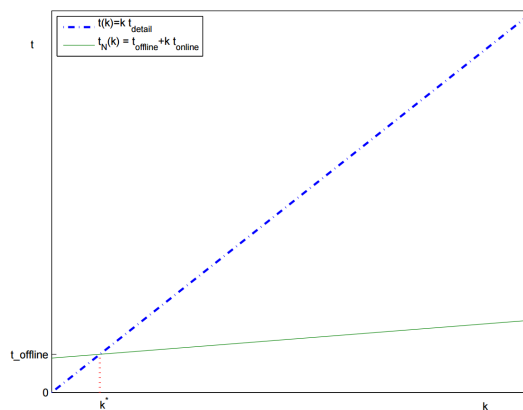


Abbildung 6: Laufzeiten mit wachsender Anzahl an Simulationen.  
(aus B. Haasdonk, Reduzierte-Basis-Methoden, Skript zur Vorlesung SS 2011, Universität Stuttgart, IANS-Report 4/11, 2011.)

**Bemerkung** (Keine Unterscheidung zwischen  $u$  und  $u_h$ ). Erinnerung: Wir unterscheiden (meistens) nicht in Notation zwischen  $u_h$  (FEM-Lösung) und  $u$  (Sobolev-Raum Lösung). Dies kann nun begründet werden:

- i) Die Online-Phase ist unabhängig von  $H = \dim(X)$ , daher kann  $H$  beliebig groß und damit  $u_h$  beliebig präzise gemacht werden durch geeignete Diskretisierung mit genügend feinem Gitter, so dass  $u$  und  $u_h$  praktisch ununterscheidbar sind ( $\|u - u_h\|$  beliebig klein aber  $(P_N(\mu))$  schnell lösbar).
- ii) In der Praxis wird Reduktionsfehler den Gesamtfehler dominieren, der (FEM-)Diskretisierungsfehler spielt untergeordnete Rolle.

$$\epsilon := \|u - u_h\| \ll \|u_h - u_N\|$$

$$\Rightarrow \|u_h - u_N\| - \epsilon \leq \underbrace{\|u - u_N\|}_{\text{theoretisch das Ideal}} \leq \overbrace{\|u_h - u_N\|}^{\text{berechenbar}} + \epsilon$$

also kontrollieren wir durch Fehlerschranken für  $\|u_h - u_N\|$  bis auf  $\epsilon$  auch den eigentlich interessanten Fehler  $\|u - u_N\|$ .

### Offline-/Online- Zerlegung für Fehlerschranken/Effektivitätsschranken

Für schnelle Berechnung der Fehlerschranken & Effektivitätsschranken benötigen wir Zerlegung für

- Duale Norm des Residuums  $\|r(\cdot; \mu)\|_{X'} = \|v_r\|$  für alle Fehlerschranken
- Duale Norm des Ausgabefunktional  $\|l(\cdot; \mu)\|_{X'}$  für  $\Delta_{N,s}(\mu)$
- Norm  $\|u_N(\mu)\|_X$  der RB-Lösung für relativen Energienormfehlerschätzer  $\Delta_N^{en,rel}$ .
- Untere/obere Schranke  $\alpha_{LB}(\mu)$  bzw.  $\gamma_{UB}(\mu)$  für Koerzivitäts- bzw. Stetigkeitskonstante für Fehlerschätzer bzw. Effektivitätsschranken.

Separierbarkeit von  $(P(\mu))$  überträgt sich auf Residuum

**Satz 3.22** (Separierbare Parameter-Abhängigkeit für  $r(\cdot; \mu)$ )

Seien  $a, f$  sep. parametrisch. Nach Riesz existieren  $v_f^q \in X$  mit  $\langle v_f^q, v \rangle = f^q(v) \forall v \in X, q = 1, \dots, Q_f$  und  $v_a^{q,n} \in X$  mit  $\langle v_a^{q,n}, v \rangle = a^q(\varphi_n, v), v \in X, q = 1, \dots, Q_a, n = 1, \dots, N$ . Setze  $Q_r := NQ_a + Q_f$  und Aufzählung von  $\{v_a^{q,n}, v_f\}$  durch

$$(v_r^1, \dots, v_r^{Q_r}) := (v_f^1, \dots, v_f^{Q_f}, v_a^{1,1}, \dots, v_a^{Q_a,1}, v_a^{1,2}, \dots, v_a^{Q_a,2}, \dots, v_a^{Q_a,N})$$

Für  $\mu \in \mathcal{P}$  sei  $u_N = \sum_{n=1}^N u_{Nn} \varphi_n$  Lösung von  $(P_N(\mu))$  und hiermit definiere

$$(\Theta_r^1(\mu), \dots, \Theta_r^{Q_r}(\mu)) := (\Theta_f^1(\mu), \dots, \Theta_f^{Q_f}(\mu), -\Theta_a^1(\mu), \dots, -\Theta_a^{Q_a}(\mu)u_{N1}, -\Theta_a^1(\mu)u_{N2}, \dots, -\Theta_a^{Q_a}(\mu)u_{N2}, \dots, -\Theta_a^{Q_a}(\mu)u_{NN})$$

Mit  $r^q(\cdot) := \langle v_r^q, \cdot \rangle \in X'$ ,  $q = 1, \dots, Q_r$  sind  $r(\cdot; \mu)$  und  $v_r(\mu)$  separierbar parametrisch via

$$r(\cdot; \mu) = \sum_{q=1}^{Q_r} \Theta_r^q(\mu) \cdot r^q(\cdot), \quad v_r(\mu) = \sum_{q=1}^{Q_r} \Theta_r^q(\mu) \cdot v_r^q \quad \forall \mu \in \mathcal{P}$$

*Beweis.* Definition und Linearität ergibt:

$$\begin{aligned} \langle v_r(\mu), v \rangle &= r(v; \mu) = f(v; \mu) - a(u_N(\mu), v; \mu) \\ &= \sum_q \Theta_f^q(\mu) f^q(v) - \sum_q \sum_n \Theta_a^q(\mu) u_{Nn} a^q(\varphi_n, v) \\ &= \underbrace{\left\langle \sum_q \Theta_f^q(\mu) v_f^q - \sum_q \sum_n \Theta_a^q(\mu) u_{Nn} v_a^q, v \right\rangle}_{\sum \Theta_r^q(\mu) v_r^q} \\ &= \sum_{q=1}^{Q_r} \Theta_r^q(\mu) r^q(v) \quad \forall v \in X \end{aligned}$$

□

Offensichtlich Berechnung von Riesz-Repräsentant notwendig, dies geschieht durch Ausnutzen der Endlichdim. von  $X = \text{span}\{\psi_i\}_{i=1}^H$  und  $K := (\langle \psi_i, \psi_j \rangle)_{i,j=1}^H$

**Satz 3.23** (Berechnung von Riesz-Repr.)

Für  $g \in X'$  erhält man Koeffizientenvektor  $\underline{v} = (v_i)_{i=1}^H \in \mathbb{R}^H$  seines Riesz-Repräsentanten  $v_g = \sum_{i=1}^H v_i \psi_i \in X$  durch lösen von

$$K \underline{v} = \underline{g} \quad (3.5)$$

mit Vektor  $\underline{g} := (g(\psi_i))_{i=1}^H \in \mathbb{R}^H$

*Beweis.* Für jedes  $u = \sum_{i=1}^H u_i \psi_i \in X$  mit Koeffizientenvektor  $\underline{u} = (u_i)_{i=1}^H$  erhalten wir

$$g(u) = g\left(\sum u_i \psi_i\right) = \sum u_i g(\psi_i) = \underline{u}^T \underline{g} \stackrel{3.5}{=} \underline{u}^T K \underline{v} = \langle u, v_g \rangle$$

□

**Bemerkung.** 3.5 ist typischerweise dünn besetzt, also mit iterativen LGS-Lösern berechenbar.

**Korollar 3.24** (Offline-/Online- für Residuen-Norm)

(Offline:) Nach Offline von  $(P_N(\mu))$  gemäß 3.21 def.  $G_r := (r^q(v_r^{q'}))_{q,q'=1}^{Q_r} \in \mathbb{R}^{Q_r \times Q_r}$  mittels Residuen-Komponenten  $r^q$  und Riesz-Repr.  $v_r^q$  aus 3.22 (Online:) Für  $\mu \in \mathcal{P}$  und RB-Lösung  $\underline{u}_N \in \mathbb{R}^N$  berechne Residuen-Koeff-Vektor  $\underline{\Theta}_r(\mu) = (\Theta_r^1(\mu), \dots, \Theta_r^{Q_r}(\mu))^T \in \mathbb{R}^{Q_r}$ . Dann gilt:

$$\|v_r(\mu)\|_X = \|r(\cdot; \mu)\|_X = \sqrt{\underline{\Theta}_r(\mu)^T \cdot G_r \underline{\Theta}_r(\mu)} \quad (3.6)$$

*Beweis.* Zunächst sehen wir  $G_r = (\langle v_r^q, v_r^{q'} \rangle)_{q,q'=1}^{Q_r}$ . Isometrie der Riesz-Abbildung & Separierbarkeit ergeben

$$\|r(\mu)\|_X^2 = \|v_r(\mu)\|_X^2 = \left\langle \sum_{q=1}^{Q_r} \Theta_r^q(\mu) v_r^q, \sum_{q'=1}^{Q_r} \Theta_r^{q'}(\mu) v_r^{q'} \right\rangle = \underline{\Theta}_r^T \cdot G_r \cdot \underline{\Theta}_r(\mu)$$

□

**Bemerkung** (Stabilisierung durch Orthonormierung von  $\{v_r^q\}$ ). Wie in `demos_chapter3(3)` gesehen, existiert eine Genauigkeitsgrenze für Fehlerschätzer, diese liegt in numerischen Auslöschungseffekten in 3.6 begründet, denn  $G_r$  ist potentiell schlecht konditioniert. Gemäß einer Idee von Behr & Rave 2014 lässt sich die Genauigkeit steigern, indem die  $\{v_r^q\}$  orthonormiert werden und 3.6 mit entsprechender Transformationsmatrix modifiziert werden.

**Korollar 3.25** (Offline-/Online- Zerlegung für  $\|l(\cdot; \mu)\|_{X'}$ )  
(Offline:) Berechne Riesz-Repr.  $v_l^q \in X$  der Ausgabekomponenten, d. h.

$$\langle v_l^q, v \rangle = l^q(v) \quad \forall v \in X, q = 1, \dots, Q_l$$

und def.  $G_l := (l^q(v_l^{q'}))_{q,q'=1}^{Q_l}$  (Online:) Zu  $\mu \in \mathcal{P}$  berechne  $\underline{\Theta}_l(\mu) := (\Theta_l^1(\mu), \dots, \Theta_l^{Q_l}(\mu))$   
und  $\|l(\cdot; \mu)\|_{X'} = \sqrt{\underline{\Theta}_l^T G_l \underline{\Theta}_l}$

*Beweis.* analog zu 3.24

□

**Korollar 3.26** (Offline-/Online für  $\|u_N(\mu)\|_X, \|u_N(\mu)\|_\mu$ )  
(Offline:) Nach der Offline-Phase von  $(P_N(\mu))$  def.

$$K_N := (\langle \varphi_i, \varphi_j \rangle)_{i,j=1}^N \in \mathbb{R}^{N \times N}$$

(Online:) Zu  $\mu \in \mathcal{P}$  berechne  $A_N(\mu)$  und  $\underline{u}_N(\mu)$  durch Online-Phase von  $(P_N(\mu))$

$$\|u_N(\mu)\|_X = \sqrt{\underline{u}_N^T K_N \underline{u}_N}$$

$$\|u_N(\mu)\|_\mu = \sqrt{\underline{u}_N^T \left( \frac{1}{2} (A_N(\mu) + A_N(\mu)^T) \right) \underline{u}_N}$$

*Beweis.*

$$\|u_N\|^2 = \left\langle \sum_n u_{Nn} \varphi_n, \sum_{n'} u_{Nn'} \varphi_{n'} \right\rangle = \sum_{n,n'} u_{Nn} u_{Nn'} \langle \varphi_n, \varphi_{n'} \rangle = \underline{u}_N^T \cdot K_N \cdot \underline{u}_N$$

analog für Energienorm mit  $A_{N,s} := \frac{1}{2}(A_N(\mu) + A_N(\mu)^T)$

□

**Bemerkung.**  $K_N$  wieder einfach aus  $K$  berechenbar (Übung).

Für Fehlerschranken fehlen noch untere Schranke  $\alpha_{LB}(\mu) \leq \alpha(\mu)$ , welche schnell berechenbar sein sollen. Falls  $a(\cdot, \cdot; \mu)$  glm. koerziv bzgl.  $\mu$  und  $\bar{\alpha} < 0$  bekannt, so ist  $\alpha_{LB}(\mu) := \bar{\alpha}$  gültige Wahlmöglichkeit. In gewissen Fällen kann eine größere und damit bessere Schranke angegeben werden.

**Satz 3.27** (“Min- $\Theta$ -Verfahren” zur Berechnung von  $\alpha_{LB}(\mu)$ )

Seien  $a^q(u, u) \geq 0 \ \forall q, u$  und  $\Theta_a^q(\mu) > 0 \ \forall \mu$

(Offline:) Sei  $\alpha(\bar{\mu})$  für ein  $\bar{\mu} \in \mathcal{P}$  verfügbar

(Online:) Setze für  $\mu \in \mathcal{P}$

$$\alpha_{LB}(\mu) := \alpha(\bar{\mu}) \cdot \min_q \frac{\Theta_a^q(\mu)}{\Theta_a^q(\bar{\mu})}$$

Dann gilt  $0 < \alpha_{LB}(\mu) \leq \alpha(\mu)$

*Beweis.* Wegen  $0 < \alpha(\bar{\mu})$  und  $0 < c(\mu) := \min_q \frac{\Theta_a^q(\mu)}{\Theta_a^q(\bar{\mu})}$  gilt  $0 < \alpha(\bar{\mu}) \cdot c(\mu) := \alpha_{LB}(\mu)$

Folgende Argumentation ähnlich zu 2.6 ii)

Für alle  $u \in X$  gilt

$$\begin{aligned} a(u, u; \mu) &= \sum_q \Theta_a^q(\mu) a^q(u, u) = \sum_q \frac{\Theta_a^q(\mu)}{\Theta_a^q(\bar{\mu})} \cdot \Theta_a^q(\bar{\mu}) a^q(u, u) \\ &\geq \sum_q \underbrace{\left( \min_{q'} \frac{\Theta_a^{q'}(\mu)}{\Theta_a^{q'}(\bar{\mu})} \right)}_{c(\mu)} \cdot \Theta_a^q(\bar{\mu}) a^q(u, u) \\ &= c(\mu) \cdot \underbrace{\sum_q \Theta_a^q(\bar{\mu}) a^q(u, u)}_{=a(u, u; \bar{\mu})} = c(\mu) a(u, u; \bar{\mu}) \\ &\stackrel{\text{glm. koerziv bzgl. } \mu}{\geq} c(\mu) \cdot \alpha(\bar{\mu}) \cdot \|u\|^2 \\ &= \alpha_{LB}(\mu) \cdot \|u\|^2 \end{aligned}$$

Also insbesondere

$$\alpha(\mu) = \inf_u \frac{a(u, u; \mu)}{\|u\|^2} \geq \alpha_{LB}(\mu)$$

□

**Bemerkung.**

- “Min- $\Theta$ ” kann für Thermischen Block angewandt werden
- obiges gilt auch für nichtsymm.  $a(\cdot, \cdot)$
- $\alpha(\bar{\mu})$  kann mittels eines hochdimensionalen Eigenwertproblems bestimmt werden:

**Satz 3.28** (Berechnung von  $\alpha(\mu)$  für  $(P(\mu))$ )

Seien  $A(\mu)$ ,  $K \in \mathbb{R}^{H \times H}$  wie in 3.19/3.20.

Setze  $A_s(\mu) := \frac{1}{2}(A(\mu) + A(\mu)^T)$ . Dann gilt

$$\alpha(\mu) = \lambda_{\min}(K^{-1}A_s(\mu))$$

wobei  $\lambda_{\min}$  den kleinsten Eigenwert bezeichnet.

*Beweis.* Sie  $K = LL^T$  (z. B. Cholesky oder Matrix-Wurzel) und verwende  $\underline{v} = L^T \underline{u}$ :

$$\begin{aligned} \alpha(\mu) &= \inf_{u \in X} \frac{a(u, u; \mu)}{\|u\|^2} = \inf_{\underline{u} \in \mathbb{R}^H} \frac{\underline{u}^T A(\mu) \underline{u}}{\underline{u}^T K \underline{u}} \\ &= \inf_{\underline{u} \in \mathbb{R}^H} \frac{\underline{u}^T A_s(\mu) \underline{u}}{\underline{u}^T K \underline{u}} \\ &= \inf_{\underline{v} \in \mathbb{R}^H} \frac{\underline{v}^T L^{-1} A_s \overbrace{L^{-T}}^{\text{inv. transp. } -1 \cdot T} \underline{v}}{\underline{v}^T L^{-1} L L^T L^{-T} \underline{v}} = \inf_{\underline{v} \in \mathbb{R}^H} \frac{\underline{v}^T L^{-1} A_s L^{-T} \underline{v}}{\underline{v}^T \underline{v}} \end{aligned}$$

Also ist  $\alpha(\mu)$  Minimum eines Rayleigh-Quotienten, also kleinster Eigenwert der symmetrischen & positiv definiten Matrix  $\bar{A}_s := L^{-1} A_s L^{-T}$

Die Matrizen  $\bar{A}_s$  und  $K^{-1} A_s$  sind ähnlich, da

$$L^T (K^{-1} A_s) L^{-T} = L^T L^{-T} L^{-1} A_s L^{-T} = L^{-1} A_s L^{-T} = \bar{A}_s$$

Also haben sie identische Eigenwerte. □

**Bemerkung.**

- Inversion von  $K$  muss verhindert werden. Daher verwende EW-Löser, welcher nur Matrix-Vektor-Multiplikation verwendet. Sobald ein Produkt  $y = K^{-1} A_s x$  erforderlich ist, löst man das System  $Ky = A_s x$ . Alternativ kann auch kleinster EW eines verallgemeinerten EWP  $A_s \underline{u} = \lambda K \underline{u}$  berechnet werden.
- Für variationelle Form des verallg. EWP für  $\infty$ -dim  $(P(\mu))$  siehe Patera & Rozza
- Für Probleme, bei denen die Voraussetzungen von Min- $\Theta$  nicht erfüllt sind, kann "Successive Constraint Method" (SCM) eine Alternative darstellen.  $\rightsquigarrow$  §4

**Satz 3.29** ("Max- $\Theta$ "-Verfahren für  $\gamma_{UB}(\mu)$ , symmetrisches  $a(\cdot, \cdot)$ )

Sei  $a$  symmetrisch, koerziv, separierbar parametrisch mit  $a^q$  positiv semidefinit und  $\Theta_a^q > 0 \forall q, u$

(Offline:) Sei  $\bar{\mu} \in \mathcal{P}$  und  $\gamma(\bar{\mu})$  berechnet

(Online:) Setze für  $\mu \in \mathcal{P}$  :  $\gamma_{UB}(\mu) := \gamma(\bar{\mu}) \max_q \frac{\Theta_a^q(\mu)}{\Theta_a^q(\bar{\mu})}$ . Dann gilt

$$\gamma(\mu) \leq \gamma_{UB}(\mu) < \infty$$

*Beweis.* Übung. □

**Bemerkung** (Komplexitäten). Durch die angegebenen Berechnungsverfahren ist vollständige Offline-/Online-Zerlegung der RB-Lösung, Fehlerschranken und Effektivitätsschranken erreicht (Offline unabh. von  $\mu$ , Online unabh. von  $H$ ). Komplexitäten für  $\Delta_N(\mu), \Delta_{N,s}(\mu)$ :

- Offline:  $\mathcal{O}(H^3 + H^2(Q_f + Q_l + NQ_a) + HQ_l^2 + H(Q_f + NQ_a)^2)$  für EWP für  $\alpha(\bar{\mu})$ , Riesz-Repräsentanten für  $f^q, l^q, a^q(\varphi_n, \cdot)$  und Matrix  $G_l$  und  $G_r$
- Online:  $\mathcal{O}((Q_f + NQ_a)^2 + Q_l^2 + Q_a)$  für Berechnung von  $\|v_r(\cdot; \mu)\|$ ,  $\|l(\cdot; \mu)\|_{X'}$  und  $\alpha_{LB}(\mu)$  durch Min- $\Theta$ . Problematisch ist quadratische Abhängigkeit von  $Q_f, Q_l, NQ_a$ , welches diese Größen in der Praxis stark einschränkt.

**demos\_chapter3(4)** Beispiel-Lauf von Reduktionsschritten in RBmatlab.

- Vorteilhafte Eigenschaften einer Basis  $\Phi_N$ : orthogonal für numerische Stabilität, Hierarchie, so dass Basisvektoren nach Relevanz geordnet sind, d.h.  $(X_{N'})_{N'=1}^N$ ,  $X_{N'} = \text{span}\{\varphi_1, \dots, \varphi_{N'}\}$  soll Sequenz von “optimalen” Räumen sein, damit durch Variation von  $N'$  eine Fehlerkontrolle erlaubt.
- Probleme (3.7), (3.8) stellen schwierige nichtlineare Optimierungsprobleme dar. Um zu praktischer Basisgenerierung zu kommen, werden verschiedene Vereinfachungen gemacht:
  - “Snapshotbasierte” Räume: Statt  $Y \subset X$  beliebig, wird  $Y = \text{span}\{u(\mu^i)\}_{i=1}^N$  mit unbekanntem  $\{\mu^i\}_{i=1}^N$  gesucht.
  - “Diskretisierung des Parameterraumes”. Statt  $\mu \in \mathcal{P}$  wird Maximum bzw. Mittelung nur über  $\mu \in \mathcal{S}_{train}$  durchgeführt, wobei  $\mathcal{S}_{train} = \{\mu^i\}_{i=1}^n \subset \mathcal{P}$  endliche Menge von Trainingsparametern  $\mu$  (z. B. Punkte eines äquidistanten Gitters oder zufällig gewählte Parameter oder mittels adaptiven Verfahren gewählt).
  - Statt eines Fehlermaßes, welches echte Lösung  $u(\mu)$  erfordert, wird häufig ein Fehlerschätzer gewählt, welcher sehr viel schneller auswertbar ist.
  - Das resultierende vereinfachte Optimierungsproblem kann approximativ minimiert werden, indem statt simultan über  $\{\mu^i\}_{i=1}^N$  zu optimieren, einzelne Basisvektoren der Reihe nach durch Optimierung bestimmt werden (“Greedy-Verfahren”)

**Definition 3.30** (Kolmogorov  $n$ -Weite)

Sei  $\mathcal{M} \subseteq X$  kompakte Teilmenge. Zu einem abgeschlossenen Unterraum  $Y \subseteq X$  nennen wir

$$d(Y, \mathcal{M}) := \sup_{v \in \mathcal{M}} \inf_{w \in Y} \|v - w\| = \sup_{v \in \mathcal{M}} \|v - P_Y v\|$$

den *Abstand* von  $Y$  zu  $\mathcal{M}$ . Für  $n \in \mathbb{N}$  nennen wir

$$d_n(\mathcal{M}) := \inf_{Y \subset X, \dim(Y)=n} d(Y, \mathcal{M})$$

die *Kolmogorov  $n$ -Weite* der Menge  $\mathcal{M}$ . Als Abschwächung definieren wir

$$\bar{d}_n(\mathcal{M}) := \inf_{Y \subset \text{span}(\mathcal{M}), \dim(Y)=n} d(Y, \mathcal{M}).$$

**Bemerkung.**

- $d_n, \bar{d}_n$  fallen monoton.
- $d_n, \bar{d}_n$  sind rein approximationstheoretische Maße, deren Abfall die Approximierbarkeit von  $\mathcal{M}$  mit linearen Unterräumen charakterisiert, unabhängig von der RB-Approximation
- Wenn wir für ein  $(\mathcal{P}(\mu))$  Konvergenz oder sogar Konvergenzrate von  $d_n(\mathcal{M})$  zeigen können, so erhalten wir ebenso Konvergenz mit mind derselben Rate via Céa 3.9 für die RB-Approximation

$$\|u(\mu) - u_N(\mu)\| \leq \frac{\gamma(\mu)}{\alpha(\mu)} \underbrace{\inf_{v \in X_N} \|u(\mu) - v\|}_{d(X_N, \mathcal{M})} \leq \frac{\bar{\gamma}}{\bar{\alpha}} d_n(\mathcal{M})$$

- Beziehung  $d_n$  zu  $\bar{d}_n$ . Es gilt trivialerweise

$$d_0(\mathcal{M}) = \bar{d}_0(\mathcal{M}) = d(0, \mathcal{M}) = \sup_{v \in \mathcal{M}} \|v\|,$$

$$d_n(\mathcal{M}) \leq \bar{d}_n(\mathcal{M}) \quad \forall n \in \mathbb{N},$$

falls  $n_0 := \dim(\text{span}(\mathcal{M})) < \infty$ ,  $d_n(\mathcal{M}) = \bar{d}_n(\mathcal{M}) = 0 \quad \forall n \geq n_0$

- Präzise Werte für  $d_n$  sind selten bekannt. Für endliche Menge oder Einheitskugeln können aber exakte Werte der Schranken für  $d_n$  angegeben werden.
- Beispiel:  $\mathcal{M} := \{v \in X \mid \|v\| \leq 1\}$  erfüllt  $d_n(\mathcal{M}) = \bar{d}_n(\mathcal{M}) = 1$  für alle  $n < \dim(X)$  und  $d_n(\mathcal{M}) = \bar{d}_n(\mathcal{M}) = 0$  für  $n \geq \dim(X)$ .
- Beispiel:  $\mathcal{M} := [1, 1]^m \subset X := \mathbb{R}^m$  erfüllt  $d_n(\mathcal{M}) = \bar{d}_n(\mathcal{M}) = \sqrt{mn}$  für alle  $n \leq m$  und  $d_n(\mathcal{M}) = \bar{d}_n(\mathcal{M}) = 0$  für  $n \geq m$ .
- Beispiel: “Müsli-Schachtel”:  $\mathcal{M} := \prod_{i \in \mathbb{N}} [-2^i, 2^i] \subseteq l_2 \Rightarrow d_n(\mathcal{M}) \leq C \cdot 2^{-n}$ , exponentielle Konvergenz

**Definition 3.31** (Gram-Matrix)

Zu  $\{u_i\}_{i=1}^n \subset X$  definieren wir die *Gram-Matrix* als  $K := (\langle u_i, u_j \rangle_X)_{i,j=1}^n \in \mathbb{R}^{n \times n}$



**Lemma 3.32** (Eigenschaften von  $K$ )

Für  $K$  Gram-Matrix von  $\{u_i\}_{i=1}^n$  gilt

- i)  $K$  ist symmetrisch und positiv semidefinit.
- ii)  $\text{Rang}(K) = \dim(\text{span}\{u_i\}_{i=1}^n)$
- iii)  $\{u_i\}_{i=1}^n$  linear unabhängig  $\Leftrightarrow K$  positiv definit

*Beweis.* Übung □

**Bemerkung** (Geometrische Information in  $K$ ).  $K$  enthält sehr viel Information der  $\{u_i\}_{i=1}^n$ , insbesondere kann man mittels  $K$  eine isometrische Einbettung in  $\mathbb{R}^n$  erzeugt werden, d. h. es existiert  $\{x_i\}_{i=1}^n \in \mathbb{R}^n$  mit

$$\langle x_i, x_j \rangle_{\mathbb{R}^n} = \langle u_i, u_j \rangle_X$$

und

$$\|x_i - x_j\|_{\mathbb{R}^n} = \|u_i - u_j\|_X$$

Sie  $K = UDU^T$  Eigenwertzerlegung mit  $D = \text{diag}(\lambda_1, \dots, \lambda_n)$ . Setze  $(x_1, \dots, x_n) := D^{\frac{1}{2}}U^T$ . Dann:

$$\begin{aligned} \langle x_i, x_j \rangle_{\mathbb{R}^n} &= [(D^{\frac{1}{2}}U^T)_{i,j}]^T [(D^{\frac{1}{2}}U^T)_{i,j}] \\ &= (UD^{\frac{1}{2}})_{i,j} (D^{\frac{1}{2}}U^T)_{i,j} \\ &= (U)_{i,j} D^{\frac{1}{2}} D^{\frac{1}{2}} (U^T)_{i,j} = (UDU^T)_{ij} = (K)_{ij} = \langle u_i, u_j \rangle_X \end{aligned}$$

und

$$\begin{aligned} \|x_i - x_j\|_{\mathbb{R}^n}^2 &= \langle x_i, x_i \rangle_{\mathbb{R}^n} - 2\langle x_i, x_j \rangle_{\mathbb{R}^n} + \langle x_j, x_j \rangle_{\mathbb{R}^n} \\ &= \langle u_i, u_i \rangle_X - 2\langle u_i, u_j \rangle_X + \langle u_j, u_j \rangle_X = \|u_i - u_j\|_X^2 \end{aligned}$$

Damit können viele lineare Operationen auf  $\{u_i\}_{i=1}^n$  durch geeignete Operation mit  $K$  ausgedrückt werden. Z.B. Normberechnung in  $\text{span}\{u_i\}_{i=1}^n$ : (siehe auch Offline/Online für Normen)

$$v = \sum_{i=1}^n v_i u_i \quad , \quad \underline{v} = (v_i)_{i=1}^n \in \mathbb{R}^n \quad \Rightarrow \quad \|v\|_X^2 = \underline{v}^T K \cdot \underline{v}$$

### 3.4 Basisgenerierung

#### Approximation durch lineare Unterräume

Motivation für Snapshot-basierte Verfahren:

- Bestimmung eines möglichst guten  $X_N$ , welches  $\mathcal{M}$  global approximiert.

- Formulierung durch Optimierungsproblem, z.B. minimiere maximalen Fehler in Energienorm

$$\min_{\substack{Y \subseteq X \\ \dim Y = N}} \max_{\mu \in \mathcal{P}} \|u(\mu) - u_N(\mu)\|_\mu \quad (3.7)$$

oder Minimum des mittleren quadratischen Projektionsfehlers

$$\min_{\substack{Y \subseteq X \\ \dim Y = N}} \int_{\mathcal{P}} \|u(\mu) - P_Y u(\mu)\|^2 d\mu \quad (3.8)$$

oder beliebiges anderes Distanzmaß.

Wir haben bereits gesehen, dass in bestimmten Fällen fehlerfreie Approximation durch geeignete RB-Räume möglich ist

**Satz 3.33** (Optimales  $X_N$  für Thermischer Block,  $B_1 = 1$ )

Sie  $p \in \mathbb{N}$ ,  $B_1 = 1$ ,  $B_2 = p$ ,  $\mu^i := (\mu_{min}, \dots, \mu_{min})^T + e_i \cdot (\mu_{max} - \mu_{min})$ ,  $i = 1, \dots, p$ . Dann ist  $X_N := \text{span}(u(\mu^i))_{i=1}^p$  optimal in dem Sinne, dass

$$\inf_{v \in X_N} \|u(\mu) - v\| = \inf_{v \in X_N} \|u(\mu) - v\|_\mu = 0 \quad \forall \mu \in \mathcal{P}$$

*Beweis.* Übung. □

**Satz 3.34** (Optimales  $X_N$  für  $Q_a = 1$ )

Sei  $Q_a = 1$  und o.B.d.A.  $a(u, v; \mu) = \Theta_a^1(\mu) a^1(u, v)$  und  $\Theta_a^1(\mu) > 0$ . Seien  $\{\mu^i\}_{i=1}^{Q_f}$  derart, dass  $\{f(\cdot; \mu^i)\}_{i=1}^{Q_f}$  linear unabhängig. Dann erfüllt der Lagrange RB-Raum  $X_N = \text{span}(u(\mu^i))_{i=1}^{Q_f}$ ,  $\dim X_N = Q_f$  und

$$\inf_{v \in X_N} \|u(\mu) - v\| = \inf_{v \in X_N} \|u(\mu) - v\|_\mu = 0 \quad \forall \mu \in \mathcal{P}$$

*Beweis.*

i) ✓

ii)  $\text{span}\{f^q\}_{q=1}^{Q_f} = \text{span}\{f(\cdot; \mu^i)\}_{i=1}^q$

“ $\supseteq$ ” ist klar weil  $f(\cdot; \mu^i) = \sum_{q=1}^{Q_f} \Theta_f^q f^q(\cdot)$

“ $=$ ” aus Dimensionsbetrachtung

$$\dim\left(\text{span}\{f(\cdot; \mu^i)\}_{i=1}^{Q_f}\right) = Q_f$$

$$\dim\left(\text{span}\{f^q\}_{q=1}^{Q_f}\right) \leq Q_f$$

$\Rightarrow$  folgt  $\dim\left(\text{span}\{f^q\}_{q=1}^{Q_f}\right) = Q_f$  Gleichheit beider Räume

iii) Zeige nun exakte Approximation in  $X_N$ .

Zu  $\mu \in \mathcal{P}$  existiert wegen ii)  $c(\mu) = (c_i(\mu))_{i=1}^{Q_f}$  mit

$$f(\cdot; \mu) = \sum_{i=1}^{Q_f} c_i(\mu) f(\cdot; \mu^i) \quad (*)$$

Damit ist dann  $u(\mu) := \sum c_i(\mu) \frac{\Theta_a^q(\mu^i)}{\Theta_a^q(\mu)} u(\mu^i)$  Lösung von  $(P(\mu))$ :

$$\begin{aligned} a(u(\mu), v; \mu) &= \Theta_a^1(\mu) a^1\left(\sum c_i(\mu) \frac{\Theta_a^1(\mu^i)}{\Theta_a^1(\mu)} u(\mu^i), v\right) \\ &= \sum c_i(\mu) \underbrace{\Theta_a^1(\mu^i) a^1(u(\mu^i), v)}_{=a(u(\mu^i), v; \mu^i)} \\ &= \sum c_i(\mu) f(v; \mu^i) \stackrel{(*)}{=} f(v; \mu) \end{aligned}$$

□

Für die folgende Aussage referenzieren wir Fink & Rheinboldt: On the Error Behavior of the Reduced Basis Technique for Nonlinear Finite Element Approximations, ZAMM, 63:21-28, 1983.

**Satz 3.35** (Lokale exponentielle Konvergenz)

Sei  $\mu^0 \in U \subset \mathcal{P} \subset \mathbb{R}$  und  $u(\mu)$  analytisch in Umgebung  $U$ . Sei  $X_{k, \mu^0}$  der Taylor-RB-Raum für  $k \in \mathbb{N}$ . Dann existiert ein  $B_\delta(\mu^0) \subset U$  und  $C > 0$ , so dass

$$\inf_{v \in X_{k, \mu^0}} \|u(\mu) - v\| \leq C |\mu - \mu^0|^{k+1} \quad \forall \mu \in B_\delta(\mu^0)$$

*Beweis.* Taylor-Entwicklung

$$\begin{aligned} u(\mu) &= \sum_{i=0}^{\infty} \frac{\partial^i}{\partial \mu^i} u(\mu^0) \frac{1}{i!} (\mu - \mu^0)^i \\ &= \underbrace{\sum_{i=0}^k \frac{\partial^i}{\partial \mu^i} u(\mu^0) \frac{1}{i!} (\mu - \mu^0)^i}_{v_k(\mu)} + \underbrace{\sum_{i=k+1}^{\infty} \frac{\partial^i}{\partial \mu^i} u(\mu^0) \frac{1}{i!} (\mu - \mu^0)^{i-(k+1)} (\mu - \mu^0)^{k+1}}_{w_k(\mu)} \end{aligned}$$

Sei  $\delta < 1$  so dass  $B_\delta(\mu^0) \subset U$  und  $C' := \sup_i \left\| \frac{\partial^i}{\partial \mu^i} u(\mu^0) \frac{1}{i!} \right\| < \infty$ . Dann gilt für  $\mu \in B_\delta(\mu)$

$$\begin{aligned} \|w_k(\mu)\| &\leq \sum_{i=k+1}^{\infty} \left\| \frac{\partial^i}{\partial \mu^i} u(\mu^0) \frac{1}{i!} \right\| \cdot |\mu - \mu^0|^{i-k+1} \leq C' \sum_{i=k+1}^{\infty} |\mu - \mu^0|^{i-(k+1)} \\ &\leq C' \frac{1}{1 - |\mu - \mu^0|} \leq C' \frac{1}{1 - \delta} =: C \end{aligned}$$

$$\inf_{v \in X_{k, \mu^0}} \|u(\mu) - v\| \leq \|u(\mu) - v_k(\mu)\| = \|w_k(\mu) \cdot (\mu - \mu^0)^{k+1}\| \leq C |\mu - \mu^0|^{k+1}$$

□

Die folgende Aussage basiert auf Maday & Patera & Turinici: Global a priori convergence theory for reduced-basis approximations of single-parameter symmetric coercive elliptic partial differential equations. C.R. Acad. Sci., Paris, Ser. 1, 335, 289-294, 2002.

**Satz 3.36** (Globale exponentielle Konvergenz,  $p = 1$ )

Sei  $\mathcal{P} = [\mu_{min}, \mu_{max}] \subset \mathbb{R}^+$  mit  $\mu_{max} > 1$  genügend groß und  $\mu_{min} = \frac{1}{\mu_{max}}$

$$a(u, v; \mu) = \mu a^1(u, v) + a^2(u, v)$$

mit  $a^1, a^2$  symmetrisch positiv semidefinit und  $f \in X'$  sei nicht parametrisch. Zu  $N \in \mathbb{N}$ ,  $N \geq 2$  seien

$$\mu_{min} = \mu^1 < \dots < \mu^N = \mu_{max}$$

logarithmisch äquidistant, d.h.

$$\ln(\mu^{i+1}) - \ln(\mu^i) = \frac{\ln(\mu_{max}) - \ln(\mu_{min})}{N-1} = \delta_N$$

und  $X_N = \text{span} \{u(\mu^i)\}_{i=1}^N$  zugehöriger Lagrange RB-Raum. Dann existiert  $N_0$  so dass für alle  $N \geq N_0$  gilt

$$\frac{\|u(\mu) - u_N(\mu)\|_\mu}{\|u(\mu)\|_\mu} \leq \mu_{max}^2 e^{\frac{-N-1}{N_0-1}} \quad \forall \mu \in \mathcal{P} \quad (3.9)$$

**Bemerkung.**

- Voraussetzungen sind z.B. für einen thermischen Block mit  $B_1 = 2$ ,  $B_2 = 1$  erfüllt, wenn  $\mu_2 = 1$  konstant gehalten wird und nur  $\mu_1 = \mu$  variiert.
- Verallgemeinerung für  $p > 1$  existiert.
- Satz 3.36 liefert sogar die exponentielle Konvergenz des Approximationsfehlers und damit der Weiten  $d_N, \bar{d}_N$ .

**Korollar 3.37** (Exponentielle Konvergenz von  $d_N, \bar{d}_N$ )

Unter den Voraussetzungen von 3.36 gilt insbesondere mit  $C > 0$  unabhängig von  $N$

$$\inf_{v \in X_N} \|u(\mu) - v\| \leq C \cdot e^{\frac{-N-1}{N_0-1}} \quad \forall \mu \in \mathcal{P}, N \geq N_0$$

also für Kulmogorov  $N$ -Weite

$$d_N(\mathcal{M}) \leq C \cdot e^{\frac{-N-1}{N_0-1}} \quad \forall N \geq N_0$$

und wegen  $X_N \subset \text{span}(\mathcal{M})$

$$\bar{d}_N(\mathcal{M}) \leq C \cdot e^{\frac{-N-1}{N_0-1}}$$

*Beweis.* Wegen Normäquivalenz und Beschränktheit von  $u$  gilt

$$\|u(\mu)\|_\mu \leq \sqrt{\gamma(\mu)} \|u(\mu)\| \leq \frac{\gamma(\mu)}{\alpha(\mu)} \|f(\mu)\| \leq C', \quad \text{mit} \quad C' := \sup_{\mu \in \mathcal{P}} \frac{\sqrt{\gamma(\mu)}}{\alpha(\mu)} \|f(\mu)\|$$

Also folgt

$$\begin{aligned} \inf_{v \in X_N} \|u(\mu) - v\| &\leq \|u(\mu) - u_N(\mu)\| \leq \frac{1}{\sqrt{\alpha(\mu)}} \|u(\mu) - u_N(\mu)\|_\mu \cdot \frac{\|u(\mu)\|_\mu}{\|u(\mu)\|_\mu} \\ &\stackrel{3.36}{\leq} \frac{\|u(\mu)\|_\mu}{\sqrt{\alpha(\mu)}} \cdot e^{\frac{-N-1}{N_0-1}} \leq \underbrace{\frac{C'}{\bar{\alpha}}}_{=:C} \cdot e^{\frac{-N-1}{N_0-1}} \end{aligned}$$

□

Ziel ist Beweis von 3.36, hierzu benötigen wir jedoch einige Notationen und Hilfsaussagen.

- Es sei  $\dim X = H$  endlich aber beliebig groß. Man kann zeigen, dass die Konstante  $N_0$  und Forderung an  $\mu_{max}$  unabhängig von  $H$  ist.
- Logarithmische Abbildung des Parametergebiets. Es sei

$$\tau(z) = \ln(z)$$

und damit  $\hat{\mu} := \tau(\mu)$ ,  $\hat{\mu}_{min} = \tau(\mu_{min})$ ,  $\hat{\mu}_{max} = \tau(\mu_{max}) = -\hat{\mu}_{min}$ ,  $\hat{\mathcal{P}} := \tau(\mathcal{P})$ ,  $\hat{u}(\hat{\mu}) := u(\tau^{-1}(\hat{\mu}))$ , also  $\hat{u}(\hat{\mu})$  Lösung von

$$e^{\hat{\mu}} a^1(\hat{u}(\hat{\mu}), v) + a^2(\hat{u}(\hat{\mu}), v) = f(v) \quad \forall v \in X \quad (3.10)$$

und dann  $u(\mu) = \hat{u}(\tau(\mu))$ .

- Es sei  $\langle u, v \rangle_X := a(u, v; \mu = 1) = a^1(u, v) + a^2(u, v)$ , dann ist (3.10) äquivalent zu

$$\langle \hat{u}(\hat{\mu}), v \rangle_X + (e^{\hat{\mu}} - 1) a^1(\hat{u}(\hat{\mu}), v) = f(v) \quad \forall v \in X \quad (3.11)$$

- Seien  $(\Upsilon_i, \lambda_i)_{i=1}^H \in (X, \mathbb{R}^+)$  Eigenfunktionen/-werte von verallgemeinertem EWP

$$a^1(\Upsilon_i, v) = \lambda_i \langle \Upsilon_i, v \rangle_X \quad \forall v \in X \quad (3.12)$$

mit  $0 \leq \lambda_1 \leq \dots \leq \lambda_H$ ,  $\|\Upsilon_i\| = 1$  ist dann  $\{\Upsilon_i\}_{i=1}^H$  ONB von  $X$ . Aus (3.12) mit  $v = \Upsilon_i$  und positiver Semidefinitheit von  $a^2$  folgt

$$1 = \langle \Upsilon_i, \Upsilon_i \rangle_X = a^1(\Upsilon_i, \Upsilon_i) + a^2(\Upsilon_i, \Upsilon_i) = \lambda_i + a^2(\Upsilon_i, \Upsilon_i)$$

$$\Rightarrow \lambda_i \in [0, 1] =: \Lambda \text{ weil } \lambda_i = 1 - a^2(\Upsilon_i, \Upsilon_i) \leq 1$$

- Aus Orthogonalität und (3.12) folgt

$$a(\Upsilon_j, \Upsilon_i; \mu) = \underbrace{\langle \Upsilon_j, \Upsilon_i \rangle}_{\delta_{ij}} + (e^{\hat{\mu}} - 1) \underbrace{a^1(\Upsilon_j, \Upsilon_i)}_{\lambda_i \delta_{ij}} = (1 - \lambda_j + \lambda_j e^{\hat{\mu}}) \delta_{ij} \quad (3.13)$$

**Lemma 3.38** (Lösungsdarstellung)

Die Lösung von (3.10), (3.11) ist explizit gegeben durch

$$\hat{u}(\hat{\mu}) = \sum_{j=1}^H f_j \Upsilon_j g(\hat{\mu}, \lambda_j) \quad (3.14)$$

mit  $f_j = f(\Upsilon_j)$  und  $g : \hat{\mathcal{P}} \times \Lambda \rightarrow \mathbb{R}^+$  definiert durch

$$g(z, \sigma) = \frac{1}{1 - \sigma + \sigma e^z} \quad (3.15)$$

*Beweis.* Einsetzen von (3.14) in (3.11) liefert für Testfunktionen  $v := \Upsilon_i$

$$\begin{aligned} & \left\langle \sum_j f_j \Upsilon_j g(\hat{\mu}, \lambda_j), \Upsilon_i \right\rangle_X + (e^{\hat{\mu}} - 1) a^1 \left( \sum_j f_j \Upsilon_j g(\hat{\mu}, \lambda_j), \Upsilon_i \right) \\ &= \sum_j f_j g(\hat{\mu}, \lambda_j) \left( \langle \Upsilon_j, \Upsilon_i \rangle_X + (e^{\hat{\mu}} - 1) a^1(\Upsilon_j, \Upsilon_i) \right) \\ &\stackrel{(3.13)}{=} \sum_j f_j g(\hat{\mu}, \lambda_j) (1 - \lambda_j + \lambda_j e^{\hat{\mu}}) \delta_{ij} \\ &= f_i g(\hat{\mu}, \lambda_i) \underbrace{(1 - \lambda_i + \lambda_i e^{\hat{\mu}})}_{=\frac{1}{g(\hat{\mu}, \lambda_i)}} = f_i = f(\Upsilon_i) \end{aligned}$$

also ist  $\hat{u}(\hat{\mu})$  Lösung von (3.10) / (3.11). □

**Bemerkung.** Im obigen Beweis wird also ausgenutzt, dass die Systemmatrix bezüglich  $\{\Upsilon_i\}_{i=1}^H$  diagonal ist.

**Lemma 3.39** (Energienorm-Darstellung in ONB)

Mit  $\mu = \tau^{-1}(\hat{\mu})$  gilt für eine Funktion  $w = \sum_{i=1}^H w_i \Upsilon_i$

$$\|w\|_\mu^2 = \sum_{i=1}^H \frac{w_i^2}{g(\hat{\mu}, \lambda_i)}$$

*Beweis.*

$$\|w\|_\mu = a(w, w; \mu) = \sum_{i,j} w_i w_j \underbrace{a(\Upsilon_i, \Upsilon_j; \mu)}_{\stackrel{(3.13)}{=} (1 - \lambda_j + \lambda_j e^{\hat{\mu}}) \delta_{ij}} = \sum \frac{w_i^2}{g(\hat{\mu}, \lambda_i)}$$

□

Wir benötigen (grobe) Schranken für  $g$  und seinen Ableitungen bezüglich  $z$ .

**Lemma 3.40** (Schranken für  $g, \frac{\partial^i}{\partial z^i} g$ )

Für alle  $z \in \hat{\mathcal{P}}, \sigma \in \Lambda = [0,1]$  gilt

i)

$$g(z, \sigma) \in \left[ \frac{1}{\mu_{max}}, \mu_{max} \right]$$

ii)

$$\frac{1}{g(z, \sigma)} \in \left[ \frac{1}{\mu_{max}}, \mu_{max} \right]$$

iii)

$$\left| \frac{\partial^i}{\partial z^i} g(z, \sigma) \right| \leq \bar{C} \cdot C \cdot j! \quad \text{mit} \quad \bar{C} = \mu_{max}, C = 2\mu_{max}^2$$

*Beweis.* i) & ii)

$$\frac{1}{g(z, \sigma)} = 1 + \sigma(e^z - 1) \stackrel{j=1}{\leq} e^{\hat{\mu}_{max}} = \mu_{max} \Rightarrow g(z, \sigma) \geq \frac{1}{\mu_{max}}$$

Für festes  $z$  minimiere  $\frac{1}{g(z, \sigma) = 1 + \sigma(e^z - 1)}$  bezüglich  $\sigma$

$$\min_{\sigma \in [0,1]} \frac{1}{g(z, \sigma)} = \begin{cases} 1 & z = 0 \quad (\sigma \text{ beliebig}) \\ 1 & z > 0 \quad (\sigma = 0) \\ e^z & z < 0 \quad (\sigma = 1) \end{cases}$$

$$\frac{1}{g(z, \sigma)} \geq \min_{\sigma, z} \frac{1}{g(z, \sigma)} = e^{\hat{\mu}_{min}} = \mu_{min} = \frac{1}{\mu_{max}} \Rightarrow g(z, \sigma) \leq \mu_{max}$$

iii) Wir zeigen per Induktion, dass  $\frac{\partial^i}{\partial z^i} g(z, \sigma)$  sich darstellen lässt als

$$\frac{\partial^i}{\partial z^i} g(z, \sigma) = \sum_{k=2}^{j+1} \beta_k^j \sigma^{(k-1)} e^{(k-1)z} g^k(z, \sigma), \quad j \geq 1 \quad (3.16)$$

mit

$$\left. \begin{aligned} \beta_2^1 &:= -1 \\ \beta_2^{j+1} &= \beta_2^j = -1 \\ \beta_k^{j+1} &= \beta_k^j(k-1) - \beta_{k-1}^j(k-1), \quad k = 3, \dots, j+1 \\ \beta_{j+2}^{j+1} &:= -(j+1)\beta_{j+1}^j \end{aligned} \right\} \quad (3.17)$$

Denn für  $j = 1$  erhält man

$$\frac{\partial}{\partial z} g(z, \sigma) = \frac{-\sigma e^z}{(1 - \sigma - \sigma e^z)^2} = -e^z \sigma g^2(z, \sigma)$$

also mit (3.16)  $\beta_2^1 = -1$ .

### Induktionsschritt

$$\begin{aligned}
\frac{\partial^i}{\partial z^i} g(z, \sigma) &= \frac{\partial}{\partial z} \left( \sum_{k=2}^{j+1} \beta_k^j \sigma^{(k-1)} e^{(k-1)z} g^k(z, \sigma) \right) \\
&= \sum_{k=2}^{j+1} \beta_k^j \sigma^{(k-1)} \left( e^{(k-1)z} \underbrace{\frac{\partial}{\partial z} g^k(z, \sigma)}_{=kg^{k-1}(z, \sigma)} + (k-1)e^{(k-1)z} g^k(z, \sigma) \right) \\
&= \sum_{k=2}^{j+1} \beta_k^j \sigma^{(k-1)} \left[ -\sigma e^{kz} k \cdot g^{k+1}(z, \sigma) + (k-1)e^{(k-1)z} g^k(z, \sigma) \right] \\
&= \sum_{k=2}^{j+1} \beta_k^j (k-1) \sigma^{(k-1)} e^{(k-1)z} g^k(z, \sigma) + \sum_{k=3}^{j+2} \beta_{k-1}^j \sigma^{(k-1)} \cdot \left( -\sigma e^{(k-1)z} (k-1) g^k(z, \sigma) \right) \\
&= \underbrace{\beta_2^j \sigma e^z g^2(z, \sigma)}_{k=2} + \sum_{k=3}^{j+1} \left( \beta_k^j (k-1) - \beta_{k-1}^j (k-1) \right) \sigma^{(k-1)} e^{(k-1)z} g^k(z, \sigma) \\
&\quad \underbrace{- \beta_{j+1}^j (j+1) \sigma^{j+1} e^{(j+1)z} g^{j+2}(z, \sigma)}_{\text{"}k=j+2\text{"}}
\end{aligned}$$

Für  $j \geq 1$  setze  $S_j := \sum_{k=2}^{j+1} |\beta_k^j|$  und zeige per Induktion, dass

$$S_j \leq 2^j \cdot j! \quad j \geq 1$$

Für  $j = 1$  ist  $S_j = 1 \leq 2^1 \cdot 1! = 2$  also Induktionsanfang. Gelte Behauptung für  $j \geq 1$ . Dann gilt:

$$\begin{aligned}
|\beta_2^{j+1}| &= 1 \\
|\beta_k^{j+1}| &< (j+1)(|\beta_k^j| + |\beta_{k-1}^j|) \quad k = 3, \dots, j+1 \\
|\beta_{j+2}^{j+1}| &= (j+1)|\beta_{j+1}^j| \\
\Rightarrow S_{j+1} &= \sum_{k=2}^{j+2} |\beta_k^{j+1}| \leq 2(j+1)S_j \stackrel{i.A.}{\leq} 2(j+1)2^j \cdot j! = 2^{j+1}(j+1)!
\end{aligned}$$



Damit folgt (iii):

$$\begin{aligned}
\left| \frac{\partial^j}{\partial z^j} g(z, \sigma) \right| &= \left| \sum_{k=2}^{j+1} \beta_k^j \sigma^{(k-1)} e^{(k-1)z} g^k(z, \sigma) \right| \\
&\leq \underbrace{\left( \sum_{k=2}^{j+1} |\beta_k^j| \right)}_{\leq 2^j \cdot j!} \underbrace{\sup_k |\sigma^{(k-1)} e^{(k-1)z} g^k(z, \sigma)|}_{\leq 1 \cdot e^{j \mu_{\max}} \cdot \mu_{\max}^{j+1} = \mu_{\max} (\mu_{\max}^2)^j} \\
&\leq (2 \mu_{\max}^2)^j \cdot j! \cdot \mu_{\max}
\end{aligned}$$

□

**Bemerkung.**  $g(\hat{\mu}, \lambda_i)$  sind gemäß 3.38 Koeffizienten für  $\hat{u}(\hat{\mu})$  in ONB Entwicklung. Entsprechend sind  $\frac{\partial^j}{\partial z^j} g(z, \sigma)$  für  $z = \hat{\mu}$ ,  $\sigma = \lambda_i$  die Koeffizienten der Sensitivitätsableitung  $\frac{\partial^j}{\partial \hat{\mu}^j} \hat{u}(\hat{\mu})$  in der ONB Entwicklung, also impliziert Lemma 3.40 eine Beschränktheit der Sensitivitätsableitungen.

**Lemma 3.41** (Darstellung von Fkt. aus  $X_N$ )

Für Koeffizientenfunktionen  $\tilde{C}_n : \hat{\mathcal{P}} \rightarrow \mathbb{R}$ ,  $n = 1, \dots, N$

$$\hat{w}_N(\hat{\mu}) := \sum_{n=1}^N \tilde{C}_n(\hat{\mu}) \hat{u}(\hat{\mu}^n)$$

mit  $\hat{\mu}^n := \ln \mu^n$  ist also  $\hat{w}_N(\hat{\mu}) \in X_N$ . Dann lässt sich  $\hat{w}_N$  darstellen als

$$\hat{w}_N(\hat{\mu}) = \sum_{i=1}^H f_i \Upsilon_i \tilde{g}_N(\hat{\mu}, \lambda_i)$$

wobei  $\tilde{g}_N(\hat{\mu}, \sigma) := \sum_{n=1}^N \tilde{C}_n(\hat{\mu}) g(\hat{\mu}^n, \sigma)$

*Beweis.* Aus Lösungsdarstellung 3.38 folgt

$$u(\mu^n) = \hat{u}(\hat{\mu}^n) = \sum_{i=1}^H f_i \Upsilon_i g(\hat{\mu}^n, \lambda_i), \quad n = 1, \dots, N$$

Also ist

$$\hat{w}_N(\hat{\mu}) = \sum_n \tilde{C}_n(\hat{\mu}) \sum_{i=1}^H f_i \Upsilon_i g(\hat{\mu}^n, \lambda_i) = \sum_{i=1}^H f_i \Upsilon_i \underbrace{\sum_{n=1}^N \tilde{C}_n(\hat{\mu}) g(\hat{\mu}^n, \sigma)}_{=\tilde{g}_N(\hat{\mu}, \sigma)}$$

□

Wir benötigen noch Lagrange-Interpolation in  $M$  aufeinanderfolgenden Punkten  $\{\hat{\mu}^i, \dots, \hat{\mu}^{i+M-1}\}$ : Sie  $h \in C^M(\hat{\mathcal{P}})$  zu interpolierende Funktion. Es bezeichne  $I_M^i : C^0(\hat{\mathcal{P}}) \rightarrow P_{M-1}(\hat{\mathcal{P}})$  Polynominterpolation zu  $\{\mu^{i+m-1}\}_{m=1}^M$  für  $M \geq 2$  und  $i \in \{1, \dots, N\}$  s. d.  $i + M \leq N + 1$ . Sei  $L_M^{i;m} \in P_{M-1}(\hat{\mathcal{P}})$  Lagrange-Polynom zu den Stützstellen, d. h.

$$L_M^{i;m}(\hat{\mu}^{i+m'-1}) = \delta_{mm'} \quad \text{für } 1 \leq m, m' \leq M$$

Dann ist der Interpolant darstellbar als

$$(I_M^i h)(\hat{\mu}) = \sum_{m=1}^M L_M^{i;m}(\hat{\mu}) h(\hat{\mu}^{i+m-1})$$

Für den Interpolationsfehler gilt in  $\hat{\mu} \in [\hat{\mu}^i, \mu^{i+M-1}]$

$$\begin{aligned} |h(\hat{\mu}) - (I_M^i h)(\hat{\mu})| &\leq \frac{\prod_{m=1}^M |\hat{\mu} - \hat{\mu}^{i+m-1}|}{\underbrace{M!}} \sup_{\hat{\mu}'} |h^{(M)}(\hat{\mu}')| \leq \frac{[(M-1)\delta_N]^M}{M!} \sup_{\hat{\mu}'} |h^{(M)}(\hat{\mu}')| \\ &\leq \frac{(\hat{\mu}^{i+M-1} - \hat{\mu}^i)^M}{M!} = \frac{[(M-1)\delta_N]^M}{M!} \end{aligned} \quad (3.18)$$

(endlich:)

*Beweis.* Satz 3.36

Idee: zeige Existenz eines  $\hat{w}_N(\hat{\mu}) \in X_N$  s. d. (mit  $\mu := \tau^{-1}(\hat{\mu})$ )

$$\frac{\|\hat{u}(\hat{\mu}) - \hat{w}_N(\hat{\mu})\|_\mu}{\|\hat{u}(\hat{\mu})\|_\mu} \leq \mu_{max}^2 \cdot e^{\frac{-(N-1)}{N_0-1}} \quad \forall N \geq N_0 \quad (3.19)$$

Denn dann folgt Behauptung via  $\|\hat{u}(\hat{\mu})\|_\mu = \|u(\mu)\|_\mu$  und

$$\|u(\mu) - u_N(\mu)\|_\mu = \inf_{v \in X_N} \|u(\mu) - v\|_\mu \leq \|u(\mu) - \hat{w}_N(\hat{\mu})\|_\mu = \|\hat{u}(\hat{\mu}) - \hat{w}_N(\hat{\mu})\|_\mu$$

Für Konstruktion eines  $\hat{w}_N(\hat{\mu})$  reicht es, die Koeffizienten  $\tilde{C}_n(\hat{\mu})$  zu definieren (siehe 3.41): Sei  $\hat{\mu} \in \mathcal{P}$  gegeben und  $M \in \{2, \dots, N\}$  wähle  $i$  s. d.  $\hat{\mu} \in [\hat{\mu}^i, \hat{\mu}^{i+M-1}] =: J_M^i$ , also  $|J_M^i| = (M-1)d_N$

Definiere nun  $\tilde{C}_n$  durch Lagrange-Polynome zu  $\{\hat{\mu}^{i+m-1}\}_{m=1}^M$ :

$$\tilde{C}_n(\hat{\mu}) = \begin{cases} 0 & \text{falls } n < i \text{ oder } n \geq i + M \\ L_M^{i;n-i+1}(\hat{\mu}) & \text{falls } i \leq n \leq i + M - 1 \end{cases}$$

Dann ist zugehöriges  $\tilde{g}_N(\hat{\mu}, \sigma)$  aus 3.41 Interpolierende im Sinne von

$$\tilde{g}_N(\hat{\mu}, \sigma) = (I_M^i g(\cdot, \sigma))(\hat{\mu})$$

denn

$$\tilde{g}_N(\hat{\mu}, \sigma) \stackrel{3.41}{=} \sum_{n=1}^N \tilde{C}_n(\hat{\mu}) g(\hat{\mu}^n, \sigma) = \sum_{n=i}^{i+n-1} L_M^{i;n-i+1}(\hat{\mu}) g(\hat{\mu}^n, \sigma) = (I_M^i g(\cdot, \sigma))(\hat{\mu})$$

also insbesondere  $\tilde{g}_N(\hat{\mu}^n, \sigma) = g(\hat{\mu}^n, \sigma)$ .

Betrachtet man die linke Seite von (3.19) für  $\hat{w}_N(\hat{\mu})$ : Mit 3.41 & 3.38 & 3.39 folgt

$$\begin{aligned} \frac{\|\hat{u}(\hat{\mu}) - \hat{w}_N(\hat{\mu})\|_\mu^2}{\|\hat{u}(\hat{\mu})\|_\mu^2} &\stackrel{3.39}{=} \frac{\sum_{i=1}^H \frac{f_i^2 (g(\hat{\mu}, \lambda_i) - \tilde{g}_N(\hat{\mu}, \lambda_i))^2}{g(\hat{\mu}, \lambda_i)}}{\sum_{i=1}^H \frac{f_i^2 g(\hat{\mu}, \lambda_i)^2}{g(\hat{\mu}, \lambda_i)}} \\ &= \frac{\sum_{i=1}^H f_i^2 g(\hat{\mu}, \lambda_i)^2 \frac{(g(\hat{\mu}, \lambda_i) - \tilde{g}_N(\hat{\mu}, \lambda_i))^2}{g(\hat{\mu}, \lambda_i)^2} \frac{1}{g(\hat{\mu}, \lambda_i)}}{\sum_{i=1}^H \frac{f_i^2 g(\hat{\mu}, \lambda_i)^2}{g(\hat{\mu}, \lambda_i)}} \\ &\leq \sup_{z, \sigma} \frac{(g(z, \sigma) - \tilde{g}_N(z, \sigma))^2}{g(z, \sigma)^2} \cdot \frac{\sum_{i=1}^H f_i^2 \frac{g(\hat{\mu}, \lambda_i)^2}{g(\hat{\mu}, \lambda_i)}}{\sum_{i=1}^H f_i^2 \frac{g(\hat{\mu}, \lambda_i)^2}{g(\hat{\mu}, \lambda_i)}} \\ &\leq \left( \sup_{z, \sigma} \frac{1}{g(z, \sigma)^2} \right) \left( \sup_{z, \sigma} (g(z, \sigma) - \tilde{g}_N(z, \sigma))^2 \right) \\ &\stackrel{3.40ii)}{\leq} \mu_{max}^2 \left( \sup_{z, \sigma} |g(z, \sigma) - \tilde{g}_N(z, \sigma)| \right)^2 \end{aligned} \quad (3.20)$$

Für Fehler rechts erhalte mittels Interpolationsfehlerabschätzung:

$$\begin{aligned} |g(z, \sigma) - \tilde{g}_N(z, \sigma)| &= |g(z, \sigma) - (I_M^i g(\cdot, \sigma))(z)| \\ &\stackrel{(3.18)}{\leq} \frac{((M-1)\delta_N)^M}{M!} \sup_{\hat{\mu}'} \underbrace{\left| \frac{\partial^M}{\partial \hat{\mu}^M} g(\hat{\mu}', \sigma) \right|}_{\leq \bar{C} C^M M! \text{ wegen 3.40 ii)}} \\ &= \frac{((M-1)\delta_N)^M}{M!} \cdot \bar{C} C^M M! \\ &= (C(M-1)\delta_N)^M \cdot \bar{C} \end{aligned} \quad (3.21)$$

Bisher:  $M$  beliebig. Finde nun  $M_{opt} \in \{2, \dots, N\}$ , welches den Fehler “klein” macht. Suche zunächst ein reelles  $\bar{M}_{opt} \in [2, N] \subset \mathbb{R}$ . Hierzu setze

$$\bar{M}_{opt} := 1 + \frac{1}{C e \delta_N}$$

Wir sehen mit Abkürzung  $\lambda := \ln \mu_{max} - \ln \mu_{min} = 2 \ln \mu_{max}$

$$\begin{aligned} \frac{1}{Ce\delta_N} \geq 1 &\iff 1 \geq Ce \frac{\ln \mu_{max} - \ln \mu_{min}}{N-1} \iff N-1 \geq C \cdot e \cdot \lambda \\ &\iff N \geq Ce\lambda + 1 \end{aligned}$$

Also ist mit Forderung  $N_0 \geq C \cdot e \cdot \lambda + 1$  ist  $\bar{M}_{opt} \geq 2$ . Weiter:

$$\begin{aligned} 1 + \frac{1}{Ce\delta_N} &\iff \frac{1}{Ce\delta_N} \iff 1 \leq (N-1)(Ce\delta_N) \\ &\iff 1 \leq (N-1)C \cdot e \frac{\lambda}{N-1} \iff 1 \leq Ce\lambda \\ &\stackrel{3.40}{\iff} 1 \leq 2\mu_{max}^2 \cdot e \cdot 2 \ln \mu_{max} \end{aligned}$$

Also ist für  $\mu_{max}$  genügend groß  $\bar{M}_{opt} \leq N$ .

Insgesamt nun also  $\bar{M}_{opt} \in [2, N]$ .

Wegen  $C(\bar{M}_{opt} - 1)\delta_N = C \frac{1}{Ce\delta_N} \cdot \delta_N = \frac{1}{e}$  folgt

$$\left( C(\bar{M}_{opt} - 1)\delta_N \right)^{\bar{M}_{opt}-1} = \left( \frac{1}{e} \right)^{\frac{1}{Ce\delta_N}} = e^{-\frac{1}{Ce\delta_N}} = e^{-\frac{N-1}{Ce\lambda}} \leq e^{-\frac{N-1}{N_0-1}}$$

falls  $Ce\lambda \leq N_0 - 1$ , d. h.  $N_0 \geq Ce\lambda + 1$  (identische Forderung an  $N_0$  wie zuvor). Setze nun  $M_{opt} := \lfloor \bar{M}_{opt} \rfloor$  größte ganze Zahl kleiner/gleich  $\bar{M}_{opt}$

$$\Rightarrow M_{opt} \in \{2, \dots, N\}$$

$$\text{wg. } M_{opt} \leq \bar{M}_{opt} \Rightarrow C(M_{opt} - 1)\delta_N \leq C(\bar{M}_{opt} - 1)\delta_N \left( = \frac{1}{e} < 1 \right)$$

$$\text{und } M_{opt} > \bar{M}_{opt} - 1$$

folgt

$$(C(M_{opt} - 1)\delta_N)^{M_{opt}} \leq (C(\bar{M}_{opt} - 1)\delta_N)^{\bar{M}_{opt}} \leq e^{-\frac{N-1}{N_0-1}} \quad (3.22)$$

Damit insgesamt

$$\begin{aligned} \frac{\|\hat{u}(\hat{\mu}) - \hat{w}_N(\hat{\mu})\|_\mu}{\|\hat{u}(\hat{\mu})\|_\mu} &\stackrel{(3.20)}{\leq} \mu_{max} \sup_{z, \sigma} |g(z, \sigma) - \tilde{g}_N(z, \sigma)| \\ &\stackrel{(3.21)}{\leq} \mu_{max} \underbrace{\bar{C}}_{=\mu_{max}} \cdot (C(M_{opt} - 1)\delta_N)^{M_{opt}} \stackrel{(3.22)}{\leq} \mu_{max}^2 e^{-\frac{N-1}{N_0-1}} \end{aligned}$$

also (3.19) und damit Satz 3.36 gezeigt. □

**Definition 3.42** (Gram-Schmidt)

Seien  $\{u_i\}_{i=1}^n \in X$  lin. unabh. Dann ist *Gram-Schmidt Basis*  $\Phi_{GR} := \{\varphi_1, \dots, \varphi_n\}$  definiert durch

$$\bar{\varphi}_m := u_m - \sum_{i=1}^{m-1} \langle u_m, \varphi_i \rangle \varphi_i \quad , \quad \varphi_m := \frac{\bar{\varphi}_m}{\|\bar{\varphi}_m\|} \quad , \quad m = 1, \dots, n$$

und  $X_{GR}$  der zugehörige *Gram-Schmidt RB-Raum*

**Lemma 3.43** (Eigenschaften von  $\Phi_{GR}$ )

- i)  $\Phi_{GR}$  ist ONB
- ii)  $\text{span}\{u_i\}_{i=1}^n = X_{GR}$

*Beweis.* i) Normiertheit klar nach Definition

Orthogonalität per Induktion:

Sei  $\langle \varphi_i, \varphi_j \rangle = 0 \quad \forall \quad j < i$

Dann gilt für  $j < i + 1$ :

$$\begin{aligned} \langle \bar{\varphi}_{i+1}, \varphi_j \rangle &= \langle u_{i+1}, \varphi_j \rangle - \sum_{k=1}^{(i+1)-1} \langle u_{i+1}, \varphi_k \rangle \underbrace{\langle \varphi_k, \varphi_j \rangle}_{\delta_{kj} \text{ sowohl für } j < i \text{ als auch } j=i} \\ &= \langle u_{i+1}, \varphi_j \rangle - \langle u_{i+1}, \varphi_j \rangle = 0 \end{aligned}$$

also auch  $\langle \varphi_{i+1}, \varphi_j \rangle = \langle \frac{\bar{\varphi}_{i+1}}{\|\bar{\varphi}_{i+1}\|}, \varphi_j \rangle = 0$

- ii) “ $\supseteq$ ” klar nach Konstruktion
- “ $=$ ” folgt durch Dimensionsbetrachtung:

$$\dim \text{span}\{\varphi_i\}_{i=1}^n = n = \dim \text{span}\{u_i\}_{i=1}^n$$

□

**Bemerkung.**

- Algorithmus liefert also ONB, garantiert Stabilität des RB-Verfahren für symmetrisches  $a(\cdot, \cdot)$  gemäß 3.7
- Es existiert nur triviale Approximationsaussage, z. B. wegen ii):

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

Für Teilbasis  $\Phi_{GR,m} := \{\varphi_1, \dots, \varphi_m\}$ ,  $m < n$  werden  $\{u_i\}_{i=1}^m$  exakt approximiert über  $\{u_i\}_{i=m+1}^n$  weiß man nichts.

- Basis hängt von Reihenfolge der  $\{u_i\}_{i=1}^n$  ab, macht also nur Sinn, wenn diese eine natürliche Reihenfolge haben.

- Gram Schmidt Orthonormierung folgt häufig als “Postprocessing” für anderweitig erzeugte Basis, z. B. Lagrange-, Greedy-Basis, etc.

**Satz 3.44** (Berechnung von  $\Phi_{GR}$  über Gram-Matrix)

Seien  $\{u_i\}_{i=1}^m \subset X$  lin. unabh.,  $K = (\langle u_i, u_j \rangle)_{i,j=1}^n$  mit Cholesky-Zerlegung  $K = LL^T$ , d. h.  $L$  untere  $\Delta$ -Matrix mit positiver Diagonalen. Definiere  $A = (a_{ij})_{i,j=1}^n := (L^T)^{-1}$ . Dann ist die Gram-Schmidt ONB  $\Phi_{GR}$  äquivalent berechenbar durch

$$\varphi_j = \sum_{i=1}^j a_{ij} u_i \quad \text{für} \quad 1 \leq j \leq n$$

*Beweis.* Übung. □

### Proper Orthogonal Decomposition (POD)

**Definition 3.45** (Korrelationsoperator)

Sei  $\{u_i\}_{i=1}^n \subset X$ . Dann definieren wir den *empirischen Korrelationsoperator*  $R \in L(X, X)$  durch

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

**Bemerkung.**

- Linearität von  $R$  ist klar, Beschränktheit folgt wegen

$$\|Ru\| \leq \frac{1}{n} \sum \|u_i\|^2 \|u\| \Rightarrow \|R\| = \sup_{u \neq 0} \frac{\|Ru\|}{\|u\|} \leq \frac{1}{n} \sum \|u_i\|^2 < \infty$$

also  $R \in L(X, X)$

- Wir nennen ein  $A \in L(X, X)$  *kompakt* falls abgeschlossenes Bild der offenen Einheitskugel, d. h.  $\overline{A(B_1(0))}$ , kompakt ist
- Wir nennen  $A \in L(X, X)$  *selbstadjungiert* (genauer Hilbertraum-selbstadjungiert), falls  $\langle Au, v \rangle = \langle u, Av \rangle \quad \forall u, v \in X$

**Satz 3.46** (Spektralsatz)

Sie  $A \in L(X, X)$  kompakt & selbstadjungiert, dann existiert endliche oder abzählbar unendliche orthonormiertes System von Eigenvektoren  $\{\varphi_i\}_{i \in I}$ ,  $I \subseteq \mathbb{N}$  zu Eigenwerten  $\{\lambda_i\}_{i \in I} \subset \mathbb{R} \setminus \{0\}$  mit

$$Au = \sum_{i \in I} \lambda_i \langle u, \varphi_i \rangle \varphi_i \quad \forall u \in X$$

Falls  $I$  unendlich, so  $\lim_{i \rightarrow \infty} \lambda_i = 0$ .

*Beweis.* z. B. Alt: “lineare Funktionalanalysis” Satz 12.12 □

**Satz 3.47** (POD-Basis)

Zu  $\{u_i\}_{i=1}^n$  mit  $R$  aus 3.45 existiert orthonormierte Menge  $\{\varphi_i\}_{i=1}^{n'}$  von  $n' \leq n$  Eigenvektoren zu reellen Eigenwerten  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{n'} > 0$  mit

$$Ru = \sum_{i=1}^{n'} \lambda_i \langle \varphi_i, u \rangle \varphi_i \quad (3.23)$$

Für  $m = 1, \dots, n'$  definieren  $\Phi_{POD} := \Phi_{POD,m} := \{\varphi_i\}_{i=1}^m$  als *POD-Basis* und  $X_{POD} := X_{POD,m} := \text{span } \Phi_{POD,m}$  als POD-Raum.

*Beweis.*  $R$  hat endlich dimensionales Bild, also  $\overline{R(B_1(0))}$  abgeschlossen, beschränkt im endlich dimensionalen Raum, also kompakt.  $R$  ist selbstadjungiert, denn  $\langle Ru, v \rangle = \frac{1}{n} \sum_{i=1}^n \langle u_u, u \rangle \langle u, v \rangle = \langle u, Rv \rangle$ . Also existiert nach Spektralsatz 3.46 entsprechend endliches ONS, das (3.23) erfüllt. Dies kann insbesondere nicht unendlich sein, wegen endlichem Bild.  $\square$

**Bemerkung.**

- Die Projektion  $X \rightarrow X_{POD}$  wird in der statistischen Datenanalyse auch Hotelling-Transformation, Principal Component Analysis (PCA) oder Karhunen-Loève-Transformation genannt.
- Bezeichnung POD, als “proper”, ist Anlehnung an das französische “valeur propre” für Eigenwert.
- Wir nennen Basisvektoren von  $\Phi_{POD}$  auch POD-Moden.

**Illustration**

- $\{\varphi_i\}_{i=1}^{n'}$  ist ONB für  $\text{span}\{u_i\}_{i=1}^n$  aber nicht eindeutig (VZ oder vertauschen bei mehrfachen Eigenwerten)
- $\varphi_1$  ist Richtung höchster Varianz von  $\{u_i\}_{i=1}^n$   
 $\varphi_2$  ist Richtung höchster Varianz von  $\{P_{x_{POD,1}}^\perp u_i\}_{i=1}^n$
- Koordinaten der Daten in der POD-Basis sind unkorreliert  $\rightarrow$  Übung.
- $\{\varphi_i\}, \{\sqrt{\lambda_i}\}$  sind die Hauptachsen bzw. Achsenabschnitte des Ellipsoids  $\{\langle u, R^{-1}u \rangle = 1\}$
- Falls  $X = \mathbb{R}^H$  und  $\{u_i\}_{i=1}^n$  Realisierungen von  $n$  unabhängigen, identisch normalverteilten Zufallsvariablen mit Verteilung  $\mathcal{N}(\mu, \Sigma) := C \cdot \exp(- (x - \mu)^T \Sigma (x - \mu))$  mit Mittelwert  $\mu = 0$ , so ist  $R \in \mathbb{R}^{H \times H}$  guter Schätzer für  $\Sigma$ , insbesondere  $R \rightarrow \Sigma$  konvergiert für  $n \rightarrow \infty$  in geeignetem Sinne.

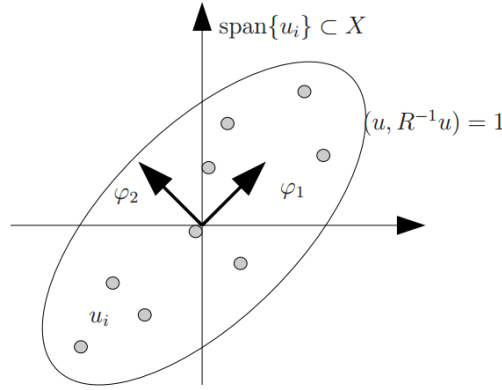


Abbildung 7: Ellipsoide aus Kovarianzoperator  
(aus B. Haasdonk, Reduzierte-Basis-Methoden, Skript zur Vorlesung SS 2011, Universität Stuttgart, IANS-Report 4/11, 2011.)

**Satz 3.48** (Berechnung von  $\Phi_{POD}$  über Gram-Matrix)

Sei  $\{u_i\}_{i=1}^n \subset X$  und  $K = (\langle u_i, u_j \rangle)_{i,j=1}^n$ . Dann sind äquivalent:

i)  $\varphi \in X$  ist Eigenvektor von  $R$  zu Eigenwert  $\lambda > 0$  mit Norm 1 und einer Darstellung

$$\varphi = \sum a_i u_i \quad \text{mit o. B. d. A. } a \in \ker(K)^\perp$$

ii)  $a = (a_i)_{i=1}^n \in \mathbb{R}^n$  ist Eigenvektor von  $\frac{1}{n}K$  zu  $\lambda > 0$  mit Norm  $\frac{1}{\sqrt{n\lambda}}$

*Beweis.* ii)  $\Rightarrow$  i)

Sei  $a$  Eigenvektor von  $\frac{1}{n}K$  zu Eigenwert  $\lambda$  mit  $\|a\| = \frac{1}{\sqrt{n\lambda}}$  also

$$\lambda a = \frac{1}{n} K a$$

Multiplikation der  $i$ -ten Komponenten mit  $u_i$  und Summieren ergibt

$$\sum_{i=1}^n u_i \lambda a_i = \sum_{i=1}^n u_i \frac{1}{n} \left( \underbrace{\sum_{j=1}^n \langle u_i, u_j \rangle a_j}_{(Ka)_i} \right)$$

Mit  $\varphi := \sum u_i a_i$  gilt also

$$\lambda \varphi = \frac{1}{n} \sum_{i=1}^n u_i \langle u_i, \varphi \rangle = R \varphi$$

Also  $\varphi$  Eigenvektor von  $R$  zu Eigenwert  $\lambda$ . Für Norm folgt

$$\|\varphi\|^2 = \langle \sum a_i u_i, \sum a_j u_j \rangle = a^T \underbrace{K a}_{n\lambda a} = n\lambda \cdot \|a\|^2 = 1$$



$K$  ist symmetrisch, also existiert vollständiges ONS von Eigenvektoren.  $\ker(K)$  wird aufgespannt von EV zu EW 0, also  $a \perp \ker(K)$ ,  $a$  EV zu  $\lambda > 0$ .

i)  $\Rightarrow$  ii):

Sie  $\varphi$  EV von  $R$  zu EW  $\lambda > 0$  und  $\|\varphi\| = 1$ . Sei  $\bar{a} \in \mathbb{R}^n$  mit  $\varphi = \sum \bar{a}_i u_i$  (existiert weil  $\varphi \in \text{Bild}(R) = \text{span}\{u_i\}_{i=1}^n$ ). Verschiebungen von  $\bar{a}$  um  $a^0 \in \ker(K)$  erhalten  $\varphi$ :

$$\begin{aligned}\varphi' &:= \sum (\bar{a}_i + a_i^0) u_i \Rightarrow \langle \varphi', u_k \rangle = \left\langle \sum_{i=1}^n \bar{a}_i u_i, u_k \right\rangle + \left\langle \sum_i a_i^0 u_i, u_k \right\rangle \\ &= \langle \varphi, u_k \rangle + \underbrace{\sum_{i=1}^n a_i^0 \langle u_i, u_k \rangle}_{Ka^0=0} \quad , \quad k = 1, \dots, n\end{aligned}$$

Also  $\varphi' = \varphi$ .

Wähle speziell  $a := \bar{a} - P\bar{a}$ ,  $P$  orthogonale Projektion auf  $\ker(K)$ .

$$\Rightarrow a \in \ker(K)^\perp, \quad P\bar{a} \in \ker(K) \quad \Rightarrow \quad \varphi = \sum \bar{a}_i u_i = \sum a_i u_i$$

i)  $\Rightarrow$  ii) o.B.d.A.  $a \in \ker(K)^\perp$

Da  $\varphi \in V$  zu  $\lambda > 0$  gilt:

$$\underbrace{\frac{1}{n} \sum_{i=1}^n \langle u_i, \sum_{j=1}^n a_j u_j \rangle u_i}_{R\varphi} = \lambda \varphi = \lambda \sum_j a_j u_j$$

Testen mit  $u_k$  liefert

$$\underbrace{\frac{1}{n} \sum_j \langle u_i, u_j \rangle \langle u_i, u_k \rangle a_j}_{(K^2 a)_k} = \lambda \underbrace{\sum_j a_j \langle u_j, u_k \rangle}_{(Ka)_k}$$

Also  $\frac{1}{n} K^2 a = \lambda Ka$  also  $Ka$  EV von  $\frac{1}{n} K$  zu EW  $\lambda$ . Dann ist schon  $a$  EV, denn  $a \in \ker(K)^\perp$ :

$$(*) \quad Ka \left( \frac{1}{n} K - \lambda \right) a = 0$$

$a \in \ker(K)^\perp$ ,  $Ka \in \ker(K)^\perp$  wegen Symmetrie  $\langle Ka, v \rangle = \langle a, Kv \rangle = 0 \quad \forall v \in \ker(K)$

$\Rightarrow (\frac{1}{n} K - \lambda)a \in \ker(K)^\perp$

aber auch wg. (\*)  $(\frac{1}{n} K - \lambda)a \in \ker(K)$

$\Rightarrow (\frac{1}{n} K - \lambda)a = 0$  also  $a$  EV von  $\frac{1}{n} K$  zu  $\lambda$ .

Wie im ersten Teil gilt

$$\begin{aligned}1 = \|\varphi\|^2 &= \sum a_i a_j \langle u_i, u_j \rangle = a^T Ka = a^T \cdot n\lambda a = n\lambda a \|a\|^2 \\ \Rightarrow \|a\| &= \frac{1}{\sqrt{n \cdot \lambda}}\end{aligned}$$

□

**Bemerkung.** Falls  $X$  endlichdimensional  $\dim(X) = H$ , kann daher POD entweder als teures EWP für  $R$  in  $X$  (Komplexität  $\mathcal{O}(H^3)$ ) oder, meist günstiger, als EWP für  $K$  (Komplexität  $\mathcal{O}(n^3)$ ) ermittelt werden.

Bezeichnung für letzteres ist auch “method of snapshots” (Sirovich, 1987) oder Kernel-PCA (Scholkopf & Smola, 2002). POD kann auch über Singulärwertzerlegung der Koeffizientenmatrix berechnet werden:

**Satz 3.49** (Berechnung für  $X = \mathbb{R}^H$  via SVD)

Sei  $X = \mathbb{R}^H$  mit  $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{\mathbb{R}^H}$ .  $U = [u_1, \dots, u_n] \in \mathbb{R}^{H \times n}$  Snapshot-Matrix mit Rang  $U = n'$  und

$$U = \Phi S V^T$$

eine verkürzte SVD, d.h.  $\Phi \in \mathbb{R}^{H \times n'}$ ,  $V \in \mathbb{R}^{n \times n'}$  orthonormale Spalten und  $S = \text{diag}(\sigma_1, \dots, \sigma_{n'}) \in \mathbb{R}^{n' \times n'}$  ( $\sigma_i$ : Singulärwerte) mit  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{n'} \geq 0$ .

Dann ist  $\Phi_{POD, n'} = \Phi$ .

*Beweis.* Sei  $\Phi = (\bar{\varphi}_1, \dots, \bar{\varphi}_n)$ . Nach Definition gilt  $Ru = \frac{1}{n} U U^T U \quad \forall u \in \mathbb{R}^H$ . Damit ist  $\bar{\varphi}_i$  EV von  $R$  zu  $i$ -ten Eigenwert  $\frac{1}{n} \sigma_i^2$ :

$$\begin{aligned} R \bar{\varphi}_i &= \frac{1}{n} U U^T \bar{\varphi}_i = \frac{1}{n} \Phi \underbrace{S V^T V}_I S \underbrace{\Phi^T \bar{\varphi}_i}_{e_i \in \mathbb{R}^{n'}} \\ &= \frac{1}{n} \Phi \underbrace{S^2 e_i}_{\sigma_i^2 e_i} = \frac{1}{n} \sigma_i^2 \bar{\varphi}_i \end{aligned}$$

Die EW  $\frac{1}{n} \sigma_i^2$  sind monoton fallend, also identisch sortiert wie EW von  $R$ , das heißt  $\lambda_i = \frac{1}{n} \sigma_i^2$  und  $\varphi_i = \bar{\varphi}_i$ .  $\square$

**Bemerkung.**

- obiges ist sehr eingänglich (“1-Zeilenbeweis”), aber algorithmisch nicht unbedingt besser, weil SVD auch durch EWP definiert (Numerik I)
- Verallgemeinerung für allg. HR  $X \rightarrow$  Übung (Blatt 5)

**Satz 3.50** (Approximationsfehler für  $X_{POD, m}$ )

Sei  $\{u_i\}_{i=1}^n \subset X$  und für  $Y \subset X$  Unterraum ist mittlerer quadratischer Fehler  $J(Y) := \frac{1}{n} \sum_{i=1}^n \|u_i - P_Y u_i\|^2$ . Dann gilt für den POD-Raum

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

mit  $\lambda_i$  EW von  $R$ .

*Beweis.* Sei  $\Psi = \{\Psi_1, \dots, \Psi_m\}$  ONB für  $Y$ . Dann folgt

$$\begin{aligned}
J(Y) &= \frac{1}{n} \sum_{i=1}^n \|u_i - P_Y u_i\|^2 = \frac{1}{n} \sum_{i=1}^n \left\| u_i - \sum_{j=1}^m \langle \Psi_j, u_i \rangle \Psi_j \right\|^2 \\
&= \frac{1}{n} \sum_{i=1}^n \|u_i\|^2 - \frac{2}{n} \sum_{i=1}^n \sum_{j=1}^m \langle u_i, \Psi_j \rangle^2 + \frac{1}{n} \sum_{i=1}^n \sum_{j,k} \langle \Psi_j, u_i \rangle \langle u_i, \Psi_k \rangle \underbrace{\langle \Psi_j, \Psi_k \rangle}_{=\sigma_{jk}} \\
&= \frac{1}{n} \sum_{i=1}^n \|u_i\|^2 - \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m \langle u_i, \Psi_j \rangle^2
\end{aligned}$$

Wegen  $u_i \in \text{Bild}(R) = X_{POD, n'}$  ist  $u_i = \sum_{j=1}^{n'} \langle \varphi_j, u_i \rangle \varphi_j$

$$\|u_i\|^2 = \sum_{j,k=1}^{n'} \langle \varphi_j, u_i \rangle \langle \varphi_k, u_i \rangle \underbrace{\langle \varphi_j, \varphi_k \rangle}_{=\sigma_{jk}} = \sum_{j=1}^{n'} \langle \varphi_j, u_i \rangle^2$$

also mittlerer quadratischer Projektionsfehler:

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

□

**Satz 3.51** (Bestapproximation durch  $X_{POD, m}$ )

Unter allen Räumen der Dimension  $m$  ist  $X_{POD, m}$  bzgl.  $J$  optimal

$$j(X_{POD, m}) = \inf_{\substack{Y \subseteq X \\ \dim Y = m}} J(Y)$$

*Beweis.* Übung.

□

**Bemerkung** (Zusammenfassung).

- POD liefert also orthonormale Basis, garantiert Stabilität des RB-Verfahrens (bei symmetrischem  $a(\cdot, \cdot)$ )

- Es existieren Approximationsaussagen bzgl. des mittleren quadratischen Projektionsfehlers, sogar Optimalität nachweisbar. Die POD Teilbasen ermöglichen Approximation aller Snapshots mit Fehlerkontrolle der abgeschnittenen Eigenwerte.
- Die POD-Basis hängt nicht von Reihenfolge der Snapshots ab.
- Die POD-Basen sind hierarchisch:

$$\Phi_{POD,m} \subseteq \Phi_{POD,m'} \quad \text{für } m \leq m'$$

- Die POD kann auch zur Erweiterung einer bestehenden ONB  $\Phi$  verwendet werden, indem  $\{\tilde{u}_i\}_{i=1}^n$ ,  $\tilde{u}_i = u_i - P_{\text{span}(\Phi)}u_i$  und eine POD Basis  $\tilde{\Phi}_{POD}$  hierfür berechnet wird. Dann ist  $\Phi \cup \tilde{\Phi}_{POD}$  eine erweiterte/neue ONB.
- Man kann POD auch als inkrementelles Verfahren mit 1D-Minimierung von  $J(Y)$  verstehen.

Sei  $\{u_i\}_{i=1}^n \subset X$ . Dann definiere

$$\bar{\varphi}_1 := \text{POD}_1(\{u_i\}_{i=1}^n) := \operatorname{arginf}_{\substack{\varphi \in X \\ \|\varphi\|=1}} \frac{1}{n} \sum_{i=1}^n \|u_i - \langle u_i, \varphi \rangle \varphi\|^2$$

$$\bar{X}_1 := \text{span}(\bar{\varphi}_1)$$

und für  $i = 2, \dots, n'$

$$\bar{\varphi}_i := \text{POD}_1(\{u_i - P_{\bar{X}_{i-1}}u_i\}),$$

$$\bar{X}_i := \text{span}(\bar{\varphi}_1, \dots, \bar{\varphi}_i)$$

Dann ist  $(\bar{\varphi}_1, \dots, \bar{\varphi}_m)$  POD-Basis (aus  $m$  Moden/Basisvektoren) (bis auf Rotation, Vorzeichen).

## Greedy-Verfahren

### Definition 3.52 (Greedy-Verfahren)

Sei  $S_{\text{train}} \subset \mathcal{P}$  "Trainingsmenge" von Parametern,  $\Delta(Y, \mu) \in \mathbb{R}^+$  für Teilräume  $Y \subset X$  und Parameter  $\mu \in \mathcal{P}$  ein "Fehlerindikator" und  $\epsilon_{\text{tol}} > 0$  eine Fehlertoleranz. Die Greedy-Basen  $\Phi_{GRE,m}$ , Greedy-Raum  $X_{GRE,m}$  und Sample-Menge  $S_m$  für  $m = 0, \dots, N$  sind iterativ definiert durch

$$S_0 = \emptyset, \quad X_{GRE,0} = \{0\}, \quad \Phi_{GRE,0} = \emptyset, \quad m := 0$$

Solange  $\epsilon_m := \max_{\mu \in S_{train}} \Delta(X_{GRE,m}, \mu) > \epsilon_{tol}$

$$\begin{aligned}\mu^{(m+1)} &:= \operatorname{maxarg}_{\mu \in S_{train}} \Delta(X_{GRE,m}, \mu) \\ S_{m+1} &:= S_m \cup \{\mu^{(m+1)}\} \\ \varphi_{m+1} &:= u(\mu^{(m+1)}) \text{ Lösung von } (P(\mu^{(m+1)})) \\ \Phi_{GRE,m+1} &:= \Phi_{GRE,m} \cup \{\varphi_{m+1}\} \\ X_{GRE,m+1} &:= X_{GRE,m} + \operatorname{span}(\varphi_{m+1}) \\ m &\leftarrow m + 1\end{aligned}$$

setze schließlich  $N := m$

**Bemerkung.**

- Erste Verwendung von Greedy-Verfahren für RB: Veroy, Prud'homme, Rovas, Patera 2003, seitdem "Standard"
- $\Phi_{GRE,m}$  ist Lagrange-RB zur Sample-Menge  $S_m$ , i. a. nicht orthonormal. Kann für numerische Stabilität mit Gram-Schmidt orthonormalisiert werden.
- Basen sind hierarchisch:  $\Phi_{GRE,m} \subset \Phi_{GRE,m'}, m \leq m'$
- In Literatur wird Suche nach  $\mu^{(1)}$  häufig umgangen, indem dieses beliebig aus  $S_{train}$  gewählt wird.
- $S_{train}$  wird häufig als strukturierte oder zufällige Menge aus  $\mathcal{P}$  mit *endlich* vielen Samples gewählt.
- Falls  $S_{train}$  zu klein, kann das RB-Modell overfitting aufweisen, d. h.

$$\sup_{\mu \in \mathcal{P}} \|u(\mu) - u_N(\mu)\| >> \sup_{\mu \in S_{train}} \|u(\mu) - u_N(\mu)\|$$

- Greedy-Verfahren ist also akkumulatives Verfahren, welches iterativ den "schlechtest"-aufgelösten Parameter  $\mu^{(m+1)}$  wählt,  $u(\mu^{(m+1)})$  berechnet, und als neuen Basisvektor hinzufügt. Insofern kann dies als approximative Lösung des Optimierungsproblems

$$\min_{\substack{Y \subset X \\ \dim Y = N}} \max_{\mu \in \mathcal{P}} \Delta(Y, \mu)$$

interpretiert werden: Statt maximieren über  $\mathcal{P} \rightsquigarrow$  maximieren über  $S_{train}$ , statt minimieren über  $X \subset X \rightsquigarrow$  iterative Sequenz von Räumen  $Y = X_{GRE,m}$ .

**Lemma 3.53** (Fehlerindikatoren, Terminieren des Verfahrens)

- i) Falls  $|S_{train}| = n < \infty$  und für alle  $\mu \in \mathcal{P}$  und  $Y \subset X$  gilt

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

so terminiert das Greedy-Verfahren mit  $N \leq n$  und

$$\max_{\mu \in S_{train}} \Delta(X_{GRE,n}, \mu) \leq \epsilon_{tol}$$

ii) Dies erfüllen z. B.

$$\Delta(Y, \mu) := \|u(\mu) - u_N(\mu)\|$$

$$\Delta(Y, \mu) := \|u(\mu) - P_y u(\mu)\|$$

$$\Delta(Y, \mu) := \Delta_N^{en}(\mu)$$

oder andere Fehlerschätzer, wobei  $X_N = Y$  gesetzt wird.

*Beweis.* Übung. □

**Korollar 3.54** (Fehlerraussage)

Für  $\Delta(Y, \mu) := \|u(\mu) - u_N(\mu)\|$  oder  $\Delta(Y, \mu) := \Delta_{N'}$  gilt für  $N' = 1, \dots, N$

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

*Beweis.* Klar nach Konstruktion. □

**Bemerkung** (Wahl der Fehlerindikatoren).

- Greedy-Verfahren hervorragendes Einsatzfeld für Fehlerschätzer, denn  $\Delta_N(\mu)$  kann sehr schnell für alle  $\mu \in S_{train}$  ausgerechnet werden, ohne dass alle  $u(\mu)$ ,  $\mu \in S_{train}$  berechnet werden müssen (im Gegensatz zu POD). Dadurch können sehr große Mengen  $S_{train}$  behandelt werden. Dies erhöht die Erwartung, dass  $\Phi_{GRE,N}$  auch für neue Parameter  $\mu \in \mathcal{P} \setminus S_{train}$  eine gute Approximation liefert.
- Wahl:  $\Delta(Y, \mu) := \|u(\mu) - P_y u(\mu)\|$ , orthogonaler Projektionsfehler  
 Motivation: Falls dies klein, so ist mit Céa auch RB-Fehler klein  
 Nachteile: Teuer auszuwerten, hochdimensionale Operation erfordert alle Snapshots  $u(\mu)$ ,  $\mu \in S_{train}$  müssen vorliegen, Größe von  $S_{train}$  hiermit eingeschränkt.  
 Vorteil: Terminieren ist garantiert. Approximationsraum ist entkoppelt von RB-Approximation, d. h. Verfahren kann angewandt werden ohne Vorliegen des RB-Verfahrens und ohne Fehlerschätzer.
- Wahl:  $\Delta(Y, \mu) := \|u(\mu) - u_N(\mu)\|$ , RB-Fehler  
 Motivation: Dies ist die ultimative Größe, welche kontrolliert werden muss, z. B. (3.7).  
 Nachteile: Wie bei Projektionsfehler: teuer, alle Snapshots  $\mu \in S_{train}$  vorberechnen,  $S_{train}$  Größe eingeschränkt.  
 Vorteile: Terminieren ist garantiert, Verfahren kann mit RB-Verfahren angewandt werden, für welche keine FS vorliegen.

- Wahl:  $\Delta(Y, \mu) = \Delta_N(\mu)$  [oder Energie-/rel. Fehlerschätzer]  
Nachteil: Falls Fehlerschätzer den Fehler stark überschätzt, kann der RB-Raum evtl. größer als nötig sein.  
Vorteile: schnell auswertbar, unabhängig von  $H$  denn Offline-Online. Es müssen nur  $N$  Snapshots berechnet werden,  $|S_{train}|$  kann sehr groß gewählt werden, Terminieren kann garantiert werden.
- Ziel-orientierte Indikatoren: Falls  $\Delta(Y, \mu)$  als Ausgabefehler  $|s(\mu) - s_N(\mu)|$  oder -Schranke  $\Delta_{N,s}$  gewählt wird, nennt man das Verfahren “goal-oriented”. Die Basis wird potentiell sehr klein und kann Ausgabe gut approximieren. Die Feldvariable  $u$  wird jedoch nicht notwendigerweise gut approximiert.
- Falls  $\Delta(Y, \mu)$  als Feldvariablen-Fehler  $\|u(\mu) - u_N(\mu)\|$ , PProjektionsfehler oder -schätzer gewählt wird, ist Verfahren nicht goal-oriented, die Basis wird größer sein, aber sowohl  $u$  als auch  $v$ , als auch beliebig andere Funktionale  $\tilde{s}$  gut approximiert.

**Bemerkung** (Reihenfolge).

- Greedy-Basis hängt meistens nicht von Reihenfolge der Parameter  $S_{train}$  ab. Nur falls zufällig das Maximum von  $\Delta(Y, \mu)$  mehrdeutig ist  $\rightsquigarrow$  Praktische Lösung: Wähle ersten Parameter, der maximales  $\Delta(Y, \mu)$  erzeugt.

**Bemerkung** (Bestimmung der Approximationsgüte/Overfitting). In Terminologie der Statistik/Maschinellen Lernens ist  $S_{train}$  eine *Trainingsmenge* und  $\epsilon_N$  aus Greedy-Verfahren der sogenannten *Trainingsfehler*.  $S_{train}$  muss  $\mathcal{P}$  gut repräsentieren, sollte möglichst groß gewählt werden. Falls  $S_{train}$  zu klein, oder unrepräsentativ für  $\mathcal{P}$  kann Overfitting auftreten.

Somit ist kleiner Trainingsfehler nicht hinreichend für gutes Modell. Modelle sollen daher nicht alleine anhand von Trainings-, sondern anhand unabhängiger Testmengen:

$$\epsilon_{test} = \max_{\mu \in S_{train}} \Delta(X_N, \mu) \quad , \text{ meistens zufällige Parametermenge}$$

**Bemerkung** (Monotonie).

- Im Allgemeinen gilt nicht  $\Delta(X_N, \mu) \geq \Delta(X_{n+1}, \mu)$
- Es kann daher vorkommen, dass  $(\epsilon_n)_{n=1}^N$  nichtmonoton ist.
- Falls Beziehung zu Bestapproximation gilt, d. h. für ein  $C > 0$  unabhängig von  $n$  gilt:

$$\Delta(X_n, \mu) \leq C \cdot \inf_{v \in X_n} \|u(\mu) - v\|$$

kann zumindest eine Beschränkung oder asymptotischer Abfall erwartet werden.

- In bestimmten Fällen kann Monotonie bewiesen werden:

**Satz 3.55** (Monotonie von  $(\epsilon_n)$ )

Das Greedy-Verfahren erzeugt monoton fallende Sequenzen  $(\epsilon_n)_{n \geq 1}$  falls:

- i)  $\Delta(Y, \mu) := \|u(\mu) - P_{X_n} u(\mu)\|$  oder
- ii)  $(P)$  ist compliant, d. h.  $l = f$  und  $a$  symmetrisch und  $\Delta(Y, \mu) = \|u(\mu) - u_N(\mu)\|_\mu$

*Beweis.* i) klar

- ii) folgt aus 3.11

□

**Bemerkung** (Konvergenz des Greedy-Verfahrens).

- Einige Jahre lang war Greedy-Verfahren ein in der Praxis gut funktionierendes Verfahren, jedoch ohne theoretische Erklärung wann/warum es funktioniert.
- Notwendiges Kriterium für Erfolg des Greedy-Verfahren: Kolmogorov  $n$ -Weite von  $\mathcal{M}$  muss (schnell) abfallen. Sei  $\Delta(Y, \mu)$  so gewählt, dass  $\Delta(Y, \mu) \geq \|u(\mu) - P_Y u(\mu)\|$

$$\Rightarrow \sup_{\mu \in \mathcal{P}} \Delta(Y, \mu) \geq \inf_{\substack{Y \subseteq X \\ \dim Y = n}} \sup_{\mu \in \mathcal{P}} \|u(\mu) - P_Y u(\mu)\| = d_n(\mathcal{M})$$

$\Rightarrow$  Falls  $\Phi_{GRE, n}$  gut, muss  $d_n(\mathcal{M})$  klein sein. Falls  $d_n(\mathcal{M})$  nicht klein, kann  $\Phi_{GRE, n}$  keine gute Approximation liefern.

- Spannend ist umgekehrte Frage, ob abfallendes  $d_n$  auch hinreichend für Gelingen des Greedy-Verfahrens.
- Antwort auf diese Fragen wurden in letzten Jahren gegeben:  
 (BMPPT 2012): Buffa, Maday, Patera, Prud'homme, Turinici: A-priori convergence of the greedy algorithm for the parameterized reduced basis method  
 M2AN, 46:595-..., 2012  
 (BCDDPW): Binev, Cohen, Dahmen, DeVore, Petrova, Wojtaszczyk: Convergence Rates for Greedy Algorithms in Reduced Basis Methods  
 SIAM J. Math. Anal., 43(3), 1455..., 2011.
- Die Hoffnung, ein Ergebnis der Form  $\epsilon_n \leq cd_1(\mathcal{M})$  zu erhalten kann (ohne weitere Annahme) leider nicht erreicht werden. Dies sieht man durch Vergleich von

$$\bar{d}_n(\mathcal{M}) = \inf_{\substack{Y \subseteq \text{span}(\mathcal{M}) \\ \dim Y = n}} d(Y, \mathcal{M}) \quad \text{und} \quad d_n(\mathcal{M})$$

Theorem 4.1 in (BCDDPW2011) besagt:

- i) Für jedes  $\mathcal{M}$  und  $n \geq 0$  gilt  $\bar{d}_n(\mathcal{M}) \leq (n+1)d_n(\mathcal{M})$



ii) Für jedes  $n > 0$  und  $\epsilon > 0$  existiert  $\mathcal{M}$ , s. d.

$$\bar{d}_n(\mathcal{M}) \geq (n - 1 - \epsilon)d_n(\mathcal{M})$$

Wegen  $\epsilon_n \geq \bar{d}_n(\mathcal{M})$  und ii) ist “direkter Vergleich” von  $\epsilon_n$  und  $d_n(\mathcal{M})$  mit  $C$  unabhängig von  $n$  nicht möglich.

- Lösung ist zusätzliche Annahmen von Raten des Abfalls von  $d_n$ , damit können ähnliche Abfallraten für  $\epsilon_n$  gezeigt werden, z. B. zeigen (BMPPT2012):

Für  $S_{train} = \mathcal{P}$  und  $\Delta(Y, \mu) = \|u(\mu) - P_Y u(\mu)\|$ :

$$\epsilon_n \leq 2^{n+1}(n+1)d_n(\mathcal{M})$$

Falls  $d_n$  schnell genug abfällt (z. B. exponentiell  $d_n(\mathcal{M}) \leq C \cdot e^{-\alpha n}$ ) so folgt dann auch exponentieller Abfall von  $\epsilon_n$  (mit anderem  $\alpha$ )

- Ein verbessertes Ergebnis (ohne Faktor  $(n+1)$ ) und ein Ergebnis für Fall algebraischer (polynomiell in  $N^{-1}$ ) Konvergenz liefert (BCDPW), welches wir in unserer Notation formulieren (ohne Beweis).

**Satz 3.56** (Greedy Konvergenzraten)

Sei  $S_{train} = \mathcal{P}$  kompakt und  $\Delta(Y, \mu)$  so gewählt, dass ex. ein  $\gamma \in (0, 1]$  mit

$$\|u(\mu^{(n+1)}) - P_{X_n} u(\mu^{(n+1)})\| \geq \gamma \sup_{u \in \mathcal{M}} \|u - P_{X_n} u\| \quad (3.24)$$

- i) (algebraische Konvergenz) Falls  $d_n(\mathcal{M}) \leq M \cdot n^{-\alpha}$  für geeignetes  $\alpha$ ,  $M > 0$  und alle  $n \in \mathbb{N}$  und  $d_0(\mathcal{M}) \leq M$  dann gilt

$$\epsilon_n \leq C \cdot M n^{-\alpha}, \quad n > 0$$

mit explizit berechenbarer Konstante  $C$ .

- ii) (exponentielle Konvergenz) Falls  $d_n(\mathcal{M}) \leq M \cdot e^{-an^\alpha}$  für  $n \geq 0$ ,  $M, a, \alpha > 0$  dann gilt

$$\epsilon_n \leq C M e^{-cn^\beta}, \quad n \geq 0$$

mit  $\beta := \frac{\alpha}{\alpha+1}$  und geeignete Konstanten  $C, c > 0$ .

**Bemerkung.** “Quasi-Optimalität des Greedy-Verfahrens: bis auf Konstante so gut wie optimale Approximation.

**Bemerkung** (“strong” vs “weak” greedy).

- Für  $\gamma = 1$  nennt man das Verfahren “strong greedy”. Wird nur durch die Wahl

$$\Delta(Y, \mu) := \|u(\mu) - P_Y u(\mu)\|$$

realisiert.

- Für  $\gamma < 1$  nennt man das Verfahren “weak greedy” d.h. statt schlechtest-approximiertes Element wird ein einigermaßen schlecht approximiertes Element gewählt zur Basisgenerierung.
- Achtung  $\gamma \neq \gamma(\mu)$  Stetigkeitskonstante

Interessant ist Frage, ob Verwendung von Fehlerschätzern Bedingung (3.24) erfüllt für geeignetes  $\gamma$ . Für  $\Delta(Y, \mu) := \Delta_N(\mu)$  kann dies positiv beantwortet werden.

**Satz 3.57** ( $\Delta_N$  liefert weak Greedy)

Das Greedy-Verfahren mit Fehlerindikator  $\Delta(Y, \mu) := \Delta_N(\mu)$  stellt weak greedy Verfahren dar mit Konstante

$$\gamma := \frac{\bar{\alpha}^2}{\bar{\gamma}^2}$$

mit  $\bar{\alpha}, \bar{\gamma}$  uniforme untere/obere Schranke für Koerzivitäts-/Stetigkeitskonstante.

*Beweis.* Lemma von Cea 3.9, Fehlerschranke 3.13 und Effektivitätsschranke 3.16 gelten für alle Räume  $X_n$ ,  $n \geq 1$  also

$$\begin{aligned} \|u(\mu^{(n+1)}) - P_{X_n} u(\mu^{(n+1)})\| &= \inf_{v \in X_n} \|u(\mu^{(n+1)}) - v\| \\ &\geq \frac{\alpha(\mu)}{\gamma(\mu)} \|u(\mu^{(n+1)}) - u_N(\mu^{(n+1)})\| \\ &\stackrel{3.16}{\geq} \frac{\alpha(\mu)}{\gamma(\mu)\eta_N(\mu)} \cdot \Delta_N(\mu^{(n+1)}) \end{aligned}$$

Behauptung folgt mit

$$\frac{\alpha(\mu)}{\gamma(\mu)\eta_N(\mu)} \stackrel{3.16}{\geq} \frac{\alpha(\mu)}{\gamma(\mu)} \frac{\bar{\alpha}}{\bar{\gamma}} \geq \frac{\bar{\alpha}^2}{\bar{\gamma}^2} =: \gamma$$

und

$$\Delta_N(\mu^{(n+1)}) = \sup_{\mu \in \mathcal{P}} \Delta_N(\mu) \stackrel{3.13}{\geq} \sup_{\mu \in \mathcal{P}} \|u(\mu) - u_N(\mu)\| \geq \sup_{\mu \in \mathcal{P}} \|u(\mu) - P_{X_n} u(\mu)\|$$

□

**Bemerkung.**

- Für thermischen Block  $B_1 = 2$ ,  $B_2 = 1$  gesehen:  $d_n$  fällt exponentiell d.h. hier liefert Greedy-Verfahren exponentielle Konvergenz.
- “Lücke” zwischen Theorie & Praxis ist jedoch noch, dass  $S_{train} \neq \mathcal{P}$  weil nur endliche Mengen  $S_{train}$  betrachtet werden können.
- In der Praxis beobachtet man jedoch auch für allgemeines  $B_1, B_2$  und solchen endlichen  $S_{train}$  konvergenz.

### Numerische Beispiele:

**demos\_chapter3(5)** Illustration von Gram-Schmidt ONB aus `demos_chapter(3)` d.h.  $B_1 = B_2 = 3$  und nur  $\mu_1$  variiert.  $\varphi_1$  ist normierter Snapshot,  $\varphi_2, \dots, \varphi_8$  weisen stärker werdende Gradienten auf mit lokalen Strukturen um Kanten von  $\Omega_1$ .

**demos\_chapter3(6)**  $B_1 = B_2 = 2$ ,  $\mu \in \mathcal{P} = [0.5, 2]^4$   
Greedy-Basis mit zufälliger Menge  $S_{train}$ ,  $|S_{train}| = 1000$ . Fehlerindikator  $\Delta(Y, \mu) = \Delta_N(\mu)$ , Gram-Schmidt ON in jeder Iteration. Testmenge  $S_{test}$ ,  $|S_{test}| = 100$ . Bestimmung von maximalem Testfehler und -Schätzer.  
 $\Rightarrow$  schöne exponentielle Konvergenz von

$$\max_{\mu \in S_{test}} \|u(\mu) - u_N(\mu)\| \quad \text{und} \quad \max_{\mu \in S_{test}} \Delta_N(\mu)$$

Schätzer ist sehr nah an echtem Fehler (gute Effektivität).

**demos\_chapter3(7)** Effekt bei steigendem  $p = B_1 \cdot B_2$   
 $B_1 = B_2 = 2, 3, 4$ ,  $\mu \in \mathcal{P} = [0.5, 2]^p$ , Greedy wie in vorigem Beispiel.  
(Achtung 10 Minuten Laufzeit)  
Illustration des Trainingsfehlers  $(\epsilon_n)_{n \geq 1}$   
 $\Rightarrow$  Exponentielle Konvergenz, aber schlechtere Exponenten für größere  $p$

### Bemerkung (Trainingsmenge-Wahl).

- Die Trainingsmenge sollte möglichst repräsentativ für  $\mathcal{P}$  sein, kann aber nicht beliebig groß sein aus Laufzeitgründen. Sollte nicht zu klein gewählt werden, um nicht Overfitting zu bewirken. Sorgfältige Wahl von  $S_{train}$  kann also entweder Qualität des RB-Modells oder die Offline-Laufzeit verbessern. Hierzu gibt es einige Ansätze & Modifikationen des Greedy-Verfahrens:
- “Multistage-Greedy”: Wähle sehr große Menge  $S_{train}$ , zerlege diese in Sequenz größerer Mengen

$$S_{train}^{(0)} \subset S_{train}^{(1)} \subset \dots \subset S_{train}^{(m)} = S_{train}$$

Erzeuge  $\Phi_{GRE}^{(0)}$  aus  $S_{train}^{(0)}$ , dann erweitere diese Basis durch Greedy-Verfahren auf  $S_{train}^{(1)}$ , etc. Effekt ist wesentliche Beschleunigung des Greedy-Verfahrens für  $S_{train}$ . Die meisten Iterationen werden nur mit kleiner Trainingsmenge durchgeführt (schnell), nur wenige Iterationen für  $S_{train}^{(m)}$  erforderlich (teuer).

Ref.: Sen: Reduced-Basis Approximation and A Posteriori Error Estimation for Many-Parameter Heat Conduction Problems, Numerical Heat Transfer, Part B: Fundamentals 54(5): 369-389, 2008.

- Randomisiertes Greedy: Statt fester Menge  $S_{train}$  der Größe  $N_{train}$  in allen Iterationen, wähle in jeder Iteration eine neue Trainingsmenge der Größe  $N_{train}$ . Dadurch

wird praktisch eine Trainingsmenge der Größe  $N \cdot N_{train}$  in der Basisgenerierung verwendet.

Ref.: [HSZ2013]: Hesthaven, Stamm, Zhang: Efficient greedy algorithms for high-dimensional parameter spaces with applications to empirical interpolation and reduced basis methods. M2AN, 2013.

- Saturierungs-Annahme (?): Unter Annahme, dass ein Fehlerindikator für ein Parameter sich in einer Sequenz von Basiserweiterungen höchstens um Faktor  $C_s$  verschlechtert, besteht folgende Beschleunigungsmöglichkeit:

Für feste Menge  $S_{train}$  wird jeder Parameter  $\mu$ , der im Laufe des Greedy-Verfahren  $\Delta_N(\mu) \leq \frac{\epsilon_{tol}}{C_s}$  erfüllt, markiert und künftige Fehlerschätzer nicht mehr berechnet, da  $\mu$  bereits präzise erfasst. [HSZ2013] mit weiteren technischen Schnörkeln.

- Adaptive Trainingsmengen-Erweiterung:  
Idee: Übertragen des adaptiven FEM-Schemas “Solve, Estimate, Mark, Refine” auf das Parametergebiet:

Initiale Trainingsmenge (grob)  $S_{train}^{(0)}$  ist Menge der Knoten eines Gitters auf  $\mathcal{P}$ . Auf  $S_{train}^{(0)}$  wird ein Greedy-Verfahren mit “early stopping” angewandt, d.h. das Greedy-Verfahren wird abgebrochen, sobald Overfitting detektiert wird, d.h.  $\frac{E_{val}}{E_{train}}$  zu groß wird, wobei  $E_{val}$ ,  $E_{train}$  den aktuell maximalen Fehlerindikator über einer Validationsmenge (zufällig) und  $S_{train}$  darstellen. Sobald Overfitting detektiert wird, werden für alle Gitterelemente Fehlerindikatoren bestimmt (z.B. Fehlerschätzer im Mittelpunkt), ein Anteil  $\Theta \in (0,1]$  der Elemente mit größten Indikatoren zur Verfeinerung markiert, das Parametergebiet verfeinert und seine Knoten ergeben erweiterte Trainingsmenge  $S_{train}^{(1)}$ . Dies wird wiederholt, bis  $\epsilon_{tol}$  erreicht wird.

Ergebnis ist gleichverteilter Fehler und sehr problemangepasste Wahl von  $S_{train}$  z.B. führt dieser Algorithmus automatisch zu Verfeinerungen in wichtigen Bereichen. Bei Diffusion z.B. in bereichen kleiner Diffusionsparameter.

Ref.: [HDO11] Haasdonk, Dihlmann, Ohlberger: A Training Set and Multiple Basis Generation Approach for Parametrized Model Reduction Based on Adaptive Grids in Parameter Space. MCMDS, 17 : 423-442, 2011.

- Greedy mit Optimierung: Statt großer Menge  $S_{train}$  wird kleines  $S_{train}$  gewählt. Jedes  $\mu \in S_{train}$  wird als Startwert eines Optimierungsproblems gewählt. Aus den  $N_{train}$  lokalen Optima wird  $\mu^{(n+1)}$  als nächster Snapshotparameter gewählt.

Ref.: Urban, Volkwein, Zeeb: Greedy Sampling using Nonlinear Optimization. Kapitel in: Quarteroni, Rozza: Reduced Order Methods for Modeling and Computational Reduction, Springer MS&A Serie, 2014.

**Bemerkung** (Partitionsansätze). Bei komplexen Problemen kann die für gewünschtes  $\epsilon_{tol}$  erforderliche Basisgröße  $N$  zu groß sein. Generell verhalten sich  $\epsilon_{tol}$  und  $N$  gegenläufig und können nicht unabhängig gewählt werden.

Idee: Partitionierung des Parametergebiet durch Bisektions- oder strukturierte Gitter. Für jedes Teilgebiet wird eine Basis mit dem Greedy-Verfahren erzeugt. Diese Basen sind jeweils kleiner als einzelne globale Basis. In der Online-Phase wird zu  $\mu \in \mathcal{P}$  das geeignete Teilgebiet bestimmt und dessen RB zur Simulation verwendet. Indem das Parametergitter adaptiv genügend fein gewählt wird, kann sowohl  $\epsilon_{tol}$  als auch die maximale Basisgröße  $N_{max}$ , d.h. Online-Laufzeit vorgeschrieben werden.

“Haken”: Offline-Phase teuer (Rechenzeit & Daten)

Ref.:

- a) Bisektionsgitter: Eftang, Patera, Ronquist: An “hp” Certified Reduced Basis Method for Parametrized Elliptic Partial Differential Equations. SJSC, 32(6) : 3170-3200, 2010.
- b) Hexaeder-Gitter: [HDO11]

### 3.5 Primal-Duale RB-Verfahren

#### Motivation

- Erinnerung: Für nichtsymmetrische, non-compliant Probleme konnten wir nur  $\Delta_{N,s}(\mu)$  bereitstellen, welcher nur linear in  $\|v_r\|$  skaliert und wir haben die Unmöglichkeit von Effektivitätsschranken für die Ausgabe gesehen (ohne weitere Annahmen).
- Stattdessen für compliant Fall skaliert  $\bar{\Delta}_{N,s}$  quadratisch mit  $\|v_r\|$  und wir konnten Effektivitätsschranken zeigen.
- Dieser Abschnitt: Verbesserte Ausgabeschätzung für allgemeine nichtsymmetrische oder nicht-compliant Probleme (aber immer noch keine Effektivitätsschranken)
- $(P(\mu))$  und  $(P_N(\mu))$  werden weiter benötigt als “primale Probleme”.

**Definition 3.58** (Duales volles Problem  $(P^{\text{du}}(\mu))$ )

Seien  $a, f, l$  wie in  $(P(\mu))$  gegeben. Dann ist für  $\mu \in \mathcal{P}$  gesucht  $u^{\text{du}}(\mu) \in X$  als Lösung von

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

#### Bemerkung.

- Offensichtlich “negatives Ausgabefunktional” als rechte Seite und Vertauschen der Test- / Lösungsargumente in  $a(\cdot, \cdot)$ .
- Wohlgestelltheit für  $a$  koerziv: Existenz & Eindeutigkeit & Stabilität via Lax-Milgram.
- Duales Problem wird nur formell als Referenz verwendet zu welchem wir den dualen Fehlerabschätzer verwenden.

- Literatur zu diesem Abschnitt:

[Ro03]: Rovas: Reduced-Basis Output Bound Methods for Parametrized Partial Differential Equations, PhD-Thesis, MIT, 2003.

### Dualer RB-Raum

- Wir nehmen an, dass wir einen dualen RB-Raum gewählt haben:

$$X_N^{\text{du}} \subset X, \quad X_N^{\text{du}} = \text{span } \phi^{\text{du}}, \quad \dim X_N^{\text{du}} = N^{\text{du}}$$

welcher duale Lösungen  $u^{\text{du}}(\mu)$  gut approximiert.

- Es ist i.A.  $X_n^{\text{du}} \neq X_N$ ,  $N^{\text{du}} \neq N$ ,  $X_N^{\text{du}}$  kann durch Greedy / POD etc. Verfahren aus Snapshots  $u^{\text{du}}(\mu)$  erzeugt werden.

#### Definition 3.59 (Primal-Duales RB-Problem ( $P'_N(\mu)$ ))

Sei ein Problem ( $P(\mu)$ ) gegeben,  $X_N$ ,  $X_N^{\text{du}}$  primaler & dualer RB-Raum. Zu  $\mu \in \mathcal{P}$  sei  $u_N(\mu) \in X_N$  *primale RB-Lösung* wie in ( $P_N(\mu)$ ), d.h.

$$a(u_N(\mu), v; \mu) = f(v; \mu) \quad \forall v \in X_N$$

$u_N^{\text{du}}(\mu) \in X_N^{\text{du}}$ . Sei *duale RB-Lösung*, d.h.  $a(v, u_N^{\text{du}}(\mu); \mu) = -l(v; \mu)$ ,  $\forall v \in X_N^{\text{du}}$ , und die RB-Ausgabe  $s'_N(\mu) \in \mathbb{R}$  gegeben durch

$$s'_N(\mu) = l(u_N(\mu); \mu) - r(u_N^{\text{du}}(\mu); \mu) \quad (3.25)$$

wobei  $r(\cdot; \mu) \in X'$  das *primale Residuum*, d.h.  $r(v; \mu) = f(v; \mu) - a(u_N(\mu), v; \mu)$ . Weiter benötigen wir das *duale Residuum*  $r^{\text{du}}(\cdot; \mu) \in X'$  definiert durch

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

#### Bemerkung.

- Wohlgestelltheit wieder klar mit Lax-Milgram.
- In FEM-Literatur existiert der Begriff “dual-weighted residual” (DWR), ähnlich wie oben das Residuum und mit dualer Lösung kombiniert in (3.25).
- Im Vergleich zu ( $P_N(\mu)$ ) haben wir  $s'_N(\mu) = s_N(\mu) - r(u_N^{\text{du}}(\mu); \mu)$ , somit stellt  $r(u_N^{\text{du}})$  ein Korrekturterm dar, der die verbesserte Ausgabeschätzung liefert.
- Wie im primalen Fall ist auch duales Residuum rechte Seite der dualen Fehlergleichung

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

- Reproduktion von Lösungen gilt analog zu  $(P_N(\mu))$ :

Falls  $u(\mu) \in X_N$ ,  $u_N^{\text{du}}(\mu) \in X_N^{\text{du}}$

$$\Rightarrow u_N(\mu) = u(\mu), \quad u_N^{\text{du}}(\mu) = u^{\text{du}}(\mu)$$

und  $s'_N(\mu) = s(\mu)$ . Letzteres sieht man:

$$s(\mu) - s'_N(\mu) = l(u) - l(\underbrace{u_N}_u) + r(\underbrace{u^{\text{du}}}_{a(u, u^{\text{du}})}) - a(\underbrace{u_N}_u, u^{\text{du}}) = 0$$

**Satz 3.60** (Beziehung zur Bestapproximation)

i) Für den dualen Fehler gilt

$$\|u^{\text{du}}(\mu) - u_N^{\text{du}}(\mu)\| \leq \frac{\gamma(\mu)}{\alpha(\mu)} \inf_{v \in X_N^{\text{du}}} \|u^{\text{du}}(\mu) - v\|$$

ii) Für den Ausgabefehler gilt

$$\begin{aligned} |s(\mu) - s'_N(\mu)| &= |a(u - u_N, u^{\text{du}} - u_N^{\text{du}})| \leq \gamma(\mu) \|u - u_N\| \|u^{\text{du}} - u_N^{\text{du}}\| \\ &\leq \frac{\gamma(\mu)^3}{\alpha(\mu)^2} \inf_{v \in X_N} \|u - v\| \cdot \inf_{v \in X_N^{\text{du}}} \|u^{\text{du}} - v\| \end{aligned} \quad (3.26)$$

*Beweis.*

i) Céa

ii)

$$\begin{aligned} s(\mu) - s'_N(\mu) &= l(u) - l(u_N) + r(u_N^{\text{du}}) \\ &= \underbrace{l(u - u_N)}_{-a(u - u_N, u^{\text{du}})} + \underbrace{f(u_N^{\text{du}})}_{a(u, u_N^{\text{du}})} - a(u_N, u_N^{\text{du}}) \\ &= -a(u - u_N, u^{\text{du}}) + a(u - u_N, u_N^{\text{du}}) \\ &= -a(u - u_N, u^{\text{du}} - u_N^{\text{du}}) \end{aligned} \quad (3.27)$$

Somit erste Gleichheit in (3.26), Ungleichheit in (3.27) dann klar wegen Stetigkeit und  $2 \times \text{Céa}$ .

□

In a-priori-Schranke (3.26) sehen wir den “multiplikativen Effekt”, somit ist  $s'_N$  tatsächlich gute Schätzung und es besteht die Hoffnung dies durch a-posteriori Schranken zu verifizieren. Zunächst ganz analog zu primalem Problem eine a-posteriori Schranke für dualen Fehler:

**Satz 3.61** (Duale a-posteriori Fehlerschranke & Effektivitätsschranke)

i)

$$\|u^{\text{du}}(\mu) - u_N^{\text{du}}(\mu)\| \leq \Delta_N^{\text{du}}(\mu) := \frac{\|r^{\text{du}}(\mu)\|_{X'}}{\alpha_{LB}(\mu)}$$

ii)

$$\eta_N^{\text{du}}(\mu) := \frac{\Delta_N^{\text{du}}(\mu)}{\|u^{\text{du}}(\mu) - u_N^{\text{du}}(\mu)\|} \leq \frac{\gamma_{UB}(\mu)}{\alpha_{LB}(\mu)} \leq \frac{\bar{\gamma}}{\bar{\alpha}}$$

*Beweis.* Genauso wie für  $(P_N(\mu))$ . □

Damit folgt gewünschte Schranke für Ausgabefehler:

**Satz 3.62** (Primal-dualer Ausgabefehlerschätzer)

Für alle  $\mu$  gilt

$$|s(\mu) - s'_N(\mu)| \leq \Delta'_{N,s}(\mu) := \frac{\|r(\mu)\|_{X'} \cdot \|r^{\text{du}}(\mu)\|_{X'}}{\alpha_{LB}(\mu)}$$

*Beweis.* Mit (3.26) folgt mit  $e := u - u_N$ ,  $e^{\text{du}} = u^{\text{du}} - u_N^{\text{du}}$

$$|s - s'_N| = |a(e, e^{\text{du}})| = |r(e^{\text{du}})| \leq \|r\|_{X'} \cdot \|e^{\text{du}}\| \leq \|r\|_{X'} \Delta_N^{\text{du}} = \frac{\|r\| \|r^{\text{du}}\|}{\alpha_{LB}} = \Delta'_{N,s}$$

□

**Bemerkung.**

- Also “multiplikativer Effekt” in  $\Delta'_{N,s}$  erreicht.
- Dies liefert ein Kriterium zur Wahl von  $N$ ,  $N^{\text{du}}$ : Wähle diese s.d.  $\|r\| \approx \|r^{\text{du}}\|$ , damit quadratischer Effekt auch numerisch realisiert wird.
- Man kann feststellen, dass ähnlich zu  $\Delta_{N,s}$  auch  $\Delta'_{N,s}$  ohne weitere Annahmen keine Effektivitätsschranke liefert:  $\Delta'_{N,s}$  kann ungleich 0 sein, während  $s - s'_N = 0$ .

Wähle  $v_l \perp v_f \in X \setminus \{0\}$ ,  $X_N = X_N^{\text{du}} \perp \{v_f, v_l\}$

$$a(u, v) := \langle u, v \rangle, \quad f(v) := \langle v_f, v \rangle, \quad l(v) := \langle v_l, v \rangle$$

$$\Rightarrow u = v_f, \quad u^{\text{du}} = -v_l, \quad u_N = 0, \quad u_N^{\text{du}} = 0, \quad e = v_f, \quad e^{\text{du}} = -v_l$$

$$\Rightarrow r \neq 0, \quad r^{\text{du}} \neq 0 \quad \Rightarrow \quad \Delta'_{N,s} \neq 0$$

$$\text{aber } s - s'_N = -a(e, e^{\text{du}}) = \langle v_f, v_l \rangle = 0.$$

- Erinnerung: Im Compliant Fall hatten wir definiert/gezeigt in 3.18

$$0 \leq s - s_N \leq \bar{\Delta}_{N,s}(\mu) := \frac{\|r\|^2}{\alpha_{LB}}$$

und Effektivitätsschranke erreicht.



- Im compliant Fall ist primal-dualer Ansatz überflüssig, denn für  $X_N = X_N^{\text{du}}$ :

$$s'_N(\mu) = s_N(\mu) \quad \text{und} \quad \Delta'_{N,s}(\mu) = \bar{\Delta}_{N,s}(\mu)$$

Mit  $l = f$  und Symmetrie folgt:  $a(v, u) = a(u, v) = f(v) = l(v)$

$$\Rightarrow u = -u^{\text{du}}, \text{ analog } u_N = -u_N^{\text{du}}$$

$$\text{also } r(v) = f(v) - a(u_N, v) = l(v) + a(u_N^{\text{du}}, v) = l(v) + a(v, u_N^{\text{du}}) = -r^{\text{du}}(v)$$

$$\Rightarrow \|r\| = \|r^{\text{du}}\| \quad \Rightarrow \quad \Delta'_{N,s} = \bar{\Delta}_{N,s}$$

$$\text{Weiter ist } r(u_N^{\text{du}}) = -r(u_N) = 0 \Rightarrow s'_N = l(u_N) - r(u_N^{\text{du}}) = l(u_N) = s_N.$$

**Bemerkung** (Basisgenerierung).

- Erste Möglichkeit: Separate Greedy-Verfahren für  $X_N, X_N^{\text{du}}$  mit *identischem*  $\epsilon_{\text{tol}}$ , um quadratischen Effekt zu bewirken.
- Zweite Möglichkeit: Kombiniertes Greedy-Verfahren zur simultanen Erzeugung von  $X_N, X_N^{\text{du}}$ , indem  $\Delta'_{N,s}$  als Fehlerschätzer gewählt wird und  $u(\mu^{(n)}), u^{\text{du}}(\mu^{(n)})$  als Anreicherung in  $X_N, X_N^{\text{du}}$  hinzugefügt werden.

**Bemerkung** (Offline-Online für  $s'_N$ ). Im Fall separierbarer Parameterabhängigkeit folgt Offline/Online für  $A_N, f_N, l_N, \|r\|^2, \|r^{\text{du}}\|^2$  analog zu §3.3. Zerlegung für Korrekturterm in  $s'_N(\mu)$  ergibt sich ähnlich aus  $u_N^{\text{du}} = \sum u_{N,n}^{\text{du}} \varphi_n^{\text{du}}$  mit dualer Basis  $\Phi^{\text{du}} = \{\varphi_1^{\text{du}}, \dots, \varphi_{N^{\text{du}}}^{\text{du}}\}$ ,  $u_N^{\text{du}} = \left(u_{N,n}^{\text{du}}\right)_{n=1}^{N^{\text{du}}}$

$$r(u_N^{\text{du}}) = \sum_{q=1}^{Q_r} \Theta_r^q(\mu) r^q(u_N^{\text{du}}) = \sum_{q=1}^{Q_r} \sum_{n=1}^{N^{\text{du}}} \Theta_r^q(\mu) u_{N,n}^{\text{du}}(\mu) \underbrace{r^q(\varphi_n^{\text{du}})}_{\langle v_r^q, \varphi_n^{\text{du}} \rangle}$$

$$\Rightarrow \text{Offline: } G_{r,\text{du}} := \left(\langle v_r^q, \varphi_n^{\text{du}} \rangle\right)_{q=1, n=1}^{Q_r, N^{\text{du}}}$$

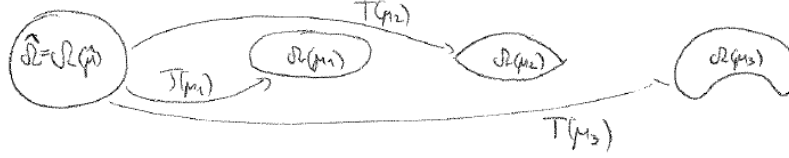
$$\text{Online: } r(u_N^{\text{du}}) = \Theta_r^q(u)^T G_{r,\text{du}} u_N$$

### 3.6 Geometrieparametrisierung

#### Motivation

- Neben Koeffizientenfunktionen in elliptischen PDEs oder Randwerten können auch Geometrieparametrisierungen behandelt werden.
- Hohe Anwendungsrelevanz: Beispiel Fahrzeugkarosserie oder Flügelprofil zur Optimierung des Widerstandsbeiwerts.
- Problem: Zu  $\mu \in \mathcal{P}$  sei Gebiet  $\Omega(\mu) \in \mathbb{R}^d$  und Lösungsraum  $X(\mu) := H_0^1(\Omega(\mu))$  parameterabhängig. Löse  $(P(\mu))$  nach  $u(\mu) \in X(\mu)$ .

- Hindernis: Snapshots  $u(\mu)$  lassen sich nicht linear kombinieren, Konstruktion eines  $X_N$  unklar.
- Idee/Lösung: Wahl eines Referenzparameter  $\hat{\mu} \in \mathcal{P}$ , Referenzgeometrie  $\hat{\Omega} := \Omega(\hat{\mu})$  und  $T(\mu) : \hat{\Omega} \rightarrow \Omega(\mu)$  invertierbare Abbildung. (“Geometrieabbildung”, “Referenzabbildung”)



- Mittels  $T(\mu)$  oder  $T^{-1}(\mu)$  sind Lösungen vergleichbar:

$$x \in \Omega(\mu) \Leftrightarrow T^{-1}(x; \mu) =: \hat{x} \in \hat{\Omega}$$

Falls  $T$  geeignete Regularität, so ist für  $u$ :

$$\hat{u}(\hat{x}; \mu) := u(T(\hat{x}; \mu); \mu) \Rightarrow \hat{u}(\cdot; \mu) \in X(\hat{\mu})$$

- Definiere  $\hat{X} := X(\hat{\mu})$ , zu  $\mu \in \mathcal{P}$  suche  $\hat{u}(\mu) \in \hat{X}$  als Lösung eines geeigneten  $(\hat{P}(\mu))$ , dann  $u(x; \mu) := \hat{u}(T^{-1}(x; \mu); \mu)$  Lösung von  $(P(\mu))$ .
- Damit ist RB-Behandlung klar. Suche  $\hat{X}_N \subset \hat{X}$  und RB-Lösung  $\hat{u}_N(\mu) \in \hat{X}_N$  wie in §3.1-3.5. Dann ist  $u_N(x; \mu) := \hat{u}_N(T^{-1}(x; \mu); \mu) \approx u(\mu)$ .

**Referenz:** [RHP08]: Rozza, Huynh, Patera: Reduced Basis Approximation and A Posteriori Error Estimation for Affinely Parametrized Elliptic Coercive Partial Differential Equations - Application to Transport and Continuum Mechanics, Archives of Computational Methods in Engineering, 15(3) : 229-275, 2008.

**Definition 3.63** (Stückweise affine Geometrietransformation)

Sei  $\Omega(\mu) \subseteq \mathbb{R}^d$  parameterabhängiges Gebiet mit Partition  $\{\Omega_k(\mu)\}_{k=1}^K$ , d.h.:

$$\Omega_k(\mu) \cap \Omega_{k'}(\mu) = \emptyset \text{ für } k \neq k', \quad \text{und} \quad \overline{\Omega(\mu)} = \bigcup_{k=1}^K \overline{\Omega_k(\mu)}$$

Wähle  $\hat{\mu} \in \mathcal{P}$  und  $\hat{\Omega} := \Omega(\hat{\mu})$ ,  $\hat{\Omega}_k := \Omega_k(\hat{\mu})$ . Wir nennen  $T(\mu) : \hat{\Omega} \rightarrow \Omega(\mu)$  stückweise affine Geometrietransformation falls ex.  $T_k(\mu) : \hat{\Omega}_k \rightarrow \Omega_k(\mu)$  affin und bijektiv, d.h.

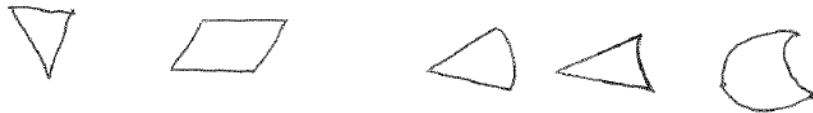
$$T_k(\hat{x}; \mu) := M_k(\mu) \cdot \hat{x} + t_k(\mu), \quad M_k(\mu) \in \mathbb{R}^{d \times d} \text{ regulär, } t_k(\mu) \in \mathbb{R}^d$$

und  $T$  ist stückweise durch  $\{T_k\}$  definiert via  $T(\mu)|_{\hat{\Omega}_k} = T_k(\mu)$  und ist für jedes  $\mu$  stetig auf  $\bigcup_{k=1}^K \partial \hat{\Omega}_k$  fortsetzbar, d.h.

$$T_k(\hat{x}; \mu) = T_{k'}(\hat{x}; \mu) \quad \forall \hat{x} \in \partial \hat{\Omega}_k \cap \partial \hat{\Omega}_{k'}$$

### Bemerkung.

- Insbesondere also  $T(\mu) \in C(\hat{\Omega}, \Omega(\mu))$ , also stetig bzgl.  $x$  (nicht notwendig bzgl.  $\mu$ ).
- Unter  $T_k$  für  $\hat{\Omega}_k \subseteq \mathbb{R}^2$  ist Bild von Dreieck ein Dreieck, entsprechend Formerhaltung von  $n$ -Eck, Parallelogramm, Ellipsen. In höheren Dimensionen analog für Simplex, Parallelepipiped, Ellipsoid.
- Als  $\hat{\Omega}_k \subseteq \mathbb{R}^2$  werden meist Dreiecke, Rechtecke, Parallelogramme oder allgemeiner “elliptic triangles” oder “curvy triangles”.



- Um  $T_k$  zu bestimmen, reicht es, in  $\Omega \subseteq \mathbb{R}^2$  3 korrespondierende nicht ko-lineare Punkte zu kennen ( $3 \times 2$  Gleichungen für 4+2 Unbekannte), in  $\Omega \subseteq \mathbb{R}^3$  entsprechend 4 nicht ko-planare Punkte ( $4 \times 3$  Gleichungen für 9+3 Unbekannte)
- Wir nennen  $\{\hat{\Omega}_k\}_{k=1}^K$  auch “Makro-Gitter” der Geometrieparametrisierung. Typischerweise wird FEM-Gitter eine Verfeinerung dieses Gitters sein.
- Stückweise affine Geometrieparametrisierungen sind kompatibel mit Sobolev/FEM-Lösungsraum:

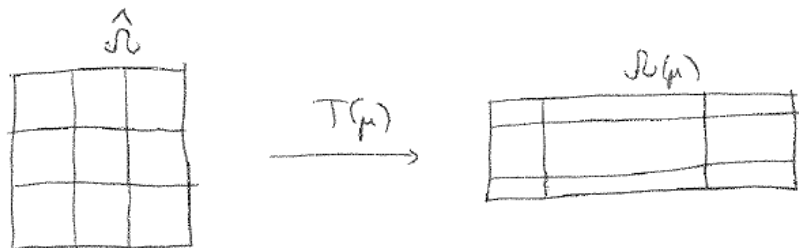
$$\hat{u} \in H_0^1(\hat{\Omega}) \quad \Leftrightarrow \quad u := \hat{u} \circ T^{-1}(\mu) \in H_0^1(\Omega(\mu))$$

Falls  $\mathcal{T}_h$  das aus  $\hat{\mathcal{T}}_h$  mit  $T(\mu)$  transformierte Gitter ist, gilt

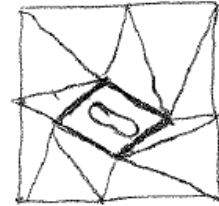
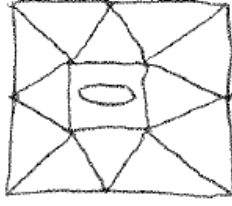
$$\hat{u}_h \in \mathbb{P}_{m,0}(\hat{\mathcal{T}}_h) \quad \Leftrightarrow \quad u_h := \hat{u}_h \circ T^{-1}(\mu) \in \mathbb{P}_{m,0}(\mathcal{T}_h)$$

weil Polynome von Grad  $m$  unter affinen Abbildungen wieder Polynome von Grad  $m$  ergeben.

### Beispiele



- i) Falls  $\hat{\Omega}$ ,  $\Omega(\mu)$ ,  $\Omega_k$  Rechtecke sind, so sind auch  $\Omega_k(\mu)$  notwendig Rechtecke, denn für  $T_k$  nur achsenparallele Streckung möglich (keine Rotation oder Scherung des mittleren Quadrates).

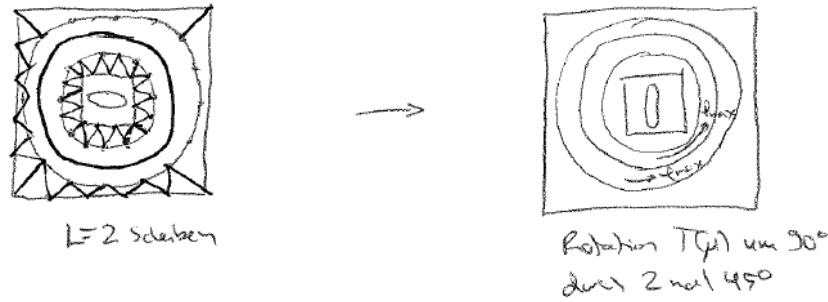


- ii) Durch Dreiecke sind auch Rotationen / Scherungen des mittleren Quadrats möglich. Man wird meist  $T(\mu)$  auch stetig bzgl.  $\mu$  wählen, z.B.  $\mu \hat{=}$  Punktkoordinate, Verschiebungsparameter, Rotationswinkel oder Skalierungsfaktor, etc. Das mögliche Intervall für Rotationswinkel des mittleren Quadrats ist durch Regularitätsanforderung der  $M_k$  (Nichtdegenerieren der  $\Omega_k$ ) beschränkt, z.B. hier keine Drehung um  $90^\circ$  möglich, weil "mittlere" Dreiecke an jeder Kante degenerieren.
- iii) Durch genügend feines Makro-Gitter und Verwendung von Zwischenschichten kann Rotationswinkel *beliebig* vergrößert werden:

Kreisring mit Radien  $r_2 < r_1$



$T(\mu)$  Rotation des inneren Kreises um Winkel  $\mu \in (-\varphi_{min}, +\varphi_{max})$  mit  $\varphi_{max} = \varphi_{max}(\frac{r_1}{r_2})$  realisierbar, analog  $\varphi_{min}(\frac{r_1}{r_2})$ . Durch genügend viele Punkte auf beiden Kreisen ist stückweise affine Geometrieparametrisierung definierbar, welche Makro-Dreiecksgitter regulär transformiert. Bei Verwendung von  $L$  solcher Ringe ist also insgesamt Rotation um  $\mu \in (-L\varphi_{min}, +L\varphi_{max})$  für inneres  $n$ -Eck realisierbar. Durch konforme Fortsetzung der Triangulierung der Scheiben auf Innen- & Außengebiet ist somit beliebige Rotation in Beispiel ii) möglich.



### Transformation auf Referenzgebiet

- Sei allgemeine elliptische PDE zweiter Ordnung gegeben. Zu  $\mu \in \mathcal{P}$  suche  $u : \Omega(\mu) \rightarrow \mathbb{R}$  als Lösung von

$$\begin{aligned} -\nabla \cdot (A(\mu) \nabla u) + \nabla \cdot (b(\mu) u) + c(\mu) u &= q(\mu) && \text{auf } \Omega(\mu) \\ u &= 0 && \text{auf } \partial\Omega(\mu) \end{aligned}$$

- Schwache Form: Zu  $\mu \in \mathcal{P}$  suche  $u \in H_0^1(\Omega(\mu))$  sodass

$$\begin{aligned} \int_{\Omega(\mu)} (A(\mu) \nabla u) \cdot \nabla v - u(b(\mu) \cdot v) + c(\mu) uv &= \int_{\Omega} qv \quad \forall v \in H_0^1(\Omega(\mu)) \\ \Leftrightarrow \underbrace{\int_{\Omega(\mu)} (\nabla u^T, u) \underbrace{\begin{pmatrix} A(\mu) & 0 \\ b(\mu)^T & c \end{pmatrix}}_{B(\mu)} \begin{pmatrix} \nabla v \\ v \end{pmatrix}}_{a(u,v;\mu)} &= \underbrace{\int_{\Omega} qv}_{f(v;\mu)} \quad \forall v \in H_0^1(\Omega(\mu)) \end{aligned}$$

- Dies ist Instanz von  $(P(\mu))$  wenn noch beliebiges Ausgabefunktional gewählt wird. Wir ignorieren die Ausgabe hier.
- Unter geeigneten Bedingungen an  $B(\mu)$  ist  $(P(\mu))$  koerzives Problem mit stetiger Linear-/Bilinearform auf  $X(\mu)$ . (siehe NumPDE 14/15)
- Transformation der Gradienten auf  $\Omega_k$ :

Sei  $\hat{u}(\hat{x}) \in H_0^1(\hat{\Omega})$ ,  $u(x) := \hat{u}(T^{-1}(x; \mu))$ ,  $x \in \Omega_k$

$$\begin{aligned} \Rightarrow \nabla_x u(x) &= (D_x u)^T = (D_{\hat{x}} \hat{u}(T^{-1}(x; \mu)) \cdot D_x T^{-1}(x; \mu))^T \\ &= (\nabla_x \underbrace{\hat{u}(T^{-1}(x; \mu))}_{\hat{x}})^T \cdot M_k^{-1})^T = M_k^{-T} \cdot \nabla_{\hat{x}} \hat{u}(\hat{x}) \end{aligned}$$

- Transformation der Bilinearform-Komponenten:

Mit  $v := \hat{v} \circ T^{-1}$

$$\begin{aligned} & \int_{\Omega_k} (\nabla u^T, u) B(\mu) \begin{pmatrix} \nabla v \\ v \end{pmatrix} dx \\ &= \int_{\hat{\Omega}_k} (\nabla_{\hat{x}} \hat{u}^T, \hat{u}) \begin{pmatrix} M_k^{-1} & 0 \\ 0 & 1 \end{pmatrix} B(\mu) \begin{pmatrix} M_k^{-T} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \nabla_{\hat{x}} \hat{v} \\ \hat{v} \end{pmatrix} |\det M_k| d\hat{x} \end{aligned}$$

analog Linearform.

**Satz 3.64** (Transformation auf Referenzgebiet,  $(\hat{P}(\mu))$ )

Zu  $\mu \in \mathcal{P}$  löst  $u(\mu) \in X(\mu)$  das Problem  $a(u(\mu), v; \mu) = f(v; \mu)$ ,  $\forall v \in X(\mu)$  mit

$$a(u, v; \mu) = \sum_{k=1}^K \int_{\Omega_k(\mu)} (\nabla u^T, u) B(\mu) \begin{pmatrix} \nabla v \\ v \end{pmatrix}, \quad f(v; \mu) = \sum_{k=1}^K \int_{\Omega_k(\mu)} q(\mu) v$$

genau dann, wenn  $\hat{u}(\hat{x}; \mu) := u(T(\hat{x}; \mu); \mu)$  löst

$$\hat{a}(\hat{u}(\mu), \hat{v}; \mu) = \sum_{k=1}^K \int_{\hat{\Omega}_k} (\nabla \hat{u}^T, \hat{u}) \hat{B}^k(\mu) \begin{pmatrix} \nabla \hat{v} \\ \hat{v} \end{pmatrix} = \sum_{k=1}^K \int_{\Omega_k} \hat{q}^k(\mu) \hat{v} =: \hat{f}(\hat{v}; \mu) \quad \forall \hat{v} \in \hat{X}$$

mit

$$\hat{B}^k(\mu) = \begin{pmatrix} M_k(\mu)^{-1} & 0 \\ 0 & 1 \end{pmatrix} B(\mu) \begin{pmatrix} M_k(\mu)^{-T} & 0 \\ 0 & 1 \end{pmatrix} |\det M_k|$$

und

$$\hat{q}^k(\mu) := q(\mu) \cdot |\det M_k(\mu)|$$

**Bemerkung** (Separierbare Parameterabhängigkeit).

- Durch Vorgabe von  $M_k(\mu) \in \mathbb{R}^{d \times d}$  ist also  $M_k^{-1}(\mu) \in \mathbb{R}^{d \times d}$  explizit bekannt, ebenso  $|\det M_k(\mu)|$ , also die Matrixeinträge/Determinante als Koeffizientenfunktion verwendbar.
- Mit  $B(\mu)$  separierbar parametrisch ist also auch  $\hat{B}^k(\mu)$  separierbar parametrisch mit  $Q_{\hat{B}} \leq (d^2 + 1)Q_B$ , mit  $q(\mu)$  separierbar parametrisch ist auch  $\hat{q}^k(\mu)$  separierbar parametrisch mit  $Q_{\hat{q}} = Q_q$ .
- Durch Ausmult. ist auch  $\hat{a}$  separierbar parametrisch

$$(\hat{a}^q(\hat{u}, \hat{v}))_{q=1}^{Q_{\hat{a}}} := \left( \int_{\hat{\Omega}_1} \frac{\partial}{\partial \hat{x}_1} \hat{u} \frac{\partial}{\partial \hat{x}_1} \hat{v} \cdot \hat{B}_{11}^{1,1}, \dots, \int_{\hat{\Omega}_1} \frac{\partial}{\partial \hat{x}_1} \hat{u} \frac{\partial}{\partial \hat{x}_1} \hat{v} \cdot \hat{B}_{11}^{1, q_{\hat{B}}}, \dots, \int_{\hat{\Omega}_k} \hat{u} \hat{v} \cdot \hat{B}_{(d+1), (d+1)}^{K, Q_{\hat{B}}} \right)$$

mit  $Q_{\hat{a}} = (d+1)^2 K \cdot Q_{\hat{B}}$  und entsprechenden Koeff.  $(\Theta_{\hat{a}}^q)_{q=1}^{Q_{\hat{a}}}$ , die sich aus  $M_k(\mu)$ ,  $\Theta_{B_{ij}^k}^q(\mu)$ , etc. ergeben.

- $Q_{\hat{a}}$  kann sehr groß sein. Diese Anzahl kann reduziert werden, falls es Koeff.  $\Theta_{\hat{a}}^q(\mu)$  gibt, die

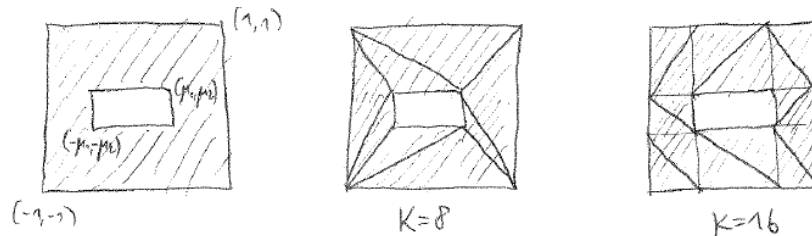
- Null sind, z. B. bei eingeschränkten Transformationen (nur Skalierung, nur Translation)  $\Rightarrow (M_k)_{1,2} = (M_k)_{2,1} = 0$

- linear abhängig sind. Falls z. B.  $\Theta_{\hat{a}}^q(\mu) = C \cdot \Theta_{\hat{a}}^{q'}(\mu)$

$$\Rightarrow \Theta_{\hat{a}}^q(\mu) \hat{a}^q(\cdot, \cdot) + \Theta_{\hat{a}}^{q'}(\mu) \hat{a}^{q'}(\cdot, \cdot) = \underbrace{\Theta_{\hat{a}}^{q'}(\mu)}_{\rightarrow 1 \text{ Komponente statt } 2} (C \hat{a}^q(\cdot, \cdot) + \hat{a}^{q'}(\cdot, \cdot))$$

also  $Q_{\hat{a}}$  reduziert.

- “Zusammenfassen” von linear abh. Termen kann automatisiert werden durch symbolische Arithmetik  $\rightsquigarrow$  rbMIT Software-Paket
- Für mögl. kleines  $Q_{\hat{a}}$  ist nicht unbedingt kleines  $K$  sinnvoll, sondern möglichst identisch transformierte Teilgebiete.



- $K = 16$  Zerlegung führt auf kleineres  $Q_{\hat{a}}$  als  $K = 8$  Zerlegung.

**Bemerkung** (Inhomogene Neumann-Randbedingung). (bzw. ähnliche Argumentation bei Fluss-Kurven-Integralen als Ausgabe)

Sei  $\partial\Omega = \Gamma_D \cup \Gamma_N$  mit  $\Gamma_N \neq \emptyset$  Neumann Rand,  $g_N : \Gamma_N \rightarrow \mathbb{R}$  Neumann RW.

Für  $v \in H_{\Gamma_D}^1(\Omega(\mu))$  lautet  $(P(\mu))$  rechts:

$$f(v; \mu) = \int_{\Gamma_N} g_N v$$

Ein Fluss-Kurvenintegral als Ausgabe z. B.

$$l(v; \mu) = \int_{\Gamma_N} (A(\mu) \Delta v) \cdot n$$

- Für Transformation dieser Integrale ist Rand-Geom. abb  $T_\Gamma : \hat{\Gamma}_N \rightarrow \Gamma_N(\mu)$  und entsprechende Jacobi-Matrix/Determinante
- Problem besteht, falls  $\Gamma_N$  gekrümmt  
 $\Rightarrow n = n(x; \mu)$  ortsabhängig.  
 $\Rightarrow$  Randintegralterm trotz separierbarem  $A(\mu)$  eventuell nicht mehr separierbar parametrisch. Explizite Referenzabbildung notwendig und “hochdimensional”. Auswertung des Integrals, d. h. ohne Offline-Online-Zerlegung.

**Bemerkung** (Fehlermaße).

- Im Raum  $\hat{X} = X(\hat{\mu})$  hat Norm  $\|(\hat{u}(\mu))\|_{H^1(\hat{\Omega})}$  keine physikalische Bedeutung für Funktionen  $u(x; \mu) := \hat{u}T^{-1}(x; \mu; \mu)$ , kann sich wesentlich von  $\|u(\mu)\|_{H^1(\Omega(\mu))}$  unterscheiden.
- Die Energie(norm?) ist plausibler, kann jedoch bei starken Größenvariationen von  $\Omega(\mu)$  unterschiedliche Größenordnungen annehmen.
- Im Greedy-Verfahren macht relativer Energienorm-Fehler/-Schätzer Sinn, da dieser Größenvariationen von  $\Omega(\mu)$  ausgleicht.

**Bemerkung** (Weitere Möglichkeiten der Geometrieparametrisierung).

- In der Literatur existieren weitere Methoden, anstelle der stückweise affinen Geometrieparametrisierung: z. B. “Free-Form-Deformation” (Gitterförmig angeordnete Kontrollpunkte für Interpolation mit Bernstein-Polynomen), “Radiale Basisfunktionen Interpolation” (beliebige Platzierung von wenigen Kontrollpunkten und RBF Interpolation der Koordinatenfunktionen), “Transfinite Mapping”
- Diese Ansätze führen meist auf nicht-separierbare Parameterabhängigkeit in  $(\hat{P}(\mu))$ . Mit Techniken aus § 4 (z. B. “empirische Interpolationen”) sind approx. separierbare Darstellungen konstruierbar.



## 4 Allgemeinere Lineare Probleme

### 4.1 Allgemeine Parameterabhängigkeit

Es folgt zunächst eine allgemeine Approximationsmethode für parametrische Funktionen, welche anschließend für RB-Behandlung von allgemeinen parametrisierten Problemen verwendet werden kann.

#### Empirische Interpolation (EI)

##### Motivation

- § 3 zeigte Relevanz der separierbaren Parameterabhängigkeit für effiziente Offline/Online-Zerlegung und Glattheit der Lösung  $u(\mu)$  bzgl.  $\mu$ .
- Gesucht: Approximationsverfahren für parametrische Funktion

$$g : \Omega \times \mathcal{P} \rightarrow \mathbb{R}$$

der Form

$$g(x; \mu) \approx I_\mu(g(\cdot; \mu))(x) = \sum_{m=1}^M \Theta_g^m(\mu) g^m(x)$$

mit skalaren Funktionen  $\Theta_g^m(\mu)$  und “kollaterale reduzierter Basis”  $Q_\mu = \{g^m\}_{m=1}^M$

- Statt allg. approx. Räume (z. B. FEM-Räume, zu hohe Dimension) oder Taylor-Ansatz (nur lokale Approx.) wird wieder Snapshot-basierter Ansatz gewählt, d. h.  $Q_M \subset \text{span}\{g(\cdot, \mu)|_{\mu \in S_{train} \subset \mathcal{P}}\}$
- Die empirische Interpolation ist eine Möglichkeit. Details finden sich in

BMNP04 Barrault, Maday, Nguyen, Patera: An ‘empirical interpolation’ method: application to efficient reduced-basis discretization of partial differential equations

MNPP07 Maday, Nguyen, Patera, Pau: A general, multipurpose interpolation procedure: the magic points

#### Definition 4.1 (Empirische Interpolation)

Sei  $G \subset C^0(\bar{\Omega}, \mathbb{R})$  Menge von zu interpolierenden Funktionen. Für  $\mu \in \mathbb{N}$ ,  $M \leq \dim(\text{span}(G))$  definiere rekursiv Interpolationspunktemenge  $T_\mu \subset \bar{\Omega}$  und die kollaterale Basis  $Q_\mu \subset \text{span}(G)$

$$\begin{aligned} M = 1 : \tilde{q}_1 &:= \arg\max_{g \in G} \|g\|_\infty \\ x_1 &:= \arg\max_{x \in \bar{\Omega}} |\tilde{q}_1(x)| \\ T_1 &:= \{x_1\} \\ q_1 &:= \frac{\tilde{q}_1}{\tilde{q}_1(x_1)} \\ Q_1 &:= \{q_1\} \end{aligned}$$

$$\begin{aligned}
M > 1 : \tilde{q}_M &:= \operatorname{argmax}_{g \in G} \|g - I_{M-1}(g)\|_\infty \\
r_M &:= \tilde{q}_M - I_{M-1}\tilde{q}_M \\
x_M &:= \operatorname{argmax}_{x \in \bar{\Omega}} |r_M(x)| \\
T_M &:= T_{M-1} \cup \{x_M\} \\
q_M &:= \frac{r_M}{r_M(x_M)} \\
Q_M &:= Q_{M-1} \cup \{q_M\}
\end{aligned}$$

wobei  $I_M : C^0(\bar{\Omega}, \mathbb{R}) \rightarrow \operatorname{span}(Q_M)$  den Interpolationsparameter zu Punkten  $T_M$  bezeichnet, d. h.  $I_M(g)(x_i) = g(x_i) \ \forall g \in C^0(\bar{\Omega}, \mathbb{R}), i = 1, \dots, M$ .

**Bemerkung.**

- In der Praxis werden obige Optimierungsprobleme zur Bestimmung von  $\tilde{q}_m, x_m$  durch einfache lineare Suche realisiert, indem endliche  $\bar{\Omega}$  und  $G$  betrachtet werden.
- Es sind Mehrdeutigkeiten von  $\tilde{q}_m$  und  $x_m$  möglich, welche durch Aufzählung der Mengen und “Wahl des ersten Auftretens” eindeutig werden.
- Die Basis  $Q_M$  ist weder orthogonal noch nodal, aber hierarchisch, d. h.  $Q_{M-1} \subseteq Q_M$ .

**Bemerkung.**

- $Q_M$  sind beschränkt

$$1 = q_m(x_m) = \|q_m\|_\infty$$

Für analytische Untersuchungen wird später die nodale Basis  $\zeta_M \subset \operatorname{span}(Q_M)$  zu Knoten  $T_M$  betrachtet, welche jedoch nicht mehr hierarchisch sind, d. h.  $\zeta_{M-1} \not\subseteq \zeta_M$

- Die Erzeugung von  $Q_M$  und  $T_M$  kann als Greedy-Minimierungsstrategie für

$$\min_{\substack{X_M \subset \operatorname{span}(G) \\ \dim(X_M)=M}} \max_{g \in G} \|g - I_M(g)\|_\infty \text{ interpretiert werden.}$$

$$\begin{aligned}
& T_M \subset \bar{\Omega} \\
& |T_M|=M
\end{aligned}$$

Es kann jedoch kein monotoner Fehlerabfall garantiert werden.

**Satz 4.2** (Berechnung der Interpolation)

Seien  $T_M = \{x_1, \dots, x_M\}$  und  $Q_M = \{q_1, \dots, q_M\}$  gemäß Def. 4.1 gegeben. Dann ist die Matrix  $\underline{Q}_M := (\underline{q}_j(x_i))_{i,j=1}^M \in \mathbb{R}^{M \times M}$  eine untere Dreiecksmatrix mit 1-Diagonale, also regulär. Sei  $g \in C^0(\bar{\Omega}, \mathbb{R}), \underline{g}_M = (g(x_i))_{i=1}^M \in \mathbb{R}^M, \underline{\alpha}_M := (\alpha_i)_{i=1}^M \in \mathbb{R}^M$  Lösung von

$$\underline{Q}_M \underline{\alpha}_M = \underline{g}_M$$

Dann ist die Interpolierte von  $g$  gerade

$$I_M(g) = \sum_{i=1}^M \alpha_i \underline{q}_i \tag{4.1}$$

*Beweis.* Seien  $i, j = 1, \dots, M$

$$i = j : (\underline{Q}_M)_{ii} = \underline{q}_i(x_i) = \frac{r_i(x_i)}{r_i(x_i)} = 1$$

$$j > i : (\underline{Q}_M)_{ij} = \underline{q}_j(x_i) = \frac{r_j(x_i)}{r_i(x_j)} = 0$$

weil  $r_j(x_i) = \tilde{q}^{(x_i)} - I_{j-1}(\tilde{q}_j)(x_i) = 0$  da  $I_{j-1}$  Interpolierende zu  $T_{j-1}$  und  $x_i \in T_{j-1}$ .

Also  $\underline{Q}_M$  untere Dreiecksmatrix mit 1-Diagonale.

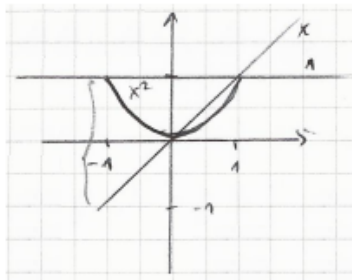
Für 4.1 zeige Übereinstimmung von beiden Seiten in allen Interpolationspunkten  $T_M$ :

Für  $i = 1, \dots, M$  ist

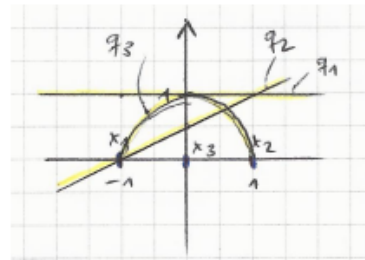
$$\sum_{j=1}^M \alpha_j \underline{q}_j(x_i) = \sum_{j=1}^M \alpha_j (\underline{Q}_M)_{ij} = (\underline{Q}_M \underline{\alpha}_M)_i = (g_M)_i = g(x_i) = I_M(g)(x_i)$$

□

**Beispiel (EI für Polynome)** Sie  $G = \{1, x, x^2\}$  Monome auf  $\bar{\Omega} = [-1, 1]$ .



**G**



**Q<sub>M</sub>**

Dann ist

$$\tilde{q}_1 := \operatorname{argmax}_{g \in G} \|g\|_\infty \text{ beliebig, z. B. } \tilde{q}_1(x) = 1$$

und

$$x_1 = \operatorname{argmax}_{x \in [-1, 1]} |\tilde{q}_1(x)| \text{ beliebig, z. B. } x_1 = -1$$

somit ergibt sich

$$\underline{q}_1(x) = \frac{\tilde{q}_1(x)}{\tilde{q}_1(x_1)} = 1$$

Dann ist

$$\tilde{q}_2 := \operatorname{argmax}_{g \in G} \|g - I_1(g)\|_\infty = x,$$

$$r_2 = \tilde{q}_2 - I_1(\tilde{q}_2) = x + 1,$$

$$x_2 := \operatorname{argmax}_{x \in [-1,1]} |r_2(x)| = 1 \text{ und}$$

$$\underline{q}_2(x) = \frac{r_2(x)}{r_2(x_2)} = \frac{1}{2}(x+1)$$

schließlich

$$\tilde{q}_3 := \operatorname{argmax}_{g \in G} \|g - I_2(g)\|_\infty = x^2$$

$$r_3 = x^2 - 1, \quad x_3 := \operatorname{argmax}_{x \in [-1,1]} |r_3(x)| = 0$$

$$\underline{q}_3(x) = \frac{r_3(x)}{r_3(x_3)} = 1 - x^2$$

| $i$ | $x^i$ | $\ x\ _\infty$ | $\ x^i - I_1(x^i)\ _\infty$ | $\ x^i - I_2(x^i)\ _\infty$ |
|-----|-------|----------------|-----------------------------|-----------------------------|
| 0   | 1     | 1              | $\ 1 - 1\ _\infty = 0$      | 0                           |
| 1   | $x$   | 1              | $\ x - (-1)\ _\infty = 2$   | 0                           |
| 2   | $x^2$ | 1              | $\ x^2 - 1\ _\infty = 1$    | $\ x^2 - 1\ _\infty = 1$    |

**Numerisches Beispiel** demos\_chapter4(1)

$$G = \{x^i\}_{i=0}^{29}$$

- $\Rightarrow$  alle  $q_i$  beschränkt,  $\|q_i\|_\infty = 1$
- Fehler fällt nicht monoton
- $\cos^{-1}(T_M)$  ist etwa äquidistant,  $T_M$  approximieren also qualitativ die Tschebyscheff-Knoten, die für polynomiale Interpolation als beste Wahl bekannt sind  $\Rightarrow T_M$  sind sogenannte "Magic Points" weil sie "auf magische Weise" wichtige Bereiche von  $\Omega$  identifizieren.

**Eigenschaften der EI** Wir untersuchen zunächst Stabilität der Interpolation. Dies geschieht durch Untersuchung der Lebesgue-Konstante (siehe Numerik I).

**Satz 4.3** (Lebesgue-Konstante)

Sie  $I_M : C^0(\Omega, \mathbb{R}) \rightarrow X_M := \operatorname{span}(\zeta_i)_{i=1}^M \in C^0(\bar{\Omega}, \mathbb{R})$  Interpolationsoperator zu den Punkten  $\{x_i\}_{i=1}^M$  und  $\{\zeta\}_{i=1}^M$  nodal, d. h.  $\zeta_i(x_j) = \delta_{ij}$ ,  $I_M(u) = \sum_{i=1}^M \zeta u(x_i)$ . Dann ist

$$\Lambda_M := \max_{x \in \bar{\Omega}} \sum_{i=1}^M |\zeta_i(x)|$$

die Lebesgue-Konstante der Interpolation.

i) Es gilt

$$\|u - I_M(u)\|_\infty \leq (1 + \Lambda_M) \inf_{v \in X_M} \|u - v\|_\infty$$

$$\forall u \in C^0(\bar{\Omega}, \mathbb{R})$$

ii) Für EI gilt

$$\Lambda_M \leq 2^M - 1$$

*Beweis.* i) Sei  $u \in C^0(\bar{\Omega}, \mathbb{R})$ ,  $x \in \bar{\Omega}$  und  $v = \sum a_i \zeta_i \in X_M$ .

Dann ist

$$\begin{aligned} |u(x) - I_M(u)(x)| &= \left| u(x) - \underbrace{\sum_{i=1}^M a_i \zeta_i(x)}_{v(x)} + \underbrace{\sum_{i=1}^M a_i \zeta_i(x)}_{v(x)} - \underbrace{\sum_{i=1}^M u(x_i) \zeta_i(x)}_{I_M(u)(x)} \right| \\ &\leq |u(x) - v(x)| + \left| \sum_{i=1}^M \zeta_i(x) \cdot (a_i - u(x_i)) \right| \end{aligned} \quad (4.2)$$

Für letzten Term gilt wegen  $\{\zeta_i\}$  nodal

$$\begin{aligned} \left| \sum_{i=1}^M \zeta_i(x) \cdot (a_i - u(x_i)) \right| &= \left| \sum_{i=1}^M \zeta_i(x) \cdot \left( \underbrace{\left( \sum_{j=1}^M a_j \zeta_j(x_i) \right)}_{a_i} - u(x_i) \right) \right| \\ &\leq \sum_{i=1}^M |\zeta_i(x)| |v(x_i) - u(x_i)| \\ &\leq \underbrace{\left( \max_x \sum_{i=1}^M |\zeta_i(x)| \right)}_{=\Lambda_M} \cdot \|u - v\|_\infty \end{aligned}$$

Also für (4.3):

$$\|u - I_M(u)\|_\infty \leq \|u - v\|_\infty + \Lambda_M \|u - v\|_\infty = (1 + \Lambda_M) \|u - v\|_\infty$$

nach Infimum über  $v \in X_M$  folgt Behauptung.

ii) → Übung.

□

### Bemerkung.

- Obige Abschätzung für  $\Lambda_M$  ist sehr pessimistisch, in der Praxis meist bessere Raten/Lebesgue-Konstanten (langsames Wachstum), jedoch ist die Schranke scharf, d. h. es ex. Beispiele mit  $\Lambda_M = 2^M - 1$ .
- Für gewisse Mengen  $G$  ist sogar exponentielle Konvergenz des Interpolationsfehlers beweisbar, wie in folgendem Satz. Dort auftretende Forderung ist “Beschränktheit durch Müslipackung”, d. h. Polytop mit exponentiell abfallenden Seitenlängen.

**Satz 4.4** (Exponentielle Konvergenz der EI)

Falls es Sequenz von exponentiell approx. Unterräumen gibt, d.h.  $Z_1 \subset Z_2 \subset \dots \subset Z_M \dots \subset \text{span}(G)$ ,  $\dim Z_M = M$  und ex.  $c > 0, \alpha > \log 4$  so dass  $\inf_{v \in Z_M} \|u - v\| \leq c \cdot e^{-\alpha M} \forall u \in G, M \in \mathbb{N}$ , dann findet der EI-Basiskonstruktionsprozess “fast so gute Räume” indem

$$\|u - I_M(u)\| \leq c \cdot e^{-(\alpha - \log 4)M}.$$

*Beweis.* siehe MNPP07 □

Neben solchen a priori-Aussagen ist auch a posteriori-Fehlerkontrolle möglich.

**Satz 4.5** (A posteriori-Fehlerschätzer für EI)

Seien  $I_M, I_{M'} : C^0(\bar{\Omega}, \mathbb{R}) \rightarrow \text{span}(G)$  EI-Operatoren für  $M' > M$ ,  $Q_M \subset Q_{M'} = \{q_i\}_{i=1}^{M'}, T_M \subset T_{M'} = \{x_i\}_{i=1}^{M'}$ .

Sei  $\underline{Q} := (\underline{q}_i(x_i))_{i,j=M+1}^{M'}$ . Für  $g \in G$  sei  $\underline{g} := (g(x_i)) - I_M(g)(x_i)_{i=M+1}^{M'}$  und  $\underline{\alpha}' := \{\alpha'_i\}_{i=M+1}^{M'} := \underline{Q}^{-1} \underline{g}'$ .

Falls  $g \in \text{span}(\underline{Q}_{M'})$ , so gelten folgende a posteriori-Fehlerschranken für den Interpolationsfehler:

$$\begin{aligned} \|g - I_M(g)\|_\infty &\leq \Delta_{M,M',\infty}(g) := \|\underline{\alpha}'\|_1 = \sum_{i=M+1}^{M'} |\alpha'_i| \\ \|g - I_M(g)\|_{L^2} &\leq \Delta_{M,M',L^2}(g) := \sqrt{(\underline{\alpha}')^T K_Q \underline{\alpha}'} \end{aligned}$$

mit  $K_Q := (\int_\Omega q_i q_j)_{i,j=M+1}^{M'}$

*Beweis.* Nach Def. der EI gilt

$$I_M(g) = \sum_{i=1}^M \alpha_i \underline{q}_i, \quad I_{M'}(g) = \sum_{i=1}^{M'} \alpha'_i \underline{q}_i$$

mit  $\underline{Q}_M \underline{\alpha}_M = \underline{g}_M$  und  $\underline{Q}_{M'} \underline{\alpha}_{M'} = \underline{g}_{M'}$  mit  $\underline{\alpha}_M = (\alpha_i)_{i=1}^M, \underline{\alpha}_{M'} = (\alpha'_i)_{i=1}^{M'}$ .

Da  $\underline{Q}_M$  oberer linker Block von  $\underline{Q}_{M'}$  und  $\underline{Q}$  unterer rechter Block von  $\underline{Q}_{M'}$ , und jeweils untere Dreiecksmatrizen gilt  $\alpha_i = \alpha'_i, i = 1, \dots, M$  und  $g - I_M(g) = I_{M'}(g) - I_M(g) =$

$$\sum_{i=M+1}^{M'} \alpha'_i \underline{q}_i$$

Für  $L^\infty$ -Norm folgt  $\|g - I_M(g)\|_\infty \leq \sum_{i=M+1}^{M'} |\alpha'_i| \underbrace{\|\underline{q}_i\|_\infty}_1 = \|\underline{\alpha}'\|_1$

Für  $L^2$ -Norm gilt  $\|g - I_M(g)\|_2 = \left( \sum_{i,j=M+1}^{M'} \alpha'_i \alpha'_j \int_\Omega \underline{q}_i \underline{q}_j \right)^{\frac{1}{2}} = \Delta_{M,M',L^2}(g)$  □

**Satz 4.6** (Erhaltungseigenschaft)

Sie  $f \in [C^0(\bar{\Omega}, \mathbb{R})]'$  und  $f(g) = 0 \forall g \in G$ .

Dann ist auch  $f(I_M(g)) = 0 \forall g \in G$ .

*Beweis.* wg. Linearität folgt  $f(\sum \beta_i g_i) = 0 \forall \beta_i \in \mathbb{R}, g_i \in G$

$$\Rightarrow f(g) = 0 \forall g \in \text{span}(G)$$

Nach Konstruktion ist  $Q_M \subseteq \text{span}(g)$ , also  $f(\underline{q}_i) = 0, i = 1, \dots, M$ , also auch  $f(I_M(g)) = f(\sum \alpha_i \underline{q}_i) = 0 \forall g \in G$   $\square$

**Bemerkung.** Anschauliche Beispiele

- Falls  $g(\bar{x}) = 0 \forall g \in G \Rightarrow I_M(g)(\bar{x}) = 0$  via  $f \equiv$  Punktauswertung in  $\bar{x}$ .

- Falls  $g \in G$  haben Mittelwert 0, d. h.  $\int_{\Omega} g(x) dx = 0 \forall g \in G$

$$\Rightarrow \int_{\Omega} I_M(g) = 0 \forall g \in G.$$

**Anwendung in RB-Methoden** Falls eine Koeffizientenfunktion nicht separierbar parametrisch ist, erzeuge mittels EI eine separierbar parametrische Approximation.

**Definition 4.7** (EI für Funktionale)

Sie  $f(v; \mu) = \int_{\Omega} g(x, \mu) v(x) dx$  parametrisch stetige Linearform,  $g(\cdot; \mu) \in C^0(\bar{\Omega}, \mathbb{R})$ .

Wähle  $S_{train} \subset \mathcal{P}$  endliche Teilmenge, setze  $G := \{g(\cdot; \mu) | \mu \in S_{train}\}$  Konstruiere EI-Basis  $Q_M = \{q^i\}_{i=1}^M$  und EI-Punkte  $T_M = \{x_i\}_{i=1}^M$  gemäß Definition 4.1. Dann ist

$$g(x; \mu) \approx \sum_{m=1}^M \Theta^m(\mu) q^m(x) \text{ also}$$

$$f(v; \mu) \approx \tilde{f}(v; \mu) := \sum_{m=1}^M \Theta^m(\mu) \underbrace{\int_{\Omega} q^m(x) v(x) dx}_{\tilde{f}^m(v)} =: \sum_{m=1}^M \Theta^m(\mu) \tilde{f}^m(v)$$

Mit Komponenten  $\{\tilde{f}^m\}_{m=1}^M$  und Koeffizienten  $(\Theta^1(\mu), \dots, \Theta^M(\mu))^T := \underline{Q}_M^{-1} g(\mu)$ , wobei  $\underline{g}(\mu)$  durch lokale Auswertungen definiert ist

$$\underline{g}(\mu) = (g(x_1; \mu), \dots, g(x_M; \mu))^T$$

**Bemerkung.**

- Analoge Vorgehensweise bei Ausgabefunktional oder Bilinearform möglich. Anschließend Offline/Online Zerlegung wie in §3 realisierbar.
- Fehlerschätzer aus §3 sind mit EI im Allgemeinen nicht rigoros, da der Interpolationsfehler zusätzlich berücksichtigt werden muss. Falls  $g(\cdot; \mu) \in \text{span}(Q_M)$  (starke Annahme) so sind Schätzer weiterhin rigoros.

- Falls Interpolationsfehler  $\neq 0$ , aber Fehlerschätzer für EI vorhanden, so kann mittels eines Hilfsproblems (hochdimensionales interpoliertes Problem)  $(\tilde{P}(\mu))$  Fehlerkontrolle für das interpolierte und reduzierte Problem realisiert werden. (wir ignorieren wieder Ausgabe im Folgenden).

**Definition 4.8** (Interpolierte Probleme  $(\tilde{P}(\mu)), (\tilde{P}_N(\mu))$ )

Sei eine Instanz von  $(P(\mu))$  gegeben mit  $a(\cdot, \cdot; \mu)$  koerziv. Seien Approximationen  $\tilde{a}(\cdot, \cdot; \mu)$  und  $\tilde{f}(\cdot; \mu)$  bilinear/linear gegeben mit  $\tilde{a}, \tilde{f}$  separierbar parametrisch und

$$\begin{aligned} |a(u, v; \mu) - \tilde{a}(u, v; \mu)| &\leq \epsilon_a \|u\| \|v\| & \forall u, v \in X, \mu \in \mathcal{P} \\ |f(v; \mu) - \tilde{f}(v; \mu)| &\leq \epsilon_f \|v\| & \forall v \in X, \mu \in \mathcal{P} \end{aligned}$$

mit möglichen kleinen Approximationsfehlern  $\epsilon_a, \epsilon_f \geq 0$ .

Wir nennen  $\tilde{u}(\mu) \in X$  Lösung des *vollen interpolierten Problems*  $(\tilde{P}(\mu))$

$$\tilde{a}(\tilde{u}(\mu), v; \mu) = \tilde{f}(v; \mu) \quad \forall v \in X$$

und  $\tilde{u}_N \in X_N$  Lösung des *reduzierten interpolierten Problems*  $(\tilde{P}_N(\mu))$

$$\tilde{a}(\tilde{u}_N(\mu), v; \mu) = \tilde{f}(v; \mu) \quad \forall v \in X_N$$

für beliebige gewählten  $X_N \subset X$ .

**Bemerkung** (Beispiele für  $\epsilon_a, \epsilon_f$ , aus EI). Sie  $X = H_0^1(\Omega)$  und  $f(v) = \int_{\Omega} q(x)v(x)dx$  mit  $q \in L^1(\Omega) \cap C^0(\Omega)$  und  $q_M$  EI von  $q$ , damit  $\tilde{f}(v) := \int_{\Omega} q_M(x)v(x)dx$

$$\Rightarrow |f(v) - \tilde{f}(v)| = \left| \int_{\Omega} (q(x) - q_M(x))v(x)dx \right| \stackrel{CS}{\leq} \|q - q_M\|_{L^2} \underbrace{\|v\|_{L^2}}_{\leq \|v\|_{H_0^1}} \stackrel{4.5}{\leq} \Delta_{\mu;2}(q) \|v\|$$

$$\Rightarrow \epsilon_f := \Delta_{\mu;2}(q) \text{ (bzw. Supremum über } \mathcal{P} \text{ falls } q \text{ parametrisch)}$$

Sei z. B.  $a(u, v) = \int_{\Omega} \kappa(x; \mu)u(x)v(x)dx$  mit  $\kappa(x; \mu) \in C^0(\bar{\Omega})$  und  $\kappa_{\mu}$  EI von  $\kappa$

$$\tilde{a}(u, v) = \int_{\Omega} \kappa_{\mu}(x; \mu)u(x)v(x)dx$$

$$\Rightarrow |a(u, v) - \tilde{a}(u, v)| \leq \|\kappa - \kappa_{\mu}\|_{\infty} \underbrace{\left| \int_{\Omega} u(x)v(x)dx \right|}_{\leq \|u\|_{L^2} \|v\|_{L^2}} \leq \|u\|_{H_0^1} \|v\|_{H_0^1} \Delta_{\mu;\infty}(\kappa)$$

$$\Rightarrow \epsilon_a := \sup_{\mu \in \mathcal{P}} \Delta_{\mu;\infty}(\kappa(\cdot; \mu))$$

**Bemerkung** (Wohlgestelltheit von  $(\tilde{P}(\mu)), (\tilde{P}_N(\mu))$ ). Falls  $\epsilon_a, \epsilon_f$  genügend klein, sind  $\tilde{a}, \tilde{f}$  stetig,  $\tilde{a}$  ist koerziv, damit ist  $(\tilde{P}(\mu))$  nach Lax-Milgram wohlgestellt (Existenz & Eindeutigkeit & Stabilität bzgl. rechter Seite). Genauer: Bei  $\epsilon_a, \epsilon_f < \infty$  folgt Stetigkeit.

$$\begin{aligned} |\tilde{f}(v)| &\leq |\tilde{f}(v) - f(v)| + |f(v)| \leq \epsilon_f \|v\| + \|f\| \cdot \|v\| = (\epsilon_f + \|f\|_{X'}) \|v\| \\ \Rightarrow \tilde{f} &\in X' \text{ mit } \|\tilde{f}\| \leq \|f\| + \epsilon_f \end{aligned}$$



Falls  $\epsilon_a < \alpha$  mit  $\alpha$  Koerzivitätskonstante von  $a(\cdot, \cdot)$ , so folgt

$$\frac{\tilde{a}(u, u)}{\|u\|^2} = \frac{a(u, u) - (a(u, u) - \tilde{a}(u, u))}{\|u\|^2} \geq \underbrace{\frac{a(u, u)}{\|u\|^2}}_{\geq \alpha} - \underbrace{\frac{|a(u, u) - \tilde{a}(u, u)|}{\|u\|^2}}_{\leq \epsilon_a} \geq \alpha - \epsilon_a > 0$$

also  $\tilde{a}$  koerziv mit Koerzivitätskonstante  $\tilde{\alpha} \geq \alpha - \epsilon_a$ .

**Bemerkung** (Schranke  $\|u - \tilde{u}\|$ ). Falls  $\epsilon_a < \alpha$  so folgt mit Störungsargument eine Schranke für  $\|u - \tilde{u}\|$  bezüglich  $\epsilon_a, \epsilon_f$

$$a(u - \tilde{u}, v) = a(u, v) - a(\tilde{u}, v) = f(v) - a(\tilde{u}, v) =: r(v, \tilde{u})$$

Lax-Milgram liefert also mit  $r(v, \tilde{u}) = f(v) - \underbrace{\tilde{f}(v) + a(\tilde{u}, v)}_0 - a(\tilde{u}, v)$ .

$$\|u - \tilde{u}\| \leq \frac{\|r(\cdot; \tilde{u})\|_{X'}}{\alpha} \leq \frac{1}{\alpha}(\epsilon_f + \epsilon_a \|\tilde{u}\|) \quad (4.3)$$

Für  $\|\tilde{u}\|$  liefert Lax-Milgram

$$\|\tilde{u}\| \leq \frac{\|\tilde{f}\|}{\tilde{\alpha}} \leq \frac{\|\tilde{f}\|}{\alpha - \epsilon_a}$$

also insgesamt in 4.3

$$\|u - \tilde{u}\| \leq \frac{1}{\alpha} \epsilon_f + \frac{\|\tilde{f}\|}{\alpha(\alpha - \epsilon_a)} \epsilon_a =: \Delta_{EI} \quad (4.4)$$

Dreiecksungleichung  $\|u - \tilde{u}_N\| \leq \|u - \tilde{u}\| + \|\tilde{u} - \tilde{u}_N\|$  und deren Schranken (4.4) bzw.  $\Delta_N$  liefern:

**Korollar 4.9** (EI-RB Fehlerschranke)

Seien  $(P(\mu))$ ,  $(\tilde{P}(\mu))$ ,  $(\tilde{P}_N(\mu))$  wie in Definition 4.8 gegeben und  $\epsilon_a < \alpha$  Dann gilt die Fehlerschranke

$$\|u(\mu) - u_N(\mu)\| \leq \Delta_{EI}(\mu) + \Delta_N(\mu)$$

mit  $\Delta_{EI}(\mu)$  aus (4.4) und  $\Delta_N(\mu)$  Standard RB-Fehlerschranke für  $\|\tilde{u} - \tilde{u}_N\|$  aus Satz 3.13 ii).

## 4.2 Inf-sup stabile Probleme

**Definition 4.10** (inf-sup Stabilität)

Eine stetige Bilineaform  $a : X_1 \times X_2 \rightarrow \mathbb{R}$  heißt *inf-sup stabil* auf  $X_1 \times X_2$  mit *inf-sup-Konstante*  $\beta$  falls

$$\beta := \inf_{v \in X_1 \setminus \{0\}} \sup_{w \in X_2 \setminus \{0\}} \frac{a(v, w)}{\|v\| \|w\|} > 0$$

Wir fassen einige Einsichten/Aussagen zu inf-sup stabilen Problemen zusammen, für Beweise siehe Skript NUMPD14/15 oder Braess FEM.

**Bemerkung** (Beziehung zwischen  $\alpha/\beta$ ).

- inf-sup Stabilität ist allgemeinerer Stabilitätsbegriff als Koerzivität.  
falls  $a$  koerziv  $\Rightarrow a$  inf-sup stabil mit  $\beta \geq \alpha$   
falls  $a$  koerziv & symmetrisch  $\Rightarrow a$  inf-sup stabil mit  $\beta = \alpha$   
 $\beta$  ist also immer “bessere” (oder gleich gute) Konstante als  $\alpha$ .
- inf-sup Stabilität vererbt sich nicht auf Teilräume dies stellt bei FEM- und RB-Behandlung ein Problem dar.

**Bemerkung** (Beispiel).

a) Divergenzgleichung: Sie  $\Omega$  Lipschitz-Gebiet und beschränkt

Für  $q \in Q_1 := L_0^2(\Omega) = \{p \in L^2(\Omega) \mid \int_{\Omega} p(x) dx = 0\}$  und  $v \in V := H_0^1(\Omega)^d$  ist  $a(q, v) = \int_{\Omega} q \operatorname{div}(v)$  inf-sup stabil auf  $Q \times V$ .

b) Stokes-Problem:

$$-\mu \Delta u + \nabla p = r \quad \text{in } \Omega, \quad \operatorname{div} u = 0 \quad \text{in } \Omega, \quad u = 0 \quad \text{auf } \Gamma$$

ergibt mit  $X_1 := X_2 := H_0^1(\Omega)^d \times L_0^2(\Omega)$  schwache Form:

Suche  $(u, p) \in X_1$  so dass  $a((u, p), (v, q)) = f((v, q)) \quad \forall (v, q) \in X_2$

mit  $a((u, p), (v, q)) := \mu \int_{\Omega} \nabla u : \nabla v - \int_{\Omega} p \operatorname{div}(v) - \int_{\Omega} q \operatorname{div}(u)$ .

Dann ist  $a$  inf-sup stabil auf  $X_1 \times X_2$ .

Der folgende Satz garantiert Wohlgestelltheit für inf-sup stabile Probleme

**Satz 4.11** (Nečas-Theorem)

Sei  $a : X_1 \times X_2$  stetig Bilinearform,  $l \in X_2'$ . Die Gleichung

$$a(u, v) = l(v) \quad \forall v \in X_2$$

hat eine eindeutige Lösung  $u \in X_1$  und ist stetig von den Daten abhängig via

$$\|u\|_{X_1} \leq \frac{1}{\beta} \|l\|_{X_2'}$$

für ein  $\beta > 0$  (unabhängig von  $l$ ) genau dann wenn eine der folgenden Bedingungen gilt:

- $a(\cdot, \cdot)$  erfüllt  $\inf_v \sup_w \frac{a(v, w)}{\|v\|_{X_1} \|w\|_{X_2}} \geq \beta > 0$   
und  $\forall w \in X_2, w \neq 0$  ex.  $v \in X_1$  mit  $a(v, w) \neq 0$ .
- $a(\cdot, \cdot)$  erfüllt  $\inf_w \sup_v \frac{a(v, w)}{\|v\|_{X_1} \|w\|_{X_2}} \geq \beta > 0$   
und  $\forall v \in X_1, v \neq 0$  ex.  $w \in X_2$  mit  $a(v, w) \neq 0$ .

*Beweis.* z. B. Satz 3.57 in NUMPDE14/15 □

**Bemerkung.**

- inf-sup Stabilität ist also “allgemeinerer” Stabilitätsbegriff, welcher Wohlgestelltheit impliziert.
- Lax-Milgram kann als Korollar von 4.11 gezeigt werden
- Aus 4.11 folgt insbesondere

$$\text{wohlgestellt} \Rightarrow \inf_{v \in X_1} \sup_{w \in X_2} \frac{a(v, w)}{\|v\| \|w\|} = \beta = \inf_{w \in X_2} \sup_{v \in X_1} \frac{a(v, w)}{\|v\| \|w\|} \quad (4.5)$$

also dasselbe  $\beta$  für beide Richtungen. Umgekehrte Richtung in (4.6) gilt auch.

**Satz 4.12** (Eigenschaften von  $\beta$ / Suprimierender Operator)

Sie  $a(\cdot, \cdot)$  stetige Bilinearform auf  $X_1 \times X_2$  und

$$a(u, v) = \langle Au, v \rangle_{X_2} \quad \forall (u, v) \in X_1 \times X_2$$

mit Operator  $A \in L(X_1, X_2)$ . Dann gilt:

i)  $A$  ist *suprimierender Operator*,

$$\sup_{v \in X_2} \frac{a(u, v)}{\|v\|} = \frac{a(u, Au)}{\|Au\|} \quad \forall u \in X_1 \setminus \{0\}$$

ii)

$$\beta = \inf_{u \neq 0} \frac{\|Au\|}{\|u\|}$$

iii)

$$\gamma = \sup_{u \neq 0} \frac{\|Au\|}{\|u\|}$$

iv)  $\forall u \in X_1$  ex.  $v \in X_2$  s. d.  $\beta \|u\| \|v\| \leq a(u, v)$

v)  $A$  ist injektiv, d. h.  $\forall v \neq 0$  gilt  $Av \neq 0$ .

*Beweis.* Satz 3.32 in NUMPDE14/15 □

**Bemerkung.** Solches  $A$  existiert wegen Riesz:  $a(u, \cdot) \in X_2'$  also existiert Riesz-Repräsentant  $Au \in X_2$  mit  $a(u, \cdot) = \langle Au, \cdot \rangle_{X_2}$ .

**Satz 4.13** (Eigenwert für  $\beta/\gamma$ )

Seien  $X_1, X_2$  Hilberträume,  $a : X_1 \times X_2$  stetige Bilinearform mit  $a(u, v) = \langle Au, v \rangle$  und  $A, A^*$  komp. Operator d. h.  $A \in K(X_1, X_2)$ ,  $A^* \in L(X_2, X_1)$  und  $\sigma := \sigma(A^*A)$  Spektrum von  $A^*A$ . Dann gilt:

i)  $\forall \lambda \in \sigma$  ex.  $u \in X_1$ :

$$\langle Au, Av \rangle_{X_2} = \lambda \langle u, v \rangle_{X_1} \quad \forall v \in X_1$$

ii)  $\beta^2 = \inf \sigma$

iii)  $\gamma^2 = \max \sigma$

*Beweis.* i) Seien  $u, \lambda$  EV/EW von  $A^*A \Rightarrow \langle A^*Au, v \rangle_{X_1} = \langle \lambda u, v \rangle_{X_1} \quad \forall v \in X_1$

$$\Rightarrow \langle Au, Av \rangle_{X_1} = \lambda \langle u, v \rangle_{X_1} \quad \forall v \in X_1$$

ii) Wegen  $A^*A$  komp. & selbstadjungiert folgt mit Spektralsatz

$$A^*Au = \sum_{i \in I} \lambda_i \langle u, \varphi_i \rangle \varphi_i$$

und  $\sigma = \{\lambda_i\}_{i \in I}$  oder  $\sigma = \{\lambda_i\}_{i \in I} \cup \{0\}$

$$\text{Nach 4.12 ii) } \beta = \inf_{u \neq 0} \frac{\|Au\|}{\|u\|} = \inf_{u \neq 0} \sqrt{\frac{\|Au\|^2}{\|u\|^2}} = \inf_{\|u\|=1} \sqrt{\langle Au, Au \rangle} = \inf_{\|u\|=1} \sqrt{\langle A^*Au, A \rangle}$$

$$\beta^2 = \inf_{\|u\|=1} \langle A^*Au, u \rangle \quad (4.6)$$

Falls  $0 \in \sigma$  mit EV  $\varphi^0 \Rightarrow 0 \leq \beta^2 = \langle A^*A\varphi^0, \varphi^0 \rangle = 0 \Rightarrow \beta^2 = 0 = \inf \sigma$ .

Falls  $0 \notin \sigma \Rightarrow \{\varphi_i\}_{i \in I}$  ist vollständiges ONS von  $X_1$

$$\Rightarrow \forall u \in X_1 \text{ mit } \|u\| = 1 \text{ ex. } u_i \in \mathbb{R} \text{ mit } u = \sum_{i \in I} u_i \varphi_i, \sum u_i^2 = 1$$

$$\stackrel{(4.6)}{\Rightarrow} \beta = \inf_{\|u\|=1} \langle A^*Au, u \rangle = \inf_{\|u\|=1} \sum \lambda_i u_i^2 \underbrace{\geq}_{\text{Konvexkombination}} \inf \{\lambda_i\} = \inf \sigma$$

$$\text{und } \beta^2 = \inf_{\|u\|=1} \sum \lambda_i u_i^2 \underbrace{\leq}_{u=\varphi_j} \sum \lambda_i u_i^2 = \lambda_j \Rightarrow \beta^2 \leq \inf \{\lambda_i\} \Rightarrow \beta^2 = \inf \sigma$$

iii) analog zu ii)

□

**Korollar 4.14** (Inf-sup Stabilität im Endlichdimensionalen)

Sei  $X_1 = \mathbb{R}^m, X_2 = \mathbb{R}^n, A \in \mathbb{R}^{n \times m}, m \leq n$  und  $a(u, v) := \langle Au, v \rangle$ . Seien  $\{\sigma_i\}_{i=1}^m$  Singulärwerte von  $A$ . Dann

$$\text{i) } \beta = \min_{i=1, \dots, m} \sigma_i$$

ii)  $a$  inf-sup stabil  $\iff A$  hat vollen Spaltenrang

*Beweis.* i)  $a$  stetig und  $A, A^T$  kompakter linearer Operator, da endlichdimensionale Bilder. Mit 4.13 folgt wegen  $\sigma(A^T A) = \{\sigma_i^2\}_{i=1}^m : \beta^2 = \inf \sigma(A^T A) = \min_{i=1, \dots, m} \{\sigma_i^2\}$ .

Weil  $\beta \geq 0, \sigma_i \geq 0 \Rightarrow \beta = \min_{i=1, \dots, m} (\sigma_i)$ .

ii)  $a$  inf-sup-stabil  $\iff \beta > 0 \stackrel{i)}{\iff} \sigma_i > 0 \forall i \iff A$  hat Rang  $m \iff A$  hat vollen Spaltenrang.

□

RB-Methode folgt mit *Petrov Galerkin-Projektion*

**Definition 4.15** ( $(P(\mu)), (P_N(\mu))$  für inf-sup stabilen Fall)

Sei  $a : X_1 \times X_2 \times \mathcal{P} \rightarrow \mathbb{R}$  stetig param. Bilinearform,  $a(\cdot, \cdot; \mu)$  inf-sup stabil mit  $\beta(\mu)$  inf-sup Konstante  $f(\cdot; \mu)$  stetig param. Linearform. Zu  $\mu \in \mathcal{P}$  ist  $u(\mu) \in X_1$  Lösung des vollen Problems

$$a(u(\mu), v; \mu) = f(v; \mu) \quad \forall v \in X_2$$

Seien  $X_{N,1} \subseteq X_1$ ,  $X_{N,2} \subseteq X_2$  mit  $N = \dim X_{N,1} = \dim X_{N,2}$  RB-Räume. Dann ist  $u_N(\mu) \in X_{N,1}$  RB-Lösung von

$$a(u_N(\mu), v; \mu) = f(v; \mu) \quad \forall v \in X_{N,2}$$

**Bemerkung.**

- Wir verzichten in 4.15 auf das Ausgabefunktional, dies kann analog zu §3.5 behandelt werden, d. h.  $l$  parametrische Linearform

$$s'_N(\mu) = l(u_N(\mu); \mu) - r(u_N^{du}; \mu) \text{ mit } u_N^{du} \in X_{N,2}^{du} \text{ Lösung von} \quad (4.7)$$

$$a(v, u_N^{du}; \mu) = -l(v; \mu) \quad \forall v \in X_{N,1}^{du} \text{ und dualen RB-Räumen } X_{N,1}^{du}, X_{N,2}^{du} \quad (4.8)$$

- Wahl von  $X_{N,1}, X_{N,2}$  ist nicht beliebig, sondern müssen geeignet gewählt werden, so dass  $(P_N(\mu))$  gute Approximation liefert (wie in §3. Wähle  $X_{N,2}$  s. d.  $(P_N(\mu))$  möglichst stabil, d. h. möglichst große diskrete inf-sup Konstante.

Wegen  $\dim X_{N,1} = \dim X_{N,2}$  vereinfacht sich Charakterisierung von inf-sup Stabilität.

**Satz 4.16** (Diskrete inf-sup Bedingung)

Für  $a : X_{N,1} \times X_{N,2} \rightarrow \mathbb{R}$  sind äquivalent

$$\text{i) } \inf_{v \in X_{N,1}} \sup_{w \in X_{N,2}} \frac{a(v, w)}{\|v\| \|w\|} = \beta_N > 0$$

$$\text{ii) } \inf_{w \in X_{N,2}} \sup_{v \in X_{N,1}} \frac{a(v, w)}{\|v\| \|w\|} = \beta_N > 0$$

$$\text{iii) } \forall w \in X_{N,2} w \neq 0 \text{ ex. } v \in X_{N,1} \text{ mit } a(v, w) \neq 0 \text{ (A surjektiv)}$$

$$\text{iv) } \forall v \in X_{N,1} v \neq 0 \text{ ex. } w \in X_{N,2} \text{ mit } a(v, w) \neq 0 \text{ (A injektiv)}$$

*Beweis.* Satz 3.46 in NUMPDE14/15

□

**Korollar 4.17** (Wohlgestelltheit von  $(P_N)$ )

Falls  $a$  eine der Bedingungen aus 4.16 erfüllt, dann hat ad Problem  $(P_N(\mu))$  aus 4.15 eindeutige Lösung  $u_N(\mu) \in X_{N,1}$  mit

$$\|u_N(\mu)\| \leq \frac{\|f\|_{X'_2}}{\beta_N(\mu)}$$

*Beweis.* klar mit Nečas □

Petrov-Galerkin Projektion erlaubt Approximationsaussage:

**Satz 4.18** (Beziehung zur Bestapproximation)

Sei  $a$  inf-sup stabil mit  $\beta_N > 0$ . Dann gilt

$$\|u(\mu) - u_N(\mu)\| \leq \frac{\gamma(\mu)}{\beta_N(\mu)} \inf_{v \in X_{N,1}} \|u(\mu) - v\|$$

*Beweis.* Satz 3.55 in NUMPDE14/15 □

**Bemerkung.**

- Es gilt wieder “Reproduktion von Lösung”:  
Falls  $u(\mu) \in X_{N,1} \Rightarrow u_N(\mu) = u(\mu)$
- Im Fall von primal-dualem Ansatz (4.7), (4.8) gilt analoge Aussage wie Satz 3.60, insbesondere multiplikativer Effekt für Ausgabefehler

$$\|u^{du} - u_N^{du}\| \leq \frac{\gamma(\mu)}{\beta_N^{du}} \inf \|u^{du} - v\|$$

$$|s(\mu) - s'_N(\mu)| \leq \gamma(\mu) \|u - u_N\| \|u^{du} - u_N^{du}\|$$

**Satz 4.19** (A-posteriori Fehlerschätzer)

Sei  $\beta_{LB}(\mu) > 0$  untere Schranke für  $\beta_N(\mu)$ . Dann gilt

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

mit  $v_r \in X'_2$  Riesz-Repräsentant des Residuums wie in §3.

*Beweis.*  $u - u_N$  ist Lösung von  $a(u - u_N, v) = r(v) \forall v \in X_2$ , also folgt mit Nečas

$$\|u - u_N\| \leq \frac{\|r\|_{X'_2}}{\beta_N(\mu)} \leq \frac{\|v_r\|_{X_2}}{\beta_{LB}} = \Delta_N$$

□

**Bemerkung.**

- Ident. zu Satz 3.16 ii) folgt Effektivitätsschranke

$$\eta_N(\mu) := \frac{\Delta_N(\mu)}{\|e\|} \leq \frac{\gamma_{LB}(\mu)}{\beta_{LB}(\mu)}$$

- Analog zu 3.61 & 3.62 gilt Schranke für duale Lösung & Ausgabe

$$\|u^{du} - u_N^{du}\| \leq \Delta_N^{du}(\mu) := \frac{\|r^{du}\|}{\beta_{LB}^{du}}$$

$$|s - s'_N| \leq \Delta'_{N,s}(\mu) : 0 \frac{\|v_r\| \cdot \|v_r^{du}\|}{\beta_{LB}(\mu)}$$

also multiplikativer Effekt erreicht.

- Zusammenfassend: (fast) alle Aussagen aus §3 gelten mit  $\alpha_{LB}$  ersetzt durch  $\beta_{LB}$ , bzw.  $\alpha$  durch  $\beta$ .
- Offline-Online-Zerlegung folgt analog zu §3.
- Einzig offene Punkte: Wahl von  $X'_{N,2}$  und Bestimmung von  $\beta_{LB}$ . Das sind kritische Komponenten: Weil inf-sup Stabilität nicht auf Teilräume vererbt wird, kann  $\beta_N(\mu)$  und damit  $\beta_{LB}(\mu)$  beliebig klein, sogar Null werden.  
 → singuläres red. System trotz  $\beta(\mu) > 0$ . Ziel ist daher Wahl von  $X_{N,2}$  s. d.  $\beta_N(\mu)$  möglichst groß uniform in  $N$ , d. h. Verhindern von  $\lim_{N \rightarrow \infty} \beta_N(\mu) \rightarrow 0$ . Idealerweise Ziel:

$$\beta_N(\mu) \geq \beta(\mu) > 0$$

**Satz 4.20** (Kriterium für  $X_{N,2}$ )

Sei  $A(\mu)$  suprimierender Operator aus 4.12. Falls für alle  $v \in X_{N,1}$  gilt, dass  $A(\mu) \in X_{N,2}$ , dann ist  $a$  inf-sup stabil auf  $X_{N,1} \times X_{N,2}$  mit  $\beta_N(\mu) \geq \beta(\mu)$ .

*Beweis.* Übung. □

**Bemerkung** (Min. Residuum Verfahren).

- Sei  $X_{N,1} = \text{span}\{\varphi_1, \dots, \varphi_N\}$ . Wähle  $X_{N,2} := \text{span}\{A(\mu)\varphi_1, \dots, A(\mu)\varphi_N\} \Rightarrow \dim X_{N,1} = \dim X_{N,2}$  wegen Injektivität von  $A$ .

Wegen Linearität gilt:  $\forall v \in X_{N,1} \Rightarrow A(\mu)v \in X_{N,2}$  also ideales RB-Verfahren gemäß 4.20 erreicht. Aber  $X_{N,2}$  ist  $\mu$ -abhängig. Dies stellt jedoch kein Problem dar, denn Offline-Online weiter möglich.

- Petrov-Galerkin-System entspricht einem Galerkin-RB-System eines koerziven, symmetrischen Problems mit separierbar parametrischen Bilinearform  $\tilde{a}$ :

$$\text{Definiere } \tilde{a}(u, v; \mu) := \langle A(\mu)u, A(\mu)v \rangle_{X_2} \quad \forall u, v \in X_1$$

$\Rightarrow \tilde{a}$  ist bilinear, stetig, symmetrisch.  $\tilde{a}$  ist koerziv

$$\frac{\tilde{a}(u,u)}{\|u\|^2} = \frac{\langle Au, Au \rangle}{\|u\|^2} = \frac{\|Au\|^2}{\|u\|^2} \geq \left( \inf_{v \neq 0} \frac{\|Av\|}{\|v\|} \right)^2 \stackrel{4.12ii)}{=} (\beta_N(\mu))^2 > 0$$

also  $\tilde{a} := \beta(\mu)^2$  Koerzivitätskonstante von  $\tilde{a}$

$\tilde{a}$  ist separierbar parametrisch  $Q_{\tilde{a}} = (Q_a)^2$  Komp.  $\{\langle A^q \cdot, A^q \cdot \rangle\}_{q,q'=1}^{Q_a}$  und Koeffizienten  $\Theta_a^q(\mu), \Theta_a^{q'}(\mu)$ .

Für Petrov-Galerkin-System gilt mit  $\overline{\varphi}_i := A(\mu)\varphi_i$

$$\begin{aligned} A_N \underline{u}_N &= \underline{f}_N \text{ mit } (A_N)_{ij} = a(\varphi_j, \overline{\varphi}_i) = a(\varphi_j, A(\mu)\varphi_i) \\ &= \langle A\varphi_j, A\varphi_i \rangle = \tilde{a}(\varphi_j, \varphi_i) \end{aligned}$$

also RB-?? für  $\tilde{a}$ . Analog  $(\underline{f}_N)_i = f(\overline{\varphi}_i; \mu) = f(A\varphi_i) = \tilde{f}(\varphi_i; \mu)$  mit  $\tilde{f} := f \circ A$

- Darstellung durch FEM-Matrizen: Seien  $\underline{A}, K, \underline{\Phi}_N, \underline{f}$  die bekannten FEM-Matrizen/Vektoren aus §3.3.

$A\varphi_j$  ist Riesz-Repräsentant von  $a(\varphi_j, \cdot)$

$\Rightarrow A\varphi_j \in X_1$  hat FEM Koeff.-Vektor  $K^{-1}(a(\varphi_j, \psi_i))_{i=1}^H = K^{-1} \underline{A} \varphi_j$

$$A_N = (\langle A\varphi_j, A\varphi_i \rangle_{i,j=1}^N = \underline{\Phi}_N^T \underline{A}^T K^{-1} \underline{A} \underline{\Phi}_N$$

analog  $\underline{f}_N = \underline{\Phi}_N^T \underline{A} K^{-1} \underline{f}$



## 5 Nichtlineare Probleme

Satz 5.1

Satz 5.2

Satz 5.3

Satz 5.4

Satz 5.5

Satz 5.6

$$a + b = c \quad (5.1)$$

$$a + b = c \quad (5.2)$$

$$a + b = c \quad (5.2')$$

$$a + b = c \quad (5.3)$$

$$a + b = c \quad (5.4)$$

$$(5.5)$$

**Bemerkung.**

- $\beta_{LB}$  im Zähler sieht zunächst seltsam aus. Weil  $\tau \in [0,1]$  ist  $(1 - \sqrt{1 - \tau}) \in [0,1]$ , also insbesondere  $\sqrt{(\cdot)}$  wohldefiniert und

$$\Delta_N \leq \frac{\beta_{LB}}{2L_{DF}} \tau = \frac{\beta_{LB}}{2L_{DF}} \cdot \frac{4\epsilon L_{DF}}{\beta_{LB}^2} = \frac{2\epsilon}{\beta_{LB}}$$

Also gewohnte Struktur: Residuum durch Stabilitätskonstante, jedoch Faktor 2.

*Beweis.* Ähnlich zu 5.4, statt festem Radius betrachte Kugel mit variablem Radius  $\alpha$ . Weil  $DF$  Lipschitz-stetig, ist  $H(x) = x - (DF|_{u_N})^{-1}H(x)$  Kontraktion auf  $B(u_N, \alpha)$  falls  $\alpha \leq \frac{\beta_{LB}}{2L_{DF}} = \frac{\beta_{LB}}{4\gamma_a} =: \hat{\alpha}$ :

Für  $x, x' \in B(u_N, \alpha)$  ist mit (5.2')

$$\begin{aligned} \|H(x) - H(x')\| &\stackrel{(5.2')}{\leq} \underbrace{\frac{\frac{1}{\beta_{LB}}}{\gamma}}_{\leq L_{DF}\alpha \leq L_{DF} \cdot \frac{\beta_{LB}}{2L_{DF}}} L(\alpha) \leq \frac{1}{\beta_{LB}} \cdot L_{DF} \cdot \frac{\beta_{LB}}{2L_{DF}} \|x - x'\| = \frac{1}{2} \|x - x'\| \end{aligned}$$

Suche nun Bedingung für  $\alpha$  sodass  $H$  Selbstabbildung auf  $B(u_N, \alpha)$ . Mit (5.2) folgt für  $x \in B(u_N, \alpha)$

$$\begin{aligned} \|H(x) - u_N\| &\leq \underbrace{\gamma}_{\leq \frac{1}{\beta_{LB}}} \left( \int_0^1 \|DF|_{u_N} - DF|_{u_N+t(x-u_N)}\|_{X;X'} dt \right) \underbrace{\|x - u_N\|}_{\leq \alpha} + \gamma \cdot \epsilon \\ &\leq \frac{1}{\beta_{LB}} L_{DF} \alpha^2 + \frac{1}{\beta_{LB}} \epsilon \end{aligned}$$

Falls also

$$\frac{L_{DF}}{\beta_{LB}} \alpha^2 + \frac{\epsilon}{\beta_{LB}} \leq \alpha \quad (5.6)$$

so ist  $H$  Selbstabbildung.

$$\begin{aligned} (5.6) \quad &\Leftrightarrow \quad \alpha^2 - \frac{\beta_{LB}}{L_{DF}} \alpha + \frac{\beta_{LB}}{L_{DF}} \cdot \frac{\epsilon}{\beta_{LB}} \leq 0 \\ &\Leftrightarrow \quad \alpha \in [\alpha_-, \alpha_+] \text{ mit } \alpha_{\pm} := \frac{\beta_{LB}}{2L_{DF}} \pm \sqrt{\frac{\beta_{LB}^2}{4L_{DF}^2} - \frac{\epsilon}{L_{DF}}} \\ &\quad \alpha_{\pm} = \hat{\alpha} \left( 1 \pm \sqrt{1 - \frac{\epsilon}{L_{DF} \hat{\alpha}^2}} \right) = \hat{\alpha} (1 \pm \sqrt{1 - \tau}) \end{aligned}$$

weil

$$\frac{\epsilon}{L_{DF} \hat{\alpha}^2} = \frac{\epsilon}{L_{DF}} \cdot \frac{4L_{DF}^2}{\beta_{LB}^2} = \frac{4L_{DF}}{\beta_{LB}^2} \epsilon = \tau \leq 1$$

also  $\alpha_{\pm}$  wohldefiniert.

Für  $\alpha \in [\alpha, \hat{\alpha}]$  ist  $H$  Selbstabbildung und Kontraktion. Für kleinstes  $\alpha = \alpha_-$  erhalte beste Schranke, also ex.  $u \in B(u_N, \alpha_-)$  mit

$$\|u - u_N\| \leq u_N = \alpha_-$$

Für Effektivitätsschranke setze  $e := u - u_N$ . Sei  $v_r \in X$  Riesz-Repräsentant des Residuums

$$\langle v_r, v, = \rangle F(u_N)(v)$$

Benötigt Fehler-Residuums-Beziehung für quadratisches Problem

$$\begin{aligned} F(u_N) &= a(u_N, u_N, \cdot) - b(u_N, \cdot) - \underbrace{f(\cdot)}_{=a(u, u, \cdot) - b(u, \cdot)} \\ &= 2a(u_N, u_N, \cdot) - 2a(u_N, u, \cdot) - a(u_N, u_N, \cdot) \\ &\quad + 2a(u_N, u, \cdot) - a(u, u, \cdot) - b(u - u_N, \cdot) \\ &= -2a(u_N, e, \cdot) - b(e, \cdot) - a(e, e, \cdot) \\ &= -DF|_{u_N}(e) - a(e, e, \cdot) \end{aligned}$$

$$\begin{aligned}
\Rightarrow \|v_r\|^2 &= \langle v_r, v_r, \rangle F(u_N)(v_r) = -DF|_{u_N}(e)(v_r) - a(e, e, v_r) \\
&\leq \gamma_{DF}(u_N)\|e\|\|v_r\| + \gamma_a\|e\|\|v_r\| \\
\Rightarrow \|v_r\| &\leq \gamma_{DF}(u_N)\|e\| + \gamma_a\|e\|^2
\end{aligned}$$

Mit  $\|v_r\| = \epsilon$  und  $\Delta_N \stackrel{(5.4)}{\leq} \frac{2\epsilon}{\beta_{LB}}$  folgt

$$\Delta_N \leq \frac{2\|v_r\|}{\beta_{LB}} \leq \frac{2}{\beta_{LB}}\gamma_{DF}\|e\| + \frac{2}{\beta_{LB}}\gamma_a \underbrace{\|e\|^2}_{\leq \Delta_N \cdot \Delta_N}$$

Wegen

$$\frac{2}{\beta_{LB}}\gamma_a\Delta_N \leq \frac{2\gamma_a}{\beta_{LB}} \cdot \frac{2\epsilon}{\beta_{LB}} = \frac{4\gamma_a\epsilon}{\beta_{LB}^2} = \frac{1}{2}\tau \leq \frac{1}{2}$$

folgt

$$\begin{aligned}
\Delta_N &\leq \frac{2}{\beta_{LB}}\gamma_{DF}\|e\| + \frac{1}{2}\Delta_N \quad \Rightarrow \quad \frac{1}{2}\Delta_N \leq \frac{2}{\beta_{LB}}\gamma_{DF}\|e\| \\
&\Rightarrow \frac{\Delta_N}{\|e\|} \leq \frac{4\gamma_{DF}(u_N)}{\beta_{LB}}
\end{aligned}$$

□

#### Bemerkung.

- Lokale Existenz und Eindeutigkeit und Fehlerschranken gilt analog für allgemeinere Nichtlinearitäten  $F$ , welche Lipschitz-stetige Ableitungen besitzen. Nur die Effektivitätsschranke in 5.6 verwendet die Struktur des quadratischen nichtlinearen Problems.
- Ausgabe-Fehlerschranken sind einfach möglich analog zu  $\Delta_{N,s}$  aus §3. Auch verbesserte Abschätzung mittels geeigneten dualen Problems ist möglich.
- Berechnung von  $\beta_{LB}$  durch SCM-ähnliche Techniken möglich, siehe [VPP03].
- Falls PDE linear, aber Ausgabe quadratisch nichtlinear, lässt sich ein erweitertes Hilfsproblem formulieren, welche linear, inf-sup-stabil, symmetrisch ist und identische Ausgabe wie Originalproblem mittels geeignetem linearen Ausgabefunktions liefert.  $\Rightarrow$  Techniken aus §3 und §4 anwendbar. Damit z.B. Fehlerfunktionale  $s(\mu) = \int_{\Omega}(u(\mu) - u_d)^2$  oder Variationen oder Energien in verschiedenen Versionen behandelbar.  
Referenz: [HP07]: “Reduced basis approximation and a posteriori error estimation for stress intensity factors”, IJNME, 72, 1219-1259, 2007.
- Falls PDE polynomiell in  $u$  der Ordnung  $p$ , lässt sich eine  $p + 1$  Multilinearform-Formulierung der schwachen Form finden und Techniken aus §5 analog anwenden. Problem wird für  $p \gg 3$  jedoch die Offline-Online-Zerlegung (...???) weil Komponenten-Tensoren der Stufe  $p + 1$ , also sind Offline Datenmengen und Assemblierungskosten  $\mathcal{O}(Q_a N^{p+1})$ .

- Falls PDE nichtpolynomiell nichtlinear, kann mit Hilfe der EI ein nichtlineares reduziertes Problem formuliert werden. Eine Variante ist die Empirische Operatorinterpolation in
  - [HOR08]: Haasdonk, Ohlberger, Rozza: A Reduced Basis Method for Evolution Schemes with Parameter-Dependent Explicit Operators, ETNA, 32:145-161, 2008.
  - [DHO12]: Drohmann, Haasdonk, Ohlberger: Reduced Basis Approximation for Nonlinear Parametrized Evolution Equations based on Empirical Operator Interpolation, SJSC, 34:A937-A969, 2012.

## 6 Zeitabhängige Probleme

### Motivation:

Anfangs-Randwertprobleme, z.B. Wärmeleitung. Suche  $u(x,t)$ ,  $x \in \Omega$ ,  $t \in [0,T]$

$$\begin{aligned} \partial_t u - \Delta u &= q && \text{in } \Omega \times (0,T) \\ u(x,t) &= g_D(x,t) && \text{auf } \partial\Omega \times (0,T) \\ u(x,0) &= u_0(x) && \text{in } \Omega \end{aligned}$$

### Numerischer Ansatz:

Zeitdiskretisierung: Wähle  $K \in \mathbb{N}$ ,  $\Delta t := \frac{T}{K}$

$$t^k := k \cdot \Delta t, \quad k = 0, \dots, K$$

Wähle  $X \subseteq Y := L^2(\Omega)$  Lösungsraum bezüglich Ortsraum, z.B.  $X = L^2(\Omega)$  oder  $X = H_0^1(\Omega)$  oder  $X = \text{span}(\mathbb{1}_{e_i})_{i=1}^H$  Finite-Volumen-, oder  $X = \text{span}(\psi_i)_{i=1}^H$  Finite-Elemente-Raum, etc.

Suche Lösungssequenz  $u = (u^k)_{k=0}^K \in (X)^{K+1}$  mit  $u^k(x) \approx u(x, t^k)$ .

Referenzen:

- Für PDE-Diskretisierung siehe NUMPDE14/15.
- Für folgende RB-Behandlung siehe auf S. 8 genanntes Tutorial.

**Bemerkung.** Statt variationeller Formulierung betrachte im Folgenden operatorbasierte Formulierung, dies erfasst auch Finite-Differenzen oder Finite-Volumen Diskretisierungen. Alles Folgende könnte auch durch Variationsformulierung ausgedrückt werden.

**Definition 6.1** (Volles Evolutionsproblem  $(E(\mu))$ )

Sei  $X$  Hilbertraum,  $\mu \in \mathcal{P}$ :  $u(\mu) = (u^k(\mu))_{k=0}^K \in (X)^{K+1}$  Lösungen von

$$\begin{aligned} \mathcal{L}_I^k(\mu) u^{k+1} &= \mathcal{L}_E^k(\mu) u^k + b^k(\mu) \\ u^0 &:= P_X(u_0) \end{aligned}$$

mit  $\mathcal{L}_I^k, \mathcal{L}_E^k \in L(X)$ ,  $P_X : Y \rightarrow X$  beliebige stetige Projektion.

### Bemerkung.

- Wir verzichten hier wieder auf Ausgabefunktionale.
- Obiges erfasst allgemeine implizite/explicite Zeitdiskretisierung wie impliziter/expliciter Euler oder Crank-Nicolson-Verfahren für parabolische oder hyperbolische DGL.
- $\mathcal{L}_I^k, \mathcal{L}_E^k, b^k$  hängen typischerweise von  $\Delta t$  ab.

### Annahmen an Operatoren:

- $\mathcal{L}_I^k, \mathcal{L}_E^k$  seien stetig mit Konstanten  $\gamma_I^k(\mu), \gamma_E^k(\mu)$  und uniform in  $t, \mu$ , d.h.  $\gamma_I^k(\mu) \leq \bar{\gamma}, \gamma_E^k(\mu) \leq \bar{\gamma}_E$ .
- $\mathcal{L}_I^k$  sei uniform koerziv bzgl.  $\mu$  und  $k$ , d.h. ex.  $\bar{\alpha}_I, \alpha_I^k(\mu)$  mit

$$0 < \bar{\alpha}_I \leq \alpha_I^k(\mu) := \inf_v \frac{\langle \mathcal{L}_I^k v, v \rangle_X}{\|v\|^2}$$

- $b^k$  seien uniform beschränkt  $\|b^k(\mu)\| \leq \bar{\gamma}_b \ \forall k, m$ .
- $\mathcal{L}_I^k, \mathcal{L}_E^k$  seien separierbar parametrisch mit zeitunabhängigen Komponenten

$$\mathcal{L}_I^k(\mu) = \sum_{q=1}^{Q_I} \Theta_{I,q}^k(\mu) \mathcal{L}_{I,q}, \quad \mathcal{L}_{I,q} \in L(X)$$

analog für  $\mathcal{L}_E, b^k, u_0$ .