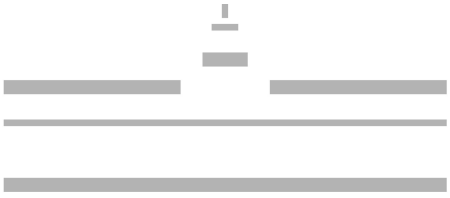

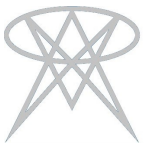

FACHBEREICH 10
MATHEMATIK UND
INFORMATIK

Modellreduktion und partielle Differentialgleichungen

Mitschrift der Tafelnotizen

Tobias Wedemeier

28. Januar 2016
gelesen von
Dr. Smetana

Hierbei handelt es sich um eine Mitschrift der Tafelnotizenvon **Dr. Smetana**, WWU Münster, aus der Vorlesung **Modellreduktion und partielle Differentialgleichungen** im Wintersemester 2015/16. Dies ist kein Skript der Vorlesung und keine eigene Arbeit des Autors.

Für Fehler in der Mitschrift wird keine Haftung übernommen. Hinweise auf Fehler sind gerne gesehen, hierfür kann man mich in der Uni ansprechen oder alternativ eine e-Mail an: *tobias.wedemeier@gmx.de* Auch ist eine Mitarbeit über Github möglich.

Wenn Teile aus der Vorlesung selber fehlen, können diese gerne an meine e-Mail versandt werden. Ich werde diese dann einarbeiten.

Inhaltsverzeichnis

1 Einletung und Motivation

1.1 Parameterabhängige PDGL

Sei $\Omega \subseteq \mathbb{R}^d$ ein polygonales Gebiet. Zu einem Parametervektor $\mu \in P \subseteq \mathbb{R}^d$ aus einer Menge von 'erlaubten' Parametern ist eine Funktion, z.B. 'Temperatur'

$$u(\mu) : \Omega \to \mathbb{R}$$

gesucht, so dass $-\nabla\big(\kappa(\mu)\nabla u(\mu)\big) = q(\mu)$ in Ω, wobei $u(\mu) = 0$ auf $\partial\Omega$, mit $\kappa(\mu) : \Omega \to \mathbb{R}$ dem 'Wärmeleitkoeffizient' und $q(\mu)$ eine 'Wärmequelle', z.B. $q(\mu) = 1$. Weiter kann eine Augabe erwünscht sein, z.B.

$$s(\mu) = \frac{1}{|\Omega_s|} \int \lim_{\Omega_s} u(x, \mu) \mathrm{d}x,$$

die mittlere Temperatur auf Ω_s.

1.2 Definition (schwache Formulierung in Hilberträumen)

Sei X ein reeller Hilbertraum. Zu $\mu \in P$ ist gesucht ein $u(\mu) \in X$ und eine Ausgabe $s(\mu) \in \mathbb{R}$, so dass

$$b\big(u(\mu), v; \mu\big) = f(v; \mu),\ s(\mu) = l\big(u(\mu); \mu\big)\ \forall v \in X$$

für eine Bilinearform $b(\cdot, \cdot; \mu) : X \times X \to \mathbb{R}$ und linearen Funktionalen $f(\cdot; \mu), l(\cdot; \mu) : X \to \mathbb{R}$.
Die schwache Formulierung für Beispiel 1.1 lautet:

$$X := H_0^1(\Omega) = \left(f \in L^2(\Omega)\ : + \frac{\mathrm{d}}{\mathrm{d}x_1} f \in L^2(\Omega),\ f|_{\partial\Omega=0} \right)$$

Dann kann man die Bilinearform über

$$b\big(u(\mu), v; \mu\big) := \int_\Omega \kappa(\mu)\nabla u(\mu)\nabla v \mathrm{d}x; f(v; \mu) := \int \lim_\Omega q(\mu) v \mathrm{d}x$$

ausdrücken und

$$s(\mu) = \frac{1}{|\Omega_s|} \int \lim_{\Omega_s} u(x; \mu)\mathrm{d}x =: l\big(u(\mu); \mu\big)$$

ABER: Für sehr wenige PDGL's können wir die Lösung analytisch bestimmen. Daher sind wir an einer numerische Approximation interessiert. Ein weit verbreitetes Diskretisierungsverfahren ist die Finite Elemente Methode. Diese Methode basiert auf obiger schwacher Formulierung.

1.3 Definition (hochdimensionales, diskretes Modell)

Sei $X_h \subseteq X$ mit $\dim(X_h) = N_h < \infty$. Der Index h bezeichnet hier die Gitterweite. Zu $\mu \in P$ ist gesucht ein $u_h(\mu) \in X_h$ und eine Ausgabe $s_h(\mu) \in \mathbb{R}$, so dass

$$b\big(u_h(\mu), v_h; \mu\big) = f(v_h; \mu),\ s_h(\mu) = l_h\big(v_h(\mu); \mu\big)\ \forall v_h \in X_h. \tag{1.1}$$

Anwendungen für die Standarddiskretisierungsverfahren sehr teuer oder zu teuer sind:

many-query context

- Parameterstudien
- Design
- Parameteridentifikation / inverse Probleme

- Optimierung
- Statistische Analyse

schnelle Simulationsantwort

- Echtzeit-Steuerung technischer Geräte
- interaktive Benutzeroberflächen

1.4 Parameterabhängige Lösungsmenge

Sei $X := \{u(\mu) \ : \ \mu \in P\} \subseteq P$ für $P \in \mathbb{R}^p$ ist die durch μ parametrisierte Lösungsmenge. X ist im Allgemeinen unendlichdimensional. $\Rightarrow$ Motivation für die Suche nach einem 'niedrigdimensionalen' Teilraum $X_N \subseteq X$ zur Approximation von M und einer Approximation $u_N(\mu) \approx u(\mu),\ u_N \in X_N$. Eine Möglichkeit eine reduzierte Basis zu generieren besteht darin geschickt Parameterwerte $\mu_1, \ldots, \mu_N \in P$ zu wählen und den Raum als $X_N := \operatorname{span}\{u(\mu_1), \ldots, u(\mu_N)\}$ zu definieren. Eine Lösung $u(\mu_i)$ für einen Parameterwert $\mu \in P$ wird auch **Snapshot** genannt.

1.5 Beispiel

Gesucht ist $u(\cdot;\mu) \in C^2([0,1])$ mit $(1+\mu)u'' = 1$ auf $(0,1)$ und $u(0) = u(1) = 1$ für den Parameter $\mu \in P := [0,1] \subseteq \mathbb{R}$.
Snapshots:
$\mu_1 = 0 \Rightarrow u_1 := u(\cdot;\mu_1) = \frac{1}{2}x^2 - \frac{1}{2}x + 1$, $\mu_2 = 1 \Rightarrow u_2 := u(\cdot;\mu_2) = \frac{1}{4}x^2 - \frac{1}{4}x + 1$ und $X_N := \operatorname{span}\{u_1, u_2\}$. Dann ist die reduzierte Lösung $u_N(\mu) \in X_N$ gegeben durch

$$u_N(\mu) = \alpha_1(\mu)u_1 + \alpha_2(\mu)u_2,$$

mit $\alpha_1 = \frac{2}{\mu+1} - 1$ und $\alpha_2 = 2 - \frac{2}{\mu-1}$. Diese erfüllt folgende Fehleraussage und ist somit exakt:

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

Da $\alpha_1 + \alpha_2 = 1$ und $0 \leq \alpha_1, \alpha_2 \leq 1$ ist M die Menge der Konvexkombinationen von u_1 und u_2.

1.6 Definition (reduziertes Modell)

Sei $X_N \subseteq X$ ein reduzierter Basisraum mit $\dim(X_N) < \infty$. Zu $\mu \in P$ ist gesucht ein $u_N(\mu) \in X_N$ und eine Ausgabe $s_N(\mu) \in \mathbb{R}$, so dass

$$b\big(u_N(\mu), v_N; \mu\big) = f(v_N;\mu),\ s_N(\mu) = l_N\big(u_N(\mu);\mu\big)\ \forall v_N \in X_N \tag{1.2}$$

1.7 Bermerkung (Begrifflichkeit)

Zusammengefasst unterscheiden wir zwischen den folgenden drei Modellen:

1) Eine partielle DGL ist ein **analytisches Modell**, welches die analytische Lösung $u(\mu) \in X$ in einem (typischerweise) ∞-dimensionalen Funktionenraum charakterisiert ist.

2) Ein **hochdimensionales, diskretes Modell** ist ein Berechnungsverfahren zur Bestimmung einer Näherung $u_h(\mu) \in X_h$, wobei X_h ein hochdimensionaler Funktionenraum ist. Beispiele sind **Finite Elemente** oder **Finite Volumenräume** und typischerweise hat X_h eine Dimension von mindestens 10^5.

3) Ein **reduziertes Modell** ist ein Berechnungsverfahren zur Bestimmung einer Näherung $u_N(\mu) \in X_N$ in einem sehr problemangepassten und daher niedrigdimensionalen Raum von typischerweise $\dim X_N < 100$.

4) **Modellreduktion** beschäftigt sich mit Modellen der Erzeugung von reduzierten Modellen aus hochdimensionalen, diskreten (oder auch analytischen) Modellen und Untersuchungen ihrer Eigenschaften.

1.8 Organisation der Vorlesung

Zentrale Fragen:

- **Reduzierte Basis:** Wie kann ein möglichst kompakter Teilraum konstruiert werden?
- **Reduziertes Modell:** Existenz von reduzierten Lösungen $u_N(\mu)$? Wie kann eine reduzierte Lösung $u_N(\mu)$ berechnet werden?
- **Effizienz:** Wie kann $u_N(\mu)$ schnell berechnet werden?
- **Stabilität:** Wie kann die Stabilität des reduzierten Modells für wachsendes N garantiert werden?
- **Approximationsgüte:** Warum können wir erwarten, dass eine relativ kleine Anzahl von Basisfunktionen ausreicht?
- **Fehlerschätzer:** Kann der Fehler des reduzierten zum vollen Modell beschränkt werden?
- **Effektivität:** Kann garantiert werden, dass der Fehlerschätzer den Fehler nicht beliebig überschätzt?

Vorläufige Gliederung (bis Weihnachten)

1) Einleitung / Moitavtion

2) Grundlagen:
   - Kurze Einführung in lineare Funktionalanalysis
   - Kurze Einführung in Finite Elemente

3) Reduzierte Basis Methoden für lineare, koerzive Probleme
   - Reduzierte Basis Verfahren
   - Offline-/ Online-Zerlegung
   - Fehlerschätzer
   - Basisgenerierung

2 Grundlagen

2.1 Lineare Funktionalanalysis in Hilberträumen

2.1.1 Lineare Operatoren

2.1 Definition (Hilbertraum)

Sei X ein reeller Vektorraum mit $(\cdot,\cdot) : X \times X \to \mathbb{R}$ ein Skalarprodukt und induzierter Norm $\|x\| := \sqrt{(x,x)}$, falls X vollständig bzgl. $\|.\|$, ist X ein (reeller) **Hilbertraum** (HR).

2.2 Beispiele (Hilbertraum)

(1) $X := \mathbb{R}^d$ mit $(x, y) := \sum_{i=1}^{d} x_i y_i$ ist ein HR.

(2) $X := L^2(\Omega)$ mit $(x, y) := \int_\Omega f(x)g(x)\mathrm{d}x$ ist ein HR.

(3) $X := C^0([0, 1])$ mit $(f, g) := \int_0^1 f(x)g(x)\mathrm{d}x$ ist kein HR.

2.3 Lemma

Seien X und Y reelle Vektorräume. Ist die Abbildung $T : X \to Y$ linear und $x_0 \in X$, so sind äquivalent:

(1) T ist stetig.

(2) T ist stetig in x_0.

(3) $\sup \lim_{\|x\|_X \leq 1} \|Tx\|_Y < \infty$.

(4) $\exists$ Konstante (mit $\|Tx\|_Y \leq c \|x\|_X \ \forall x \in X$)

2.4 Definition (Lineare Operatoren)

Seinen X und Y reelle Vektorräume. Wir definieren

$$L(X;Y) := (T : X \to Y \ ; \ T \text{ ist linear und stetig}) .$$

Abbildungen in $L(X;Y)$ nennen wir **lineare Operatoren**. Nach Lemma 2.3 (3) ist für jeden Operator $T \in L(X;Y)$ die **Operatornorm** von T definiert durch

$$\|T\|_{L(X;Y)} := \sup_{\|x\|_X \leq 1} \|Tx\|_Y < \infty,$$

oder in kurz $\|T\|$. Es ist $L(X) := L(X;X)$.

2.5 Definition (Spezielle lineare Operatoren)

(1) $X' := L(X;\mathbb{R})$ ist der **Dualraum** von X. Die Elemente von X' nennen wir auch **lineare Funktionale**.

(2) Die Menge der kompakten (linearen) Operatoren von X nach Y ist definiert durch

$$K(X;Y) := \left(T \in L(X;Y) \ ; \ T(\overline{B_1(0)}) \text{ kompakt}\right) .$$

(3) Eine lineare Abbildung $P : X \to X$ heißt (lineare) **Projektion**, falls $P^2 = P$.

(4) Für $T \in L(X;Y)$ ist $\ker(T) := (x \in X \ ; \ Tx = 0)$ der **Nullraum** oder **Kern** von T. Aus der Stetigkeit von T folgt, dass $\ker(T)$ ein abgeschlossener Unterraum ist. Der **Bildraum** von T ist $\mathrm{bild}(T) := (Tx \in Y \ ; \ x \in X)$.

(5) Ist $T \in L(X;Y)$ bijektiv, so ist $T^{-1} \in L(Y;X)$. Dann heißt T (linear, stetiger) **Isomorphismus**.

(6) $T \in L(X;Y)$ heißt **Isometrie**, falls

$$\|Tx\|_Y = \|x\|_X \ \forall x \in X$$

2.6 Beispiel

Sei $g \in L^2(\Omega)$. Dann ist nach der Hölderungleichung durch

$$T_g f := \int_\Omega f(x)g(x)\mathrm{d}x$$

ein Funktional $T_g \in L^2(\Omega)'$ definiert.

2.7 Satz (Projektionssatz)

Sei X ein Hilbertraum und $A \subseteq X$ nicht leer, abgeschlossen und konvex. Dann gibt es genau eine Abbildung $P : X \to A$ mit

$$\|x - Px\|_X = \mathrm{dist}(x, A) = \inf \lim_{y \in A} \|x - y\|_X \ \forall x \in X.$$

Die Abbildung $P : X \to A$ heißt orthogonale Projektion von X auf A.
Beweis: [Alt, Satz 2.2, S.96]

2.8 Folgerung

Ist $A \subseteq X$ nicht-leer, abgeschlossen und Unterraum, so ist P linear und $Px \in A$ charakterisiert durch $(x - Px, a)_X = 0 \ \forall a \in A$. Falls $\dim(A) = n < \infty$ und $(\varphi_i)_{i=1}^h$ Orthonormalbasis von A, gilt

$$Px = \sum_{i=1}^{n} (x, \varphi_i)_X \varphi_i.$$

2.9 Satz (Riesz'scher Darstellungssatz)

Ist X Hilbertraum, so ist $J : X \to X'$ definiert durch

$$J(v)(w) := (v, w)_X \ \forall v, w \in X$$

eine stetige, lineare, bijektive Isometrie. Insbesondere existiert zu $l \in X'$ ein eindeutiger **Riesz Repräsentant** $V_l := J^{-1}(l) \in X$ mit $l(.) = (v_l, .)_X$.

Beweis:
C-S-Ungleichung: $|J(v)(w)| \leq \|v\|_X \|w\|_X$. Dann folgt: $J(v) \in X'$ mit

$$\|J(v)\|_{X'} = \sup_{w \in X \setminus \{0\}} \frac{|J(v)(w)|}{\|w\|_X} = \sup_{w \in X \setminus \{0\}} \frac{|(v, w)_X|}{\|w\|_X} \leq \|v\|_X \Rightarrow J \text{ stetig.}$$

Da $|J(v)(v)| = \|v\|_X^2$ folgt:

$$\sup_{w \in X \setminus \{0\}} \frac{|J(v)(w)|}{\|w\|_X} \geq \frac{|J(v)(v)|}{\|v\|_X} = \frac{\|v\|_x^2}{\|v\|_X} = \|v\|_X .$$

Also ist J eine Isometrie und insbesondere ist J injektiv.
Zeige J surjektiv: Sei $l \in X'$, $l \neq 0$, Kern(l) ist abgeschlossener Teilraum, also existiert $P : X \to \ker(l)$ orthogonale Projektion nach Satz 2.7. Sei $v_0 \in X$mit $l(v_0) = 1$. Setze $v_1 := v_0 - Pv_0 \Rightarrow l(v_1) = l(v_0) = 1$ und $v_1 \neq 0$. Mit Folgerung 2.8:

$$\Rightarrow (w, v)_X = 0 \ \forall x \in \ker(l) \Rightarrow v_1 \perp \ker(l).$$

Für $v \in X$ gilt

$$v \underbrace{v - l(v)}_{\in \ker(l)} \cdot v_1 + l(v) \cdot v_1$$

und $v - l(v)v_1 \in \ker(l)$ wegen

$$l(v - l(v)v_1) = l(v) - l(v)l(v_1) = 0.$$

Also ist

$$\begin{aligned}(v_1, v)_X &= (\underbrace{v_1, v - l(v)v_1}_{=0,\ \text{da}\ \ker(l) \perp v_1})_X + (v_1, l(v)v_1)_X \\ &= l(v) \|v_1\|_X^2 \\ &\Rightarrow l(v) = \left(\frac{v_1}{\|v_1\|_X^2}, v\right)_X = J\left(\frac{v_1}{\|v_1\|_X^2}\right)(v). \\ &\Rightarrow l \in \mathrm{bild}(J) \Rightarrow J \text{ bijektiv.}\end{aligned}$$

□

2.10 Folgerung / Beispiel:

Mit Hilfe des Rieszschen Darstellungssatz können wir damit $L^2(\Omega)$' - den Dualraum von $L^2(\Omega)$ - charakterisieren. Wie in 2.6 definieren wir für $g \in L^2(\Omega)$ das Funktional

$$T_g f := \int_\Omega f(x)g(x)\mathrm{d}x.$$

Definition 2.11 (Bilinearformen)

Seien X_1, X_2 Hilberträume, $b : X_1 \times X_2 \to \mathbb{R}$ eine Bilinearform.

(1) Falls

$$\gamma := \sup_{u \in X_1 \setminus \{0\}} \lim \quad \sup_{v \in X_2 \setminus \{0\}} \lim \quad \frac{b(u,v)}{\|u\|_{X_1} \|v\|_{X_2}} < \infty$$

so ist b stetig mit Stetigkeitskonstante γ.

(2) Falls $X = X_1 = X_2$, definieren

$$b_s(u,v) = \frac{1}{2}b(u,v) + b(v,u),\ b_a = \frac{1}{2}b(u,v) - b(v,u)\ \forall u, v \in X$$

den symmetrischen bzw. antisymmetrischen Anteil von $b = b_s + b_a$.

(3) Falls $X = X_1 = X_2$, b stetig und

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

heißt b Koerziv mit Stetigkeitskonstante α.

2.12 Bemerkung

(1) $\alpha \in \mathbb{R}$ ist wohldefiniert, denn mit Stetigkeit folgt

$$\frac{b(u,u)}{\|u\|_X^2} \geq -\gamma \frac{\|u\|_X \|u\|_X}{\|u\|_X^2} = -\gamma.$$

(2) b ist koerziv bzgl. $\alpha \Leftrightarrow b_s$ ist koerziv bzgl. α.

2.13 Satz (Operatoren und Bilinearformen)

Seien X_1, X_2 Hilberträume.

(1) Zu $B \in L(X_1, X_2)$ existiert eine eindeutig definierte stetige Bilinearform $b : X_1 \times X_2 \to \mathbb{R}$ mit

$$b(u,v) = (Bu, v)_{X_2} \ \forall u \in X_1, v \in X_2. \tag{2.1}$$

(2) Zu $b : X_1 \times X_2 \to \mathbb{R}$ stetige Bilinearform existiert eindeutiges $B \in L(X_1, X_2)$ welches (2.1) erfüllt.

Beweis:

(1) b definiert durch (2.1) ist bilinear wegen Bilinearität von $(.,.)$ und Linearität von B. Stetigkeit:

$$b(u,v) = (Bu,v)_{X_2} \overset{\text{C.S.}}{\leq} \|B\| \|u\|_{X_1} \|v\|_{X_2}$$

daraus folgt $\gamma \leq \|B\| < \infty$.

(2) Sei $u \in X_1$ fest. Dann ist $b(u,.) : X_2 \to \mathbb{R}$ linear und stetig:

$$\sup_{v \in X_2 \setminus \{0\}} \frac{b(u,v)}{\|v\|_{X_2}} \leq \sup_{v \in X_2 \setminus \{0\}} \frac{\|u\|_{X_1 \|v\|_{X_2}}}{\|v\|_{X_2}} \cdot \gamma = \gamma \|u\|_{X_1} < \infty.$$

Daraus folgt $b(u,.) \in X_2'$ und es existiert nach Satz 2.9 ein eindeutiger Riesz-Repräsentant $v_u \in X_2$ mit $b(u,.) = (v_u,.)$. Definiere $B : X_1 \to X_2$ durch $Bu := v_u \in X_2$. Hiermit (2.1) und Eindeutigkeit klar. Linearität damit klar.
Stetigkeit:

$$\|bu\|^2 = (Bu, Bu) = (v_u, Bu)_{X_2} = b(u, Bu) \leq \gamma \|u\|_{X_1} \|Bu\|_{X_2}$$
$$\Rightarrow \|Bu\|_{X_2} \leq \gamma \|u\|_{X_1} \Rightarrow \sup_{u \in X_1 \setminus \{0\}} \frac{\|Bu\|_{X_2}}{\|u\|_{X_1}} \leq \gamma.$$

□

2.14 Satz von Lax-Milgram

Sei x HR, $b : X \times X \to \mathbb{R}$ koerzive, stetige Bilinearform mit Koerzivitätskonstante α. Dann existiert ein eindeutiger Operator $B \in L(X)$ mit

$$b(u,v) = (Bu, v) \ \forall u, v \in X.$$

Ferner gilt: B ist bijektiv, $B^{-1} \in L(X)$ mit

$$\|B\| \leq \gamma \text{ und } \left\|B^{-1}\right\| \leq \frac{1}{\alpha}.$$

2.1.2 Sobolevräume

2.15 Bemerkung (Motivation Sobolevräume)

Wie in 1.1 motiviert, eignet sich die sogenannte Schwache Formulierung (s. 1.2) einer PDgl besonders gut um Existenz und Eindeutigkeit von Lösungen zu untersuchen. Die dazu geeigneten Räume sind die **Sobolevräume**.

2.16 Definition ($L^p_{log}(\Omega)$)

Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet. Dann ist der Raum $L^p_{log}(\Omega)$ definiert durch

$$L^p_{log}(\Omega) := \{u \in L^p(K) \mid \forall K \subset \Omega,\ K \text{ kompakt}\}\,.$$

2.17 Definition (schwache Ableitung)

Sei $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$ ein Multiindex. Eine Funktion $u \in L^1_{log}(\Omega)$ besitzt eine schwache Ableitung $u_\alpha \in L^1_{log}(\Omega)$, wenn für alle Testfunktionen $\varphi \in C_0^\infty(\Omega)$ gilt

$$\int_\Omega u D^\alpha \varphi = (-1)^{|\alpha|} \cdot \int_\Omega u^{(\alpha)} \varphi,$$

mit $D^\alpha = D_1^{\alpha_1} \cdots D_d^{\alpha_d},\ |\alpha| = \alpha_1 + \cdots + \alpha_d$. Wir schreiben dann auch $u^{(\alpha)} = D^\alpha u$ für die schwache Ableitung.

2.18 Lemma

Falls $u \in C^{|\alpha|(\bar{\Omega})}$ und $|\alpha| \geq 1$, gilt: $D^\alpha = u^{(\alpha)}$, d.h. klassische und schwache Ableitung stimmen überein.

2.19 Beispiel

Sei $\Omega = (-1, 1)$ und $u(x) = |x|$. Dann ist $u'(x) = -1(x \leq 0), 1(x > 0)$ die schwache Ableitung von u.

Beweis:
Es gilt für beliebige $\varphi \in C_0^\infty(\Omega)$:

$$Foto$$

2.20 Beispiel

Im Gegensatz zu $|x|$ ist $v(x) = -1(x \leq 0)1(x > 0)$ auf $\Omega = (-1, 1)$ nicht schwach differenzierbar.

2.21 Definition (Sobolevräume)

Seien $m \in \mathbb{N}_0,\ p \in [1, \infty]$ und $u \in L^p_{log}(\Omega)$. Wir nehmen an, dass alle schwachen partiellen Ableitungen $D^\alpha u$ existieren für $|\alpha| \leq m$. Dann definieren wir die **Sobolevnormen** $\|u\|_{H^{m,p}(\Omega)}$, durch

$$\|u\|_{H^{m,p}(\Omega)} = \left(\sum_{|\alpha| \leq m} \|D^\alpha u\|^p_{L^p(\Omega)} \right)^{\frac{1}{p}} \quad \text{falls } 1 \leq p < \infty$$

und für $p = \infty$ als

$$\|u\|_{H^{m,p}(\Omega)} := \max \lim_{|\alpha| \leq m} \|D^\alpha u\|_{L^\infty(\Omega)}\,.$$

Schließlich definieren wir die **Sobolevräume** $H^{m,p}(\Omega)$ durch

$$H^{m,p}(\Omega) := \left\{ u \in L^p_{log}(\Omega) \mid \|u\|_{H^{m,p}(\Omega)} < \infty \right\}.$$

2.22 Bemerkung

Anstelle von $H^{m,p}(\Omega)$ werden die Sobolevräume in der Literatur auch oft mit $W^{m,p}(\Omega)$ bezeichnet.

2.23 Beispiel

Seien $\Omega = B_{\frac{1}{2}}(0) \subset \mathbb{R}^2$ und $u(x) = \ln|\ln|x||,\ x \in \Omega$. Dann gilt: $u \in H^{1,2}(\Omega)$, aber $u \notin C^0(\Omega)$. D.h. Funktionen in $H^{1,p}(\Omega)$ sind in mehreren Raumdimensionen nicht notwendigerweise stetig.

2.24 Satz (Vollständigkeit von Sobolevräumen)

Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet. Dann ist $H^{m,p}(\Omega)\ 1 \leq p \leq \infty,\ m \in \mathbb{N}_0$ mit der in 2.21 definierten Norm ein Banachraum,$H^{m,p}(\Omega)$ ist ein Hilbertraum mit dem Skalarprodukt

$$(u,v)_{H^{m,p}(\Omega)} := \sum_{|\alpha| \leq m} (D^\alpha u, D^\alpha v)_{L^2(\Omega)}.$$

Da wir uns mit Randwertproblemen befassen wollen, ist es notwendig zu klären in welchem Sinne wir bei Sobolevräumen von Randwerten reden können. Da die Funktionen zunächst nur bis auf Nullmengen definiert sind und der Rand eines Gebietes eine Nullmenge darstellt, auf der man L^p-Funktionen beliebig abändern kann. In der folgenden Definition klären wir zunächst was wir unter Nullrandwerten im schwachen Sinne verstehen wollen.

2.25 Definition (schwache Nullrandwerte)

Für $1 \leq p \leq \infty$ und $m \in \mathbb{N}$ definieren wir die Sobolevräume mit Nullrandwerten $H_0^{m,p}(\Omega)$ durch

$$H_0^{m,p}(\Omega) := \overline{C_0^m(\Omega)}^{\|.\|_{H^{m,p}(\Omega)}}.$$

2.26 Satz

Für $1 \leq p < \infty$ ist $H_0^{m,p}(\Omega)$ ein abgeschlossener Teilraum von $H^{m,p}(\Omega)$ und damit ein Banachraum.

Dass aus der Definition von $H_0^{m,p}(\Omega)$ tatsächlich folgt, dass solche Funktionen Randwerte besitzen, drückt der folgende Satz aus.

2.27 Satz (Spursatz)

Sei $\Omega \subset \mathbb{R}^d$ ein Lipschitz-Gebiet und $1 \leq p < \infty$. Dann gibt es einen linearen **Spuroperator** $\tau : H^{1,p}(\Omega) \to L^p(\partial\Omega)$, so dass für $u \in H^{m,p}(\Omega) \cap C^\infty(\bar{\Omega})$ gilt:

$$\tau u = u|_{\partial\Omega}.$$

Insbesondere gilt für $u \in H_0^{1,p}(\Omega) : \tau u = 0$.

Beweis:
Im Buch von Alt oder von Evans.

2.28 Satz (2. Soblev'scher Einbettungssatz)

Sei $1 \leq p < \infty$, dann gilt:

$$H^{1,p}((a,b)) \hookrightarrow C^0([a,b]),$$

d.h. dass (möglicherweise nach Änderung von Funktionswerten auf einer Nullmenge) Funktionen in $H^{1,p}((a,b))$ stetig sind.

Sei nun $\Omega \subset \mathbb{R}^d$ ein Gebiet und $1 \leq p < \infty$. Dann gilt

$$H_0^{2,p}(\Omega) \hookrightarrow C^0(\Omega), \text{ falls } 2 - \frac{d}{p} > 0.$$

Ist Ω ein Lipschitz-Gebiet, so gilt diese Aussage auch für Sobolevräume ohne Nullrandwerte.

Beweis:
Im Buch von Alt.

2.29 Satz (Poincaré-Friedrichs Ungleichung)

Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet mit Durchmesser $D := \operatorname{diam}(\Omega)$ und $1 \leq p < \infty$. Dann gibt es eine Konstante $c_p \leq 2D$, so dass für alle $v \in H_0^{1,p}(\Omega)$ gilt:

$$\|v\|_{L^p(\Omega)} \leq c_p \|\nabla v\|_{L^p(\Omega)}.$$

Beweis:
Siehe Buch von Dziuk.

2.1.3 Schwache Formulierung elliptischer Randwertprobleme

Wir betrachten zunächst die stationäre Wärmeleitgleichung. Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet mit glattem Rand und seien $q \in C^0(\Omega)$ und $\kappa \in C^1(\Omega)$ mit $\kappa \geq \kappa_1 > 0$, $\kappa_1 \in \mathbb{R}$ Konstante. Gesucht ist eine Funktion $u \in C^2(\Omega) \cap C^0(\bar{\Omega})$, die sogenannte klassische Lösung, so dass

$$\begin{aligned} -\nabla(\kappa \nabla u) &= q \text{ in } \Omega, \\ u &= 0 \text{ auf } \partial\Omega. \end{aligned} \tag{2.2}$$

Mit Hilfe von schwachen Ableitungen und den Sobolevräumen können wir nun den klassischen Lösungsbegriff verallgemeinern:

2.30 Defintion (schwache Formulierung der stationären Wärmeleitgleichnung)

Seien $\Omega \subset \mathbb{R}^d$ ein Lipschitz-Gebiet, $q \in L^2(\Omega)$ und $\kappa \in L^\infty(\Omega)$ mit $0 < \kappa_1 \leq \kappa$ für eine Konstante $\kappa_1 \in \mathbb{R}$ gegeben. Dann heißt $u \in H_0^{1,p}(\Omega)$ schwache Lösung des Randwertproblems der stationären Wärmeleitgleichung (2.2), falls für alle Testfunktionen $v \in H_0^1(\Omega)$ gilt

$$\int_\Omega \kappa(x) \nabla u(x) \nabla v(x) \mathrm{d}x = \int_\Omega q(x) v(x) \mathrm{d}x.$$

2.31 Satz (Existenz und Eindeutigkeit von Lösungen)

Unter den Voraussetzungen von Def. 2.30 gibt es genau eine schwache Lösung $u \in H_0^1(\Omega)$ des Randwertproblems der stationären Wärmeleitgleichung.

Beweis:
Zunächst wird durch $l(v) := \int_\Omega q(x) v(x) \mathrm{d}x$ ein lineares Funktional in $(H_0^1(\Omega))'$ definiert, denn

$$\|l(v)\|_{(H_0^1(\Omega))'} = \sup_{v \in (H_0^1(\Omega)) \setminus \{0\}} \frac{(q, v)_{L^2(\Omega)}}{\|v\|_{H^1(\Omega)}} \leq \|q\|_{L^2(\Omega)} \infty.$$

Ferner wird wegen der Poincaré-Friedrichs Ungleichung durch

$$(w, v)_{H_0^1(\Omega)} := \int_\Omega \nabla w(x) \nabla v(x) \mathrm{d}x$$

ein Skalarprodukt auf dem Hilbertraum $H_0^1(\Omega)$ definiert. Daher existiert nach dem Riesz'schen Darstellungsatz 2.9 ein eindeutiger Riesz-Repräsentant w_l mit $l(.) = (w_l, .)_{H_0^1(\Omega)}$.
Um den Beweis zu schließen, müssen wir noch nachweisen, dass die Bilinearform $b : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ definiert durch

$$b(w, v) := \int_\Omega \kappa(x) \nabla w(x) \nabla v(x) \mathrm{d}x$$

die Voraussetzungen des Satzes von Lax-Milgram erfüllt. Wir müssen also zeigen, dass die Bilinearform b stetig und koerziv ist.
Stetigkeit:

$$b(w, v) \leq \|\kappa\|_{L^\infty(\Omega)} \|w\|_{H_0^1(\Omega)} \|v\|_{H_0^1(\Omega)} .$$

Koerzivität:

$$b(w, v) =\geq \kappa_1 \|w\|^2_{H_0^1(\Omega)} .$$

Damit existiert ein bijektiver Operator $B \in L(H_0^1(\Omega))$ mit $b(u, v) = (Bu, v)_{H_0^1(\Omega)}$ und wir definieren die eindeutige Lösung $u \in H_0 h\Omega)$ des Randwertproblems als $u := B^{-1} w_l$, wobei w_l der eindeutige Riesz-Repräsentant mit $l(.) = (w_l, .)_{H_0^1(\Omega)}$ war. □

2.32 Bemerkung

Mit der gleichen Beweistechnik lassen sich auch allgemeinere PDgl'en behandeln, wie zum Beispiel das Randwertproblem in Divergenzform

$$\begin{aligned} -\nabla(A(x)\nabla u) + b(x)\nabla u + c(x)u &= q \text{ in } \Omega, \\ u &= 0 \text{ auf } \partial\Omega. \end{aligned}$$

$A(x) \in C^1(\Omega, \mathbb{R}^{d\times d})$, $b(x) \in C^0(\Omega, \mathbb{R}^d)$, $c(x) \in C^1(\Omega)$, wobei die Koeffizienten gewisse Anforderungen erfüllen müssen damit die Koerzivität der entsprechenden Bilinearform nachgewiesen werden können.

2.33 Bemerkung (Reduktion auf Nullrandwerte)

Zur Betrachtung von allgemeinen Dirichletrandwerten, kann man wie folgt vorgehen. Seien $g_D \in H^1(\Omega)$ und $q \in L^2(\Omega)$, $\kappa \in L^\infty(\Omega)$, $\kappa \geq \kappa_1 > 0$. Dann ist $u \in H^1(\Omega)$ schwache Lösung von $-\nabla(\kappa(x)\nabla u(x)) = q$ in Ω und $u = g$ auf $\partial\Omega$. Wenn gilt $\tilde{u} := u - g_D \in H_0^1(\Omega)$ und für alle $v \in H_0^1(\Omega)$ gilt (2.3). Dabei ist zu bemerken, dass mit der Definition von $\tilde{u}$ (2.3) äquivalent ist zu

$$\int_\Omega \kappa(x)\nabla\tilde{u}\nabla v(x)\mathrm{d}x = \int_\Omega q(x)v(x)\mathrm{d}x - \int_\Omega \nabla g_D(x)\nabla v(x)\mathrm{d}x.$$

Die Existenz und Eindeutigkeit einer Lösung folgt dann daraus, dass durch $l(v) := \int_\Omega q(x)v(x)\mathrm{d}x - \int_\Omega \nabla g_D(x)\nabla v(x)\mathrm{d}x$ ein lin. Funktional in $(H_0^1(\Omega))'$ definiert wird.

2.34 Definition (schwache Formulierung)

Seien X reeller Hilbertraum, $b : X \times X \to \mathbb{R}$ eine stetige und koerzive Bilinearform mit Stetigkeitskonstante γ und Koerzivitätskonstante α und $l \in X'$. Dann bezeichnen wir mit $u \in X$ die eindeutige Lösung des Problems

$$b(u, v) = l(v) \ \forall v \in X. \tag{2.3}$$

2.1.4 Regularität

Zur Motivation betrachte in einer Raumdimension die Dgl. $u''(x) = q(x)$ mit einer stetigen Funktion $q(x)$. Dann folgt mit dem Hauptsatz der Differential- und Integralrechnung, dass bereits $u \in C^2$ gelten muss.

2.35 Satz (H^2-Regularität)

Sei Ω ein Gebiet mit glattem Rand (es gelte $\partial\Omega$ ist in C^2) oder ein konvexes Lipschitz-Gebiet. Ferner seien $q \in L^2(\Omega)$ und $\kappa \in C^1(\bar{\Omega})$. Dann gilt für die eindeutige schwache Lösung $u \in H_0^1(\Omega)$ der stationären Wärmeleitungsgleichung (2.2) dass $u \in H^2(\Omega)$ und dass eine Konstante $c > 0$ existiert, so dass die folgende Abschätzung gilt:

$$\|u\|_{H^2(\Omega)} \leq c \|q\|_{L^2(\Omega)}$$

Beweis:
Für glatten Rand: Buch von Evans.

2.36 Bemerkung

Betrachtet man nicht konvexe Lipschitz-Gebiete, so kann man im Allgemeinen keine Lösung $u \in H^2(\Omega)$ erwarten.

2.2 Ritz-Galerkin Verfahren und abstrakte Fehlerabschätzungen

In diesem Abschnitt wollen wir uns mit der Approximation der Lösung der schwachen Formulierung von (2.4) befassen.

2.37 Definition (Ritz-Galerkin Verfahren)

Seien X, b wie in Definition 2.34 und $X_m \subset X$ mit $\dim(X_m) = m$ ein Unterraum. Dann ist die **Ritz-Galerkin Approximation** $u_m \in X_m$ definiert durch

$$b(u_m, v_m) = l(v_m) \; \forall v_m \in X_m.$$

2.38 Bemerkung

Die Existenz und Eindeutigkeit von u_m folgt unmittelbar aus dem Satz von Lax-Milgram 2.14, da der Unterraum X_m wieder ein Hilbertraum mit dem aus X geerbten Skalarprodukt ist.

2.39 Satz (Abstrakte Fehlerabschätzung/Lemma von Céa)

Seien X, X_m, b, u und u_m wie in den Definitionen 2.34 und 2.37 definiert. Dann gilt die abstrakte Fehlerabschätzung

$$\|u - u_m\|_X \leq \frac{\gamma}{\alpha} \inf_{v_m \in X_m} \|u - v_m\|_X .$$

Außerdem gilt die Galerkin-Orthogonalität

$$b(u - u_m, v_m) = 0 \; \forall v_m \in X_m.$$

Beweis:
Wir zeigen zunächst die Galerkin-Orthogonalität: Dazu sei $v_m \in X_m$ und es folgt mit $X_m \subset X$:

$$b(u - u_m, v_m) = b(u, v_m) - b(u_m, v_m) = l(v_m) - l(v_m) = 0.$$

Mit der Stetigkeit und Koerzivität von b folgt weiter

$$\begin{aligned} \alpha \|u - u_m\|_X^2 &\leq b(u - u_m, u - u_m) = b(u - u_m, u - v_m) \\ &\leq \gamma \|u - u_m\|_X \|u - v_m\|_X \\ &\Rightarrow \|u - u_m\|_X \leq \frac{\gamma}{\alpha} \|u - v_m\|_X \end{aligned}$$

Gehe auf beiden Seiten der Ungleichung zum Infimum über, dann folgt die Behauptung. □

2.40 Bemerkung

Die abstrakte Fehlerabschätzung zeigt, dass der Fehler zwischen Ritz-Galerkin Approximationen und exakter Lösung abgeschätzt werden kann durch die Bestapproximation in dem Teilraum X_m. Die weitere numerische Analyse beruht somit allein auf der Approximationstheorie. Insbesondere bestimmt im wesentlichen der Teilraum X_m die Approximationsgüte.

2.41 Beispiel (mögliche Wahl von Teilräumen)

Betrachten wir konkret die Stationäre Wärmeleitungsgleichung (2.2) oder allgemeiner ein elliptisches Problem in Divergenzform mit $X = H_0^1(\Omega)$ auf einem Gebiet $\Omega \subset \mathbb{R}^d$, so sind neben den Finiten Elemente Verfahren, die wir im nächsten Abschnitt betrachten wollen,vor allem folgende Wahlen von Teilräumen gebräuchlich:

- Polynomräume $X_M := \mathbb{P}^{k(m)}(\Omega) \cap \{v_m \in C^0(\bar{\Omega}) \mid + v_m = 0 \text{ auf } \partial\Omega\}$, wobei $\mathbb{P}^{k(m)}(\Omega)$ der Raum der Polynome mit Grad $\leq k(m)$ über Ω ist. Die zugehörigen Verfahren nennt man **Spektralverfahren**.
- $X_m := \operatorname{span}\{u_i \in X \mid Lu_i = \lambda u_i,\ i = 1, \ldots, m\}$ wobei u_i die i-te Eigenfunktion des zugrundeliegenden Differentialoperators L ist.
- $X_m := \operatorname{span}\{u_i \in X \mid \Delta u_i = \lambda u_i,\ i = 1, \ldots, m\}$, wobei u_i die i-te Eigenfunktion des Laplaceoperators Δ ist.

2.42 Folgerung (Matrix-Vektor von Ritz-Galerkin Verfahren)

Seien X, X_m, b, u un du_m wie in den Definitionen 2.34 und 2.37 definiert und sei zudem X_m endlichdimensional, mit Dimension $m := \dim X_m$. Ist dann $\Phi := \{\varphi_1, \ldots, \varphi_m\}$ eine Basis von X_m so folgt mit der Darstellung $u_m = \sum_{i=1}^m U_i \varphi_i$ aus der Definition von u_m

$$b\left(\sum_{i=1}^m U_i \varphi_i, \varphi_j\right) = l(\varphi_j),\ j = 1, \ldots, m.$$

Durch Ausnutzen der Linearität von b im 1. Argument folgt weiter:

$$\sum_{i=1}^m b(\varphi_i, \varphi_j) U_i = l(\varphi_j),\ j = 1, \ldots, m.$$

Definieren wir also die Matrix $\mathbf{S} \in \mathbb{R}^{m \times m}$ durch $\mathbf{S}_{ji} := b(\varphi_i, \varphi_j),\ i, j = 1, \ldots, m$ und die Vektoren $\mathbf{u}, \mathbf{l} \in \mathbb{R}^m$ durch $\mathbf{u}_i := U_i,\ \mathbf{l}_i := l(\varphi_i),\ i = 1, \ldots, m$, so ist u_m genau dann Lösung des Ritz-Galerkin Verfahren, wenn u das folgende lineare Gleichungssystem löst: $\mathbf{Su} = \mathbf{l}$.

2.3 Finite Elemente Verfahren

Finite Elemente Verfahren sind Spezialfälle von Ritz-Galerkin Verfahren für eine bestimmte Klasse von Teilräumen $X_h \subset X$, wobei X_h der **Finite Elemente Raum** ist. Die Konstruktion von x_h im Falle von Finite Elemente (FE) Verfahren beruht auf einer Zerlegung des Gebietes Ω in nicht überlappende Teilgebiete, die selbst wiederum einfache geometrische Objekte sind. Die einfachste Klasse von Finiten Elementen sind Lagrange Elemente, auf welche wir uns in dieser Vorlesung beschränken werden. Ferner betrachten wir nur Teilräume X_h welche auf einer simplizialen Zerlegung des Gebietes Ω beruhen. In zwei Raumdimensionen besteht das Rechengitter aus Dreiecken, in drei Raumdimensionen aus Tetraedern. Eingeschränkt auf einen Simplex wird eine Funktion aus X_h, dann ein Polynom mit Grad $\leq k,\ k \in \mathbb{N}$ sein. Für andere FE siehe z.B. das Buch von Brenner und Scott.

2.43 Definition (Simplex)

Seien $s \in \{1, \dots, d\}$ und $a_0, \dots, a_s \in \mathbb{R}^d$ Punkte, so dass $(a_j - a_0)_{j=1,\dots,s}$ linear unabhängig sind. Dann heißt

$$T := \left\{ x \in \mathbb{R}^d \mid x = \sum_{i=0}^{s} \lambda_i a_i,\ 0 \le \lambda_i,\ \sum_{i=0}^{s} \lambda_i = 1 \right\}$$

nicht-degeneriertes s-dimensionaler Simplex im $\mathbb{R}^d$. Die Punkte $a_0, \dots, a_s$ heißen Ecken des Simplex. Ist $r \in \{0, \dots, s\}$ und $\tilde{a}_0, \dots, \tilde{a}_r \in \{a_0, \dots, a_s\}$, so heißt

$$\tilde{T} := \left\{ x \in \mathbb{R}^d \mid x = \sum_{i=0}^{s} \lambda_i \tilde{a}_i,\ 0 \le \lambda_i,\ \sum_{i=0}^{s} \lambda_i = 1 \right\}$$

r-dimensionales Seitensimplex von T. Die nulldimensionalen Seitensimplexe heißen Ecken, die eindimensionalen Kanten. Wir bezeichnen mit T_0 den Simplex zu den Punkten $a_0 = e_0 = (0, \dots, 0), a_i = e_i,\ i = 1, \dots, d$, T_0 heißt d-dimensionaler Einheitssimplex. Der **Durchmesser** von T ist gegeben durch $h(T) := \operatorname{diam}(T) = \max_{i,j=1,\dots,s} |a_i - a_j|$. Mit

$$\rho(T) := 2 \sup\{R \mid B_R(x_0) \subset T\}$$

bezeichnen wir den **Inkugeldurchmesser** von T und mit

$$\delta(T) := \frac{h(T)}{\rho(T)}$$

den Quotienten aus h und ρ.

2.44 Definition (Baryzentrische Koordinaten)

Die baryzentrischen Koordinaten $\lambda_0, \dots, \lambda_s \in [0, 1]$ eines Punktes $x \in T$ des s-dim. Simplex T sind die Lösung des linearen Gleichungssystems

$$x = \sum_{i=0}^{s} \lambda_i a_i,\ \sum_{i=0}^{s} \lambda_i = 1.$$

Der Schwerpunkt x_s von T ist definiert durch $x_s := \frac{1}{s+1} \sum_{i=0}^{s} a_i$ und hat die baryzentrischen Koordinaten $\lambda_i := \frac{1}{s+1}$. Für die Eckpunkte a_k von T sind die baryzentrischen Koordinaten gegeben durch $\lambda_k = 1,\ \lambda_i = 0,\ i \neq k$. Die baryzentrischen Koordinaten sind eindeutig bestimmt.

2.45 Lemma (Referenzabbildung)

Jedes s-dimensionale Simplex T im $\mathbb{R}^s$ ist affin äquivalent zum Einheitssimplex T_0 der gleichen Dimension. Die eindeutige affine Abbildung $F : T_0 \to T,\ F(x) = Ax + b,\ A \in \mathbb{R}^{s \times s},\ b \in \mathbb{R}^s,\ \det A \neq 0$ mit $F(e_j) = a_j,\ j = 0, \dots, s$ heißt **Referenzabbildung**. F ist invertierbar und es gelten die Abschätzungen

$$\|\nabla F\| = \|A\| \le \frac{h(T)}{\rho(T_0)},\ \left\|\nabla(F^{-1})\right\| = \left\|A^{-1}\right\| \le \frac{h(T_0)}{\rho(T)}$$

sowie

$$c \cdot \rho(T)^s \le |\det(\nabla F)| = |\det A| = \frac{|T|}{|T_0|} \le C \cdot h(T)^s,\ c, C > 0.$$

Beweis:
Im Buch von Dziuk.

2.46 Definition (Zulässige Triangulierung)

Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet und

$$\mathbb{T}_h := \left\{ T_j \mid j = 1, \dots, m,\ T_j \text{ ist } d\text{-dim. Simplex im } \mathbb{R}^d \right\}.$$

$\mathbb{T}_h$ heißt zulässige Triangulierung der Feinheit h und Gute ρ von Ω, falls gilt:

$$\bar{\Omega} = \bigcup_{j=1}^{m} T_j,\ \partial\Omega = \bigcup_{j=1}^{m} \tilde{T}_j,$$

wobei $\tilde{T}_j$ Flächen der Simplexe T_j sind. Für je zwei $T_1, T_2 \in \mathbb{T}_h$ mit $S := T_1 \cap T_2$ gilt $S = \emptyset$ oder S ist $(d-k)$-dim. Seitensimplex von T_1 und T_2 für ein $k \in \{1, \dots, d\}$. Mit $h := \max_{j=1,\dots,m} h(T_j)$ und $\rho := \min_{j=1,\dots,m} \rho(T_j)$.

Zur Definition von Finite Elemente Räumen basierend auf einer Triangulierung $\mathbb{T}_h$ müssen wir nun lediglich lokale Funktionenräume auf dem Simplexen $T \in \mathbb{T}_h$ angeben und festlegen wie solche lokalen Funktionen global zusammengesetzt werden. Ein Tripel bestehend aus einem geometrischen Objekt T, einer lokalen Basis Φ und lokalen Freiheitsgraden δ, wollen wir um folgenden **Element** nennen.

2.47 Definition (lineares simpliziales Lagrange Element)

Sei $T \subset \mathbb{R}^d$ ein d-dim. Simplex. Sei $\delta := \{a_k \mid k = 0, \dots, d\}$ die Menge der Ecken von T. Dann ist durch Angabe von Werten in den Punkten $a_k \in \delta$ eindeutig eine lineare Funktion $p \in \mathbb{P}^1(T)$ definiert. Durch $\Phi := \{\varphi_i \mid \varphi_i(a_k) = \delta_{ik}\ i, k = 1, \dots, d\}$ ist eine modale Basis von $\mathbb{P}^1(T)$ gegeben. Wir nennen das Tripel (T, Φ, δ) **lineres simpliziales Lagrange Element**. Die Basisfunktionen $\varphi_i \in \Phi$ werden **Formfunktionen** oder im Englischen **Shapefunctions** genannt und δ ist die Menge der **modalen Variablen**. Zur Wohldefiniertheit kann man im Buch von Dziuk nachschauen.

2.48 Beispiel (lineares Lagrange Element für $d = 2$)

Wir betrachten das Einheitsdreieck T_0 mit Eckpunkten $a_0^0 = (0,0), a_1^0 = (1,0), a_2^0 = (0,1)$. Die Formfunktionen sind dann gegeben durch

$$\varphi_0^0(x,y) = 1 - x - y,\ \varphi_1^0(x,y) = x,\ \varphi_2^0(x,y) = y.$$

Sind $p(a_0^0), p(a_1^0), p(a_2^0)$ Funktionswerte einer linearen Funktion $p \in \mathbb{P}^1(T_0)$, so ist p gegeben durch

$$p(x,y) = \sum_{i=0}^{2} p(a_i^0)\varphi_i^0(x,y).$$

Für ein beliebiges Dreieck $T \subset \mathbb{R}^2$ erhält man das Lagrange Element mit Hilfe der Referenzabbildung $F : T_0 \to T$ aus Lemma 2.45.

2.49 Bemerkung

Das Beispiel 2.48 zeigt, dass es ausreicht ein Finites Element auf einer Referenzgeometrie zu definieren. Durch die Referenzabbildung erhält man dann die entsprechende Klasse von Elementen auf beliebigen Geometrien im Raum.

Ein Finites Element legt lediglich ein lokalen Funktionenraum auf einen Simplex T fest, um zu einem Unterraum von $H_0^1(\Omega)$ zu gelangen, müssen wir zusätzlich festlegen auf welche Weise die lokalen Funktionen global zusammengesetzt werden.

2.50 Definition (linearer Finite Elemente Raum S_h^1)

(1) Sei $\Omega \subset \mathbb{R}^d$ und $\mathbb{T}_h$ eine zulässige Triangulierung von Ω. Wir definieren den Raum der **linearen Finite Elemente** auf simplizialen Gittern S_h^1 durch

$$S_h^1 := \left\{ v_h \in C^0(\Omega) \mid v_h|_T \in \mathbb{P}^1(\Omega),\ T \in \mathbb{T}_h \right\}.$$

(2) Sind $\bar{a}_j,\ j = 1, \ldots, N_h$ die Ecken der Triangulierung $\mathbb{T}_h$, so ist eine Funktion $v_h \in S_h^1$ durch die Vorgabe von Funktionswerten in den Ecken $v_h(\bar{a}_j)$ eindeutig definiert. Insbesondere gilt $\dim(S_h^1) = \mathcal{N}_h$. Eine Basis von S_h^1 ist durch die Funktionen

$$\bar{\varphi}_i \in S_h^1,\ \bar{\varphi}_i(\bar{a}_j) = \delta_{ij},\ i, j = 1, \ldots, \mathcal{N}$$

gegeben. Diese Basis heißt **Knotenbasis** oder **modale Basis**.

(3) Ist (T_0, Φ, δ) das lineare Lagrange Element auf dem Einheitssimplex T_0 und $v_h \in S_h^1$ gegeben durch

$$v_h(x) := \sum_{i=1}^{\mathcal{N}_h} v_n(\bar{a}_i)\bar{\varphi}_i(x),$$

so gilt für beliebige Simplexe $T \in \mathbb{T}_h$ mit Ecken $a_0, \ldots, a_d$

$$v_h|_T(x) = \sum_{i=1}^{d} v_h(a_i)\varphi_i^0(T^{-1}(x)),$$

wobei $F : T_0 \to T$ die Referenzabbildung und $\varphi_i^0 \in \Phi$ die Formfunktionen von T_0 sind.

2.51 Definition (lineares Finite Elemente Verfahren)

Sei $\Omega \subset \mathbb{R}^d$ und $\mathbb{T}_h$ zulässige Triangulierung von Ω. Seien $X := H_0^1(\Omega)$ und $X_h := S_{h,0}^1 := S_h^1 \cap \left\{ v \in C^1(\Omega) \mid v = 0 \text{ auf } \partial\Omega \right\}$. Weiter seien eine stetige und koerzive Bilinearform $b : X \times X \to \mathbb{R}$ und ein $l \in X'$ gegeben. Dann ist $X_h \subset X$ ein Teilraum und $u_h \in X_h$ heißt Lösung des **linearen Finite Elemente Verfahrens** für das Problem aus 2.34, falls gilt:

$$b(u_h, v_h) = l(v_h)\ \forall v_h \in X_h.$$

2.52 Satz (A priori Fehlerabschätzung)

Sei $\Omega \subset \mathbb{R}^d,\ d \leq 3$ ein Lipschitz-Gebiet und $\mathbb{T}_h$ eine zulässige Triangulierung von Ω mit $\sigma(T) \leq \sigma < \infty,\ \sigma \in \mathbb{R},\ \forall T \in \mathbb{T}_h$. Seien X, X_h, b, u_h wie in Definition 2.51 und $u \in X$ wie in Definition 2.34. Liegt nun $u \in H_2(\Omega)$ so gibt es eine Konstante $c > 0, c \in \mathbb{R}$, die nur von d, σ, Ω abhängt, so dass gilt:

$$\|u - u_h\|_{H^1(\Omega)} \leq ch\,|u|_{H^2(\Omega)}\,, \tag{2.4}$$

wobei $|u|_{H^2(\Omega)} := \left\|D^2 u\right\|_{L^2(\Omega)}$.

Beweis:
Für Poissonproblem siehe Buch von Dziuk.

2.53 Bemerkung (a priori $\leftrightarrow$ a posteriori Fehlerabschätzung)

Satz 2.52 macht eine Aussage über die Konvergenz des linearen FE Verfahrens. Da auf der rechten Seite der Ungleichung (2.4) aber der Term $|u|_{H^2(\Omega)}$ auftaucht, ist (2.4) nicht geeignet um den tatsächlichen Wert des Approximationsfehlers abzuschätzen. Zu diesem Zweck leitet man A posteriori Fehlerabschätzungen her, bei denen der Fehler ausschließlich durch berechenbare Größen abgeschätzt wird. Für eine Übersicht über A posteriori Fehlerabschätzungen für FE Verfahren verweisen wir auf das Buch von Verfürth.

2.54 Bemerkung

Betrachten wir Definition 2.51 des linearen FE Verfahrens, so stellen wir zunächst fest, dass wir in der schwachen Formulierung exakte Integrale bestimmen müssen, was im Allgemeinen nicht realisierbar ist. in der Praxis verwendet man Quadraturformeln. Ferner schränkt uns Definition 2.51 auf polygonal berandete Gebiete ein. Um auch Gebiete mit glattem Rand behandeln zu können , kann man z.B. eine Gebietsapproximation durchführen bei dem alle Ecken auf dem Rand des polygonal berandeten approximierenden Gebiets $\partial\Omega_h$ auch auf $\partial\Omega$ liegen. Hier ist dann X_h nicht Teilraum von X und das Lemma von Céa nicht anwendbar. Allerdings kann man in beiden Fällen unter gewissen Voraussetzungen zeigen, dass die zusätzlichen Approximationsfehler die Konvergenzordnung des FE Verfahrens nicht beeinflussen. Für weitere Details siehe z.B. das Buch von Dziuk oder Brenner und Scott.

3 Reduzierte Basis Methoden für lineare, koerzive Probleme

3.1 Parameterabhängigkeit

3.1 Definition (parametrische Formen)

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

(1) $f : X \times \mathcal{P} \to \mathbb{R}$ eine parametrische stetige Linearform oder ein parametrisches stetiges lineares Funktional, falls $\forall \mu \in \mathcal{P} : f(., \mu) \in X$.

(2) Wir nennen $b : X_1 \times X_2 \times \mathcal{P} \to \mathbb{R}$ eine parametrische stetige koerzive Bilinearform, falls $\forall \mu \in \mathcal{P} : b(., ., \mu) : X_1 \times X_2 \to \mathbb{R}$ bilinear stetig und koerziv ist. Wir bezeichnen die Stetigkeitskonstante mit $\gamma(\mu)$ und die Koerzivitätskonstante mit $\alpha(\mu)$.

3.2 Bemerkung

Eine parametrische stetige Bi-/Linearform ist nicht unbedingt stetig bzgl. μ. Betrachte dazu das Beispiel $X = \mathbb{R}, \mathcal{P} = [0, 1], f : X \times \mathcal{P} \to \mathbb{R}$ mit

$$f(x, \mu) := \begin{cases} x, \text{ falls } \mu < \frac{1}{2} \\ \frac{1}{2}x, \text{ sonst.} \end{cases}$$

3.3 Defintion (Parametrische Beschränktheit, Stetigkeit)

(1) Wir nennen eine parametrische stetige Linearform f beschränkt bzw. Bilinearform b **gleichmäßig beschränkt** bzgl. μ, falls $\gamma_0, \gamma_1 \in \mathbb{R}^+$ existieren so dass

$$\sup_{\mu \in \mathcal{P}} \|f(., \mu)\|_X \leq \gamma_0 \text{ bzw. } \sup_{\mu \in \mathcal{P}} \gamma(\mu) \leq \gamma_1.$$

(2) Wir nennen b **glm. koerziv** bzgl. μ, falls ein $\alpha_0 > 0$ existiert so dass

$$\inf_{\mu \in \mathcal{P}} \alpha(\mu) \geq \alpha_0 > 0.$$

(3) Wir nennen f bzw. b Lipschitz-stetig bzgl. μ, falls ein $L_f \in \mathbb{R}^+$ bzw. ein $L_b \in \mathbb{R}^+$ existiert, so dass für alle $\mu_1, \mu_2 \in \mathcal{P}$ gilt

$$|f(u, \mu_1) - f(u, \mu_2)| \leq L_f \|u\|_X \|\mu_1 - \mu_2\| \ \forall u \in X,$$

bzw.

$$|b(u, v, \mu_1) - b(u, v, \mu_2)| \leq L_b \|u\|_{X_1} \|v\|_{X_2} \|\mu_1 - \mu_2\| \ \forall u \in X_1, v \in X_2.$$

3.4 Lemma (Energienorm)

Sei X HR, $b : X \times X \times \mathcal{P} \to \mathbb{R}$ parametrische , koerzive, stetige Bilinearform. Dann ist für $\mu \in \mathcal{P}$ durch

$$(((u, v)))_\mu := b_s(u, v; \mu)$$

ein Skalarprodukt auf X und durch

$$|||u|||_\mu := \sqrt{(((u, u)))_\mu}$$

die **Energienorm** definiert. Diese ist äquivalent zur X-Norm und es gilt

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

Beweis:
Skalarprodukt klar wegen Bilinearität, Stetigkeit und Koerzivität. Normäquivalenz folgt aus Stetigkeit und Koerzivtät von b_s:

$$\alpha(\mu) \|v\|_X^2 \leq b_s(v, v; \mu) \leq \gamma(\mu) \|v\|_X^2 .$$

3.5 Definition (Parametrische schwache Formulierung; Parametrisches Variationsproblem ($P(\mu)$))

Sei X HR, $\mathcal{P} \subset \mathbb{R}^p$ beschränkt, $b : x \times X \times \mathcal{P} \to \mathbb{R}$ parametrische, stetige, koerzive Bilinearform, $f, l : X \times \mathcal{P} \to \mathbb{R}$ parametrische stetige Linearform. Zu $\mu \in \mathcal{P}$ bezeichnet $u(\mu) \in X$ die eindeutige Lösung des parametrischen Variationsproblems

$$b(u(\mu), v; \mu) = f(v, \mu) \ \forall v \in X, \tag{3.1}$$

mit Ausgabe $s(\mu) = l(u(\mu), \mu)$.

3.6 Bemerkung

Existenz und Eindeutigkeit der Lösung $u(\mu)$ folgen mit dem Satz von Lax Milgram.

3.7 Definition (schwache Formulierung der parametrischen, stationären Wärmeleitungsgleichung)

Seien $\Omega \subset \mathbb{R}^d$ Lipschitz-Gebiet, $\mathcal{P} \subset \mathbb{R}^p$ beschränkt, $q(\mu) \in L^2(\Omega)$ und $\kappa(\mu) \in L^\infty(\Omega)$ mit $0 < \kappa_1 \leq \kappa(\mu)$ für alle $\mu \in \mathcal{P}$ und Konstante $\kappa_1 \in \mathbb{R}^+$. Dann heißt $u(\mu) \in H_0^1(\Omega)$ schwache Formulierung des RWP der parametrischen, stationären WLG aus 1.1, falls gilt

$$\int_\Omega \kappa(x; \mu) \nabla u(x; \mu) \nabla v(x) \mathrm{d}x = \int_\Omega q(x, \mu) v(x) \mathrm{d}x \ \forall v \in H_0^1(\Omega).$$

3.8 Folgerung (Existenz und Eindeutigkeit von Lösungen)

Unter den Voraussetzungen von Definition 3.7 gibt es für jedes $\mu \in \mathcal{P}$ genau eine schwache Lösung $u(\mu) \in H_0^1(\Omega)$ des RWP der parametrischen, stationären WLG aus 1.1.

Beweis:
Analog zum Beweis von Satz 2.31.

3.9 Definition ((lineares) FE Verfahren für parametrsiche Variationsprobleme ($P_h(\mu)$))

Sei X HR, $\mathcal{P} \in \mathbb{R}^p$ beschränkt, $b : X \times X \times \mathcal{P} \to \mathbb{R}$ parametrische, stetige, koerzive Bilinearform, $f, l : X \times \mathcal{P} \to \mathbb{R}$ parametrische, stetige Linearform. Sei ferner $\mathbb{T}_h$ eine zulässige Triangulierung des Rechengebietes $\Omega \subset \mathbb{R}^d$ und $X_h \subset X$ eine zugehöriger (linearer) Finite Elemente Raum, wobei X_h Unterraum von X. Zu $\mu \in \mathcal{P}$ heißt $u_h(\mu) \in X_h$ Lösung des (linearen) FE Verfahrens für das parametrische Variationsproblem, falls gilt

$$b(u_h(\mu), v_h; \mu) = f(v_h; \mu) \ \forall v_h \in X_h, \ s_h(\mu) = l(u_h(\mu); \mu).$$

3.10 Bemerkung

Das Verfahren aus Definition 3.9 ist nach Bemerkung 1.6 ein 'hochdimensionales, diskretes' Modell.

3.2 Reduzierte Basisverfahren

3.11 Definition (Reduzierte Basis, Reduzierte Basis Räume)

Sei $S_N := \{\mu^1, \ldots, \mu^N\} \subset \mathcal{P}$ eine Menge von Parametern mit (oBdA) linear unabhängigen Lösungen $\{u(\mu^i)\}_{i=1}^N$ von $(P_h(\mu))$. Dann ist $X_N := \operatorname{span}\{u(\mu^i)\}_{i=1}^N$ ein N-dimensionaler **Lagrange Reduzierte Basis-Raum**. Eine Basis $\Phi_N := \{\phi_1, \ldots, \phi_N\} \subset X_h$ eines Reduzierte Basis-Raumes ist eine Reduzierte Basis (RB).

3.12 Bemerkung

Es existieren weitere Arten von RB-Räumen. Im weiteren Verlauf der Vorlesung werden wir z.B. noch **POD-Räume** kennenlernen. Auch die POD-Räume werden aus sogenannten Snapshots, d.h. Lösungen $u_h(\mu^i),\ 1 \leq i \leq k$ mit $k \gg N$, erzeugt.

3.13 Definition (RB-Modell $(P_N(\mu))$, symmetrischer Fall)

Sei ein Problem $P(\mu)$ und ein diskretes Modell $(P_h(\mu))$ gegeben und zusätzlich gelte b symmetrisch und $f = l$ ("compliant"). Sei $X_h \subset X$ ein RB-Raum. Zu $\mu \in \mathcal{P}$ ist die RB-Lösung $u_N(\mu) \in X_N$ und die RB-Ausgabe $s_N(\mu) \in \mathbb{R}$ gesucht, so dass

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

und

$$s_N(\mu) = l(u_N(\mu); \mu).$$

3.14 Bemerkung

Falls b nicht symmetrisch oder $f \neq l$ ist obiges immer noch sinnvoll, aber es bestehen bessere Möglichkeiten $s_N(\mu)$ mittels eines dualen Problems zu bestimmen.

3.15 Bemerkung

Da $X_N \subset X_h \subset X$, X_N Teilraum von X_h, ist das RB-Modell ein Ritz-Galerkin Verfahren.

3.16 Folgerung (Existenz, Eindeutigkeit, Stabilität, Wohlgestelltheit)

Zu $\mu \in \mathcal{P}$ existiert eine eindeutige RB-Lösung $u_N(\mu) \in X_N$ und RB-Ausgabe $S_N(\mu)$ von $(P_N(\mu))$. Diese sind beschränkt durch $\|u_N(\mu)\|_X \leq \frac{1}{\alpha(\mu)} \|f(.;\mu)\|_X$ und $|s_N(\mu)| \leq \frac{1}{\alpha(\mu)} \|f(.;\mu)\|_{X'} \|l(.;\mu)\|_{X'}$.

Beweis:
Existenz und Eindeutigkeit von $u_N(\mu)$ folgt mit dem Satz von Lax-Milgram, wobei

$$\alpha_N(\mu) := \inf_{u \in X_N} \frac{b(u, u:\mu)}{\|u\|_X^2} \geq \inf_{u \in X} \frac{b(u, u:\mu)}{\|u\|_X^2} = \alpha(\mu) > 0.$$

Dann ist auch $s_N(\mu) = l(u_N(\mu); \mu)$ eindeutig und die Stabilität folgt mit

$$\|u_N(\mu)\|_X = \left\|B^{-1}(\mu) v_f(\mu)\right\|_X \leq \left\|B^{-1}(\mu)\right\|_X \|v_f(\mu)\|_X \leq \frac{1}{\alpha(\mu)} \|f(.;\mu)\|_X .$$

Hierbei ist $B(\mu)$ der eindeutige invertierbare Operator aus dem Satz von Lax-Milgram und $v_f(\mu)$ der Riesz-Repräsentant von $f(.;\mu) \in X_N'$.

$$|s_N(\mu)| = |l(u_N(\mu);\mu)| \leq \|l(.;\mu)\|_{X'} \|u_N(\mu)\|_X \leq \frac{1}{\alpha(\mu)} \|f(.;\mu)\|_{X'} \|l(.;\mu)\|_{X'}$$

□

3.17 Folgerung (Galerkin-Projektion, Galerkin-Orthogonalität)

Zu $\mu \in \mathcal{P}, X_h, X_N$ HR mit Energieskalarprodukt $(((.,.)))_\mu, P_\mu : X_h \to X_N$ die orthogonale Projektion aus Satz 2.7, $u_h(\mu), u_N(\mu)$ Lösungen von $(P_h(\mu))$ bzw.$(P_N(\mu))$ und der Fehler $e_N(\mu) = u_h(\mu) - u_N(\mu)$. Dann gilt

(1) $u_N(\mu) = P_\mu(u_h(\mu))$ "Galerkin-Projektion"

(2) $(((e_N(\mu), v_N)))_\mu = 0 \ \forall v_N \in X_N$ "Galerkin-Orthogonalität"

Beweis:
Lemma 3.4 impliziert $(X_h, (((.,.)))_\mu)$ HR und $X_N = \operatorname{span}\{\Phi_i\}_{i=1}^N$ endlichdimensional, also abgeschlossen ist. Daher ist P_μ nach Satz 2.7 wohldefiniert.
Mit Folgerung 2.8 folgt

$$\begin{aligned}
&(((P_\mu(u_h(\mu) - u_h(\mu), \Phi_i)))_\mu = 0, \ i = 1, \dots, N \\
\Leftrightarrow &b(P_\mu(u_h(\mu)) - u_h(\mu), \Phi_i; \mu) = 0, \ i = 1, \dots, N \\
\Leftrightarrow &b(P_\mu(u_h(\mu)), \Phi_i; \mu) = b(u_h(\mu), \Phi_i; \mu), \ i = 1, \dots, N \\
\Leftrightarrow &b(P_\mu(u_h(\mu)), \Phi_i; \mu) = f(\Phi_i; \mu), \ i = 1, \dots, N
\end{aligned}$$

Da $u_N(\mu)$ eindeutig folgt $P_\mu(u_n(\mu)) = u_N(\mu)$ daraus folgt (1). (2) folgt entweder aus 2.8 oder Satz 2.39.

3.18 Folgerung

Sei $\mu \in \mathcal{P}, u_h(\mu), u_N(\mu)$ Lösungen von $(P_h(\mu))$ bzw. $(P_N(\mu))$. Falls $u_h(\mu) \in X_N \Rightarrow u_N(\mu) = u_h(\mu)$.

Beweis:
Da $u_h(\mu), u_N(\mu) \in X_N \Rightarrow e_N(\mu) := u_h(\mu) - u_N(\mu) \in X_h$ und $(((e_N(\mu), v_N)))_\mu = 0 \ \forall v_N \in X_N$ nach Folgerung 3.17 (2). Damit gilt $(((e_N(\mu), l_N(\mu))))_\mu = 0 \Rightarrow e_N(\mu) = 0 \Rightarrow u_N(\mu) = u_h(\mu)$. □

3.19 Satz (abstrakte Fehlerabschätzung; Relation zur Bestapproximation)

Sei $\mu \in \mathcal{P}$ und $u_h(\mu), s_h(\mu)$ bzw. $u_N(\mu), s_N(\mu)$ Lösungen von $(P_h(\mu))$ bzw. $(P_N(\mu))$. Dann gilt:

(1) Der Fehler der (μ-abhängigen Energienorm) erfüllt

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

(2) Der Fehler in der (μ-unabhängigen) X-Norm erfüllt

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

mit $\gamma(\mu), \alpha(\mu)$ Stetigkeits- bzw. Koerzivitätskonstante.

(3) Für den Ausgabefehler gilt (wegen $f = l$)

$$0 \leq s_h(\mu) - s_N(\mu) = |||u_h(\mu) - u_N(\mu)|||_\mu^2 = \inf_{v \in X_N} |||u_h(\mu) - v|||\mu^2 \leq \gamma(\mu) \inf_{v \in X_N} \|u_h(\mu) - v\|_X^2 .$$

Beweis:

(1) Nach Folgerung 3.17 ist $u_N(\mu)$ orthogonale Projektion, also Bestapproximation

$$|||u_h(\mu) - u_N(\mu)|||_\mu \overset{3.17(1)}{=} |||u_h(\mu) - P_\mu(u_N(\mu))|||_\mu \overset{2.7}{=} \inf_{v \in X_h} |||u_h(\mu) - v|||_\mu.$$

(2) Mit der Normäquivalenz 3.4 folgt

$$\sqrt{\alpha_h(\mu)}\,\|u_h(\mu)-u_N(\mu)\|_X \overset{3.4}{\leq} |||u_h(\mu)-u_N(\mu)|||_\mu \overset{(1)}{=} \inf_{v\in X_N} |||u_h(\mu)-v|||_\mu \overset{3.4}{\leq} \sqrt{\gamma_h(\mu)} \inf_{v\in X_h} \|u_h(\mu)-v\|_X ,$$

wobei $\alpha_h(\mu) := \inf_{v\in X_h} \frac{b(v,v;\mu)}{\|v\|_X^2}$ und $\gamma_h(\mu) := \sup_{u,v\in X_h} \frac{b(u,v;\mu)}{\|u\|_X\|v\|_X}$. Wie in Beweis von Folgerung 3.16 folgt $\alpha_h(\mu) \geq \alpha(\mu)$ und $\gamma_h(\mu) \leq \gamma(\mu)$ $\forall \mu \in \mathcal{P}$ und damit die Behauptung.

(3)

$$\begin{aligned} s_h(\mu) - s_N(\mu) &\overset{Def}{=} l(u_h(\mu);\mu) - l(u_N(\mu);\mu) \overset{l=f}{=} f(u_h(\mu)) - f(u_N(\mu)) \\ &\overset{(P_h(\mu)}{=} b(u_h(\mu), u_h(\mu)-u_N(\mu);\mu) \\ &= b(u_h(\mu), u_h(\mu)-u_N(\mu);\mu) - b(u_h(\mu)-u_N(\mu), u_N(\mu);\mu) \\ &= b(u_h(\mu)-u_N(\mu), u_h(\mu)-u_N(\mu);\mu) \end{aligned}$$

Damit folgt

$$s_h(\mu) - s_N(\mu) = |||u_h(\mu)-u_N(\mu)|||_\mu^2 \overset{(1)}{=} \inf_{v\in X_h} |||u_h(\mu)-v|||_\mu^2 \overset{3.4}{\leq} \gamma(\mu) \inf_{v\in X_N} \|u_h(\mu)-v\|_X^2$$

Insbesondere gilt auch $s_h(\mu) - s_N(\mu) = |||u_h(\mu)-u_N(\mu)|||_\mu^2 \geq 0$.

□

3.20 Bemerkung

(1) $s_N(\mu)$ ist also untere Schranke für $s_h(\mu)$.

(2) Der Ausgabefehler ist im Allgemeinen sehr klein, da das Quadrat des RB-Fehlers eingeht.

(3) Mit dem Lemma von Céa (Satz 2.39) erhalten wir für nicht notwendigerweise symmetrische Bilinearformen $\|u_h(\mu)-u_N(\mu)\|_X \leq \frac{\gamma(\mu)}{\alpha(\mu)} \inf_{v\in X_N} \|u_h(\mu)-v\|_X$. Damit ist Satz 3.19 eine Verschärfung für symmetrische Bilinearformen.

3.21 Korollar (Monotoner Fehlerabfall in der Energienorm)

Sei $(X_N)_{N=1}^{N_{\max}}$ Folge von RB-Räumen mit $X_N \subseteq X_{N'}$ für $N \leq N' \leq N_{\max}$ ("Hierarchische Räume") und $e_N(\mu) = u_h(\mu) - u_N(\mu)$ für $\mu \in \mathcal{P}$. Dann ist die Folge $(|||e_N(\mu)|||)_{N=1}^{N_{\max}}$ monoton fallend.

Beweis:

$$|||e_N(\mu)|||\mu = \inf_{v\in X_N} |||u_h(\mu)-v|||_\mu \geq \inf_{v\in X_{N'}} |||u_h(\mu)-v|||_\mu = |||e_{N'}(\mu)|||_\mu.$$

□

3.22 Bemerkung

(1) 'Worst Case' ist eine Stagnation des Fehlers (unrealistisch, da jeder neue Basisvektor orthogonal zu $e_N(\mu)$ sein müsste). In der Praxis ist bei geschickter Basiswahl exponentielle Konvergenz zu beobachten.

(2) Monotonie gilt nicht notwendigerweise für andere Normen trotz Normenäquivalenz:

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

mit c, C Konstanten unabhängig von N. Fehlernorm $\|e_N(\mu)\|$ kann gelegentlich anwachsen, bleibt aber in einem 'Korridor' um $|||e_N(\mu)|||_\mu$.
"Beweis":

$$\|e_{N'}(\mu)\| \leq C|||e_{N'}(\mu)|||_\mu \leq C|||e_N(\mu)|||_\mu \leq \frac{C}{c}\|e_N(\mu)\|$$

3.23 Folgerung (Fehlerabschätzung für den Fehler zwischen exakter Lösung und RB-Lösung)

Sei $\mu \in \mathcal{P}$ und $u(\mu), s(\mu), u_h(\mu), s_h(\mu)$ bzw. $u_N(\mu), s_N(\mu)$ Lösung von $(P(\mu)), (P_h(\mu))$ bzw. $(P_N(\mu))$, wobei X_h linearer Finite Elemente Raum. Zusätzlich gelte b symmetrisch und $f = l$ ("compliant"). Liegt nun $u \in H^2(\Omega)$, so gibt es eine Konstante $c > 0$ die nur von d, σ und Ω abhängt, so dass gilt:

1. Der Fehler in der (μ-unabhängigen) X-Norm erfüllt

$$\|u(\mu) - u_h(\mu)\|_X \leq \sqrt{\frac{\gamma(\mu)}{\alpha(\mu)}}\left(ch\,|u(\mu)|_{H^2(\Omega)} + \inf_{v\in X_N}\|u_h(\mu) - v\|_X\right)$$

2. Für den Ausgabefehler gilt:

$$0 \leq s(\mu) - s_N(\mu) \leq \gamma(\mu)\left(c^2h^2\,|u(\mu)|_{H^2(\Omega)} + \inf_{v\in X_N}\|u_h(\mu) - v\|_X^2\right)$$

Beweis: Analog zum Beweis von Satz 3.19 unter Verwendung von Satz 2.52.

3.24 Bemerkung

Den Fehleranteil $\|u(\mu) - u_h(\mu)\|_X$ nennt man **Diskretisierungsfehler** und den Anteil $\|u_h(\mu) - u_N(\mu)\|$ nennt man **Modellfehler**. Um eine gute Approximation der exakten Lösung $u(\mu)$ und der Ausgabe $s(\mu)$ zu erhalten, müssen beide Fehleranteile klein sein.

3.25 Satz (Lipschitzstetigkeit)

Falls b und f gleichmäßig beschränkt und Lipschitz-stetig bzgl. μ und b gleichmäßig koerziv bzgl. μ, so sind auch die Lösungen $u_N(\mu)$ und $s_N(\mu)$ von $(P_N(\mu))$ Lipschitz-stetig bzgl. μ.

3.26 Satz (Gleichungssytem und numersische Stabilität)

Sei $\Phi_N = \{\phi_1, \ldots, \phi_N\}$ eine reduzierte Basis von X_N. Für $\mu \in \mathcal{P}$ definieren wir $\mathbb{B}_N(\mu) \in \mathbb{R}^{N\times N}$ und $\mathbb{F}_N(\mu) \in \mathbb{R}^N$ durch

$$(\mathbb{B}_N(\mu))_{nm} := b(\phi_m, \phi_n; \mu),\ (\mathbb{F}_N(\mu)_n := f(\phi_n; \mu)$$

und

$$\mathbb{U}_N(\mu) = (U_1^N(\mu), \ldots, U_N^N(\mu)) \in \mathbb{R}^N$$

als Lösung von

$$\mathbb{B}_N(\mu)\mathbb{U}_N(\mu) = \mathbb{F}_N(\mu). \tag{3.2}$$

(1) Dann ist $u_N(\mu) = \sum_{n=1}^N U_n^N(\mu)\phi_n$ und $s_N(\mu) = \mathbb{F}_N^T(\mu)\mathbb{U}_N(\mu)$ Lösung von $(P_N(\mu))$.

(2) Falls Φ_N orthogonal, so ist die Kondition von (3.2) unabhängig von beschränkt durch

$$\mathrm{cond}_2(\mathbb{B}_N(\mu)) = \|\mathbb{B}_N(\mu)\|_2 \left\|\mathbb{B}_N^{-1}(\mu)\right\|_2 \leq \frac{\gamma(\mu)}{\alpha(\mu)}.$$

Beweis:
(1) klar.
(2): Wegen Symmetrie con $\mathbb{B}_N(\mu)$ ist

$$\mathrm{cond}_2(\mathbb{B}_N(\mu)) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|} \tag{3.3}$$

mit betragsmäßig größten/kleinstem Eigenwert $\lambda_{\max}, \lambda_{\min}$ von $\mathbb{B}_N(\mu)$. Sei $\mathbb{U}_{\max} = (U_n)_{n=1}^N \in \mathbb{R}^N$ Eigenvektor zu $\lambda_{\max}$ und $u_{\max} := \sum_{n=1}^N U_n \phi_n$. Dann gilt

$$\begin{aligned}
\lambda_{\max} \|\mathbb{U}_{\max}\|_2^2 &= \lambda_{\max} \mathbb{U}_{\max}^T \cdot \mathbb{U}_{\max} = \mathbb{U}_{\max}^T \mathbb{B}_N(\mu) \mathbb{U}_{\max} \\
&= \sum_{n,m=1}^N U_n U_m b(\phi_n, \phi_m; \mu) = b(\sum_{n=1}^N U_n \phi_n, \sum_{m=1}^N U_m \phi_m; \mu) \\
&= b(u_{\max}, u_{\max}; \mu)
\end{aligned}$$

Aus der Orthogonalität folgt

$$\|u_{\max}\|_X^2 =$$

3.27 Bemerkung

Im Gegensatz zu großen aber dünn besetzten Matrizen bei FEM ist (3.2) klein aber voll besetzt, weil ϕ_i im Allgemeinen keinen disjunkten Träger haben.

3.28 Folgerung

Sei $\bar{\varphi}_i,\ i = 1, \ldots, \mathcal{N}_h$ die Knotenbasis von X_h wie in Def. 2.50 definiert. Für $\mu \in \mathcal{P}$ definieren wir $\mathbb{B}_h(\mu) \in \mathbb{R}^{\mathcal{N}_h \times \mathcal{N}_h}$ und $\mathbb{F}_h(\mu) \in \mathbb{R}^{\mathcal{N}_h}$ durch

$$(\mathbb{B}_h(\mu))_{ij} := b(\bar{\varphi}_j, \bar{\varphi}_j; \mu),\ (\mathbb{F}_h(\mu))_i := f(\bar{\varphi}_i; \mu),\ 1 \leq i, j \leq \mathcal{N}_h. \tag{3.4}$$

Indem wir dnun die RB-Basisfunktionen in der Knotenbasis darstellen

$$\Phi_n = \sum_{i=1}^{\mathcal{N}_h} \phi_h^i \bar{\varphi}_i,\ n = 1, \ldots, N \tag{3.5}$$

können wir eine Transformationsmatrix $\mathbb{V} \in \mathbb{R}^{\mathcal{N}_h \times N}$ definieren deren Spalten die Koeffizienten der RB-Basisfunktionen in (3.5) enthalten:

$$\mathbb{V}_{in} := \phi_n^i,\ 1 \leq i \leq \mathcal{N}_h,\ 1 \leq n \leq N. \tag{3.6}$$

Dann gilt für $\mathbb{B}_N(\mu)$ und $\mathbb{F}_N(\mu)$ aus Satz 3.25:

$$\mathbb{B}_N(\mu) = \mathbb{V}^T \mathbb{B}_h(\mu) \mathbb{V} \text{ und } \mathbb{F}_N(\mu) = \mathbb{V}^T \mathbb{F}_h(\mu).$$

Beweis:

$$\begin{aligned}(\mathbb{V}^T\mathbb{B}_h(\mu)\mathbb{V})_{mn} &= \sum_{r,s=1}^{\mathcal{N}_h}(\mathbb{V}^T)_{mr}(\mathbb{B}_h(\mu))_r s(\mathbb{V}_{sn}) \\ &= \sum_{r,s=1}^{\mathcal{N}_h}\phi_m^r b(\bar{\varphi}_s,\bar{\varphi}_r;\mu)\phi_n^s \\ &= b\left(\sum_{s=1}^{\mathcal{N}_h}\phi_n^s\bar{\varphi}_s,\sum_{r=1}^{\mathcal{N}_h}\phi_m^r\bar{\varphi}_r;\mu\right) \\ &= b(\phi_n,\phi_m;\mu) = (\mathbb{B}_N(\mu))_{mn}.\end{aligned}$$

□

3.3 Offline/Online Zerlegung des RB-Modells

3.29 Bemerkung (Komplexitätsbetrachtungen)

Da $\mathbb{B}_h(\mu)$ aus (3.4) dünn besetzt ist, erfordert die Berechnung von $u_h(\mu)$ $\mathcal{O}(\mathcal{N}_h^2)$ Rechenschritte. Da $\mathbb{B}_N(\mu)$ vollbesetzt ist das lineare Gleichungssystem (LGS) (3.2) in $\mathcal{O}(N^3)$ Rechenschritten lösbar. Daher ist nur für $N \ll \mathcal{N}_h$ da RB-Modell ein Gewinn. Genauere Betrachtung der Berechnung einer reduzierten Lösung $u_N(\mu)$:

(1) N Snapshots, also N Lösungen $u_h(\mu)$ von $(P_h(\mu))$ berechnen: $\mathcal{O}(N\mathcal{N}_h^2)$. ('Offline')

(2) N^2 Auswertungen von $b(\phi_m,\phi_n;\mu)$: $\mathcal{O}(N^2\mathcal{N}_h)$.

(3) N Auswertungen von $f(\phi_n;\mu)$: $\mathcal{O}(N\mathcal{N}_h)$.

(4) Die Lösung des LGS (3.2): $\mathcal{O}(N^3)$. ('Online')

Damit lohnt sich das RB-Modell für ein einzelnes $\mu \in \mathcal{P}$ oder wenige $\mu \in \mathcal{P}$ nicht. Wenn wir $(P_N(\mu))$ aber für viele verschiedene Parameter $\mu \in \mathcal{P}$ lösen müssen, wie zum Beispiel in einem *many-query* Kontext lohnt sich das RB-Modell, wenn man eine sogenannte **Offline/Online Zerlegung** durchführt. In der einmalig durchgeführten **Offline-Phase** werden μ-unabhängige, hochdimensionale Größen in $\mathcal{O}(\mathcal{N}_h^m), m \in \mathbb{N}$ Rechenschritten, typischerweise teuer vorberechnet. In der vielfach durchgeführten **Online-Phase** werden die Offline-Daten kombiniert um μ-abhängige Größen wie das reduzierte LGS (3.2) zu assemblieren. Die RB-Lösung $u_N(\mu)$ und $s_N(\mu)$ können dann schnell berechnet werden, wobei die Anzahl der Rechenschritte idealerweise $\mathcal{O}(N^k), k \in \mathbb{N}$ ist, d.h. unabhängig von $\mathcal{N}_h$.
Vor diesem Hintergrund können wir Schritt 1 klar der Offline-Phase und Schritt 4 der Online-Phase zu ordnen. Schritt 2 und 3 lassen sich direkt keiner der beiden Phasen klar zu ordnen, da sie sowohl teure als auch Parameter-abhängige Operationen benötigen. Um eine klare Trennung auch von Schritt 2 und 3 zu erreichen benötigen wir eine spezielle Struktur der Bilinearform $b(.,.;\mu)$ und Linearform $f(.;\mu)$.

3.30 Definition

Seinen X, X_1, X_2 HR, $\mathcal{P}$ beschränkte Parametermenge.

(1) Eine Funktion $v : \mathcal{P} \to X$ nennen wir **affin parametrisch**, falls Funktionen $v^q \in X$ und Koeffizientenfunktionen $\theta_v^q : \mathcal{P} \to \mathbb{R}$ für $q = 1,\dots,Q_v$ existieren, so dass

$$v(x;\mu) := \sum_{q=1}^{Q_v}\theta_v^q(\mu)v^q(x).$$

(2) Eine parametrische stetige Linearform $f : X \times \mathcal{P} \to \mathbb{R}$ bzw. stetige Bilinearform $b : X_1 \times X_2 \times \mathcal{P} \to \mathbb{R}$ ist **affin parametrisch**, falls $f^q \in X'$ und $\theta_f^q : \mathcal{P} \to \mathbb{R}$ für $q = 1, \dots, Q_f$ bzw. $b^q : X_1 \times X_2 \to \mathbb{R}$ und $\theta_B^q : \mathcal{P} \to \mathbb{R}$ für $q = 1, \dots, Q_b$ existieren, so dass

$$f(v;\mu) = \sum_{q=1}^{Q_f} \theta_f^q(\mu) f^q(v) \ \forall v \in X \text{ bzw.}$$

$$b(u,v;\mu) = \sum_{q=1}^{Q_b} \theta_b^q(\mu) b^q(u,v) \ \forall u \in X_1, v \in X_2.$$

3.31 Folgerung (Offline/Online-Zerlegung von $(P_N(\mu))$)

Sei $(P_N(\mu))$ gegeben und b, f affin parametrisch. Dann erlaubt $(P_N(\mu))$ die folgende Offline/Online-Zerlegung:

Offline-Phase: Nach Berechnung einer reduzierten Basis $\Phi_N := \{\phi_1, \dots, \phi_N\}$ assemblieren wir die parameterunabhägige Matrizen und Vektoren $\mathbb{B}_N^q \in \mathbb{R}^{N \times N}$ und $\mathbb{F}_N^q \in \mathbb{R}^N$, definiert durch

$$(\mathbb{B}_N^q)_{nm} := b^q(\phi_m, \phi_n),\ 1 \le m, n \le N,\ 1 \le q \le Q_b,$$

$$(\mathbb{F}_N^q)_n := f^q(\phi_n),\ 1 \le n \le N,\ 1 \le q \le Q_f.$$

Online-Phase Für einen gegebenen Parametervektor $\mu \in \mathcal{P}$ werten wir die parameterabhängigen Koeffizientenfunktionen $\theta_b^q(\mu), \theta_f^q(\mu)$ für $1 \le q \le Q_b, Q_f$ aus und assemblieren die Matrix und den Vektor

$$\mathbb{B}_N(\mu) := \sum_{q=1}^{Q_b} \theta_b^q(\mu) \mathbb{B}_N^q \text{ bzw. } \mathbb{F}_N(\mu) := \sum_{q=1}^{Q_f} \theta_f^q(\mu) \mathbb{F}_N^q,$$

welche mit der Matrix und dem Vektor aus dem LGS (3.2) aus Satz 3.26 übereinstimmen. Dieses LGS kann dann nach $u_N(\mu)$ und $s_N(\mu)$ gelöst werden.

3.32 Bemerkung

Die Matrizen $\mathbb{B}_N^q \in \mathbb{R}^{N \times N}$ und die Vektoren $\mathbb{F}_N^q \in \mathbb{R}^N$ können mit dem in Folgerung 3.28 beschriebenen Verfahren einfach aus den entsprechenden FE-Matrizen und den FE-Vektoren mit der in (3.6) definierten Transformationsmatrix assembliert werden.

3.33 Bemerkung (Rechenaufwand/Laufzeit)

(1) Für die Berechnung der Snapshots und der anschließenden Assemblierung von $\mathbb{B}_N^q, \mathbb{F}_N^q$ benötigen wir $\mathcal{O}(N\mathcal{N}_h^2 + N^2\mathcal{N}_h Q_b + N\mathcal{N}_h Q_f)$ Rechenschritte. In der Online-Phase kann dann die Assemblierung und das Lösen von LGS (3.2) in $\mathcal{O}(N^2 Q_b + N Q_f + N^3)$ Rechenschritten erfolgen. Insbesondere hängt die Komplexität der Online-Phase nicht von $\mathcal{N}_h$ ab.

(2) Die Offline/Online Zerlegung lässt sich auch in einem Laufzeitdiagramm veranschaulichen. $t_{hoch}, t_{offline}, t_{online}$ bezeichnen die Laufzeit dür eine Lösung des hochdimensionalen, diskreten Problems $(P_h(\mu))$, die Offline- und die Online-Phase von $(P_N(\mu))$. Wir nehmen an, dass diese Zeiten für unterschiedliche Parameter jeweils dieselben sind und erhalten dadurch einen linearen Zusammenhang zwischen der gesamten benötigten Laufzeit und der Anzahl k von Berechnungen der Lösungen $u_h(\mu), u_N(\mu)$. Die Gesamtlaufzeit für k hochdimensionale Lösungen ist $t_h(k) = k \cdot t_{hoch}$, während das reduzierte Modell eine Laufzeit von $t_N(k) = t_{offline} + k \cdot t_{online}$ benötigt. Wie bereits in 3.29 erwähnt lohnt sich ein RB-Modell bei mehr als $k* := \frac{t_{offline}}{t_{hoch} - t_{online}}$ benötigten Approximationen von $u(\mu)$.

3.4 A posteriori Fehlerschätzer

3.4.1 A posteriori Fehlerschranken und Effektivität

3.34 Lemma (Fehler-Residuum Beziehung)

Für $\mu \in \mathcal{P}$ definieren wir mittels der RB-Lösung $u_N(\mu)$ das Residuum $r(.;\mu) \in X_h'$ durch

$$r(v;\mu) := f(v;\mu) - b(u_N(\mu), v;\mu) \forall v \in X_h \tag{3.7}$$

und den zugehörigen Riesz-Repräsentanten $R(\mu) \in X_h$ als Lösung von

$$(R(\mu), v)_X = r(v;\mu) \ \forall v \in X_h. \tag{3.8}$$

Dann erfüllt der Fehler $e_N(\mu) := u_h(\mu) - u_N(\mu)$

$$b(e_N(\mu), v;\mu) = r(v;\mu) \ \forall v \in X_h. \tag{3.9}$$

Beweis:

$$\begin{aligned} b(e_N(\mu), v;\mu) &= b(u_h(\mu) - u_N(\mu), v;\mu) \\ &= b(u_h(\mu), v;\mu) - b(u_N(\mu), v;\mu) \\ &= f(v;\mu) - b(u_N(\mu), v;\mu) = r(v;\mu) \end{aligned}$$

□

3.35 Satz (A posteriori Fehlerschätzer)

Für $\mu \in \mathcal{P}$ seinen $u_h(\mu), s_h(\mu)$ Lösungen von $(P_h(\mu))$ und $u_N(\mu), s_N(\mu)$ Lösungen von $(P_N(\mu))$. Ferner sei $\alpha_{LB}(\mu) > 0$ eine berechenbare untere Schranke für die Koerzivitätskonstante $\alpha_h(\mu)$ von $b(.,.;\mu)$ und $R(\mu)$ der Riesz-Repräsentant des Residuums aus Lemma 3.34. Dann erfüllen die A posteriori Fehlerschätzer , definiert durch

$$\Delta_N^{en}(\mu) := \frac{\|R(\mu)\|_X}{\sqrt{\alpha_{LB}(\mu)}} \text{ und } \Delta_N^{s}(\mu) := \frac{\|R(\mu)\|_X^2}{\alpha_{LB}(\mu)}, \tag{3.10}$$

die folgenden Ungleichungen

$$|||u_h(\mu) - u_N(\mu)|||_\mu = |||e_N(\mu)|||_\mu \leq \Delta_N^{en}(\mu) \tag{3.11}$$

$$s_h(\mu) - s_N(\mu) \leq \Delta_N^{s}(\mu). \tag{3.12}$$

Beweis:
Testen von Gleichung (3.9) mit $e_N(\mu)$ ergibt:

$$\begin{aligned} |||e_N(\mu)|||_\mu &\overset{Def}{=} b(e_N(\mu), e_N(\mu);\mu) \overset{(3.9)}{=} r(e_N(\mu);\mu) \\ &\overset{(3.8)}{=} (R(\mu), e_N(\mu))_X \overset{C.S}{\leq} \|R(\mu)\|_X \|e_N(\mu)\|_X \\ &\overset{3.4}{=} \frac{1}{\sqrt{\alpha_h(\mu)}} \|R(\mu)\|_X |||e_N(\mu)|||_\mu \leq \frac{1}{\sqrt{\alpha_{LB}(\mu)}} \|R(\mu)\|_X |||e_N(\mu)|||_\mu \\ &\Rightarrow |||e_N(\mu)|||_\mu \leq \Delta_N^{en}(\mu) \Rightarrow (3.11) \end{aligned}$$

Aus Satz 3.19 folgt

$$s_h(\mu) - s_N(\mu) = |||e_n(\mu)|||_\mu^2 \leq (\Delta_N^{en}(\mu))^2 = \Delta_N^{s}(\mu).$$

□

3.36 Folgerung

Durch $\hat{s}_N(\mu) := s_N(\mu) + \Delta_N^s(\mu)$ ist eine obere Schranke für $s_h(\mu)$ gegeben, das heißt es gilt

$$s_h(mu) \leq \hat{s}_N(\mu).$$

Beweis:
Folgt direkt aus (3.12).

3.37 Bemerkung

Das Beschränken des Fehlers durch das Residuum ist eine Standardtechnik zum Herleiten von A posteriori Fehlerschätzern für FEM. Da in diesem Fall dein Schätzer für den Fehler $|||u(\mu) - u_h(\mu)|||_\mu$ gesucht wird, ist X unendlich-dimensional und die Norm $\|r(.;\mu)\|_{X'}$ kann nicht berechnet werden. Im Fall von RB Methoden ist $\|r(.;\mu)\|_{X_h'}$ mit Hilfe des Riesz-Repräsentanten berechenbar.

3.38 Bemerkung

Da $\Delta_N^{en}(\mu)$ und $\Delta_N^s(\mu)$ unter den Voraussetzungen von Satz 3.35 obere Schranken für die Fehler sind, werden sie auch als rigorose Fehlerschranken bezeichnet. Bei A posteriori Fehlerschätzern für FEM treten häufig Konstanten in den Abschätzungen auf, welche nicht entsprechend nach oben/unten durch berechenbare Konstanten beschränkt werden können, so dass in diesen Fällen die A posteriori Fehlerschätzer den Fehler auch unterschätzen können; sie also keine rigorose Fehlerschranken zu sein brauchen. Mit Hilfe des Fehlerschätzers können wir die Dimension des RB-Raumes so bestimmen, dass der Approximationsfehler kleiner als eine vorgegebene Toleranz ist. Um ein möglichest effizientes Verfahren zu erhalten ist es daher wünschenswert, dass der Quotient $\frac{\Delta_N^{en}(\mu}{|||e_N(\mu)|||_\mu}$ möglichst nahe an 1 ist. Er ist ≥ 1 wegen Satz 3.35. Diesen Quotienten werden wir im Folgenden weiter untersuchen.

3.39 Satz (Effektivitäten der Fehlerschätzer)

Wir definieren die Effektivitäten $\eta_N^{en}(\mu)$ und $\eta_N^s(\mu)$ der Fehlerschätzer $\Delta_N^{en}(\mu)$ und $\Delta_N^s(\mu)$, definiert in (3.10), durch

$$\eta_N^{en}(\mu) := \frac{\Delta_N^{en}(\mu)}{|||u_h(\mu) - u_N(\mu)|||_\mu} \text{ und } \eta_N^s(\mu) := \frac{\Delta_N^s(\mu)}{s_h(\mu) - s_N(\mu)}. \tag{3.13}$$

Unter den Voraussetzungen von Satz 3.35 gilt dann

$$\eta_N^{en}(\mu) \leq \sqrt{\frac{\gamma(\mu)}{\alpha_{LB}(\mu)}} \text{ und} \tag{3.14}$$

$$\eta_N^s(\mu) \leq \frac{\gamma(\mu)}{\alpha_{LB}(\mu)}. \tag{3.15}$$

Beweis:
Zunächst folgt aus der Definition de Riesz-Repräsentanten und des Residuums in Lemma 3.34:

$$\begin{aligned}\|R(\mu)\|_X^2 = (R(\mu), R(\mu))_X &\overset{(3.8)}{=} r(R(\mu);\mu) \\ &\overset{(3.9)}{=} b(e_N(\mu), R(\mu);\mu \overset{C.S}{\leq} |||e_N(\mu)|||_\mu |||R(\mu)|||_\mu \\ &\overset{3.4}{\leq} |||e_N(\mu)|||_\mu \sqrt{\gamma(\mu)}\, \|R(\mu)\|_X .\end{aligned}$$

$$\Rightarrow \|R(\mu)\|_X \leq |||e_N(\mu)|||_\mu \sqrt{\gamma(\mu)} \tag{3.16}$$

$$\eta_N^{en}(\mu) = \frac{\Delta_N^{en}(\mu)}{|||e_N(\mu)|||_\mu} \overset{Def}{=} \frac{\|R(\mu)\|_X}{\sqrt{\alpha_{LB}(\mu)}|||e_N(\mu)|||_\mu}$$
$$\overset{(3.16)}{\leq} \sqrt{\frac{\gamma(\mu)}{\alpha_{LB}(\mu)}} \frac{|||e_N(\mu)|||_\mu}{|||e_N(\mu)|||_\mu}$$
$$\Rightarrow (3.14)$$

Aus Satz 3.19 folgt dann:

$$\eta_N^s(\mu) \overset{Def}{=} \frac{\Delta_N^s(\mu)}{s_h(\mu)s_N(\mu)} \overset{Def/3.19}{=} \frac{(\Delta_N^{en}(\mu))^2}{|||e_N(\mu)|||_\mu^2} \overset{(3.14)}{\leq} \frac{\gamma(\mu)}{\alpha_{LB}(\mu)} \Rightarrow (3.15).$$

3.40 Folgerung

Falls $u_h(\mu) = u_N(\mu)$ dann gilt automatisch $\Delta_N^{en}(\mu) = \Delta_N^s(\mu) = 0$.

Beweis:
Folgt direkt aus Satz 3.39, kann aber auch unabhängig davon wie folgt eingesehen werden: Da $0 = b(0, v; \mu) = b(e_N(\mu), v; \mu) \overset{(3.9)}{=} r(v; \mu) \overset{(3.8)}{=} (R(\mu), v)_X$ für alle $v \in X$ gilt, folgt $\|R(\mu)\|_X = 0$ und damit $\Delta_N^{en}(\mu) = \Delta_N^s(\mu) = 0$.

3.41 Bemerkung

Folgerung 3.40 ist insbesondere dann relevant, wenn für einen Fehlerschätzer Schranken für die Effektivität (noch) nicht verfügbar sind.

3.42 Folgerung (Fehlerschätzer für die X-Norm)

Unter den Voraussetzungen von Satz 3.35 gilt für den Fehlerschätzer $\Delta_N(\mu) := \frac{\|R(\mu)\|_X}{\alpha_{LB}(\mu)}$, dass

$$\|u_h(\mu) - u_N(\mu)\|_X \leq \Delta_N(\mu).$$

Ferner gilt für die Effektivität des Fehlerschätzers $\eta_N(\mu) := \frac{\Delta_N(\mu)}{\|u_h(\mu) - u_N(\mu)\|_X}$ die folgende Schranke

$$\eta_N(\mu) \leq \frac{\gamma(\mu)}{\alpha_{LB}(\mu)}.$$

Beweis:
Analog zu den Beweisen von Satz 3.35/3.39 unter Verwendung von Lemma 3.4. □

Zusätzlich zu absoluten Fehlerschätzern wollen wir schließlich noch relative Fehlerschätzer herleiten und die zugehörigen Effektivitäten untersuchen.

3.34 Satz (Relative Fehlerschätzer)

Wir definieren die relativen Fehlerschätzer

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

für den relativen Fehler in der Energienorm,

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

für den relativen Fehler in der X-Norm und

$$\Delta_N^{s,rel}(\mu) := \frac{\|R(\mu)\|_X^2}{\alpha_{LB}(\mu)s_N(\mu)}$$

für den relativen Ausgabefehler. Dann gilt unter den Voraussetzungen von Satz 3.35 und falls $\Delta_N^{en,rel}(\mu) \leq 1, \Delta_N^{rel}(\mu) \leq 1$

$$\frac{|||u_h(\mu) - u_N(\mu)|||_\mu}{|||u_h(\mu)|||_\mu} \leq \Delta_N^{en,rel}(\mu), \tag{3.17}$$

$$\frac{||u_h(\mu) - u_N(\mu)||_X}{||u_h(\mu)||_X} \leq \Delta_N^{rel}(\mu), \tag{3.18}$$

$$\frac{s_h(\mu) - s_N(\mu)}{s_h(\mu)} \leq \Delta_N^{s,rel}(\mu). \tag{3.19}$$

Beweis:
Falls $\Delta_N^{en,rel}(\mu) \leq 1$, so gilt

$$\left|\frac{|||u_h(\mu)|||_\mu - |||u_N(\mu)|||_\mu}{|||u_N(\mu)|||_\mu}\right| \leq \frac{|||u_h(\mu) - u_N(\mu)|||_\mu}{|||u_N(\mu)|||_\mu} \overset{(3.11)}{\leq} \frac{\|R(\mu)\|_X}{\alpha_{LB}(\mu)|||u_N(\mu)|||_\mu} = \frac{\Delta_N^{en,rel}(\mu)}{2} \leq \frac{1}{2}. \tag{3.20}$$

Aus (3.20) folgt:

$$|||u_N(\mu)|||_\mu - |||u_h(\mu)|||_\mu \leq \frac{1}{2}|||u_N(\mu)|||_\mu$$

$$\Rightarrow \frac{1}{2}|||u_N(\mu)|||_\mu \leq |||u_h(\mu)|||\mu. \tag{3.21}$$

Damit gilt:

$$\frac{|||e_N(\mu)|||_\mu}{|||u_h(\mu)|||_\mu} \overset{(3.11)}{\leq} \frac{\|R(\mu)\|_X}{\sqrt{\alpha_{LB}(\mu)}|||u_h(\mu)|||_\mu} \overset{(3.21)}{\leq} \frac{\|R(\mu)\|_X}{\sqrt{\alpha_{LB}(\mu)}|||u_N(\mu)|||_\mu} \cdot 2 = \Delta_N^{en,rel}(\mu).$$

Daraus folgt (3.17). Und (3.18) folgt analog zu (3.17). Schließlich gilt

$$\frac{s_h(\mu) - s_N(\mu)}{s_h(\mu)} \overset{(3.12)}{\leq} \frac{\Delta_N^s(\mu)}{s_h(\mu)} \overset{3.19}{\leq} \frac{\Delta_N^s(\mu)}{s_N(\mu)} = \Delta_N^{s,rel}(\mu).$$

□

3.44 Satz (Effektivitäten der relativen Fehlerschätzer)

Wir definieren die Effektivitäten der relativen Fehlerschätzer $\eta_N^{en,rel}(\mu), \eta_N^{rel}(\mu), \eta_N^{s,rel}(\mu)$ wie folgt:

$$\eta_N^{en,rel}(\mu) := \frac{\Delta_N^{en,rel}(\mu)}{|||e_N(\mu)|||_\mu/|||u_h(\mu)|||_\mu}, \ \eta_N^{rel}(\mu) := \frac{\Delta_N^{rel}(\mu)}{||e_N(\mu)||_X/||u_h(\mu)||_X}, \ \eta_N^{s,rel}(\mu) := \frac{\Delta_N^{s,rel}(\mu)}{(s_h(\mu) - s_N(\mu)/s_h(\mu)}.$$

Dann gilt unter den Voraussetzungen von Satz 3.35, falls $\Delta_N^{en,rel}(\mu) \leq 1, \Delta_N^{rel}(\mu) \leq 1, \Delta_N^{s,rel}(\mu) \leq 1$:

$$\Delta_N^{en,rel}(\mu) \leq 3\sqrt{\frac{\gamma(\mu)}{\alpha_{LB}(\mu)}}, \tag{3.21b}$$

$$\Delta_N^{rel}(\mu) \leq 3\frac{\gamma(\mu)}{\alpha_{LB}(\mu)} \tag{3.22}$$

$$\eta_N^{s,rel}(\mu) \leq 2\frac{\gamma(\mu)}{\alpha_{LB}(\mu)}. \tag{3.23}$$

Beweis:
Wie in Beweis von Satz 3.43 impliziert $\Delta_N^{en,rel}(\mu) \leq 1$ dass

$$\left|\frac{|||u_h(\mu)|||_\mu - |||u_N(\mu)|||_\mu}{|||u_N(\mu)|||_\mu}\right| \leq \frac{1}{2}.$$

Damit gilt $|||u_h(\mu)|||_\mu - |||u_N(\mu)|||_\mu \leq \frac{1}{2}|||u_N(\mu)|||_\mu$ und damit

$$|||u_h(\mu)|||_\mu \leq \frac{3}{2}|||u_N(\mu)|||_\mu. \tag{3.24}$$

Damit folgt:

$$\eta_N^{en,rel}(\mu) \overset{Def}{=} \frac{2\left\|R(\mu)\right\|_X}{\sqrt{\alpha_{LB}(\mu)}|||u_h(\mu)|||_\mu}\frac{|||u_h(\mu)|||_\mu}{|||e_N(\mu)|||_\mu}$$
$$\overset{(3.16)}{\leq} 2\frac{\sqrt{\gamma(\mu)}\left\|e_N(\mu)\right\|_X}{\sqrt{\alpha_{LB}(\mu)}|||u_h(\mu)|||_\mu}\frac{|||u_h(\mu)|||_\mu}{|||e_N(\mu)|||_\mu} \overset{(3.24)}{\leq} 3\sqrt{\frac{\gamma(\mu)}{\alpha_{LB}(\mu)}} \Rightarrow (3.21b).$$

(3.22) folgt analog zu (3.21b). Schließlich gilt

$$\eta_N^{s,rel}(\mu) = \frac{\Delta_N^s(\mu)/s_N(\mu)}{(s_h(\mu)-s_N(\mu))/s_h(\mu)} = \eta_N^s(\mu)\frac{s_h(\mu)}{s_N(\mu)} \overset{(3.15)}{\leq} \frac{\gamma(\mu)}{\alpha_{LB}(\mu)}\frac{s_h(\mu)}{s_N(\mu)}.$$

Der letzte Faktor ist beschränkt, da

$$\frac{s_h(\mu)}{s_N(\mu)} = 1 + \frac{s_h(\mu)-s_N(\mu)}{s_N(\mu)} \overset{(3.19)}{\leq} 1 + \Delta_N^{s,rel}(\mu) \leq 2 \Rightarrow (3.23).$$

□

3.45 Bemerkung

Durch "Tauschen"des Skalarprodukts/der Norm auf X_h können die Fehlerschätzer und Effektivitäten verbessert werden, ohne die Ausgabe des reduzierten Modells oder die reduzierte Lösung zu verändern. Wähle $\bar{\mu} \in \mathcal{P}$ fest und betrachte auf X_h das Skalarprodukt $(((.,.)))_{\bar{\mu}}$ mit induzierter Norm $|||.|||_{\bar{\mu}}$. Wegen Lemma 3.4 Normäquivalenz folgt aus der Stetigkeit und Koerzivität der Bilinearform b auf X_h bzgl. der X-Norm die Stetigkeit und Koerzivität auf X_h bzgl. der $|||.|||_{\bar{\mu}}$-Norm und umgekehrt. Gleiches gilt für die Stetigkeit von l und f. Dann gilt $\alpha_h(\bar{\mu}) := \inf_{v\in X_h}\frac{b(v,v;\bar{\mu})}{|||v|||_{\bar{\mu}}^2} = 1$ und $\gamma_h(\bar{\mu}) := \sup_{u,v\in X_h}\frac{b(u,v;\bar{\mu})}{|||u|||_{\bar{\mu}}|||v|||_{\bar{\mu}}} = \sup_{u,v\in X_h}\frac{(((u,v)))_{\bar{\mu}}}{|||u|||_{\bar{\mu}}|||v|||_{\bar{\mu}}} \leq 1$. Für einen Fehlerschätzer welcher auf der $|||.|||_{\bar{\mu}}$-Norm des Riesz-Repräsentanten $R(\mu)$ basiert gilt also $\eta(\bar{\mu}) = 1$. Er ist in diesem Sinne optimal. Nimmt man an, dass $\alpha_h(\mu)$ und $\gamma_h(\mu)$ stetig von Parameter μ abhängen, kann man erwarten auch in einer Umgebung von $\bar{\mu}$ sehr effektive Fehlerschätzer zu erhalten.

3.4.2 Offline/Online-Zerlegung des Fehlerschätzers

Damit wir in der Online-Phase verifizieren können, dass der Approximationsfehler unter einer vorgegeben Toleranz liegt, ist es wichtig, dass wir auch den Fehlerschätzer Offline/Online zerlegen können. Die Erkenntnis, dass sich die affine Parameterabhängigkeit der Bilinearform b und der Linearform f auf das

Residuum und die Norm des Riesz-Repräsentanten überträgt ist hierbei von zentraler Bedeutung. Mit Lemma 3.34 und der afiinen Parameterabhängigkeit von b und f folgt:

$$\begin{aligned}(R(\mu), v) = r(v;\mu) &= f(v;\mu) - b(u_N(\mu), v;\mu) \\ &= \sum_{q=1}^{Q_f} \theta_f^q(\mu) f^q(v) - \sum_{q=1}^{Q_b}\sum_{n=1}^{N} \theta_b^q(\mu) U_n^N(\mu) b^q(\phi_n, v).\end{aligned}$$

Nach dem Riesz'schen Darstellungssatz 2.9 existieren $R_f^q \in X - h$ mit

$$(R_f^q, v)_X = f^q(v)\ \forall v \in X_h,\ 1 \le q \le Q_f \tag{3.25}$$

und $R_b^{q,n} \in X_h$ mit

$$(R_b^{q,n}, v)_X = b^q(\phi_n, v)\ \forall v \in X_h,\ 1 \le q \le Q_b,\ 1 \le n \le N. \tag{3.26}$$

Daher können wir weiter umformen:

$$\begin{aligned}(R(\mu), v)_X &= \sum_{q=1}^{Q_f} \theta_f^q(\mu)(R_f^q, v)_X - \sum_{q=1}^{Q_b}\sum_{n=1}^{N} \theta_b^q(\mu) U_n^N(\mu)(R_b^{q,n}, v)_X \\ &= \left(\sum_{q=1}^{Q_f} \theta_f^q(\mu) R_f^q - \sum_{q=1}^{Q_b}\sum_{n=1}^{N} \theta_b^q(\mu) U_n^N(\mu) R_b^{q,n}, v\right)_X \quad \forall v \in X_h \\ &\Rightarrow R(\mu) = \sum_{q=1}^{Q_f} \theta_f^q(\mu) R_f^q - \sum_{q=1}^{Q_b}\sum_{n=1}^{N} \theta_b^q(\mu) U_n^N(\mu) R_b^{q,n}.\end{aligned}$$

Wir fassen dieses Resultat im folgenden Lemma zusammen.

3.46 Lemma (Affine Parameterabhängigkeit von $R(\mu)$

Seien b, f affin parametrisch und $R_f^q, R_b^{q,n} \in X_h$ definiert wie in (3.25)/(3.26) . Sei $Q_R := Q_f + N \cdot Q_b$ und R_R^q für $1 \le q \le Q_R$ eine Aufzählung von $R_f^q, R_b^{q,n}$:

$$(R_R^1, \dots, R_R^{Q_R}) := (R_f^1, \dots, R_f^{Q_f}, R_b^{1,1}, \dots, R_b^{Q_b,1}, R_b^{1,2}, \dots, R_b^{Q_b,2}, \dots, R_b^{1,N}, \dots, R_b^{N,Q_b}).$$

Für $\mu \in \mathcal{P}$ sei $u_N(\mu) = \sum_{n=1}^N U_n^N(\mu)\phi_n$ die Lösung von $(P_N(\mu))$. Dann definieren wir $\theta_R^q : \mathcal{P} \to \mathbb{R},\ 1 \le q \le Q_R$ durch

$$(\theta_R^1(\mu), \dots, \theta_R^{Q_R}(\mu)) := (\theta_f^1(\mu), \dots, \theta_f^{Q_f}(\mu), -\theta_b^1(\mu)U_1^N(\mu), \dots, -Q_b^{Q_b}(\mu), \dots, -\theta_b^{Q_b}(\mu)U_N^N(\mu)).$$

Dann ist der Riesz-Repräsentant $R(\mu) \in X_h$ des Residuums affin parametrisch:

$$R(\mu) = \sum_{q=1}^{Q_R} \theta_R^q(\mu) R_R^q.$$

3.47 Lemma

Sei $g \in X_h'$ und $\bar{\varphi}_i,\ 1 = 1, \dots, N_h$ die Knotenbasis von X_h wie in Def 2.50 definiert. Dann definieren wir die (Skalarprodukt-)Matrix $\mathbb{X}_h \in \mathbb{R}^{N_h \times N_h}$ durch

$$(\mathbb{X})_{ij} := (\bar{\varphi}_j, \bar{\varphi}_i)_X,\ 1 \le i, j \le N_h.$$

Indem wir den Riesz-Repräsentanten $v_g \in X_h$ zu $g \in X_h'$ in der Knotenbasis darstellen:

$$v_g = \sum_{i=1}^{N_h} V_g^i \bar{\varphi}_i$$

erhalten wir den Koeffizientenvektor $\mathbb{V}_g = (V_g^1, \ldots, V_g^{N_h}) \in \mathbb{R}^{N_h}$ und damit v_g durch Lösen des linearen Gleichungssystems

$$\mathbb{X}_h \mathbb{V}_g = \mathbb{G}, \text{ wobei } \mathbb{G} := (g(\bar{\varphi}_1), \ldots, g(\bar{\varphi}_{N_h}) \in \mathbb{R}^{N_h}.$$

Beweis:
Betrachte dazu eine beliebige Testfunktion $w = \sum_{i=1}^{N_h} W_i \bar{\varphi}_i \in X_h$ mit Koeffizientenvektor $\mathbb{W} = (W_1, \ldots, W_{N_h})$. Dann gilt:

$$\begin{aligned} g(w) &= \sum_{i=1}^{N_h} W_i g(\bar{\varphi}_i) = \mathbb{W}^T \mathbb{G} = \mathbb{W}^T \mathbb{X}_h \mathbb{V}_g \\ &= \left(\sum_{i=1}^{N_h} V_g^i \bar{\varphi}_i, \sum_{j=1}^{N_h} W_j \bar{\varphi}_j \right)_X = (v_g, w)_X. \end{aligned}$$

□

Da $R(\mu)$ affin parametrisch ist, können wir die Berechnung der entsprechenden Norm in eine Offline- und eine Online-Phase zerlegen.

3.48 Satz (Offline/Online-Zerlegung von $\|R(\mu)\|$)

Offline-Phase: Nach der Offline-Phase des RB-Modells wie in Folgerung 3.31 beschrieben, assemblieren wir die Matrix $\mathbb{K} \in \mathbb{R}^{Q_R \times Q_R}$ definiert durch

$$\mathbb{K}_{ij} := (R_R^j, R_R^i)_X, \ 1 \leq i, j \leq Q_R$$

unter der Verwendung der Matrix $\mathbb{X}_h$.
Online-Phase: Für gegebenes μ und berechnete RB-Lösung $u_N(\mu)$ bestimmen wir den Koeffizientenvektor $\hat{\theta}_R(\mu) := (\theta_R^1(\mu), \ldots, \theta_R^{Q_R}(\mu))^T \in \mathbb{R}^{Q_R}$ und erhalten

$$\|R(\mu)\|_X = \sqrt{\hat{\theta}_R(\mu)^T \mathbb{K} \hat{\theta}_R(\mu)}.$$

Beweis:

$$\|R(\mu)\|_X^2 = \left(\sum_{q=1}^{Q_R} \theta_R^q(\mu) R_R^q, \sum_{q'=1}^{Q_R} \theta_R^{q'}(\mu) R_R^{q'} \right)_X = \hat{\theta}_R^T(\mu) \mathbb{K} \hat{\theta}_R(\mu).$$

Ferner benötigen wir für die relativen Fehlerschätzer die Norm von $u_N(\mu)$.

3.49 Satz (Offline/Online Zerlegung der Norm von $U_N(\mu)$)

Offline-Phase: Im Fall von $\Delta_N^{rel}(\mu)$ assemblieren wir nach der in Folgerung 3.31 beschriebenen Offline-Phase des RB-Modells die reduzierte (Skalarprodukt-)Matrix $\mathbb{X}_N \in \mathbb{R}^{N \times N}$ definiert durch

$$(\mathbb{X}_N)_{nm} := (\phi_m, \phi_n)_X.$$

Online-Phase: Für gegebenes μ und wie in Folgerung 3.31 beschrieben berechnetes $u_N(\mu) = \sum_{n=1}^N U_n^N(\mu)\phi_n$ mit Koeffizientenvektor $\mathbb{U}_N(\mu) = (U_1^N(\mu), \dots, U_N^N(\mu))$ berechnen wir $\|u_N(\mu)\|_X = \sqrt{\mathbb{U}_N^T(\mu)\mathbb{X}_N\mathbb{U}_N(\mu)}$ für $\Delta_N^{rel}(\mu)$ oder $|||u_N(\mu)|||_\mu = \sqrt{\mathbb{U}_N^T(\mu)\mathbb{B}_N(\mu)\mathbb{U}_N(\mu)}$ für $\Delta_N^{en,rel}(\mu)$, wobei $\mathbb{B}_N(\mu)$ definiert wie in Satz 3.26.

Beweis:

$$|||u_N(\mu)|||_\mu^2 = b(u_N(\mu), u_N(\mu); \mu) = \sum_{n,m=1}^N U_n^N(\mu)U_m^N(\mu)b(\phi_n(\mu), \phi_m(\mu); \mu) = \mathbb{U}_N^T(\mu)\mathbb{B}_N(\mu)\mathbb{U}_N(\mu).$$

Wie in Folgerung 3.28 können wir mit Hilfe der Transformationsmatrix $\mathbb{V} \in \mathbb{R}^{N_h \times N}$, definiert durch

$$\mathbb{V}_{in} := \phi_n^i,\ 1 \le i \le N_h,\ 1 \le n \le N,$$

$\mathbb{X}_N = \mathbb{V}^T\mathbb{X}_N\mathbb{V}$ folgern. Der Beweis für $\|u_N(\mu)\|_X$ geht dann analog.
Schließlich benötigen wir noch eine untere Schranke für die Koerzivitätskonstante $\alpha_{LB}(\mu)$. Falls b gleichmäßig koerziv bzgl. μ mit $\inf_{\mu\in\mathcal{P}} \alpha_h(\mu) \ge \alpha_0$ und α_0 analytisch bestimmbar oder berechenbar, kann α_0 gewählt werden. Für manche Randwertprobleme können wir sogar $\alpha(\mu)$ analytisch bestimmen und dann $\alpha_{LB}(\mu) = \alpha(\mu)$ wählen.
Im allgemeinen Fall lässt sich unter bestimmten Voraussetzungen die sogenannte **Min-Theta-Methode** anwenden.

3.50 Satz (Min-Theta-Methode zur Berechnung von $\alpha_{LB}(\mu)$)

Sei b koerziv und affin parametrisch mit $b^q(v, v) \ge 0\ \forall v \in X_h$ und $\theta_b^q(\mu) > 0\ \forall \mu \in \mathcal{P},\ q = 1, \dots, Q_b$. Sei $\mu \in \mathcal{P}$ fest und $\alpha_h(\mu)$ bekannt. Dann gilt

$$0 < \alpha_{LB}(\mu) \le \alpha_h(\mu)\ \forall \mu \in \mathcal{P}$$

mit der unteren Schranke

$$\alpha_{LB}(\mu) := \alpha_h(\bar{\mu}) \cdot \min_{q=1,\dots,Q_b} \frac{\theta_b^q(\mu}{\theta_b^q(\bar{\mu})}.$$

Beweis:
Da $\alpha_h(\mu) > 0$ und $0 < c(\mu) := \min_{q=1,\dots,Q_b} \frac{\theta_b^q(\mu}{\theta_b^q(\bar{\mu})}$, gilt

$$0 < \alpha_h(\mu)c(\mu) = \alpha_{LB}(\mu).$$

Für alle $v \in X_h$ gilt dann:

$$\begin{aligned} b(v, v; \mu) &= \sum_{q=1}^{Q_b} \theta_b^q(\mu)b^q(v, v) = \sum_{q=1}^{Q_b} \frac{\theta_b^q(\mu)}{\theta_b^q(\bar{\mu})}\theta_b^q(\bar{\mu})b^q(v, v) \\ &\ge \sum_{q=1}^{Q_b} \left(\min_{q=1,\dots,Q_b} \frac{\theta_b^q(\mu}{\theta_b^q(\bar{\mu})} \right) \theta_b^q(\bar{\mu})b^q(v, v) \\ &= c(\mu)b(v, v; \bar{\mu}) \ge c(\mu)\alpha_h(\bar{\mu}) \|v\|_X^2 = \alpha_{LB}(\mu) \|v\|_X^2 . \end{aligned}$$

Also insbesondere:

$$\alpha_h(\mu) := \min_{v\in X_h} \frac{b(v, v; \mu)}{\|v\|_X^2} \ge \alpha_{LB}(\mu).$$

□

Fpr die Min-Theta-Methode benötigen wir eine Auswertung von $\alpha_h(\mu)$ für $\mu = \bar{\mu}$ in der Offline-Phase. Dazu ist es im Allgemeinen notwendig ein hochdimensionales Eigenwertproblem zu lösen.

3.51 Satz (Berechnung von $\alpha_h(\mu)$ für das hochdimensionale, diskrete Modell)

Sei $\mathbb{B}_h(\mu)$ definiert wie in (3.4) und $\mathbb{X}_h$ definiert wie in Lemma 3.47. Dann ist

$$\alpha_h(\mu) = \lambda_{\min}(\mathbb{X}_h^{-1}\mathbb{B}_h(\mu)), \tag{3.27}$$

wobei $\lambda_{\min}$ der kleinste Eigenwert der Matrix $\mathbb{X}_N^{-1}\mathbb{B}(\mu)$ bezeichnet.
Beweis:
Da $\mathbb{X}_h$ symmetrisch und positiv definit, lässt sich $\mathbb{X}_h$ mittels Cholesky-Zerlegung als $\mathbb{X}_h = \mathbb{L}\mathbb{L}^T$ darstellen. Mittels Substitution erhalten wir ($y \neq 0$):

$$\alpha_h(\mu) = \inf_{y \in X_h} \frac{b(y,y;\mu)}{\|y\|_X^2} = \inf_{\mathbb{Y} \in \mathbb{R}^{N_h}} \frac{\mathbb{Y}^T \mathbb{B}_h(\mu)\mathbb{Y}}{\mathbb{Y}^T \mathbb{X}_h \mathbb{Y}}$$
$$\text{mit } \mathbb{W} := \mathbb{L}^T\mathbb{Y} = \inf_{\mathbb{W}} \in \mathbb{R}^{N_h} \frac{(\mathbb{L}^{-T}\mathbb{W})^T \mathbb{B}_h(\mu)(\mathbb{L}^{-T}\mathbb{W})}{(\mathbb{L}^{-T}\mathbb{W})^T \mathbb{L}\mathbb{L}^T(\mathbb{L}^{-T}\mathbb{W})} = \inf_{\mathbb{W}} \in \mathbb{R}^{N_h} \frac{\mathbb{W}^T\mathbb{L}^{-1}\mathbb{B}_h(\mu)\mathbb{L}^{-T}\mathbb{W}}{\mathbb{W}^T\mathbb{W}}.$$

Damit minimiert $\alpha_h(\mu)$ den Rayleigh-Quotienten zur symmetrischen Matrix $\tilde{\mathbb{B}}_h(\mu) := \mathbb{L}^{-1}\mathbb{B}_h(\mu)\mathbb{L}^{-T}$.
Satz von Courand-Fischer (Minimum-Maximums Prinzip)
Ist $A \in \mathbb{R}^{n\times n}$ symmetrisch mit aufsteigend sortierten Eigenwerten $\lambda_1 \leq \cdots \leq \lambda_n$ und bezeichnet X_i die Menge der i-dimensionalen Untervektorräume von $\mathbb{R}^n$, $i = 1, \ldots, n$, dann hat der i-te Eigenwert von A die Darstellung:

$$\lambda_i = \min_{X \in X_i} \max_{x \in X, x \neq 0} \frac{(x, Ax)_2}{(x,x)} = \max_{X \in X_{n-i+1}} \min_{x \in X, x \neq 0} \frac{(x, Ax)_2}{(x,x)_2}.$$

Nach diesem Prinzip ist $\alpha_h(\mu)$ damit der kleinste Eigenwert von $\tilde{\mathbb{B}}_h(\mu)$. Da die Matrizen $\tilde{\mathbb{B}}_h(\mu)$ und $\mathbb{X}_h^{-1}\mathbb{B}_h(\mu)$ ähnlich sind, wegen

$$\mathbb{L}^T(\mathbb{X}_h^{-1}\mathbb{B}_h(\mu))\mathbb{L}^{-T} = \mathbb{L}^T\mathbb{L}^{-T}\mathbb{L}^{-1}\mathbb{B}_h(\mu)\mathbb{L}^{-T} = \tilde{\mathbb{B}_h(\mu)}$$

und daher identische Eigenwerte haben, folgt damit die Behauptung.

3.52 Bemerkung

Da insbesondere die großen Matrizen $\mathbb{X}_h$ nicht invertiert werden sollte, verwendet man in der Praxis entweder einen Eigenwertlöser, welcher nur mittels Matrix-Vektor Produkten operiert: Sobald das Produkt $\mathbb{Y} = \mathbb{X}_h^{-1}\mathbb{B}_h(\mu)\mathbb{W}$ berechnet werden muss, löst man stattdessen $\mathbb{X}_h\mathbb{Y} = \mathbb{B}_h(\mu)\mathbb{W}$. Alternativ können auch Löser für generalisierte Eigenwertprobleme der Form $\mathbb{B}_h(\mu)\mathbb{Y} = \lambda\mathbb{X}\mathbb{Y}$ verwendet werden.

5.53 Bemerkung

Auch im kontinuierlichen Fall kann man ein Eigenwertproblem betrachten und lösen um $\alpha(\mu)$ zu bestimmen. Siehe dazu das Buch von Haardonk.

3.54 Bemerkung

Für Probleme in denen man die Min-Theta-Methode nicht anwenden kann, kann man stattdessen die **Successive Constraint Methode** verwenden, siehe Buch von Huynh, Rozza, Sen und Patera.

3.55 Bemerkung

Damit können wir die Offline/Online Zerlegung des Fehlers wie folgt zusammenfassen: **Offline:**
Berechnung des Riesz-Repräsentanten und Assemblierung der Matrix $\mathbb{K}$ aus Satz 3.48 in

$$\mathcal{O}((Q_f + NQ_b)N_h^2 + (Q_f + NQ_b)^2 N_h) \text{ Rechenschritten.}$$

Für relativen Fehlerschätzer möglicherweise Assemblierung von X_N aus Satz 3.49. Falls die Min-Theta Methode verwendet wird, muss noch $\alpha_h(\bar{\mu})$ in $\mathcal{O}(N_h\sigma)$ mit $\sigma \geq 2$ einmalig berechnet werden, wie in Satz 3.51 beschrieben.

Online:
Wie in Satz 3.48 beschrieben bestimmen wir $\|R(\mu)\|_X$ in $\mathcal{O}((Q_f+NQ_b)^2)$ Rechenschritten. Für den rel. Fehlerschätzer berechnen wir zusätzlich entweder $|||u_N(\mu)|||_\mu$ in $\mathcal{O}(N^2)$ Rechenschritten oder $\|u_N(\mu)\|_X$ in $\mathcal{O}(N^2)$ Rechenschritten. Schließlich können wir für die Min-Theta Methode $\alpha_{LB}(\mu)$ in $\mathcal{O}(Q_b)$ Rechenschritten bestimmen.
Die Online Phase zur Bestimmung des Fehlerschätzers ist damit unabhängig von N_h. Da die reduzierte Lösung in $\mathcal{O}(N^3)$ Rechenschritten berechnet werden kann, ist für große Q_b die Berechnung des Fehlerschätzers aber möglicherweise der Teil, der Online Phase, welcher die meiste zeit benötigt.

3.56 Bemerkung (Relevanz des Fehlerschätzers)

Fehlerschätzer werden in den RB Methoden sowohl in der Online Phase zur 'Certification' der Approximation, als auch in der Offline Phase zur Basisgenerierung eingesetzt, wie wir im nächsten Kapitel sehen werden.

4 Basiskonstruktion

Ziel:

(1) Bestimmung eines 'möglichst guten' RB-Raumes $X_N = \operatorname{span}\{u_h(\mu)\} \subset X_h$ mit Basis ϕ_N, welche $M := \{u_h(\mu) \mid \mu \in \mathcal{P}\}$ global approximiert.

(2) Optimales X_N: formalisierbar durch Minimierung eines Funktionals, z.B. minimiere den maximalen Energiefehler:

$$\min_{Y \subset X_h, \dim Y = N} \left(\max_{\mu \in \mathcal{P}} |||u_h(\mu) - u_N(\mu)|||_\mu \right) \tag{4.1}$$

oder Minimierung der mittleren Projektionsfehlers

$$\min_{Y \subset X_h, \dim Y = N} \int_P \|u_h(\mu) - P_Y u_h(\mu)\|_X^2 \, \mathrm{d}\mu, \tag{4.2}$$

wobei P_Y die orthogonale Projektion auf Y bezeichnet.

(3) Gute Basis ϕ_N: orthogonal für numerische Stabilität; Hierachie, so dass Basisvektoren nach Relevanz geordnet sind, d.h.

$$X_N := \operatorname{span}\{\varphi_1, \ldots, \varphi_{N'}\} \text{ Sequenz von optimalen Räumen,}$$

für $1 \leq N' \leq N$, damit N' Variation einer Fehlerkontrolle erlaubt.

Was wir formal unter einem optimalen Unterraum verstehen wollen, können wir mit Hilfe der sogenannten Kolmogorov-n-Weite beschreiben.

4.1 Definition (Kolmogorov-n-Weite, optimaler Unterraum)

Sei A eine kompakte Teilmenge in einem HR X. Zu einem abgeschlossenen Unterraum $X_n \subset X$ mit $\dim X_n = n$ nennen wir

$$\begin{aligned} E(X_n, A) &= \sup_v \in X \inf_{w \in X_n} \|v - w\|_X \\ &= \sup_{v \in A} \|v - P_{X_n} v\|_X \end{aligned}$$

Abstand von X_n zu A. Für $n \in \mathbb{N}$ nennen wir

$$d_n(A, X) := \inf_{X_n \subset X, \dim X_n = n} E(X_n, A) \tag{4.3}$$

die **Kolmogorov-n-Weite** der Menge A. Ein Unterraum $X_n \subset X$ mit $\dim X_n = n$, welcher die Kolmogorov-n-Weite minimiert heißt optimaler Unterraum für $d_n(A, X)$.

4.2 Bemerkung

(1) Die Kolmogorov-n-Weite ist also ein Maß für die Bestapproximation durch lineare Unterräume.

(2) Es gilt trivialerweise $d_0(A, X) = \sup \|v\|_X$ und falls $n_0 := \dim span(A) < \infty$, so gilt $d_n(A, X) = 0 \forall n \geq n_0$.

(3) Für $A = M \subset X_h$ mit $u_h(\mu)$ Lösung von $(P_h(\mu))$ und $u_N(\mu)$ Lösung von $(P_N(\mu))$ gilt also

$$\begin{aligned}\|u_N(\mu) - u_h(\mu)\|_X &\leq \sqrt{\frac{\gamma(\mu)}{\alpha(\mu)}} \inf_{v \in X_N} \|u_N(\mu) - v\| \\ &\leq E(X_n, M) \sup_{\mu \in P} \sqrt{\frac{\gamma(\mu)}{\alpha(\mu)}}.\end{aligned}$$

Falls also $E(X_n, M)$ und $\frac{\gamma(\mu)}{\alpha(\mu)}$ beschränkt, so ist der RB-Fehler klein.

(4) Präzise Werte für d_n sind selten bekannt. Für endliche Mengen der Einheitskugeln können aber exakte Werte der Schranken für d_n angegeben werden [Pinkus, 85].

(5) **Beispiel:** $A := \{v \in X \mid \|v\| \leq 1\}$ erfüllt $d_n(A, x) = 1$ für alle $n \leq \dim X$ und $d_n(A, X) = 0$ für $n \geq \dim X$.

(6) **Beispiel:** $A := [-1, 1]^m \subset X := \mathbb{R}^m$ erfüllt $d_n(A, X) = \sqrt{m - n}$ für alle $n \leq m$ und $d_n(A, X) = 0$ für $n \geq m$. Da die am Anfang des Kapitels genannten Optimierungsprobleme sehr komplex sind, treffen wir Vereinfachungen:

(i) **Diskretisierung des Parameterraums:**
Wähle endliche TM $S_{\text{train}} := \{\mu^i\}_{i=1}^{n_{\text{train}}} \subset P$ von Trainingsparametern, welche Trainingssnapshots $M_{\text{train}} := \{u_h(\mu) \mid \mu \in S_{\text{train}}\} \subset M$ definieren.

(ii) Schränke Optimierungsproblem auf S_{train} ein. Problem (4.2) wird dann in

$$\min_{Y \subset X_h, \dim Y = N} \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} \|u_h(\mu_i) - P_Y u_h(\mu_i)\|_X^2 \tag{4.4}$$

überführt.

Wir werden in Abschnitt 4.2 sehen, dass die sogenannten POD-Räume Optimierungsproblem (4.4) lösen. Zunächst beschäftigen wir uns jedoch mit dem Greedy-Algorithmus, welcher auf einer weiteren Approximation des Optimierungsproblems

$$\min_{Y \subset X_h, \dim Y = N} \max_{\mu \in S_{\text{train}}} |||u_h(\mu) - u_N(\mu)|||_\mu \tag{4.5}$$

beruht.

4.1 Greedy-Algorithmus

Da die Optimierung in (4.5) über die Unterräume immer noch ein sehr komplexes Optimierungsproblem ist, geht man stattdessen iterativ vor:
Wir konstruieren einen Approximationsraum, indem wir iterativ neue Basisfunktionen hinzufügen. Hierbei wird die neue Basisfunktion gerade so gewählt, dass sie den Ausdruck

$$\max_{\mu \in S_{\text{train}}} |||u_h(\mu) - u_N(\mu)|||_\mu$$

minimiert. Dies ist das Grundprinzip des **Greedy-Algorithmus**, welcher Schrittweise sowohl die Sample-Menge S_N als auch die Basis Φ_N erweitert. Ein wichtiger Bestandteil ist ein Fehlerindikator $\Delta(Y, \mu) \in \mathbb{R}^+$, welcher den zu erwartenden Approximationsfehler für den Parameter μ schätzt, wenn man $Y = X_N$ als den Approximationsraum wählt. Schließlich st $N_{\max}$ eine vorgegebene maximale Dimension des Raumes X_N.

4.3 Definition (Greedy-Algorithmus)

Sei $S_{\text{train}} \subset P$ eine gegebene Trainingsmenge von Parametern und $\varepsilon_{tol} > 0$ eine gegebene Fehlertoleranz. Setze $X_0 := 0, S_0 := \emptyset, \Phi_0 := \emptyset$ und definiere iterativ:

```
for n in range(1,N_max):
    Finde
     μ^(n) := argmax_{μ∈S_train} Δ(X_{n-1}, μ)
         S_n := S_{n-1} ∪ {μ^(n)} φ_n = u_h(μ^(n)) Φ_n := Φ_{n-1} ∪ {φ_n} X_n := span{Φ_n}
    if   max_{μ∈S_train} Δ(X_n, μ) ≤ ε_tol  :
         break
N:= n
return  S_N, Φ_N
```

4.4 Bemerkung

Angenommen $N_{\max} = |S_{\text{train}}|$. Wenn für alle $\mu \in P$ und Unterräume $Y \subset X_n$ gilt, dass

$$u_h(\mu) \in Y \Rightarrow \Delta(Y, \mu) = 0, \tag{4.6}$$

dann endet der Greedy-Algorithmus nach höchstens $N \leq |S_{\text{train}}|$. Dies liegt daran, dass (4.6) sicher stellt, dass kein Parameterwert in S_{train} zweimal gewählt wird.

4.5 Bemerkung

(1) Der Greedy-Alg. erzeugt eine hierachische Basis.

(2) Φ_N erzeugt vom Greedy-Alg. ist eine Lagrange-RB Basis zur Samplemenge S_N, im Allgemeinen also nicht orthogonal. Um ein numerisch möglichst stabiles Verfahren zu erhalten, orthonomalisiert man die Basis mit Gram-Schmidt.

4.6 Bemerkung (Wahl des Fehlerindikators)

Es gibt verschiedene Möglichkeiten für die Wahl des Fehlerindikators $\Delta(Y, \mu)$ im Greedy-Alg.

(1) **Projektionsfehler:** In manchen Fällen ist es sinnvoll im Greedy-Alg. den Bestapproximationsfehler zu verwenden:

$$\Delta(Y, \mu) := \inf_{v \in Y} |||u_h(\mu) - v|||_\mu = |||u_h(\mu) - P_Y u_h(\mu)|||_\mu.$$

Dieser Fehlerindikator ist teuer, da seine Auswertung Operationen erfordert, welche mit N_h skalieren. Daher kann S_{train} im Allgemeinen nur relativ klein gewählt werden. Ferner müssen alle Snapshots $u_h(\mu)$ verfügbar sein, was aus Speichergründen zusätzlich die Größe von S_{train} begrenzt. Der Vorteil dieser Fehlerindikatoren ist allerdings, dass weder ein RB Modell, noch ein a posteriori Fehlerschätzer für die Basisergänzung notwendig sind, was insbesondere für komplexe Probleme von Relevanz ist. Ferner schleißen wir aus dem Céa-Lemma, dass der so konstruierte RB-Raum gute Approximationseigenschaften hat. Wir bezeichnen diese Version des Greedy-Alg. als 'strong greedy'.

(2) **RB-Fehler zwischen $u_h(\mu)$ und $u_N(\mu)$:** Falls wir ein RB-Modell, aber kein Fehlerschätzer zur Verfügung haben, können wir als Fehlerindikator

$$\Delta(Y, \mu) = |||u_h(\mu) - u_N(\mu)|||_\mu$$

verwenden. Auch dieser Fehlerindikator ist teuer zu berechnen und benötigt alle Snapshots $u_h(\mu)$, was beides die Größe der Trainingsmenge S_{train} limitiert. Der Vorteil dieses Indikators ist, dass wir das Fehlermaß aus dem Optimierungsproblem in (4.5) betrachten.

(3) **A posteriori Fehlerschätzer:** Falls ein Fehlerschätzer zur Verfügung steht verwendet man im Allgemeinen

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

oder ein (relativen) Fehlerschätzer (für eine andere Norm). Da Aufgrund der Offline/Online Zerlegung die Auswertung von $\Delta(Y,\mu) = \Delta_N^{en}(\mu)$ sehr billig ist und wir während dieses Greedy-Alg. nur N-mal das hochdimensionale diskrete Modell $(P_h(\mu))$ lösen müssen, kann S_{train} sehr groß gewählt werden. Daher erwarten wir unter gewissen Voraussetzungen an den Fehlerschätzer $\Delta_N^{en}(\mu)$, dass wir mit diesem Fehlerindikator die Lösung des Optimierungsproblems, über den konstanten Parameterraum P, (4.1) besser approximieren können. Diese Version des Greedy-Alg. bezeichnen wir mit 'weak greedy'.

Alle drei Fehlerindikatoren erfüllen (4.6). Dies gilt trivialerweise für den Projektionsfehler, folgt aus der Reproduzierbarkeit von Lösungen in Folgerung 3.18 für den RB-Fehler und schließlich aus Folgerung 3.40 für den Fehlerschätzer.

4.7 Bemerkung

Alternativ kann man auch Fehlerindikatoren für die Ausgabe wie $|s_h(\mu) - s_N(\mu)|$ oder $\Delta_N^s(\mu)$ im Greedy-Alg. verwenden. Diese werden häufig auch als 'goal-oriented error indicators' bezeichnet. Letztere führen im Allgemeinen zu einem RB-Raum von niedriger Dimension, als die Indikatoren in Bemerkung 4.6, welcher $s_h(\mu)$, aber nicht notwendigerweise $u_h(\mu)$, gut approximiert. Im Gegensatz dazu resultieren die Indikatoren aus Bemerkung 4.6 in einer größeren Basis, welche sowohl $u_h(\mu)$, als auch $s_h(\mu)$ gut approximiert.

4.1.1 Konvergenzraten des Greedy-Algorithmus

Bis vor einigen Jahren wurde der Greedy-Alg. als heuristischer Algorithmus angesehen, welcher in der Praxis in vielen Fällen zwar ausgezeichnet funktioniert, dem aber die theoretische Grundlage fehlt. Dann wurden aber sehr nützliche exponentielle [Buffer, Meday, Pattera et al] und dann für algebraische Konvergenz [Biniv, Cohen et al] in HR und schließlich in [DeVore, Petrova et al] für Banachräume.

4.8 Satz (Konvergenzraten für Greedy-Alg.[BCDDPW,11])

Sei $M_e := \{u(\mu) \mid \mu \in P\}$, P kompakt und der Fehlerindikator $\Delta(Y,\mu)$ so gewählt, dass für ein geeignetes $\tau \in (0,1]$ gilt:

$$\left|\left|\left|u(\mu^{(n+1)}) - P_{X_n} u(\mu^{(n+1)})\right|\right|\right|_\mu \geq \tau \cdot \max_{u \in M_e} |||u - P_{X_n} u|||_\mu \,. \tag{4.7}$$

Sei ferner $\sigma_n := \max\limits_{u \in M_e} |||u - P_{X_n} u|||_\mu$. Dann gilt:

(i) **Algebraische Konvergenz:**
Falls $d_n(M_e, X) \leq Mn^{-\beta}$ für Konstanten $M, \beta \in \mathbb{R}^+$ unabhängig von τ und für alle $n \in \mathbb{N}$ und $d_0(M_e, X) \leq M$, dann gilt:

$$\sigma_n \leq C \cdot Mn^{-\beta}$$

mit geeigneter (berechenbarer) Konstante $C > 0$.

(ii) **Exponentielle Konvergenz:**
Falls $d_n(M_e, X) \leq Me^{-an^\beta}$ für $n \geq 0$ und für Konstanten $M, \beta, a \in \mathbb{R}^+$, dann gilt:

$$\sigma_n \leq C \cdot Me^{-cn^\rho}, +n \geq 0$$

mit $\rho := \frac{\beta}{\beta+1}$ mit geeignetem (berechenbaren) Konstanten $c, C > 0$.

Beweis: siehe Paper.

4.9 Bemerkung

(1) Lässt sich M_e gut durch einen linearen Unterraum approximieren, so liefert der Greedy-Alg. einen Approximationsraum, welcher nur leicht schlechter ist, als der optimale Unterraum.

(2) In der Praxis verwenden wir im Greedy-Alg. die Menge $M := \{u_h(\mu) \mid \mu \in P\}$. In [BCDPW,11] wurde gezeigt, dass sich in diesem Fall die Konvergenzraten für kleine Diskretisierungsfehler nicht verschlechtern.

4.10 (strong/weak greedy)

Für $\tau = 1$ heißt der Algorithmus strong greedy und für $\tau < 1$ weak greedy. Für $\Delta(Y, \mu) = \Delta_n^{en}(\mu)$ gilt (4.7) im Diskreten; wegen

$$\begin{aligned}
\left|\left|\left|u_h(\mu^{(n+1)}).P_{X_h}u_h(\mu^{(n+1)})\right|\right|\right|_\mu &= \inf_{v\in X_h}\left|\left|\left|u_h(\mu^{(n+1)}) - v\right|\right|\right|_\mu \\
&\overset{3.17/3.19}{=} \left|\left|\left|u_h(\mu^{(n+1)}) - u_n(\mu^{(n+1)})\right|\right|\right|_\mu \overset{3.39}{=} \frac{\Delta_n^{en}(\mu^{(n+1)})}{\eta_n^{en}(\mu^{(n+1)})} \\
&\overset{DefGreedy}{=} \frac{1}{\eta_n^{en}(\mu^{(n+1)})} \cdot \max_{\mu\in P}\Delta_n^{en}(\mu) \\
&\overset{3.35}{\geq} \frac{1}{\eta_n^{en}(\mu^{(n+1)})} \cdot \max_{\mu\in P}|||u_h(\mu) - u_n(\mu)|||_\mu \\
&\geq \frac{1}{\eta_n^{en}(\mu^{(n+1)})} \cdot \max_{\mu\in P}|||u_h(\mu) - P_{X_h}u_h(\mu)|||_\mu \\
&\overset{3.39}{\geq} \sqrt{\frac{\alpha_{LB}(\mu^{(n+1)})}{\gamma(\mu^{(n+1)})}} \cdot \max_{\mu\in P}|||u_h(\mu) - P_{X_h}u_h(\mu)|||_\mu .
\end{aligned}$$

Für glm. beschränkte und koerzive Bilinearformen $b(.,.;\mu)$ erhalten wir dann einen weak greedy, mit $\tau = \sqrt{\frac{\alpha_0}{\gamma_0}} \in (0,1)$, wobei α_0, γ_0 definiert in Definition 3.3.

4.1.2 Praktische Realisierung

4.11 Bemerkung (Wahl von S_{train})

Für niedrig dimensionale Parameterräume, d.h. $|P| \leq 5$, wird S_{train} häufig zufällig gewählt.

4.12 Bemerkung (Overfitting)

Die Folge von Fehlern $e_n := \max_{\mu\in S_{\text{train}}}\Delta(X_n, \mu)$, welche durch den Greedy-Alg. produziert wird, kann sehr stark von S_{train} abhängen. Wir können aufgrund von möglichem 'Overfitting' nicht notwendigerweise vom Approximationsverhalten des RB-Modells auf S_{train} auf das Approximationsverhalten auf P schließen; es kann sogar

$$\sup_{\mu\in P}\Delta(X_n, \mu) \gg e_n$$

gelten. Daher solte die Approximationsgüte des RB-Modells immer auf einer unabhängigen Testmenge $S_{test} \subset P$ validiert werden.

4.13 Bemerkung (adaptive Verfeinerung der Trainingsmenge)

Um sowohl eine große Trainingsmenge(hoher Rechenaufwand), als auch eine zu klein/schlecht gewählte Trainingsmenge (Overfitting) zu vermeiden, bietet es sich an die Trainingsmenge adaptiv zu verfeinern

(siehe [Haasdonk, Dihlmann, Ohlberger, 2012]). Hierbei wird die Trainingsmenge einem Gitter zu geordnet, wobei die Punkte der Trainingsmenge entweder als die Knoten [HO, 07] oder zufällig innerhalb der Gitterelemente gewählt werden [Efftang, Patera, 10]. Wie bei adaptiven FEM wird dann ein (lokaler) Fehlerschätzer auf den Gitterelementen ausgewertet und die Elemente mit dem größten Fehler zur Verfeinerung markiert. Durch Verfeinerung der markierten Elemente werden neue Punkte zur Trainingsmenge hinzugenommen. Auf diese Weise kann die Trainingsmenge an das betrachtete Problem angepasst und 'schwierige' Bereiche der Parametermenge, z.B. kleine Werte der Diffusionskonstanten, können identifiziert werden. So kann die Trainingsmenge entsprechend verfeinert werden. Außerdem besteht die Möglichkeit Overfitting vorzubeugen.

4.14 Bemerkung (Zerlegung des Parameterraums)

Im Greedy-Alg. aus Definition 4.3 wählt man im Allgemeinen $N_{\max}$ sehr groß, so dass die gewünschte Toleranz erreicht werden kann. Es wäre wünschenswert, wenn man sowohl $N_{\max}$ und damit die Online-Laufzeit, als auch die Toleranz kontrollieren könnte. Dies kann mit einer Zerlegung des Parameterraums, dem sogenannten hp-RB-Ansatz [Eftang, Petera, 10] erreicht werden. Zunächst wird der Parameterraum adaptiv in Teilgebiete zerlegt (h-Adaptivität). Anschließend werden für jedes Teilgebiet lokale reduzierte Basen generiert. Falls auf einem Teilgebiet nicht die Toleranz und $N \leq N_{\max}$ erreicht werden kann, wird dieses Teilgebiet erneut verfeinert/zerlegt und es werden neue lokale Basen generiert (p-Adaptivität). In der Online Phase muss dann lediglich für einen Parameterwert μ das zugehörige Teilgebiet und (lokale) RB-Modell identifiziert werden. Die da durch erzielte Kontrolle über Genauigkeit und Online-Laufzeit führt allerdings zu einer teureren und speicherintensiven Offline Phase.

4.2 Proper Orthogonal Decomposition (POD)

4.2.1 Exkurs Spektraltheorie/Motivation

4.15 Lemma (Adjungierter Operator im HR-Sinne)

Seien X, Y HR. Zu $A \in L(X,Y)$ existiert ein eindeutiger Operator $A^* \in L(Y,X)$ mit

$$(Ax, y)_Y = (x, A * y)_X \ \forall x \in X, y \in Y,$$

der sogenannte **adjungierte Operator**.

Beweis:
Für jedes feste $y \in Y$ ist $x \mapsto (Ax, y)_Y$ stetig und linear. Nach dem Riesz'schen Darstellungssatz existiert ein eindeutiges $z \in X$ mit $(Ax, y)_Y = (x, z)_X \ \forall x \in X$. Setze nun $A^* y = z$. Die Linearität ist klar und die Stetigkeit von $A*$ folgt aus der Stetigkeit von A, wegen

$$\begin{aligned} \|A^* y\|_X = \|z\|_X &= \sup_{x \in X \setminus \{0\}} \frac{(Ax, y)_Y}{\|x\|_X} \\ &\overset{C.S.}{\leq} \|y\|_Y \sup_{x \in X \setminus \{0\}} \frac{\|Ax\|_X}{\|x\|_X} = \|y\|_Y \|A\| . \end{aligned}$$

Eindeutigkeit folgt aus der Eindeutigkeit des Riesz-Repräsentanten.

4.16 Definition (selbstadjungiert)

Sei $A \in L(X)$, A heißt **selbstadjungiert**, falls $A = A^*$.

4.17 Beispiele

(a) Sei $X = \mathbb{R}^n$. Wird $A \in L(X)$ durch die Matrix $(a_{ij})_{ij}$ dargestellt, so wird A^* durch $(a_{ji})_{ij}$ dargestellt. Definition 4.16 ist also eine Verallgemeinerung des Begriffes der selbstadjungierten (symmetrischen) Matrix aus der Linearen Algebra.

(b) Sei $X = L^2((0,1)), u \in L^2((0,1)^2)$ und $A_u \in L(X)$ der Integraloperator

$$(A_u, v)(x) = \int_0^1 u(x,\mu)v(\mu)\mathrm{d}\mu.$$

Dann ist $A_u^* = A_{u^*}$ mit $u^*(x,\mu) = u(\mu,x)$. Dies kann als kontinuierliches Analogon von (a) aufgefasst werden.

(c) A^*A und AA^* sind stets selbstadjungiert.

Beweis:

(b)

$$\begin{aligned}(A_u v, w)_X &= \int_0^1 (A_u v)(x)v(x)\mathrm{d}x = \int_0^1\int_0^1 u(x,\mu)v(\mu)w(x)\mathrm{d}\mu\mathrm{d}x \\ &= \int_0^1 v(\mu)\int_0^1 u(x,\mu)w(x)\mathrm{d}x\mathrm{d}\mu = \int_0^1 v(\mu)(A_u^* w)(\mu)\mathrm{d}\mu.\end{aligned}$$

(b)

$$(A^*Av, w)_X = (Av, Aw)_Y = (v, A^*Aw)_X,\ AA^* \text{ analog.}$$

□

Erinnerung:
Ziel: Approximiere $u(x,\mu)$ durch $\sum_{n=1}^N U_n^N(\mu)\phi_n(x),\ X_N = \mathrm{span}\{\phi_1,\dots,\phi_N\}$. Als Minimierer von

$$\min_{Y\subset X, \dim Y = N} \int_0^1\int_0^1 \left| u(x,\mu) - \sum_{n=1}^N U_n^N(\mu)\phi_n(x)\right|^2 \mathrm{d}x\mathrm{d}\mu.$$

Ein solches Problem wurde erstmals von Schmidt 1907 betrachtet. Zur Lösung nutzte er aus, dass der Integraloperator A_u aus 4.17 (b) kompakt ist (s. Alt, Satz 8.15, S.353f) und damit auch die Operatoren $A_u^*A_u$ und $A_uA_u^*$ kompakt sind, denn:

4.18 Satz

Seien X, Y, Z Banachräume. Sind $T \in L(X,Y)$ und $S \in L(Y,Z)$ und ist T oder S kompakt, so ist ST kompakt.

Beweis:
Ein Operator $T \in L(X,Y)$ ist genau dann kompakt, wenn für jede beschränkte Folge $(x_n)_{\mathbb{N}} \subset X$, die Folge $(T(x_n))_{\mathbb{N}} \subset Y$ eine konvergente Teilfolge enthält. Sei also $(x_n)_{\mathbb{N}}$ eine beschränkte Folge, sei ferner zunächst S kompakt. Da $T \in L(X,Y)$ ist auch $(Tx_n)_{\mathbb{N}}$ beschränkt und $(STx_n)_{\mathbb{N}}$ besitzt eine konvergente Teilfolge. Ist S stetig, T kompakt und damit $(Tx_n)_{\mathbb{N}}$ konvergent,so ist auch $(STx_n)_{\mathbb{N}}$ konvergent. □

4.19 Satz (Hilbert-Schmidt Theorem)

Sei X reeller, seperabler HR und $A \in K(X)$ selbstadjungiert. Dann existiert eine vollständige Orthonormalbasis $\{\psi_k\}_{k=1}^{\infty}$ von X, so dass

$$A\psi_k = \lambda_k \psi_k$$

und $\lambda_k \to 0, k \to \infty$.

Beweis: Werner, Theorem VI 3.2, S. 209f. □

Da $A_u^* A$ (und analog $A_u A_u^*$) nicht negativ, weil $(A_u^* A_u v)_X \geq 0 \ \forall v \in L^2((0,1))$, erhalten wir das folgende Resultat:

4.20 Satz

Sei A_u wie in 4.17 (b) definiert. Der Operator $R_u := A_u A_u^*$ mit

$$R_u v = \int_0^1 (v, u(\mu))_{L^2((0,1))} u(\mu) \mathrm{d}\mu$$

ist nicht negativ, selbstadjungiert und kompakt. Ferner existiert eine vollständige Orthonormalbasis $\{\phi_k\}_{k=1}^{\infty}$ für $L^2((0,1))$ und eine Folge $\{\lambda_k\}_{k=1}^{\infty}$ nicht-negativer reeller Zahlen, so dass

$$R_u \phi_k = \lambda_k \phi_k, \ \lambda_1 \geq \lambda_2 \geq \cdots \geq 0$$

und $\lambda_k \to 0$ für $k \to \infty$.

Beweis:
Folgt aus Satz 4.19. □

Definiere nun

$$\psi_k(\mu) := (A_u^* \phi_k)(\mu) = \int_0^1 u(x,\mu)\phi_k(x)\mathrm{d}x,$$

mit ϕ_k aus Satz 4.20. Es lässt sich zeigen, dass $\{\phi_k\}_{k=1}^{\infty}$ Eigenfunktionen von $A_u^* A_u$. Die sogenannte 'Hilbert-Schmidt Zerlegung' von $u(x,\mu)$ ist dann gegeben durch:

$$u(x,\mu) = \sum_{k=1}^{\infty} \psi_k(\mu)\phi_k(x) \text{ fast überall.}$$

Schmidt zeigte weiterhin, dass

$$\min \left\{ \int_0^1 \int_0^1 \left| u(x,\mu) - \sum_{n=1}^{N} U_n^N(\mu)\phi_n(x) \right|^2 \mathrm{d}x\mathrm{d}\mu \mid U_n^N(\mu), \phi_n(x) \in L^2((0,1)) \right\}$$
$$= \int_0^1 \int_0^1 \left| u(x,\mu) - \sum_{n=1}^{N} \psi_n(\mu)\phi_n(x) \right|^2 \mathrm{d}x\mathrm{d}\mu = \sum_{k=N+1}^{\infty} \lambda_k.$$

4.2.2 Kontinuierliche POD

4.21 Satz (POD Basis)

Sei X ein HR mit $H_0^1(\Omega) \subset X \subset H^1(\Omega), P = [a_i, b_i]^p, i = 1, \ldots, p$, wir definieren den Operator

$$Rv := \int_P (v, u(\mu))_X u(\mu)\mathrm{d}\mu \text{ für } v \in X, u(\mu) \in X.$$

Ferner nehmen wir an, dass die betrachtete parameterabhängige Differentialgleichung erlaubt zu zeigen, dass R kompakt (dazu reicht z.B. dass $u(\mu)$ stetig von μ abhängt). Dann existiert eine vollständige Orthonormalbasis $\{\varphi_i\}_{i=1}^{\infty}$ von X und eine Folge $\{\lambda_i\}_{i=1}^{\infty}$ nicht-negativer reeller Zahlen, so dass

$$R\varphi_i = \lambda_i\varphi_i, \lambda_1 \geq \lambda_2 \geq \cdots \geq 0 \text{ und } \lambda_i \xrightarrow{i\to\infty} 0. \tag{4.8}$$

Sei N', so dass $\lambda_{N'} > 0$. Für $1 \leq N \leq N'$ definieren wir $\Phi_N := \{\varphi_1, \ldots, \varphi_N\}$ als POD-Basis und $X_N := \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$ als POD-Raum.

Beweis:
Wie oben zeigt man, dass R nicht-negativ und selbstadjungiert. Aufgrund der angenommenen Kompaktheit folgt die Aussage, dann aus dem Hilbert-Schmidt Theorem. □

Es lässt sich nun zeigen, dass der POD-Raum den mittleren Projektionsfehler minimiert, also Minimierer des Minimierungsproblems (4.2) ist.

4.22 Satz (Optimalität der POD-Basis)

Für $l \in \mathbb{N}$ definieren wir die Abbildung $J : \underbrace{X \times \cdots \times X}_{l-\text{mal}} \to \mathbb{R}$ durch

$$J(\phi_1, \ldots, \phi_l) = \int_P \left\| u(\mu) - \sum_{i=1}^{l} (u(\mu), \phi_i)_X \phi_i \right\|_X^2 \mathrm{d}\mu. \tag{4.9}$$

Seien $\{\lambda_i\}_{i=1}^{\infty}$ und $\{\phi_i\}_{i=1}^{\infty}$ die Eigenwerte und Eigenfunktionen von R. Dann lösen für alle $l \in \mathbb{N}$ die ersten L Eigenfunktionen $\phi_1, \ldots, \phi_L \in X$ das Minimierungsproblem

$$\min J(\psi_1, \ldots, \psi_l) \text{ so dass } (\psi_j, \psi_i) = \delta_{ij},\ 1 \leq i, j \leq l. \tag{4.10}$$

Ferner gilt:

$$J(\phi_1, \ldots, \phi_l) = \sum_{i=l+1}^{\infty} \lambda_i \text{ für jedes } l \in \mathbb{N}. \tag{4.11}$$

Beweis:
Der Beweis zu Teil 1 der Aussage beruht darauf, dass das Eigenwert-Problem (4.8) die notwendige Optimalitätsbedingung erster Ordnung (nOb1) für (4.10) ist. Den Beweis findet man zum Beispiel in [Holmes,L,B, 1996]. Für (4.11): Da $\{\phi_i\}_{i=1}^{\infty}$ Orthonormalbasis (ONB) von X gilt

$$u(\mu) = \sum_{i=1}^{\infty} (u(\mu), \phi_i)_X \phi_i. \qquad \text{(s. Alt, Satz 7.7)}$$

Damit folgt:

$$\begin{aligned}
\int_P \left\| u(\mu) - \sum_{i=1}^{\infty} (u(\mu), \phi_i)_X \phi_i \right\|_X^2 \mathrm{d}\mu &= \int_P \left\| \sum_{i=1}^{\infty} (u(\mu), \phi_i)\phi_i - \sum_{i=1}^{\infty} (u(\mu), \phi_i)_X \phi_i \right\|_X^2 \mathrm{d}\mu \\
&= \int_P \left\| \sum_{i=l+1}^{\infty} (u(\mu), \phi_i)_X \phi_i \right\|_X^2 \mathrm{d}\mu \\
&= \int_P \sum_{i,j=l+1}^{\infty} (u(\mu), \phi_i)_X (u(\mu), \phi_j)_X \underbrace{(\phi_i, \phi_j)_X}_{\text{ONB, } =\delta_{ij}} \mathrm{d}\mu \\
&= \int_P \sum_{i=l+1}^{\infty} (u(\mu), \phi_i)_X^2 \mathrm{d}\mu \\
&= \int_P \sum_{i=l+1}^{\infty} ((u(\mu), \phi_i)_X u(\mu), \phi_i)_X \mathrm{d}\mu = \sum_{i=l+1}^{\infty} (R\phi_i, \phi_i)_X \\
&= \sum_{i=l+1}^{\infty} (\lambda_i \phi_i, \phi_i)_X = \sum_{i=l+1}^{\infty} \lambda_i.
\end{aligned}$$

□

4.23 Folgerung

Seien $\{\lambda_i\}_{i=1}^{\infty}$ und $\{\phi_i\}_{i=1}^{\infty}$ die Eigenwerte und Eigenfunktionen von R. Dann gilt:

(1) $(R\phi_i, \phi_i)_X = \int_P (u(\mu), \phi_i)_X (u(\mu), \phi_j)_X \mathrm{d}\mu = \delta_{ij}\lambda_i$ für alle $i, j \in \mathbb{N}$, dass heißt die POD-Koeffizienten sind unkorreliert.

(2) Sei $\{X_i\}_{i=1}^{\infty}$ beliebige Orthonormalbasis von X. Für jedes $l \in \mathbb{N}$ gilt

$$\sum_{i=1}^{l} \int_P |(u(\mu), \phi_i)_X|^2 \mathrm{d}\mu \geq \sum_{i=1}^{l} \int_P |(u(\mu), X_i)_X|^2 \mathrm{d}\mu.$$

Daher erfassen die ersten l POD-Basisfunktionen mehr Energie als die ersten l Funktionen jeder anderen Basis von X.

Beweis:

(1) Zunächst gilt:

$$(R\phi_i, \phi_j)_X = \left(\int_P (u(\mu), \phi_i)_X u(\mu) \mathrm{d}\mu, \phi_j \right)_X = \int_P (u(\mu), \phi_i)_X (u(\mu), \phi_j)_X \mathrm{d}\mu.$$

Außerdem gilt:

$$(R\phi_i, \phi_j)_X = (\lambda_i \phi_i, \phi_j)_X = \lambda_i \delta_{ij}.$$

(2) siehe [Holmes,LB,1996].

4.2.3 POD im Diskreten Setting

4.24 Satz (POD Basis)

Sei X_N HR von Dimension N und $\{u_i^N\}_{i=1}^n \subset X_N$. Dann definieren wir den empirischen Korrelationsoperator

$$R_N : X_N \to X_N \text{ durch } R_N u^N := \frac{1}{n}\sum_{i=1}^{n}(u_i, v)_{X_N} u_i,\ v \in X_N.$$

Es ist $R_N \in K(X_N)$ und es existieren orthonormale Funktionen $\{\phi_i\}_{i=1}^{n'}, n' \leq n$ und reelle Zahlen $\{\lambda_i\}_{i=1}^{n'}$ mit $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{n'} > 0$, so dass

$$R_N \phi_i = \lambda_i \phi_i,\ i = 1, \ldots, n'.$$

Für $1 \leq M \leq n'$ definieren wir die POD Basis $\Phi_M := \{\phi_1, \ldots, \phi_M\}$ und den POD-Raum $X_M := \operatorname{span}\{\phi_1, \ldots, \phi_M\}$.

Beweis:
Analog zu oben zeigt man, dass R_N linear, beschränkt, selbstadjungiert und nicht-negativ. Da R_N endlich dimensionales Bild hat, ist R_N ferner kompakt. Damit folgt aus dem Hilbert-Schmidt Theorem die Behauptung, wobei das Orthonormalsystem aus den ϕ_i nicht unendlich sein kann, da R_N endliches Bild hat.

4.25 Beispiel (POD-Basis zur Modellreduktion parameter abhängiger PDgl)

Betrachte dazu den HR X_n, die Snapshotmenge $M_{\text{train}} = \{u_h/\mu_i \mid \mu_i \in S_{\text{train}}\}$ und $R_n v = \frac{1}{n_{\text{train}}}\sum_{i=1}^{n_{\text{train}}}(u_h(\mu_i), v)_{\mathbb{R}} u_h(\mu_i)$. Die POD-Basis wird dann über das entsprechende Eigenwertproblem wie in Satz 4.24 bestimmt.

Weiß nicht ob das korrekt ist.

4.26 Beispiel (POD in der statistischen Datenanalyse)

Die Projektion auf den POD-Raum wird in der statistischen Datenanalyse auch Hotoling-Transformation, Prinicpal Component Analysis (PCA) oder Karkunev-Loewe Transformation genannt. Hierbei betrachtet man einen Datensatz,welcher als Menge von n Punkten in N-dimensionalen Raum veranschaulicht werden kann, dass heißt $X_N = \mathbb{R}^N$ und

$$R_N v = \frac{1}{n}\sum_{j=1}^{n}\left|(u_j^N, v)_{\mathbb{R}^N}\right| u_j^N \text{ für } v \in \mathbb{R}^N.$$

Keine Ahnung ob die Namen so geschrieben werden..

4.27 Bemerkung (Eigenschaften der POD-Basis)

(1) $\{\phi_i\}_{i=1}^N$ ist orthonormale Basis, aber nicht eindeutig.

(2) Die POD-Basen sind hierachisch, dass heißt $X_{N'} \subset X_N$ für $N' \leq N$.

(3) Die POD hängt nicht von der Reihenfolge der Daten ab (im Gegensatz zu einer Basis welche aus einer Gram-Schmidt Orthogonalisierung hervorgeht).

(4) Maximierung der Varianz: ϕ_1 ist die Richtung höchster Varianz von $\{u_i^N\}_{i=1}^n$, ϕ_2 ist die Richtung höchster Varianz von $\{P_{X_1^c} u_i^N\}_{i=1}^n$, usw.

(5) Die Koordinaten der Daten bzgl. der POD-Basis sind unkorreliert. Wir wollen nun exemplarisch für $X_N = \mathbb{R}^N$ die Optimalität der POD-Basis nachweisen, dass heißt wir wollen zeigen, dass die POD-Basis das folgende nichtlineare Optimierungsproblem löst:

$$\min_{\psi_1, \ldots, \psi_l \in \mathbb{R}^N} \frac{1}{n}\sum_{j=1}^{n}\left\| u_j^N - \sum_{i=1}^{l}(u_j^N, \psi_i)_{\mathbb{R}^N}\psi_i \right\|_{\mathbb{R}^N}^2, \tag{4.12}$$

so dass $(\psi_i, \psi_j)_{\mathbb{R}^N} = \delta_{ij}$ für $1 \leq i, j \leq l$. Dazu benötigen wir einige Definitionen und Resultate aus der nichtlinearen Optimierung. Problem (4.12) ist ein Optimierungsproblem der folgenden allgemeinen Form:

$$\min J(x), \text{ so dass } g(x) = 0, \tag{Op}$$

wobei $J : \mathbb{R}^N \to \mathbb{R}$ das Zielfunktional und $g : \mathbb{R}^k \to \mathbb{R}^m,\ m \leq k$ der Nebenbedingung.

4.28 Definition (zulässige Lösung)

Ein Punkt $x \in \mathbb{R}^n$ heißt zulässig, falls $g(x) = 0$ gilt. Die Menge zulässiger Lösungen ist definiert als

$$E(Op) = \{x \in \mathbb{R}^n \mid g(x) = 0\}.$$

4.29 Definition (lokale Lösung)

Der Punkt $\bar{x}$ heißt **lokale Lösung** von (Op), falls $\bar{x} \in E(Op)$ gilt und $J(\bar{x}) \leq J(x)\ \forall U(\bar{x}) \cap E(Op)$, wobei $U(\bar{x}) \subset \mathbb{R}^N$ offene nichtleere Umgebung von $\bar{x}$.

4.30 Definition (regulärer Punkt)

Ein Punkt $\bar{x} \in E(Op)$ heißt **regulärer Punkt** bzgl. der Nebenbedingung $g(x) = 0$, falls die Gradienten $\{\nabla g(\bar{x})\}_{i=1}^n \subset \mathbb{R}^N$ linear unabhängig sind.

4.31 Satz (Notwendige Optimalitätsbedingung 1. Ordnung)

Seien J und g stetig differenzierbar. Sei ferner $\bar{x}$ eine lokale Lösung von (Op) und ein regulärer Punkt für $g(x) = 0$. Dann existiert ein eindeutiger Lagrange-Multiplikator $\bar{\lambda} = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$, welcher

$$\nabla J(\bar{x}) + \sum_{i=1}^{m} \lambda_i \nabla g(\bar{x}) = \nabla J(\bar{x}) + \nabla g(\bar{x})^T \bar{\lambda} = 0 \tag{4.13}$$

löst.

Beweis: [Nocedal, Wright,2006] □

Nun haben wir alle Hilfsmittel zusammen um den folgenden Satz zu zeigen:

4.32 Satz

Sei $U = [u_1^N, \ldots, u_n^N] \in \mathbb{R}^N$ und $\Phi_N = \{\phi_1, \ldots, \phi_N\}$ POD-Basis definiert in 4.24. Dann löst die POD-Basis das folgende Optimierungsproblem:

$$\max_{\psi_1,\ldots,\psi_l \in \mathbb{R}^N} \frac{1}{n} \sum_{i=1}^{l} \sum_{j=1}^{n} \left|(u_j^N, \psi_i)_{\mathbb{R}^N}\right|^2 \text{ s.d. } (\psi_i, \psi_j)_{\mathbb{R}^N} = \delta_{ij}\ \forall 1 \leq i, j \leq l. \tag{Op^l}$$

Ferner gilt

$$\operatorname{argmax}(Op^l) = \sum_{i=1}^{l} \lambda_i,$$

wobei λ_i die Eigenwerte aus Satz 4.24 sind.

Beweis:

Zunächst betrachten wir das Optimierungsproblem für $l = 1$

$$\max_{\psi \in \mathbb{R}^N} \frac{1}{n} \sum_{j=1}^{n} \left|(u_j^N, \psi)_{\mathbb{R}^N}\right|^2 \text{ s.d. } \|\psi\|_{\mathbb{R}^N}^2 = 1. \tag{Op^1}$$

Um die Lösung von (Op^1) zu bestimmen, betrachten wir die notwendige Optimalitätsbedingung 1. Ordnung (nOB1). Dazu führen wir zunächst die Funktion $g : \mathbb{R}^N \to \mathbb{R}$ ein, welche durch $g(\psi) = 1 - \|\psi\|^2_{\mathbb{R}^N}$ für $\psi \in \mathbb{R}^N$ definiert ist. Die Nebenbedingung in (Op^1) kann dann als $g(\Psi) = 0$ formuliert werden:

$$\nabla g(\psi) = 2\psi^T \text{ ist linear unabhängig, falls } \psi \neq 0.$$

Aufgrund der Nebenbedingung $\|\psi\|^2_{\mathbb{R}^N} = 1$ gilt für eine Lösung von (Op^1), dass $\psi \neq 0$. Damit ist jede Lösung von (Op^1) ein regulärer Punkt. Sei $\mathcal{L} : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ das Lagrange-Funktional zu (Op^1), das heißt

$$\mathcal{L}(\psi, \lambda) := \frac{1}{n} \sum_{j=1}^{n} \left|(u_j^N, \psi)_{\mathbb{R}^N}\right|^2 + \lambda(1 - \|\psi\|^2_{\mathbb{R}^N}),$$

für $(\psi, \lambda) \in \mathbb{R}^N \times \mathbb{R}$. Angenommen $\psi \in \mathbb{R}^N$ ist eine Lösung von (Op^1). Da ψ regulärer Punkt existiert nach Satz 4.31 ein eindeutiger Lagrange-Multiplikator $\lambda \in \mathbb{R}$, welcher die nOB1 erfüllt:

$$\nabla \mathcal{L}(\psi, \lambda) \overset{!}{=} 0 \text{ in } \mathbb{R}^N \times \mathbb{R}.$$

Wir berechnen den Gradienten von $\mathcal{L}$ bezüglich ψ:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \psi_i}(\psi, \lambda) &= \frac{\partial}{\partial \psi_i} \left(\frac{1}{n} \sum_{j=1}^{n} \left| \sum_{k=1}^{N} (u_{kj}\psi_k) \right|^2 + \lambda \left(1 - \sum_{k=1}^{N} \psi_k^2 \right) \right) \\
&= 2 \cdot \frac{1}{n} \sum_{j=1}^{n} \left(\sum_{k=1}^{N} u_{kj}\psi_k \right) u_{ij} - 2\lambda\psi_i \\
&= \frac{2}{n} \sum_{n=1}^{N} \left(\underbrace{\sum_{j=1}^{n} u_{ij} u_{jk}^T \psi_k}_{=(uu^T)_{ik}} \right) - 2\lambda\psi_i.
\end{aligned}$$

Daher folgt:

$$\nabla_\psi \mathcal{L}(\psi, \lambda) = 2 \left(\frac{1}{n} uu^T \psi - \lambda\psi \right) = 2(R_N\psi - \lambda\psi) \overset{!}{=} 0 \text{ in } \mathbb{R}^N \times \mathbb{R}. \tag{4.14}$$

Gleichung (4.14) ergibt das folgende Eigenwertproblem

$$\frac{1}{n} uu^T \psi = \lambda\psi \text{ in } \mathbb{R}^N, \tag{4.15}$$

welches gerade mit dem Eigenwertproblem aus Satz 4.24 übereinstimmt. Aus $\frac{\partial \mathcal{L}}{\partial \lambda}(\mathcal{L}, \lambda) \overset{!}{=} 0$ in $\mathbb{R}$ schließen wir die Nebenbedingung

$$\|\psi\| = 1. \tag{4.15b}$$

Der erste POD-Vektor ϕ, definiert in Satz 4.24 löst damit (4.15). Ferner gilt:

$$\begin{aligned}
\frac{1}{n} \sum_{j=1}^{n} \left|(u_j^N, \phi_1)_{\mathbb{R}^N}\right|^2 &= \left(\frac{1}{n} \sum_{j=1}^{n} (u_j^N, \phi_1)_{\mathbb{R}^N} u_j^N, \phi_1 \right)_{\mathbb{R}^N} \\
&= (R_N\phi_1, \phi_1) = \lambda_1 \|\phi_1\|^2_{\mathbb{R}^N} = \lambda_1.
\end{aligned}$$

Schließlich zeigen wir, dass ϕ_1 (Op^1) löst. Sei dazu $\xi \in R^N$ ein beliebiger Vektor mit $\|\xi\|_{\mathbb{R}^N} = 1$. Da die Eigenvektoren $\{\phi_i\}_{i=1}^N$ von (4.15) eine ONB für $\mathbb{R}^N$ sind, gilt

$$\xi = \sum_{i=1}^{N} (\xi, \phi_i)_{\mathbb{R}^N} \phi_i.$$

Es folgt:

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n} \left|(u_j^N, \xi)_{\mathbb{R}^N}\right|^2 &= \frac{1}{n}\sum_{j=1}^{n} \left| \left(u_j^N, \sum_{i=1}^{N} (\xi, \phi_i)_{\mathbb{R}^N} \phi_i \right)_{\mathbb{R}^N} \right|^2 \\
&= \frac{1}{n}\sum_{j=1}^{n} \sum_{i,k=1}^{N} (u_j^N, (\xi, \phi_i)\phi_i)(u_j^N, (\xi, \phi_k)\phi_k) \\
&= \frac{1}{n}\sum_{j=1}^{n} \sum_{i,k=1}^{N} \Big((u_j^N, \phi_i)(u_j^N, \phi_k)(\xi, \phi_i)(\xi, \phi_k) \Big) \\
&= \sum_{i,k=1}^{N} \left(\left(\underbrace{\frac{1}{n}\sum_{j=1}^{n} (u_j^N, \phi_i) u_j^N}_{=R_N \phi_i}, \phi_k \right) (\xi, \phi_i)(\xi, \phi_k) \right) \\
&= \sum_{i,k=1}^{N} \underbrace{(\lambda_i \phi_i, \phi_k)}_{=\lambda_i \delta_{ik}} (\xi, \phi_k)(\xi, \phi_k) \\
&= \sum_{i=1}^{N} \lambda_i \left|(\xi, \phi_i)\right|^2 \\
&\leq \lambda_1 \sum_{i=1}^{N} \left|(\xi, \phi_i)\right|^2 = \lambda_1 \left\|\xi\right\|^2 = \lambda_1 \\
&= \frac{1}{n}\sum_{j=1}^{n} \left|(u_j^N, \phi_1)_{\mathbb{R}^N}\right|^2.
\end{aligned}$$

Damit löst ϕ_1 (Op^1) und $\operatorname{argmax}(Op^1) = \lambda$.
Suchen wir nach einem zweiten Vektor, welcher orthogonal auf ϕ_1 steht, normiert ist und die Daten $\{u_i^N\}_{i=1}^n$ so gut wie möglich beschreibt, so müssen wir das folgende Optimierungsproblem lösen:

$$\max_{\psi \in \mathbb{R}^N} \frac{1}{n}\sum_{j=1}^{n} \left|(u_j^N, \psi)\right|^2, \text{ so dass } \|\psi\| = 1 \text{ und } (\psi, \phi_1) = 0,$$

wobei $\|.\| = \|.\|_{\mathbb{R}^N}$ und $(.,.) = (.,.)_{\mathbb{R}^N}$. Die Definition der POD-Basis zeigt, dass ϕ_2 eine Lösung von (Op^2) ist und $\operatorname{argmax}(Op^2) = \lambda_2$ gilt. Im Detail löst ϕ_2 die nOB1 (4.15) und für

$$\tilde{\psi} \sum_{i=2}^{N} (\tilde{\psi}, \phi_i)\phi_i \in \operatorname{span}\{\phi_1\}^{\perp}$$

gilt:

$$\frac{1}{n}\sum_{j=1}^{n} \left|(u_j^N, \tilde{\psi})\right|^2 \leq \lambda_2 = \frac{1}{n}\sum_{j=1}^{n} \left|(u_J^N, \phi_2)\right|^2.$$

Schließlich zeigen wir mittels vollständiger Induktion, dass die in 4.24 definierte POD-Basis das Optimierungsproblem (Op^l) löst. Für allgemeine l betrachten wir dazu das Lagrange-Funktional: $\mathcal{L} : \underbrace{\mathbb{R}^N \times \cdots \times \mathbb{R}^N}_{l-\text{mal}} \times \mathbb{R}^l \times \mathbb{R}^l \to \mathbb{R}$ definiert durch

$$\mathcal{L}(\psi_1, \ldots, \psi_l, \Lambda) = \frac{1}{n} \sum_{i=1}^{l} \sum_{j=1}^{n} \left|(u_j^N, \psi_i)\right|^2 + \sum_{i,j=1}^{l} \lambda_{ij}(\delta_{ij} - (\psi_i, \psi_j)),$$

für $\psi_1, \ldots, \psi_l \in \mathbb{R}^N$ und $\Lambda = ((\lambda_{ij})) \in \mathbb{R}^{l \times l}$. Die nOB1 für (Op^l) lauten dann

$$\frac{\partial \mathcal{L}}{\partial \psi_k}(\psi_1, \ldots, \psi_l, \Lambda)\delta\psi_k = 0 \ \forall \delta\psi_k \in \mathbb{R}^N, \ k \in \{1, \ldots, l\}. \tag{4.16}$$

Aus

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \psi_k}(\psi_1, \ldots, \psi_l, \Lambda)\delta\psi_k &= 2 \cdot \frac{1}{n} \sum_{j=1}^{n} (u_j^N, \psi_k)(u_j^N, \delta\psi_k) - \sum_{i=1}^{l} \delta_{ik}(\psi_i, \psi_k) - \sum_{i=1}^{l} \lambda_{ki}(\delta\psi_k, \psi_i) \\
&= 2 \cdot \frac{1}{n} \sum_{j=1}^{n} (u_j^N, \psi_k)(u_j^N, \delta\psi_k) - \sum_{i=1}^{l} (\lambda_{ik} + \lambda_{ki})(\psi_i, \delta\psi_k) \\
&= \left(2 \cdot \frac{1}{n} \sum_{j=1}^{n} (u_j^N, \psi_k) u_j^N - \sum_{i=1}^{l} (\lambda_{ik} + \lambda_{ki})\psi_i, \delta\psi_k \right)
\end{aligned}$$

und (4.16) schließen wir

$$\frac{1}{n} \sum_{j=1}^{n} (u_j^N, \psi_k) u_j^N = \frac{1}{2} \sum_{i=1}^{l} (\lambda_{ik} + \lambda_{ki})\psi_i \text{ in } \mathbb{R}^N \ \forall k \in \{1, \ldots, l\}, \tag{4.17}$$

oder bzw.

$$\frac{1}{n} U U^T \psi_k = \frac{1}{2} \sum_{i=1}^{l} (\lambda_{ik} + \lambda_{ki})\psi_i \text{ in } \mathbb{R}^N \forall k \in \{1, \ldots, l\}. \tag{4.18}$$

Wir zeigen nun mittels vollständiger Induktion, dass das Eigenwert-Problem aus 4.24 gerade den nOB1 für (Op^l) entspricht. Für $l = 1$ gilt: $k = 1$. Dann folgt aus (4.17):

$$\frac{1}{n} \sum_{j=1}^{n} (u_j^N, \psi_1) u_j^N = \lambda_1 \psi_1 \text{ in } \mathbb{R}^N \text{ mit } \lambda_1 = \lambda_{11}.$$

Wir nehmen nun an, dass für $l \geq 1$ die nOB1 durch

$$\frac{1}{n} \sum_{j=1}^{n} (u_j^N, \psi_k) u_j^N = \lambda_k \psi_k \text{ in } \mathbb{R}^N \ \forall k \in \{1, \ldots, l\} \tag{4.19}$$

gegeben sind. Wir wollen nun zeigen, dass die nOB1 für $l + 1$ durch

$$\frac{1}{n} \sum_{j=1}^{n} (u_j^N, \psi_k) u_j^N = \lambda_k \psi_k \text{ in } \mathbb{R}^N \ \forall k \in \{1, \ldots, l+1\} \tag{4.20}$$

gegeben sind. Aus (4.17) schließen wir

$$\frac{1}{n}\sum_{j=1}^{n}(u_j^N,\psi_{l+1})u_j^N=\frac{1}{2}\sum_{i=1}^{l+1}(\lambda_{i,l+1}+\lambda_{l+1,i})\psi_i \text{ in } \mathbb{R}^N. \tag{4.21}$$

Da $\{\psi_i\}_{i=1}^{l+1}$ POD-Basis (orthonormal als Lösung von (4.19)) und da R_N selbstadjungiert folgt

$$\begin{aligned}0&=\lambda_j(\psi_{l+1},\psi_j)=(\psi_{l+1},R_N\psi_j)=(R_N\psi_{l+1},\psi_j)\\&=\frac{1}{2}\sum_{i=1}^{l+1}(\lambda_{i,l+1}+\lambda_{l+1,i})(\psi_i,\psi_j)=\frac{1}{2}(\lambda_{j,l+1}+\lambda_{l+1,j}).\end{aligned}$$

Damit folgt

$$\lambda_{i,l+1}=\lambda_{l+1,i} \text{ für jedes } i\in\{1,\dots,l\}\,. \tag{4.22}$$

Einsetzen von (4.22) in (4.21) ergibt

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n}(u_j^N,\psi_{l+1})u_j^N&=\frac{1}{2}\sum_{i=1}^{l}\underbrace{(\lambda_{i,l+1}+\lambda_{l+1,i})}_{=0}\psi_i+\lambda_{l+1,l+1}\psi_{l+1}\\&=\lambda_{l+1,l+1}\psi_{l+1}.\end{aligned}$$

Mit $\lambda_{l+1}=\lambda_{l+1,l+1}$ folgt dann (4.20). Zusammengefasst sind die nOB1 für (Op^l) durch das Eigenwert-Problem

$$R_N\psi_i=\lambda_i\psi_i \text{ für } i=1,\dots,l \tag{4.23}$$

gegeben. Die in 4.24 definierte POD-Basis $\Phi_l=\{\phi_1,\dots,\phi_l\}$ löst (4.23). Der Beweis das $\{\phi_i\}_{i=1}^{l}$ eine Lösung von (Op^l) ist und das $\operatorname{argmax}(Op^l)=\sum_{i=1}^{l}\lambda_i$ mit den Eigenwerten aus Satz 4.24 gilt, folgt analog zum Beweis für (Op^1). □

4.33 Folgerung (Optimalität der OPD-Basis)

Seien die Voraussetzungen von Satz 4.32 erfüllt und gilt $\operatorname{rg}(U)=d$, wobei $U=[u_1^N,\dots,u_n^N]$. Sei $\hat{\Psi}_d=[\psi_1,\dots,\psi_d]\in\mathbb{R}^{N,d}$ Matrix mit paarweise orthonormalen Spalten und die Darstellung der Spalten von U in der Basis $\{\psi_i\}_{i=1}^{d}$ sei gegeben durch

$$U=\hat{\Psi}_d\cdot C_d,\ \text{wobei } (C_d)_{ij}:=(\psi_i,u_j^N)\ 1\le i\le d,\ 1\le j\le n.$$

Dann gilt für jedes $l\in\{1,\dots,d\}$:

$$\left\|U-\hat{\Phi}_lB_l\right\|_F\le\left\|U-\hat{Psi}_lC_l\right\|_F, \tag{4.24}$$

wobei $\hat{\Phi}_l=[\phi_1,\dots,\phi_d]$ mit ϕ_i POD-Basis aus 4.24 und $(B_l)_{ij}=(\phi_i,u_j^N)$. Hierbei bezeichnet $\|.\|_F$ in (4.24) die **Frobeniusnorm**, welche durch

$$\|A\|_F=\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{n}|A_ij|^2}=\sqrt{\operatorname{spur}(A^TA)} \text{ für } A\in\mathbb{R}^{N\times n}$$

gegeben ist. Ferner bezeichnet die Matrix $\hat{\Psi}_d$ die $l\le d$ ersten Spalten von $\hat{\Psi}$, B_l die ersten l-Zeilen von B und analog für $\hat{\Phi}_l$ und C_l.

Beweis:
Da die Spalten von $\hat{\Psi}_d$ orthonormal sind, folgt

$$\begin{aligned}\left\|u - \hat{\Psi}_l C_l\right\|_F^2 &= \left\|\hat{\Psi}_d (C_d - C_l^0)\right\|_F^2 \\ &\overset{\Psi^T\Psi=\mathrm{id}}{=} \left\|C_d - C_l^0\right\|_F^2 \\ &= \sum_{i=l+1}^{d}\sum_{j=1}^{n} |(C_l)_{ij}|^2,\end{aligned}$$

wobei $C_l^0 \in \mathbb{R}^{d\times n}$ aus $C \in \mathbb{R}^{d\times n}$ durch Ersetzen der letzten $d-l$ Reihen durch 0 hervor geht. Da wegen

$$\begin{aligned}\frac{1}{n} u u^T v = R_N v &= R_N \left(\sum_{i=1}^{N} (v,\phi_i)\phi_i\right) \\ &= \sum_{i=1}^{N} (v,\phi_i) R_N \phi_i = \sum_{i=1}^{N} \lambda_i (v,\phi_i)\phi_i \\ &= \sum_{i=1}^{d} \lambda_i (v,\phi_i)\phi_i,\end{aligned}$$

folgt, dass $u_j^N \in \mathrm{bild}(R_N) = \mathrm{span}\{\phi_i\}_{i=1}^d$. Daher gilt $U = \hat{\Phi}_d B_d$ und damit

$$\begin{aligned}\frac{1}{n}\left\|U - \hat{\Phi}_l B_l\right\|_F^2 &= \frac{1}{n}\left\|\hat{\Phi}_D (B_d - B_L^0)\right\|_F^2 = \left\|B_d - B_L^0\right\|_F^2 \\ &= \frac{1}{n}\sum_{i=l+1}^{d}\sum_{j=1}^{n} |(B_d)_{ij}|^2 = \frac{1}{n}\sum_{i=l+1}^{d}\sum_{j=1}^{n} \left|(u_j^N,\phi_i)\right|^2 \\ &= \frac{1}{n}\sum_{i=l+1}^{d}\sum_{j=1}^{n} \left((u_j^N,\phi_i)u_j^N,\phi_i\right) \\ &= \sum_{i=l+1}^{d} (R_N\phi_i,\phi_i) = \sum_{i=l+1}^{d}\lambda_i.\end{aligned} \tag{4.25}$$

Damit gilt:

$$\frac{1}{n}\|U\|_F^2 = \frac{1}{n}\left\|\hat{\Psi}_d C_d\right\|_F^2 = \frac{1}{n}\|C_d\|_F^2 = \frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n}|(C_d)_{ij}|^2$$

und

$$\frac{1}{n}\|U\|_F^2 = \frac{1}{n}\left\|\hat{\Phi}_d B_d\right\|_F^2 = \frac{1}{n}\|B_d\|_F^2 = \frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n}|(B_d)_{ij}|^2 = \sum_{i=1}^{d}\lambda_i.$$

Da die Vektoren $\phi_1, \dots, \phi_l$ aus Satz 4.32 des Optimierungsproblems (Op^l) lösen, folgt:

$$\begin{aligned}\frac{1}{n}\left\|U - \hat{\Phi}_l B_l\right\|_F^2 = \sum_{i=l+1}^{d} \lambda_i = \sum_{i=1}^{d}\lambda_i - \sum_{i=1}^{l}\lambda_i &\overset{(4.13)}{=} \frac{1}{n}\|U\|_F^2 - \frac{1}{n}\sum_{i=1}^{l}\sum_{j=1}^{n}\left|(u_j^N, \phi_i)\right|^2 \\ &\overset{\phi_i \text{ lösen } (Op^l)}{\leq} \frac{1}{n}\|U\|_F^2 - \frac{1}{n}\sum_{i=1}^{l}\sum_{j=1}^{n}\left|(u_j^N, \psi_i)\right|^2 \\ &= \frac{1}{n}\sum_{i=1}^{d}\sum_{j=1}^{n}|(C_d)_{ij}|^2 - \frac{1}{n}\sum_{i=1}^{l}\sum_{j=1}^{n}|(C_d)_{ij}|^2 \\ &= \frac{1}{n}\sum_{i=l+1}^{l}\sum_{j=1}^{n}|(C_d)_{ij}|^2 = \frac{1}{n}\left\|U - \hat{\Psi}_l C_l\right\|_F^2.\end{aligned}$$

□

4.34 Bemerkung

Folgerung 4.33 liefert direkt, dass

$$\frac{1}{n}\sum_{j=1}^{n}\left\|u_j^N - \sum_{k=1}^{l}(u_j^N, \phi_k)\phi_k\right\|^2 \leq \frac{1}{n}\sum_{j=1}^{n}\left\|u_j^N - \sum_{k=1}^{l}(u_j^N, \psi_k)\psi_k\right\|^2$$

für jede andere Menge $\{\psi_i\}_{i=1}^l$ paarweise orthonormaler Vektoren. Damit folgt aus Folgerung 4.33, dass die in 4.24 definierte POD-Basis auch das folgende Optimierungsproblem löst:

$$\min_{\psi_1,\dots,\psi_l \in \mathbb{R}^N} \frac{1}{n}\sum_{j=1}^{n}\left\|u_j^N - \sum_{k=1}^{l}(u_j^N, \psi_k)\psi_k\right\|^2, \text{ so dass } (\psi_i, \psi_j) = \delta_{ij} \text{ für } 1 \leq i, j \leq l.$$

Die nOB1 sind identisch mit denen von (Op^l).

4.35 Satz (Berechnung der POD-Basis über die Gram-Matrix)

Sei X_N HR von Dimension N und $\{u_i^N\}_{i=1}^n \subset X_N$. Die Matrix $\mathbb{K} \in \mathbb{R}^{n\times n}$ definiert durch

$$\mathbb{K}_{ij} := (u_i^N, u_j^N) \tag{4.26}$$

heißt **Gram-Matrix**. Dann sind äquivalent:

(1) $\phi \in X_N$ ist Eigenvektor zu R_N aus Satz 4.24 zum Eigenwert $\lambda > 0$ mit Norm1 und einer Darstellung

$$\phi = \sum_{i=1}^{n} a_i u_i^N \text{ mit } a \in \ker(\mathbb{K})^\perp.$$

(2) $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ ist Eigenvektor von $\frac{1}{n}\mathbb{K}$ zu $\lambda > 0$, mit Norm $\frac{1}{\sqrt{n\lambda}}$.

Beweis: [Haasdonk, Satz 5.7, S.53f]. □

4.36 Bemerkung

Die POD kann daher entweder als teures EW-Problem fürR_N in X_N (Komplexität $\mathcal{O}(N^3)$) oder, meist günstiger, als EW-Problem für $\mathbb{K}$ (Komplexität $\mathcal{O}(n^3)$) ermittelt werden. Letzteres wir auch **'Method of snapshots'** genannt.

4.37 Lemma (Berechnung der POD-Basis mittels der Singulärwertzerlegung für $X_N = \mathbb{R}^N$)

Sei $X_N 0\mathbb{R}^N, [u_1^N, \dots, u_n^N] = U \in \mathbb{R}^{N\times n}$ Snapshot-Matrix mit Rang d und $U = \Psi D V^T$ eine Singulärwertzerlegung mit $\Psi \in \mathbb{R}^{N\times d}$ mit orthonormalen Spalten, $D = \operatorname{diag}(\sigma_1, \dots, \sigma_d) \in \mathbb{R}^{d\times d}$ und $V \in \mathbb{R}^{n\times d}$ mit orthonormalen Spalten. Falls $\sigma_1 > \sigma_2 > \dots > 0$ echt fallend, so ist $\Psi = \Phi$ bis auf Vorzeichen.

Beweis:
Sei $\Psi = [\psi_1, \dots, \psi_d]$, ψ_i Eigenvektor von R_N wegen

$$R_N\psi_i = \frac{1}{n}UU^T\psi_i = \frac{1}{n}\Psi D \underbrace{V^TV}_{=\mathrm{id}} D \underbrace{\Psi^T\psi_i}_{e_i} = \frac{1}{n}\Psi DDe_i = \frac{1}{n}\sigma_i^2\psi_i.$$

Die Eigenwerte $\frac{1}{n}\sigma_i^2$ sind monoton fallend, also identisch sortiert wie Spektralzerlegung von R_N, das heißt $\frac{1}{n}\sigma_i^2 = \lambda_i$ und $\psi_i = \phi_i$ oder $\psi_i = -\phi_i$.

5 approximationstheorie

im letzten Kapitel haben wir gesehen, dass der Greedy-Algorithmus quasi-optimale rB-räume erzeugt, indem Sinne, dass die konstruierten Räume entweder zur gleichen oder einer leicht schlechteren Konvergenzrate, wie die optimalen Räume im Sinne von Kolmogorov, führen. Ferner konnten wir zeigen, dass die POD-Räume den Projektionsfehler minimieren und in diesemm Sinne optimal sind.
Es ist zu beachten, dass wir aus der (Quasi-)Optimalität der RB-Räume nicht schließen können, dass die RB-Räume tatsächlich zu einer schnellen Konvergenz der RB-Approximation führen. Tatsächlich kann die Konvergenzrate der optimalen Räume unter gewissen Umständen sehr schlecht sein. In diesem Kapitel wollen wir der Frage nach gehen unter welchen Umständen wir eine schnelle und idealerweise eine exponentielle Konvergenz der RB-Räume erwarten können.

Zum Beispiel, wenn der Rand parametrisiert wird.

5.1 Bemerkung (Von welchen Faktoren hängt die Approximationsgüte der RB-Approximation ab?)

Unter anderem beeinflussen die folgenden Faktoren die Approximationsgüte des RB-Raums:

(1) **Differenzierbarkeit und Regularität** der Lösungsabbildung $\gamma : P \to X,\ \mu \mapsto u(\mu)$ bzgl. der Parameter. Von besonderer Bedeutung ist der Fall,wenn γ analytisch ist.

(2) **Parametrische Komplexität**, welche hier durch Q_b und Q_f, also die Anzahl der Summanden in der affin parametrischen Darstellung von b unf f charakterisiert ist. Die parametrische Komplexität ist insbesondere dann kritisch, wenn Q_b, Q_f sehr groß oder gar unendlich sind. Im letzterem Fall kann eine Approximation mit endlich vielen Summanden bestimmt werden (s. Kapitel 6). Für lineare, elliptische Probleme mit einer regulären Lösungsabbildung führt ein schneller Abfall der μ-abhängigen Koeffizienten $\theta_b(\mu)$ und $\theta_f(\mu)$ oder die Existenz einiger weniger dominanter Terme zu einer schnellen Konvergenz der RB-Approximation.

(3) **Art des betrachteten Problems:** Für $P = 1$für nicht lineare Probleme wie die Navier-Stokes Gleichungen kann es schwierig sein schnell konvergierende RB-Räume zu konstruieren, während dies für elliptische Probleme für moderates P im Allgemeinen gut möglich ist. Probleme mit hochdimensionalen Parameterräumen stellen dagegen selbst für elliptische PDgl's eine Herausforderung dar. Schließlich wird im Allgemeinen die Approximationsgüte von RB-Räumen deutlich schlechter, wenn man von elliptischen oder parabolischen zu rein hyperbolischen Problemen übergeht.

Satz 5.2 (Globale Exponentielle Konvergenz im Spezialfall [Maday, Patera, Turicini, 2002])

Seien $P = [\mu_{\min}, \ldots, \mu_{\max}] \subset \mathbb{R}^+$ mit $\mu_{\min} = \frac{1}{\sqrt{\mu_r}}, \mu_{\max} = \sqrt{\mu_r}, \mu_r \in \mathbb{R}^+$ mit $1 \leq \mu_r < \infty$,

$$b(u, v; \mu) = \mu b^1(u, v) + b^2(u, v) \text{ mit } b^q(u, v) = \int_\Omega \nabla u \nabla v \ \forall u, v \in X, \ 1 \leq q \leq 2$$

und f nicht parametrisch mit $\ln \mu_r > \frac{1}{2e}$ und $N_{krit} := 1 + \lceil 2e \ln \mu_r \rceil$. Zu $N \in \mathbb{N}, N \geq 2$, seien $\mu_{\min} = \mu_1 < \cdots < \mu_N = \mu_{\max}$ gegeben durch

$$\mu_i := e^{\{\ln \mu_{\min} + (i-1)\frac{\ln \mu_r}{N-1}\}}$$

und sind daher logarithmisch äquidistant, dass heißt

$$\ln(\mu_{i+1} - \mu_i) = \frac{\ln(\mu_{\max}) - \ln(\mu_{\min})}{N-1}.$$

Schließlich sei $X_N := \operatorname{span}\{u_h(\mu_i)\}_{i=1}^N$ der zugehörige Lagrange RB-Raum. Dann gilt:

$$\frac{|||u_h(\mu) - u_N(\mu)|||_\mu}{|||u_h(\mu)|||_\mu} \leq e^{\frac{N-1}{N_{krit}-1}} \ \forall \mu \in P, \ N \geq N_{krit}.$$

Beweis: siehe [Patera, Rozza, 2007] □

5.1 Reguläre Lösungsabbildungen

5.3 Satz (Lipschitz-Stetigkeit)

Seien die Bilinearform $b(.,.;\mu)$ und die Linearform $f(.;\mu)$ Lipschitz-stetig bzgl. μ und b glm. koerziv bzgl. μ. Außerdem seien b und f glm. beschränkt bzgl. μ (s. Def 3.3). Dann ist die Lösung $u(\mu)$ von $P(\mu)$ definiert in Definition 3.5 Lipschitz-stetig bzgl. μ, dass heißt es existiert ein $L_u > 0$, so dass:

$$\|u(\mu_1) - u(\mu_2)\|_X \leq L_u \|\mu_1 - \mu_2\|_2 \ \forall \mu_1, \mu_2 \in P. \tag{5.1}$$

Beweis:
Nach Definition 3.5 lösen $u(\mu_1)$ und $u(\mu_2)$

$$b(u(\mu_1), v; \mu_1) = f(v; \mu_1) \ \forall v \in X; b(u(\mu_2), v; \mu_2) = f(v; \mu_2) \ \forall v \in X.$$

Subtraktion der beiden Gleichungen und Addition von Null ergibt:

$$b(u(\mu_1), v; \mu_1) - b(u(\mu_2), v; \mu_1) + b(u(\mu_2), v; \mu_1) - b(u(\mu_2), v; \mu_2) = f(v; \mu_1) - f(v; \mu_2).$$

Wegen der Lipschitz-Stetigkeit von b und f erhalten wir:

$$\begin{aligned} b(u(\mu_1) - u(\mu_2), v; \mu_1) &= f(v; \mu_1) - f(v; \mu_2) - b(u(\mu_2), v; \mu_1) + b(u(\mu_2), v; \mu_2) \\ &\leq L_f \|v\|_X \|\mu_1 - \mu_2\|_2 + L_b \|u(\mu_2)\|_X \|v\|_X \|\mu_1 - \mu_2\|_2 . \end{aligned}$$

Teste mit $v = u(\mu_1) - u(\mu_2)$ ergibt:

$$\alpha(\mu_1) \|u(\mu_1) - u(\mu_2)\|_X \leq L_f \|\mu_1 - \mu_2\|_2 + L_b \|u(\mu_2)\|_X \|\mu_1 - \mu_2\|_2 .$$

Das Ausnutzen der A priori Abschätzung $\|u(\mu_2)\|_X \leq \frac{\|f(.;\mu_2)\|_X}{\alpha(\mu_2)}$ ergibt schließlich:

$$\|u(\mu_1) - u(\mu_2)\|_X \leq \left(\frac{l_f}{\alpha(\mu_1)} + \frac{L_b \|f(.;\mu_2)\|_{X'}}{\alpha(\mu_1)\alpha(\mu_2)} \right) \|\mu_1 - \mu_2\|_2 .$$

Wegen der glm. Beschränktheit und glm. Koerzivität folgt mit $L_u = \frac{L_f}{\alpha_0} + \frac{L_b \gamma_0}{\alpha_0^2}$ die Behauptung. □

5.4 Bemerkung

Die Aussage aus Satz 5.3 gilt analog für $u_h(\mu)$ als Lösung von $(P_h(\mu))$ und wie bereits in Satz 3.25 bemerkt für $u_N(\mu)$ als Lösung von $(P_N(\mu))$.

5.5 Bemerkung

Sind b und f affin parametrisch,so folgt die Lipschitz-Stetigkeit von b und f bzgl. μ aus der Lipschitz-Stetigkeit der Funktionen $\theta_b(\mu)$ und $\theta_f(\mu)$.

5.6 Definition (Fréchet-Ableitung)

Seien X, Y Banachräume und sei $f : X \to Y$ eine Funktion. Dann ist f **Fréchet-differenzierbar** im Punkt $x_0 \in X$ genau dann, wenn eine stetige lineare Abbildung $A : X \to Y$ existiert, so dass

$$f(x_0 + h) - f(x_0) = Ah + o(\|h\|),\ h \to 0.$$

Wenn diese Abbildung existiert, nennen wir sie **Fréchet-Ableitung** von f in x_0 und bezeichnen sie mit $f'(x_0) =: A$.

5.7 Definition (Partielle Ableitung)

Seien X, Y, Z Banachräume. Wir betrachten eine Funktion $f : X \times Y \to Z, (x, y) \mapsto f(x, y)$, wenn wir $y_0 \in Y$ festhalten, ist $F(.) := f(., y_0) : X \to Z$ eine Funktion in einer Variablen $x \in X$ Die partielle Ableitung von f nach x ist dann analog zu den partiellen Ableitungen von Funtionen im $\mathbb{R}^n$ definiert:

$$\frac{\partial f}{\partial x}(x_0, y_0) \equiv F'(x_0).$$

Analog kann man $x_0 \in X$ festhalten und mit Hilfe von $G(.) := f(x_0, .) : Y \to Z$ die partielle Ableitng von f nach y definieren:

$$\frac{\partial f}{\partial y}(x_0, y_0) \equiv G'(y_0).$$

5.8 Bemerkung (Höhere Ableitungen)

Höhere Ableitungen werden durch iterative Anwendung der obigen Definitionen definiert.

5.9 Satz

Seien $b : X \times X \times P \to \mathbb{R}$ und $f : X \times P \to \mathbb{R}$ k-mal stetig differenzierbar nach μ. Seien ferner $b(., .; \mu)$ koerziv und stetig für alle $\mu \in P$ und $f.; \mu$ stetig für alle $\mu \in P$. Dann ist die Lösungsabbildung γ k-mal stetig differenzierbar.

Beweisidee: Lax-Milgram, Satz über implizite Funktionen für Banachräume (siehe z.B. [Garlet, 2013], [Riuzicka, 2004]). □

5.10 Satz exponentielle Konvergenz der RB Approximation für die parametrische, stationäre WLG

Wir betrachten die parametrische, stationäre WLG mit den Voraussetzungen aus Definition 3.7 und den folgenden zusätzlichen Annahmen:

(i) $P = \prod_{i=1}^{p} P_i$ mit $P_i = [\mu_i^{\min}, \ldots, \mu_i^{\max}] \subset \mathbb{R}$.

(ii) $\kappa(x;\mu)$ und $q(x;\mu)$ sind unendlich oft stetig differenzierbar bzgl. μ und es gilt: Für alle $\mu \in P$ und für alle $i = 1, \ldots, p$ existiert ein $0 < \gamma_i < +\infty$, so dass

$$\left\| \frac{1}{\kappa(.;\mu)} \frac{\partial^m \kappa(.;\mu)}{\partial \mu_i^m} \right\|_{L^\infty(\Omega)} \leq \gamma_i^m \cdot m!$$

und

$$\frac{1}{1 + \|q(.;\mu)\|_{L^2(\Omega)}} \left\| \frac{\partial^m q(.;\mu)}{\partial \mu_i^m} \right\|_{L^2(\Omega)} \leq \gamma_i^m \cdot m!.$$

Dann gilt die folgende Abschätzung:

$$d_n(M, X_n) \leq C \cdot \sum_{j=1}^{p} e^{-r_j m_j},$$

mit $r_j, m_j > 0$ und Konstante $C > 0$.

Beweis: [Quarterone, Manzoni, Megri, 2015]. □

5.2 Exponentielle Konvergenz im Falle geringer parametrischer Komplexität

5.11 Satz (Exponentielle Konvergenz im Falle dominanter Terme in der affinen Darstellung der Bilinearform [Lassik, Manzoni, Quarteroni, Rozza, 2013])

Seien $b(u, v; \mu) = \theta_b^1(\mu) b^1(u, v) + \theta_b^2(\mu) b^2(u, v)$ und $f(v;\mu) = \sum_{q=1}^{Q_f} \theta_f^q(\mu) f^q(v)$ und seien $\mathbb{B}_h^q \in \mathbb{R}^{N_h \times N_h}, q = 1, 2,\ \mathbb{F}_h^q \in \mathbb{R}^{N_h}, q01, \ldots, Q_f$ definiert durch

$$(\mathbb{B}_h^q)_{ij} := b^q(\bar{\varphi}_j, \bar{\varphi}_i),\ (\mathbb{F}_h^q)_i := f^q(\bar{\varphi}_i),\ 1 \leq i, j \leq N_h,$$

wobei $\bar{\varphi}_i, 1 \leq i \leq N_h$ die Knotenbasis aus Definition 2.50. Wir nehmen an, dass

(i) $\mathbb{B}_h^1$ ist invertierbar.

(ii) $\rho\left(\frac{\theta_b^2(\mu)}{\theta_b^1(\mu)} (\mathbb{B}_h^1)^{-1} \mathbb{B}_h^2\right) < 1$ für alle $\mu \in P$, wobei ρ der Spektralradius ist.

(iii) $\exists \varepsilon > 0 : \left|\frac{\theta_b^2(\mu)}{\theta_b^1(\mu)}\right| \leq \frac{1-\varepsilon}{\left\|(\mathbb{B}_h^1)^{-1} \mathbb{B}_h^2\right\|_{\mathbb{X}_h, \mathbb{X}_h^{-1}}}$, wobei $(\mathbb{X}_h, \mathbb{X}_h^{-1})$ Matrixnorm, definiert durch

$$\|\mathbb{B}\|_{\mathbb{X}_h, \mathbb{X}_h^{-1}} := \sup_{\mathbb{V} \in \mathbb{R}^{N_h}} \frac{\|\mathbb{B}\mathbb{V}\|_{\mathbb{X}_h^{-1}}}{\|\mathbb{V}\|_{\mathbb{X}_h}},$$

mit $\|\mathbb{V}\|_{\mathbb{X}_h} = \sqrt{V^T \mathbb{X}_h V}$ und die $(\mathbb{X}_h, \mathbb{X}_h^{-1})$-Norm die $L(X_h, X_h^{-1})$-Norm realisiert.

Dann gilt die folgende Abschätzung:

$$d_n(M, X_n) \leq C \cdot e^{-an} \text{ für } C, a > 0.$$

Beweis:
Sei $\mathbb{U}_h(\mu) \in \mathbb{R}^{N_h}$ Lösung des LGS

$$\mathbb{B}_h(\mu)\mathbb{U}_h(\mu) = \mathbb{F}_h(\mu),$$

wobei $\mathbb{B}_h(\mu), \mathbb{F}_h(\mu)$ definiert in Folgerung 3.28. Dann gilt:

$$\mathbb{U}_h(\mu) = \left(I + \frac{\theta_b^2(\mu)}{\theta_b^1(\mu)} (\mathbb{B}_h^1)^{-1} \mathbb{B}_h^2\right)^{-1} (\theta_b^1(\mu) \mathbb{B}_h^1)^{-1} \left(\sum_{q=1}^{Q_f} \theta_f^q(\mu) \mathbb{F}_h^q\right).$$

Ausnutzen von Annahme (ii) und der Neumannreihe: $\sum_{k=0}^{\infty} \mathbb{B}^k = (I - \mathbb{B})^{-1}$ liefert:

$$\mathbb{U}_h(\mu) = \sum_{k=0}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_b^2(\mu))^k \theta_f^q(\mu)}{(\theta_b^1(\mu))^{k+1}} \left((\mathbb{B}_h^1)^{-1} \mathbb{B}_h^2 \right)^k (\mathbb{B}_h^1)^{-1} \mathbb{F}_h^q.$$

Nun führen wir die Basisvektoren

$$\mathbb{A}_{k,q} = \left((\mathbb{B}_h^1)^{-1} \mathbb{B}_h^2 \right)^k (\mathbb{B}_h^1)^{-1} \mathbb{F}_h^q$$

ein und schreiben

$$\mathbb{U}_h(\mu) = \sum_{k=0}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_b^2(\mu))^k \theta_f^q(\mu)}{(\theta_b^1(\mu))^{k+1}} \mathbb{A}_{k,q}. \tag{5.2}$$

Ferner bezeichnen wir mit

$$X_n^{\mathbb{A}} := \operatorname{span} \{ \mathbb{A}_{k,q} \mid k = 0, \dots, m-1,\ q = 1, \dots, Q_f \} \subset \mathbb{R}^{N_h}$$

den n-dimensionalen Unterraum, welcher durch die ersten n Basisvektoren aufgespannt wird, wobei $n = Q_f \cdot m$.
Dann gilt:

$$\begin{aligned}
d_n(M, X_h) &= \inf_{X_n \subset X_h, \dim X_n = n} \sup_{\mu \in P} \inf_{\mathbb{V} \in \mathbb{R}^{N_h}} \|\mathbb{U}_h(\mu) - \mathbb{V}\|_{\mathbb{X}_h} \\
&\leq \sup_{\mu \in P} \inf_{\mathbb{V}_n \in X_n^{\mathbb{A}}} \|\mathbb{U}_h(\mu) - \mathbb{V}_n\|_{\mathbb{X}_h} \\
&\leq \sup_{\mu \in P} \left\| \mathbb{U}_h(\mu) - \sum_{k=0}^{m-1} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_b^2(\mu))^k \theta_f^q(\mu)}{(\theta_b^1(\mu))^{k+1}} \mathbb{A}_{k,q} \right\|_{\mathbb{X}_h} \\
&= \sup_{\mu \in P} \left\| \sum_{k=m}^{\infty} \sum_{q=1}^{Q_f} \frac{(-1)^k (\theta_b^2(\mu))^k \theta_f^q(\mu)}{(\theta_b^1(\mu))^{k+1}} \mathbb{A}_{k,q} \right\|_{\mathbb{X}_h} \\
&\leq \sup_{\mu \in P} \sum_{k=m}^{\infty} \sum_{q=1}^{Q_f} \left| \frac{(\theta_b^2(\mu))^k \theta_f^q(\mu)}{(\theta_b^1(\mu))^{k+1}} \right| \|\mathbb{A}_{k,q}\|_{\mathbb{X}_h} \\
&= \sup_{\mu \in P} \sum_{k=m}^{\infty} \sum_{q=1}^{Q_f} \left| \frac{(\theta_b^2(\mu))^k \theta_f^q(\mu)}{(\theta_b^1(\mu))^{k+1}} \right| \left\| \left((\mathbb{B}_h^1)^{-1} \mathbb{B}_h^2 \right)^k (\mathbb{B}_h^1)^{-1} \mathbb{F}_h^q \right\|_{\mathbb{X}_h} \\
&\leq Q_f \sup_{\mu,q} \left\{ \left| \frac{\theta_f^q(\mu)}{\theta_b^1(\mu)} \right| \left\| (\mathbb{B}_h^1)^{-1} \mathbb{F}_h^q \right\|_{\mathbb{X}_h} \right\} \cdot \sum_{k=m}^{\infty} \left| \frac{(\theta_b^2(\mu))^k}{(\theta_b^1(\mu))^k} \right| \left\| (\mathbb{B}_h^1)^{-1} \mathbb{B}_h^2 \right\|_{\mathbb{X}_h, \mathbb{X}_h}^k \\
&\overset{(ii)}{\leq} Q_f \sup_{\mu,q} \left\{ \left| \frac{\theta_f^q(\mu)}{\theta_b^1(\mu)} \right| \left\| (\mathbb{B}_h^1)^{-1} \mathbb{F}_h^q \right\|_{\mathbb{X}_h} \right\} \cdot \sum_{k=m}^{\infty} (1-\varepsilon)^{k-m+m} \\
&\overset{l=k-m}{=} Q_f \sup_{\mu,q} \left\{ \left| \frac{\theta_f^q(\mu)}{\theta_b^1(\mu)} \right| \left\| (\mathbb{B}_h^1)^{-1} \mathbb{F}_h^q \right\|_{\mathbb{X}_h} \right\} \cdot (1-\varepsilon)^m \sum_{l=0}^{\infty} (1-\varepsilon)^l \\
&= Q_f \sup_{\mu,q} \left\{ \left| \frac{\theta_f^q(\mu)}{\theta_b^1(\mu)} \right| \left\| (\mathbb{B}_h^1)^{-1} \mathbb{F}_h^q \right\|_{\mathbb{X}_h} \right\} \cdot (1-\varepsilon)^m \frac{1}{1-(1-\varepsilon)} \\
&= \frac{Q_f}{\varepsilon} \sup_{\mu,q} \left\{ \left| \frac{\theta_f^q(\mu)}{\theta_b^1(\mu)} \right| \left\| (\mathbb{B}_h^1)^{-1} \mathbb{F}_h^q \right\|_{\mathbb{X}_h} \right\} \cdot \exp\left(\frac{\ln(1-\varepsilon) n}{Q_f} \right)
\end{aligned}$$

Indem wir $a = -\frac{\ln(1-\varepsilon)n}{Q_f}$ und $C = \frac{Q_f}{\varepsilon} \sup_{\mu,q} \left\{ \left| \frac{\theta_f^q(\mu)}{\theta_b^1(\mu)} \right| \left\| (\mathbb{B}_h^1)^{-1} \mathbb{F}_h^q \right\|_{\mathbb{X}_h} \right\}$ wählen, folgt die Behauptung. □

5.12 Bemerkung

Gilt $b^2(u,v) = 0$ und damit $\mathbb{B}_h^2 = 0$ in Satz 5.11, so erhalten wir die folgende Darstellung für $\mathbb{U}_h(\mu)$:

$$\mathbb{U}_h(\mu) = \sum_{q=1}^{Q_f} \frac{\theta_f^q(\mu)}{\theta_b^1(\mu)} \mathbb{A}_{0,q}.$$

Der Greedy-Algorithmus würde dann eine Basis mit $N \approx Q_f$ konstruieren. Wir erhalten schließlich noch die allgemeine Aussage:

5.13 Satz

Sei $u(\mu)$ eine Lösung von $(P(\mu)), M_e := \{u(\mu) \mid \mu \in P\}$ und b affin parametrisch. Dann gilt:

$$d_n(M_e, X) \leq C \cdot e^{-cn^{\frac{1}{Q_b}}},$$

mit Konstanten $c, C > 0$.

Beweis: [Ohlberger, Rave, 2015]. □

6 Empirische Interpolation

6.1 Die empirische Interpolationsmethode

Unterkapitel 3.3 zeigt, dass affin parametrische (Bi-)Linearformen unabdingbar für eine Offlin-/Onlinezerlegung des RB-Modells sind. Gesucht ist daher ein Approximationsverfahren für parametrische Funktionen $g : \Omega \times P \to \mathbb{R}$ der Form

$$g(x;\mu) \approx I_m[(g(.;\mu))](x) = \sum_{m=1}^{M} \theta_g^m(\mu) z_m(x), \tag{6.1}$$

mit Skalarenfunktionen $\theta_g^m(\mu)$ und einer sogenannten 'kollateralen Reduzierten Basis' $Z_m = \{z_m\}_{m=1}^M$. Da wir in Unterkapitel 3.3 und in Kapitel 5 gesehen haben, dass die Anzahl der Summanden in der affin parametrischen Darstellung sowohl den Rechenaufwand in der Offline-/Online-Phase, als auch das Konvergenzverhalten der RB-Approximation beeinflussen kann. Daher ist es wichtig eine Approximation zu finden, für die M in (6.1) möglichst klein ist. Daher sind zum Beispiel Chebyshev oder Legendre Polynome als kollaterale Basis nur bedingt geeignet.

Polynome schlecht für komplexe Gebiete.

Stattdessen wählen wir wieder einen 'Snapshot basierten Ansatz', dass heißt $Z_M \subset \operatorname{span}\{g(.;\mu) \mid \mu \in S_{\text{train}} \subset P\}$. Die in [Barrault, Meday, Nguyen, Patera, 2004] eingeführte **Empirische Interpolationmethode** ist eine Möglichkeit eine solche Approximation zu bestimmen.

6.1 Definition (Empirische Interpolationsmethode)

Sei $G_{\text{train}} := \{g(.;\mu) \mid \mu \in S_{\text{train}}\} \subset G := \{g(.;\mu) \mid \mu \in P\} \subset C^0(\bar{\Omega})$ mit $S_{\text{train}} \subset P$ Menge zu interpolierender Funktionen und sei ε eine vorgegebene Toleranz und $M_{\max}$ eine vorgegebene Anzahl von Iterationsschritten. Für $M \leq M_{\max}, M \leq \dim(\operatorname{span}(G_{\text{train}}))$ definieren wir rekursiv die **Interpolationspunktmenge** $T_M := \{t^1, \ldots, t^M\} \subset \bar{\Omega}$ und die kollaterale Reduzierte Basis $Z_M := \{z_1, \ldots, z_M\} \subset \operatorname{span}(G_{\text{train}})$ wie folgt:

Initialisiere: $M = 0;\ e_0 = \varepsilon + 1;\ I_0[g(.;\mu)](x) \equiv 0;\ \mu^1 = \operatorname{argmax}_{\mu \in S_{\text{train}}} \|g(.;\mu)\|_{C^0(\bar{\Omega})}$; $T_0 = \emptyset;\ Z_0 = \emptyset$.
`while` $M < M_{\max}$ `and` $e_M > \varepsilon$:
$M \leftarrow M + 1$
$r_M(x) = g(x;\mu^M) - I_{M-1}[g(.;\mu)](x)$
$t^M = \operatorname{argmax}_{x \in \bar{\Omega}} |r_M(x)|$
$T_M = T_{M-1} \cup \{t^M\}$
$z_M(x) = r_M(x)/r_M(t^M)$
$Z_M = Z_{M-1} \cup \{z_M\}$
$e_M = \max_{\mu \in S_{\text{train}}} \|g(.;\mu)\|_{C^0(\bar{\Omega})}$
$\mu^{M+1} = \operatorname{argmax}_{\mu \in S_{\text{train}}} \|g(.;\mu)\|_{C^0(\bar{\Omega})}$
`end`

Hierbei bestimmen wir die Koeffizienten $\sigma_j^{M-1}(\mu^M)$ von $I_{M-1}[g(.;\mu^M)] := \sum_{j=1}^{M-1} \sigma_j^{M-1}(\mu^M) z_j$ durch lösen des LGS

$$\sum_{j=1}^{M-1} \sigma_j^{M-1}(\mu^M) z_j(t^i) = g(t^i;\mu^M),\ 1 \le i \le M-1. \tag{6.2}$$

Für allgemeine $\mu \in P$ definieren wir

$$I_M[g(.;\mu)] := \sum_{j=1}^{M} \sigma_j^M(\mu) z_j, \tag{6.3}$$

wobei

$$\sum_{j=1}^{M} \sigma_j^M(\mu) z_j = g(t^i;\mu),\ 1 \le i \le M. \tag{6.4}$$

Schließlich definieren wir die Matrix $\mathbb{B}^M \in \mathbb{R}^{M \times M}$ durch

$$B_{ij}^M = z_j(t^i),\ 1 \le i, j \le M. \tag{6.5}$$

Wir zeigen nun, dass die Konstruktion der Interpolationspunkte $\{t^i\}_{i=1}^M$ wohldefiniert ist und das $\{z_m\}_{m=1}^M$ linear unabhängig sind. Dazu benötigen wir das folgende Resultat:

6.2 Lemma

Unter der Annahme, dass für $W_{M-1} := \operatorname{span}\{z_1, \ldots, z_{M-1}\}$ $\dim(W_{M-1}) = M-1$ gilt und $\mathbb{B}^{M-1}$ invertierbar ist, gilt

$$I_{M-1}[v] = v\ \forall v \in W_{M-1},$$

wobei

$$I_{M-1}[v] = \sum_{j=1}^{M-1} \sigma_j^{M-1} z_j,$$

und σ_j^{M-1} Lösung von $\sum_{j=1}^{M-1} \sigma_j^{M-1} z_j(t^i) = v(t^i),\ 1 \le i \le M-1$. Mit anderen Worten ist die Interpolation also exakt für alle $v \in W_{M-1}$.

Beweis:
Da nach der Annahme $Z_{M-1} = \{z_1, \ldots, z_{M-1}\}$ Basis von W_{M-1} läst sich jedes $v \in W_{M-1}$ darstellen als

$$v(x) = \sum_{j=1}^{M-1} \gamma_j^{M-1} z_j(x),$$

mit entsprechenden Koeffizienten $\gamma_j^{M-1} \in \mathbb{R}$.
Insbesondere gilt in den Interpolationspunkten $t^1, \ldots, t^{M-1}$:

$$v(t^i) = \sum_{j=1}^{M-1} \gamma_j^{M-1} z_j(t^i),\ 1 \leq i \leq M-1.$$

Da $B_{ij}^{M-1} = z_j(t^i)$, folgt aus der Invertierbarkeit von $\mathbb{B}^{M-1}$, dass $\sigma_j^{M-1} = \gamma_j^{M-1}$ für $j = 1, \ldots, M-1$ und daher $I_{M-1}[v] = v$. □

6.3 Satz

Sei $M_{\max} < \dim(\operatorname{span}(G_{\text{train}}))$. Dann gilt für jedes $M \leq M_{\max}$, dass $W_M := \operatorname{span}\{z_1, \ldots, z_M\}$ die Dimension M hat. Ferner ist die Matrix $\mathbb{B}^M$ eine untere Dreiecksmatrix mit 1 auf der Diagonalen und daher invertierbar. Schließlich sind die Interpolationspunkte $t^1, \ldots, t^M$ paarweise disjunkt.

Beweis:
Durch vollständige Induktion:
Es gilt offensichtlich, dass $W_1 = \operatorname{span}\{z_1\}$ Dimension 1 hat und dass $\mathbb{B}^1 = 1$ invertierbar ist.
Wir nehmen an, dass $W_{M-1} = \operatorname{span}\{z_1, \ldots, z_{M-1}\}$ die Dimension $M-1$ hat und die Matrix $\mathbb{B}^{M-1}$ invertierbar ist. Es ist dann zu zeigen, dass

(i) $\dim W_M = M$ mit $W_M := \operatorname{span}\{z_1, \ldots, z_M\}$,

(ii) $\mathbb{B}^M$ ist invertierbar.
(Die Aussage über die Disjunktheit der Interpolationspunkte folgt dann direkt.)

Beweis (i):

$$\begin{aligned} e_{M-1} &= \max_{\mu \in S_{\text{train}}} \|g(.;\mu) - I_M[g(.;\mu)]\|_{C^0(\bar{\Omega})} \\ &\geq \max_{\mu \in S_{\text{train}}} \inf_{v \in W_{M-1}} \|g(.;\mu) - v\|_{C^0(\bar{\Omega})} \\ &\geq \inf_{W \subset C^0(\bar{\Omega}), \dim W = M_{\max}} \max_{\mu \in S_{\text{train}}} \inf_{v \in W} \|g(.;\mu) - v\|_{C^0(\bar{\Omega})} \\ &= d_{M_{\max}}(G_{\text{train}}, C^0(\bar{\Omega})) > 0, \end{aligned}$$

da $M_{\max} < \dim(\operatorname{span}(G_{\text{train}}))$ vorausgesetzt ist. Falls nun $\dim W_M \neq M$ gilt, so folgt $g(.;\mu^M) \in W_{M-1}$ und damit nach Lemma 6.2, dass $e_{M-1} = 0$ ↯. Dies zeigt, dass $\dim W_M = M$ und

$$\begin{aligned} \|r_M\|_{C^0(\bar{\Omega})} &= \left\|g(.;\mu^M) - I_{M-1}[g(.;\mu^M)]\right\|_{C^0(\bar{\Omega})} \\ &= \max_{\mu \in S_{\text{train}}} \|g(.;\mu) - I_{M-1}[g(.;\mu)]\|_{C^0(\bar{\Omega})} \\ &= e_{M-1} > 0. \end{aligned}$$

Beweis (ii): Es gilt:

$$B_{ij} = z_j(t^i) = \frac{r_j(t^i)}{r_j(t^j)} = \frac{g(t^i;\mu^j) - \sum_{l=1}^{j-1} \sigma_l^{j-1} z_l(t^i)}{r_j(t^j)}.$$

Damit folgt:

(1) $B_{ij}^M = 1$ für $i = j$.

(2) $B_{ij}^M = 0$ für $i < j$, da nach Konstruktion

$$g(t^i;\mu^j) = \sum_{l=1}^{j-1} \sigma_l^{j-1} z_l(t^i) \text{ für } 1 \leq i \leq j-1.$$

(3) $\left|B_{ij}^M\right| \leq 1$, da $t^j = \operatorname{argmax}_{x\in\bar{\Omega}} |r_j(x)|$.

Daher ist $\mathbb{B}^M$ untere Dreiecksmatrix mit 1 auf der Diagonalen und daher invertierbar. Dies zeigt auch, dass $t^M \neq t^i$ für $i = 1,\ldots,M-1$. Denn angenommen, dass $t^M = t^i$ für ein $i \in \{1,\ldots,M-1\}$, dann folgt aus der Definition von t^M:

$$\begin{aligned}\max_{x\in\bar{\Omega}} |r_M(x)| &= \left|r_M(t^M)\right| = \left|r_M(t^i)\right| \\ &= \left| g(t^i;\mu^M) - \sum_{j=1}^{M-1} \sigma_j^{M-1}(\mu^M) z_j(t^i) \right| \\ &= 0 \lightning .\end{aligned}$$

□

6.2 Praktische Implementierung

Da die Berechnung von t^M in der in Definition 6.1 definierten Empirischen Interpolationsmethode (EIM) sehr teuer ist, bestimmt man im Allgemeinen nicht das Maximum von $|r_M(x)|$ über $\bar{\Omega}$, sondern über eine endliche Teilmenge $\Omega_h := \{x^k\}_{k=1}^{N_h}$. Im Kontext von Finiten Elementen können die x^k zum Beispiel Quadraturpunkte oder die Ecken des Gitters sein. Die Kosten zur Berechnung des Maximums von $|r_M(x)|$ skalieren dann linear in der Anzahl der Punkte x^k. In diesem Setting können wir auch eine algebraische Version der EIM angeben.
Dazu führen wir zunächst die Abbildung $\mathbb{G} : P \to \mathbb{R}^{N_h}$ ein, welche wie folgt definiert ist

$$(\mathbb{G}(\mu))_i := g(x^i;\mu),\ i = 1,\ldots,N_h,\ \mu \in P.$$

Ferner bezeichnen wir mit $\mathbb{Z} \in \mathbb{R}^{N_h\times M}$ die Matrix

$$(\mathbb{Z})_{ik} := z_k(x^i),\ i = 1,\ldots,N_h,\ k = 1,\ldots,M.$$

Deren Spalten die diskreten Repräsentanten der kollateralen Basisfunktionen enthalten.
Schließlich bezeichnen wir mit $J_M := \{i_1,\ldots,i_M\}$ eine Menge von Interpolationsindizes, so dass $\{t^1,\ldots,t^M\} = \{x^{i_1},\ldots,x^{i_M}\}$ und führen die Matrix $\mathbb{P} = [e_{i_1},\ldots,e_{i_M}] \in \mathbb{R}^{N_h\times M}$ ein. Hierbei bezeichnen e_{i_j} die Einheitsvektoren, welche in der i_j-ten Zeile eine 1 haben. Dann gilt zum Beispiel:

$$\mathbb{P}^T\mathbb{G}(\mu) = \Big(g(t^1;\mu),\ldots,g(t^M;\mu)\Big)^T$$

und analog für die kollaterale Basen. Die diskrete empirische Interpolierende $I_M[\mathbb{G}(\mu)]$ ist dann gegeben durch

$$I_M[\mathbb{G}(\mu)] = \mathbb{Z}\sigma(\mu), \tag{6.6}$$

wobei $\sigma(\mu) = (\sigma_1(\mu),\ldots,\sigma_M(\mu))^T \in \mathbb{R}^M$ Lösung des folgenden linearen Gleichungssystems:

$$\mathbb{P}^T\mathbb{G}(\mu) = (\mathbb{P}^T\mathbb{Z})\sigma(\mu). \tag{6.7}$$

Die diskrete empirische Interpolierende können wir damit auch wie folgt schreiben:

$$I_M[\mathbb{G}(\mu)] = \mathbb{Z}(\mathbb{P}^T\mathbb{Z})^{-1}\mathbb{P}^T\mathbb{G}(\mu). \tag{6.8}$$

Der zur **Offline-Phase** der EIM gehörende Algorithmus kann dann wie folgt formuliert werden, wobei $\varepsilon, M_{\max}$ wie in Definition 6.1:
Input: $S_{\text{train}}, \mathbb{G}, \varepsilon, M_{\max}$
Output: $\mathbb{Z}, J_M$
Initialisiere: $M = 0, e_0 = \varepsilon + 1, \mu^1 = \text{argmax}_{\mu\in S_{\text{train}}} \|\mathbb{G}(\mu)\|_\infty, \mathbb{R} = \mathbb{G}(\mu^1), \mathbb{Z} = [], J_0 = \emptyset, \mathbb{P} = []$
`while` $M < M_{\max}$ `and` $e_M > \varepsilon$: $M \leftarrow M + 1$
$i_M = \text{argmax}_{i=1,\dots,N_h} |\mathbb{R}_i|$
$\bar{\mathbb{Z}} = \frac{\mathbb{R}}{\mathbb{R}_{i_m}}$, $\mathbb{Z} \leftarrow [\mathbb{Z}, \bar{\mathbb{Z}}_M]$
$J_M = J_{M-1} \cup \{i_m\}$, $\mathbb{P} \leftarrow [\mathbb{P}, \bar{e}_{i_m}]$
$e_M = \max_{\mu\in S_{\text{train}}} \left\|\mathbb{G}(\mu) - \mathbb{Z}(\mathbb{P}^T\mathbb{Z})^{-1}\mathbb{P}^T\mathbb{G}(\mu)\right\|_\infty$
$\mu_{M+1} = \text{argmax}_{\mu\in S_{\text{train}}} \left\|\mathbb{G}(\mu) - \mathbb{Z}(\mathbb{P}^T\mathbb{Z})^{-1}\mathbb{P}^T\mathbb{G}(\mu)\right\|_\infty$
$\mathbb{R} = \mathbb{G}(\mu^{M+1}) - \mathbb{Z}(\mathbb{P}^T\mathbb{Z})^{-1}\mathbb{P}^T\mathbb{G}(\mu^{M+1})$
`end`

6.4 Bemerkung

Zur Lösung des LGS (6.7) werden $\mathbb{O}(M^2)$ Rechenschritte benötigt, da $\mathbb{B}^M = \mathbb{P}^T\mathbb{Z}$ eine untere Dreiecksmatrix ist. Der obige Algorithmus erfordert in jeder Iteration die Auswertung von $\mathbb{G}(\mu)$ für alle $\mu \in S_{\text{train}}$. Ist diese Auswertung teuer kann es Sinn ergeben vor dem Beginn der `while`-Schleife einmal die Matrix

$$\mathbb{S} = [\mathbb{G}(\mu^1), \dots, \mathbb{G}(\mu^{n_{\text{train}}})] \in \mathbb{R}^{N_h \times n_{\text{train}}}$$

zu assemblieren und zu speichern. Dies ist je nach verfügbarem Speicher allerdings nur für moderate N_h und N_{train} möglich.
Schließlich verfahren wir in der **Online-Phase** für die EIM wie folgt:
Input: $\mu, \mathbb{B} = (\mathbb{P}^T\mathbb{Z}), T_M$
Output: $\sigma(\mu)$

- Werte $g(.;\mu)$ in den Interpolationspunkten aus und assembliere dadurch

$$(g(t^1;\mu), \dots, g(t^M;\mu))^T = \mathbb{B}\sigma(\mu).$$

- Löse $(g(t^1;\mu), \dots, g(t^M;\mu))^T = \mathbb{B}\sigma(\mu)$.

6.5 Bemerkung

Da $(\mathbb{P}^T\mathbb{Z})$ eine untere Dreiecksmatrix ist benötigen wir in der Online-Phase $\mathbb{O}(M^2)$ Rechenschritte. Insbesondere ist die Komplexität unabhängig von N_h.

6.6 Bemerkung

ALternativ kann Z_M in einem ersten Schritt mittels POD bestimmt werden. Die Interpolationspunkte werden anschließend so bestimmt, dass sie das Residuum maximieren. Da die Basisfunktionen POD-Basisfunktionen sind und nicht iterativ gleich dem skalierten Residuum gewählt werden, sind sie zwar orthonormal, der Rechenaufwand zur Bestimmung der Koeffizienten $\sigma(\mu)$ beträgt aber $\mathbb{O}(M^3)$. Diesen Ansatz findet man in der Literatur unter 'Discret Empirial Interpolation Method' [Chaturantabut, Sorensen, 2010].

6.3 Fehlerabschätzungen

Für analytische Untersuchungen führen wir die nodale Basis $\xi_M \subset \mathrm{span}(Z_M)$ ein, wobei $\xi_m(t^i) = \delta_{im},\ 1 \leq i, m \leq M$. Die Existenz und Eindeutigkeit dieser nodalen Basis sichert das folgende Lemma.

6.7 Lemma

Für jedes M-Tupel $(a_i)_{i=1,\dots,M}$ reeller Zahlen existiert genau ein Element $w \in W_M$, so dass $w(t^i) = a_i$ für alle $i = 1, \dots, M$.

Beweis:
Folgt direkt aus der Invertierbarkeit von $\mathbb{B}^M$. □

Darauf basierend erhalten wir die folgende A priori Fehlerabschätzung für die EIM.

6.8 Satz (A priori Fehlerabschätzung)

Sei $I_M : C^0(\bar{\Omega}) \to W_M = \mathrm{span}\{\xi_i\}_{i=1}^M \subset C^0(\bar{\Omega})$ Interpolationsoperator zu den Punkten $\{t^i\}_{i=1}^M \subset \bar{\Omega}$ und $\{\xi_i\}_{i=1}^M$ nodale Basis, dass heißt $\xi_i(t^j) = \delta_{ij},\ 1 \leq i, j \leq M,\ I_M[g(.;\mu)] = \sum_{i=1}^M g(t^i;\mu)\xi_i$. Dann ist

$$\Lambda_M := \max_{x \in \bar{\Omega}} \sum_{i=1}^M |\xi_i(x)|$$

die **Lebesgue-Konstante** der Interpolation. Es gilt:

(1)
$$\|g(.;\mu) - I_M[g(.;\mu)]\|_{C^0(\bar{\Omega})} \leq (1 - \Lambda_M) \inf_{w \in W_M} \|g(.;\mu) - w\|_{C^0(\bar{\Omega})}$$

für $g(.;\mu) \in C^0(\bar{\Omega})$.

(2) Für die Lebesgue-Konstante gilt die Abschätzung

$$\Lambda_M \leq 2^M - 1.$$

Beweis:

1. Sei $g(.;\mu[(1)]) \in C^0(\bar{\Omega}), x \in \bar{\Omega}$ und $v \in W_M$. Dann gilt

$$\begin{aligned}
|g(x;\mu) - I_M[g(.;\mu)](x)| &\leq |g(x;\mu) - v(x)| + |v(x) - I_M[g(.;\mu)](x)| \\
&\overset{6.2}{=} |g(x;\mu) - v(x)| + |I_M[v](x) - I_M[g(.;\mu)](x)| \\
&= |g(x;\mu) - v(x)| + \left| \sum_{j=1}^M (v(t^j) - g(t^j;\mu))\xi_j(x) \right| \\
&\leq \|g(.;\mu) - v\|_{C^0(\bar{\Omega})} + \max_{x \in \bar{\Omega}} \sum_{j=1}^M |\xi_j(x)| \, \|g(.;\mu) - v\|_{C^0(\bar{\Omega})} \\
&= \|g(.;\mu) - v\|_{C^0(\bar{\Omega})} + \Lambda_M \, \|g(.;\mu) - v\|_{C^0(\bar{\Omega})}
\end{aligned}$$

2. Zunächst bemerken wir,dass aus $z_m(x) = \sum_{j=1}^M z_m(t^j)\xi_j(x)$ die Darstellung

$$z_m(x) = \sum_{j=1}^M \mathbb{B}_{jm}^M \xi_j(x)$$

folgt. Da $\mathbb{B}^M_{mm} = 1$, folgt

$$\begin{aligned}|\xi_m(x)| &= \left| z_m(x) - \underbrace{\sum_{j=1}^{m-1} \mathbb{B}^M_{jm}\xi_j(x)}_{=0} - \sum_{j=m+1}^{M} \mathbb{B}^M_{jm}\xi_j(x) \right| \\ &= \left| z_m(x) - \sum_{j=m+1}^{M} \mathbb{B}^M_{jm}\xi_j(x) \right| \\ &\leq 1 + \sum_{j=m+1}^{M} |\xi_j(x)|\,,\ 1 \leq m \leq M-1.\end{aligned}$$

Da $|\xi_M(x)| = |z_M(x)| \leq 1$ gilt zum Beispiel

$$|\xi_{M-1}(x)| \leq 1+1;\ |\xi_{M-2}(x)| \leq 1+1+2;\ |\xi_{M-3}(x)| \leq 1+1+2+4$$

und damit

$$|\xi_{M+1-m}(x)| \leq 2^{m-1},\ 1 \leq m \leq M.$$

Daher gilt

$$\begin{aligned}\sum_{m=1}^{M} |\xi_m(x)| &= \sum_{m=1}^{M} |\xi_{M+1-m}(x)| \leq \sum_{m=1}^{M} 2^{m-1} \\ &= \sum_{m=0}^{M-1} 2^m = 2^M - 1 \text{ (geometrische Reihe)}\end{aligned}$$

□

6.9 Bemerkung

Obige Abschätzung ist sehr pessimistisch; in der Praxis werden meist sehr viel bessere Lebesgue-Konstanten beobachtet. Allerdings ist die Schranke scharf, dass heißt es existieren Beispiele mit $\Lambda_M = 2^M - 1$ (vgl. [Maday, Nguyen, Patera, Pan, 2009]).
Die Abschätzung in 6.8 macht aber eine Aussage über die Stabilität der EI in dem Sinne, dass wir Konvergenz der Approximation für $M \to \infty$ erhalten, falls der Bestapproximationsfehler schneller als 2^M fällt, dass für langsame Konvergenzraten die EI aber für große M instabil werden kann.

6.10 Satz (Exponentielle Konvergenz des Interpolationsfehlers [Maday, Nguyen, Patera, Pan, 2009])

Falls eine Folge von endlich dimensionalen Unterräumen $Y_1 \subset Y_2 \subset \cdots \subset Y_M \subset \cdots \subset \operatorname{span}(G)$ mit $\dim Y_M = M$ gibt und eine Konstante $c > 0$ und ein $\alpha > \ln(4)$ existiert, so dass gilt

$$\sup_{\mu \in P} \inf_{v \in Y_M} \|g(.;\mu) - v\|_{C^0(\bar{\Omega})} \leq c \cdot e^{-(\alpha - \ln(4))M}.$$

Beweis: siehe [MNPP, 2009].

6.11 Satz A posteriori Fehlerabschätzung für die EIM

Seien $I_M, I_{M'} : C^0(\bar{\Omega}) \to \mathrm{span}(G_{\mathrm{train}})$ EI-Operatoren für $M' > M,\ Z_M \subset Z_{M'} := \{z_i\}_{i=1}^{M'},\ T_M \subset T_{M'} := \{t^i\}_{i=1}^{M'}$. Ferner seien für $g(.;\mu) \in G$ die Koeffizienten $\sigma_j^{M'}(\mu)$ von $I_{M'}[g(.;\mu)] = \sum_{j=1}^{M'} \sigma_j^{M'}(\mu) z_j$ Lösung des LGS

$$\sum_{j=1}^{M'} \sigma_j^{M'}(\mu) z_j(t^i) = g(t^i;\mu),\ 1 \le i \le M'.$$

Falls $g(.;\mu) \in W_{M'} := \mathrm{span}\{z_1, \dots, z_{M'}\}$, so gilt die folgende A posteriori Fehlerabschätzung:

$$\|g(.;\mu) - I_M[g(.;\mu)]\|_{C^0(\bar{\Omega})} \le \Delta_M^{M'}(\mu), \tag{6.9}$$

wobei

$$\Delta_M^{M'}(\mu) := \sum_{i=M+1}^{M'} \left|\sigma_i^{M'}(\mu)\right|. \tag{6.10}$$

Beweis:
Nach der Definition der EIM gilt:

$$I_M[g(.;\mu)] = \sum_{i=1}^{M} \sigma_i^M z_i \text{ mit } \sum_{i=1}^{M} \sigma_i^M z_i(t^j) = g(t^j;\mu),\ j = 1, \dots, M.$$

Da $\mathbb{B}^M$ linke untere Dreiecksmatrix und oberer linker Block von $\mathbb{B}^{M'}$, folgt $\sigma_j^M(\mu) = \sigma_j^{M'}(\mu)$ für $j = 1, \dots, M,\ \mu \in P$.

$$\begin{aligned} g(x;\mu) - I_M[g(.;\mu)](x) &= I_{M'}[g(.;\mu)](x) - I_M[g(.;\mu)](x) \\ &= \sum_{j=1}^{M'} \sigma_j^{M'} z_j - \sum_{j=1}^{M} \sigma_j^M z_j \\ &= \sum_{j=M+1}^{M'} \sigma_j^{M'} z_j. \end{aligned}$$

Wegen $\|z_j\|_{C^0(\bar{\Omega})} = 1$ folgt dann die Aussage. □

6.12 Bemerkung

Da im Allgemeinen nicht $g(.;\mu) \in W_{M'}$ gilt, ist $\Delta_M^{M'}(\mu)$ im Allgemeinen keine rigorose obere Schranke für den Interpolationsfehler.
Falls letztere sehr schnell gegen 0 konvergiert für $M \to \infty$, erwarten wir, dass die Effektivität $\Delta_M^{M'}(\mu) / \|g(.;\mu) - I_M[g(.;\mu)]\|_C$ aber zumindest nahe bei 1 liegt. In praktischen Anwendungen zeigt es sich, dass eine Wahl von $M' - M \approx 10$ häufig einen verlässlichen Schätzer liefert. Eine rigorose, aber sehr teure Fehlerschranke wurde in [Eftang, Grepl, Patera, 2010] hergeleitet.

6.4 Anwendungen für lineare RB Methoden

Wir betrachten wieder das Modellproblem der parametrischen stationären WLG aus Definition3.7 und nehmen zusätzlich an, dass $\kappa(.;\mu), q(.;\mu) \in C^0(\bar{\Omega})$ für alle $\mu \in P$. Falls die Datenfunktionen $\kappa(.;\mu)$ und $q(.;\mu)$ nicht affin parametrisch sind, so können wir sie mit empirischen Interpolierenden approximieren.

$$\kappa(x;\mu) \approx I_{M_\kappa}[\kappa(.;\mu)](x) = \sum_{j=1}^{M_\kappa} \sigma_j^\kappa(\mu) z_j^\kappa$$

und

$$q(x;\mu) \approx I_{M_q}[q(.;\mu)](x) = \sum_{j=1}^{M_q} \sigma_j^q(\mu) z_j^q.$$

Die entsprechende Bilinearform und Linearform definieren wir dann wie folgt:

$$b_M(u,v;\mu) := \int_\Omega I_{M_\kappa}[\kappa(.;\mu)]\nabla u \nabla v, \ \forall u,v \in X, \tag{6.11}$$

$$f_M(v;\mu) := \int_\Omega I_{M_q}[q(.;\mu)]v, \ \forall v \in X, \tag{6.12}$$

wobei wir zur Vereinfachung bei f und b den Doppelindex durch M ersetzen.

6.13 Definition (FEM für parametrische Variationprobleme mit EIM Approximation)

Seien die Voraussetzungen aus Definition 3.9 erfüllt. Ferner seien $b_M : X \times X \times P \to \mathbb{R}$ und $f_M : X \times P \to \mathbb{R}$ Bilinearform und Linearform, welche aus b und f durch eine Approximation der Datenfunktionen mittels EIM hervorgegangen sind. Zusätzlich nehmen wir an, dass $b(.,.;\mu)$ stetig auf $X_h \times X_h$ und koerziv auf X_h für alle $\mu \in P$, das heißt

$$\exists \alpha_h^M(\mu) : b_M(v,v;\mu) \geq \alpha_h^M(\mu) \|v\|_X^2 \ \ \forall v \in X_h$$

und das $f(.;\mu)$ stetig auf X_h. Zu $\mu \in P$ heißt $u_h^M(\mu) \in X_h$ Lösung des FEM für das parametrische Variationsproblem mit EIM Approximation, falls gilt:

$$b_M(u_h^M, v;\mu) = f(v;\mu), \ \forall v \in X_h. \tag{6.13}$$

6.14 Bemerkung

Die Koerzivität von b muss sich nicht notwendigerweise auf b_M übertragen, kann aber im diskreten Fall leicht verifiziert werden.

6.15 Definition (RB Modell mit EIM Approximation)

Seien die Voraussetzungen von Definition 3.13 und 6.13 erfüllt. Zu $\mu \in P$ heißt $u_N^M(\mu) \in X_N$ RB-Lösung des RB Modells mit EIM Approximation, falls gilt:

$$b_M(u_N^M(\mu), v;\mu) = f_M(v;\mu) \ \forall v \in X_N.$$

6.16 Bemerkung

Ähnlich zu Satz 3.19 und Satz 3.35 erhält man A priori und A posteriori Fehlerabschätzungen, weelche allerdings zusätzliche Terme für die Abschätzung des Interpolationsfehlers enthalten. Für Details siehe zum Beispiel [Quateroni, Manzoni, Negri, 2015] und zu A posteriori Fehlerschätzern insbesondere [Tonn, 2011, Kap. 4.5].

6.17 Bemerkung

Falls die Koeffizientenfunktionen keine Punktauswertungen erlauben, kann die Generalized Empirial Interpolation verwendet werden [Maday, Mala, 2013] oder gegebenenfalls zunächst eine Projektion in den Finite Elemente Raum vorgenommen werden.

7 Lokalisierte Modellreduktion

Ziele:

- ermögliche topologische Flexibilität in der Online-Phase
- erlaube lokale Änderungen der Geometrie oder der Parametermenge in der Online-Phase (Online-adaptive Modellreduktion)
- reduziere Anzahl der Parameter

Ansatz:

- Zerlege das Rechengebiet in Teilgebiete, wende in den Teilgebieten RB Methoden an und erhalte dadurch eine reduzierte Lösung auf dem globalen Begiet Ω.
- Stichwort: Legos

7.1 Einführung in Gebietszerlegungsmethoden

In diesem Unterkapitel betrachten wir das folgende Modellproblem: Finde u, so dass

$$-\Delta u = q \text{ in } \Omega;\ \nabla u \cdot n = 0 \text{ auf } \Sigma_N;\ u = 0 \text{ auf } \Sigma_D, \tag{7.1}$$

wobei $q \in L^2(\Omega), \partial\Omega = \bar{\Sigma}_N \cup \bar{\Sigma}_D$ und n äußere Normale auf $\Omega \subset \mathbb{R}^2$. Zunächst zerlegen wir ω in zwei nicht-überlappende Teilgebiete Ω_1 und Ω_2, so dass $\bar{\Omega} = \bar{\Omega}_1 \cup \bar{\Omega}$. Das Interface bezeichnen wir mit $\Gamma := \bar{\Omega}_1 \cap \bar{\Omega}$.

7.1.1 Das Modellproblem auf zerlegtem Gebiet und die Steklov-Poincaré Interface Gleichung

Wir bezeichnen mit u_i die **Restriktion** der Lösung u von (7.1) auf $\Omega_i, i = 1, 2$ und mit n_i die äußere Normale auf $\partial\Omega_i \cap \Gamma$. Dann kann das Modellproblem (7.1) äquivalent auf dem zerlegten Gebiet wie folgt formuliert werden:

$$\begin{aligned} -\Delta u_1 &= q \text{ in } \Omega_1;\ \nabla u_1 \cdot n = 0 \text{ auf } \Sigma_N \cap \partial\Omega_1;\ u = 0 \text{ auf } \Sigma_D \cap \partial\Omega_1, \\ -\Delta u_2 &= q \text{ in } \Omega_2;\ \nabla u_2 \cdot n = 0 \text{ auf } \Sigma_N \cap \partial\Omega_2;\ u = 0 \text{ auf } \Sigma_D \cap \partial\Omega_2, \\ u_1 &= u_2 \text{ auf } \Gamma;\ \nabla u_1 \cdot n_1 = -\nabla u_2 \cdot n_2 \text{ auf } \Gamma \text{ 'transmission conditions'}. \end{aligned} \tag{7.2}$$

Beachte das die Äquivalenz im schwachen Sinne zu verstehen ist ($\rightarrow$ 7.1.2). Wir können (7.2) weiter zu einem Problem mit einer gesuchten Funktion auf dem Interface Γ umformulieren. Bezeichne dazu mit λ den unbekannten und gesuchten Wert von u auf Γ. Wir betrachten die beiden Probleme

$$\begin{aligned} -\Delta w_i = q \text{ in } \Omega_i;\ \nabla w_i \cdot n = 0 \text{ auf } \Sigma_N \cap \partial\Omega_i; \\ w_i = 0 \text{ auf } \Sigma_D \cap \partial\Omega_i;\ w_i = \lambda \text{ auf } \Gamma,\ i = 1, 2. \end{aligned} \tag{7.3}$$

Als nächstes zerlegen wir die w_i in Funktionen, welche die Daten, also q repräsentieren und Funktionen, welche die gesuchte Funktion $\lambda = u|_\Gamma$ repräsentieren.
Betrachte dazu $w_i = u_i^0 + u_i^q$, wobei u_i^0 und $u_i^q, 1 = 1, 2$, als Lösungen der folgenden Probleme definiert sind:

$$-\Delta u_i^0 = 0 \text{ in } \Omega_i;\ \nabla u_i^0 = 0 \text{ auf } \Sigma_N \cap \partial\Omega_i;\ u_i^q = 0 \text{ auf } \Sigma_D \cap \partial\Omega_i;\ u_i^0 = \lambda \text{ auf } \Gamma \tag{7.4}$$

und

$$-\Delta u_i^q = 0 \text{ in } \Omega_i;\ \nabla u_i^q = 0 \text{ auf } \Sigma_N \cap \partial\Omega_i;\ u_i^q = 0 \text{ auf } \Sigma_D \cap \partial\Omega_i;\ u_i^q = \lambda \text{ auf } \Gamma, \tag{7.5}$$

mit $i = 1, 2$.
Da u_i^0 **harmonische Fortsetzung** von λ nach Ω_i bezeichnen wir u_i^0 von nun an mit $H_i\lambda$. Ferner schreiben wir $R_i q$ für die Repräsentanten der Daten u_i^q. Indem wir (7.3) mit (7.4) vergleichen stellen wir fest, dass

$$w_i = u_i \text{ für } i = 1, 2 \Leftrightarrow \nabla w_1 n_1 = -\nabla w_2 n_2 \text{ auf } \Gamma. \tag{7.6}$$

Letztere Bedingung bedeutet, dass die gesuchte Funktion $\lambda = u|\Gamma$ die **Steklov-Poincaré Interface Gleichung** erfüllen muss:

$$S\lambda = \chi \text{ auf } \Gamma, \tag{7.7}$$

wobei

$$\chi := -\nabla R_1 q n_1 - \nabla R_2 q n_2 = -\sum_{i=1}^{2} \nabla R_i q n_i \tag{7.8}$$

und der **Steklov-Poincaré Operator** S für Funktionen η auf Γ formal wie folgt definiert ist:

$$S_\eta := \nabla H_1 \eta n_1 + \nabla H_2 \eta n_2 = \sum_{i=1}^{2} \nabla H_i \eta n_i. \tag{7.9}$$

7.1.2 Schwache Formulierung des Modellproblems auf zerlegtem Gebiet

7.1 Definiton

Sei $q \in L^2(\Omega), \Omega, \Sigma_N, \Sigma_D$ wie in Bild 7.1 und $\partial\Omega = \bar{\Sigma}_N \cup \bar{\Sigma}_D$. Dann heißt $u \in X := \left\{v \in H^1(\Omega) \mid v = 0 \text{ auf } \Sigma_D\right\}$ schwache Lösung des Modellproblems (7.1), falls gilt

$$b(u, v) = f(v)\ \forall v \in X,$$

mit

$$b(u, v) := \int_\Omega \nabla u \nabla v,\ f(v) := \int_\Omega qv\ \forall u, v \in X.$$

Um eine schwache Formulierung des Modellproblems auf zerlegtem Gebiet herleiten zu können benötigen wir zunächst eine adäquate Charakterisierung des Spurraums auf Γ mittels Sobolevräumen mit gebrochenem Exponenten.

7.2 Definition

Seien Γ und $\Omega_i \subset \mathbb{R}^2$ wie in Bild 7.2, $i = 1, 2$. Für $v \in L^2(\Omega)$ definieren wir die **slobedeckij-Halbnorm** als

$$[v]_{\frac{1}{2},2,\Gamma} := \left(\int_\Omega \int_\Omega \frac{|v(x) - v(y)|^2}{\|x - y\|^3} \mathrm{d}x\mathrm{d}y\right)^{\frac{1}{2}},$$

$H^{\frac{1}{2}}(\Omega)$ nicht immer gleich definiert und auch nicht auf allen Gebieten äquivalent.

den Raum $H^{\frac{1}{2}(\Omega)}$ als

$$H^{\frac{1}{2}}(\Gamma) := \left\{v \in L^2(\Gamma) \mid [v]_{\frac{1}{2},2,\Gamma} < \infty\right\} \tag{7.10}$$

mit zugehöriger Norm

$$\|v\|_{H^{\frac{1}{2}}(\Gamma)} := \left(\|v\|^2_{L^2(\Omega)} + [v]_{\frac{1}{2},2,\Gamma}\right)^{\frac{1}{2}}. \tag{7.11}$$

Ferner definieren wir durch

$$(u, v)_{H^{\frac{1}{2}}(\Gamma)} := (u, v)_{L^2(\Omega)} + \int_\Omega \int_\Omega \frac{(u(x) - u(y))(v(x) - v(y))}{\|x - y\|^3} \mathrm{d}x\mathrm{d}y \tag{7.12}$$

ein Skalarprodukt $H^{\frac{1}{2}}(\Gamma)$.

7.3 Satz

Unter den Voraussetzungen aus Definition 7.2 ist $H^{\frac{1}{2}}(\Gamma)$ mit der in (7.11) definierten Norm und dem in (7.12) definierten Skalarprodukt eine Hilbertraum.

Beweis: [Adams, 1975], [Grisvard, 2011]. □

7.4 Spursatz

Seien $\Gamma, \Omega_i, i = 1, 2$ wie in Bild 7.2. Dann gilt für Ω_1 (und analog für Ω_2):

(1) Es existiert eine eindeutige lineare, stetige Abbildung

$$\mathcal{T} : H^1(\Omega_1) \to H^{\frac{1}{2}}(\Gamma), \text{ so dass } \mathcal{T}v = v|_\Gamma$$

für alle $v \in H^1(\Omega) \cap C^0(\Omega_1)$. Ferner gilt die folgende Abschätzung

$$\|\mathcal{T}v\|_{H^{\frac{1}{2}}(\Gamma)} \leq c_t \cdot \|v\|_{H^1(\Omega_1)}, \tag{7.13}$$

mit der **Spurkonstanten** $c_t > 0$.

(2) Es existiert eine lineare, stetige Abbildung $\mathcal{F} : H^{\frac{1}{2}}(\Gamma) \to H^1(\Omega)$, so dass

$$\mathcal{T}\mathcal{F}\varphi = \varphi \ \forall \varphi \in H^{\frac{1}{2}}(\Gamma). \tag{7.14}$$

Beweis: [Adams, 1975], [Grisvard, 2011]. □

7.5 Bemerkung

Betrachtet man auf dem ganzen Rand von Ω homogene Dirichletrandwerte und gilt wie in Bild 7.2 $\partial\Omega \cap \Gamma \neq \emptyset$, so muss auf Γ ein andere Spurraum als $H^{\frac{1}{2}}(\Gamma)$ betrachtet werden. Für Details siehe zum Beispiel [Quateroni, Valli 2005].

Nun können wir die schwache Formulierung des Modellproblems auf zerlegtem Gebiet herleiten.

7.6 Lemma

Seien $q \in L^2(\Omega), \Omega, \Omega_i, \Sigma_N, \Sigma_D, \Gamma$ wie in Bild 7.2, $\partial\Omega = \bar{\Sigma_D} \cup \bar{\Sigma_N}$ und seien

$$b_i(u_i, v_i) := \int_{\Omega_i} \nabla u_i \nabla v_i \ \forall u_i, v_i \in H^1(\Omega_i),$$

$$f_i(v_i) := \int_{\Omega_i} q|_{\Omega_i} v_i \ \forall v_i \in H^1(\Omega_i),$$

$$X_i := \left\{ v_i \in H^1(\Omega_i) \mid v_i = 0 \text{ auf } \Sigma_D \cap \Omega_i \right\},$$

$$X_i^0 := \left\{ v_i \in H^1(\Omega_i) \mid v_i = 0 \text{ auf } (\Sigma_D \cap \Omega_i) \cup \Gamma \right\}, \ i = 1, 2.$$

Dann lässt sich die schwache Formulierung des Modellproblems (7.1) aus Definition 7.1 äquivalent wie folgt formulieren: Finde $u_1 \in X_1, u_2 \in X_2$, so dass

$$\begin{aligned} b_1(u_1, v_1) &= f_1(v_1) \ \forall v_1 \in X_1^0 \\ b_2(u_2, v_2) &= f_2(v_2) \ \forall v_2 \in X_2^0 \\ u_1 &= u_2 \text{ auf } \Gamma \\ b_2(u_2, \mathcal{F}_2\eta) &= f_2(\mathcal{F}_2\eta) + f_1(\mathcal{F}_1\eta) - b_1(u_1, \mathcal{F}_1\eta) \ \forall \eta \in H^{\frac{1}{2}}(\Gamma), \end{aligned} \tag{7.15}$$

wobei $\mathcal{F}_i : H^{\frac{1}{2}}(\Gamma) \to X_i$ Fortsetzungsoperatoren wie in Satz 7.4.

Beweis:
Sei zunächst u die schwache Lösung aus Definition 7.1. Wir setzen $u_i := u|_{\Omega_i}, i = 1, 2$ und folgern zunächst, dass $u_i \in X_i$ wegen $u \in X$. Ferner gelten $(7.15)_1$ und $(7.15)_2$, da wir zum Beispiel alle Testfunktionen $v_1 \in X_1^0$ durch 0 auf Ω_2 fortsetzen können und dadurch zulässige Testfunktionen in X erhalten (folgt aus Satz 7.4, weil die Spur auf Γ Null ist).
Weiterhin liegt für jedes $\eta \in H^{\frac{1}{2}}(\Gamma)$ die Funktion

$$\mathcal{F}_\eta := \begin{cases} \mathcal{F}_1\eta & \text{in } \Omega_1, \\ \mathcal{F}_2\eta & \text{in } \Omega_2 \end{cases} \tag{7.16}$$

in X, was $b(u, \mathcal{F}_\eta) = f(\mathcal{F}_\eta)$ und weiter $(7.15)_4$ impliziert.
Für $(7.15)_3$ zeigen wir zunächst: aus $u_i \in X_i$ und $u_1 = u_2$ auf Γ folgt $u \in X$. Wir wissen: $u_i \in X_i$ und $u \in X$. Wie wir sehen werden impliziert die Annahme $u_1 \neq u_2$ auf Γ, dass $u \notin H^1(\Omega)$ und damit einen Widerspruch zu $u \in X$.
Gelte also $u_i \in X_i \subset H^1(\Omega_i)$ und $u_1 = u_2$ auf Γ.

$$\begin{aligned}
\int_\Omega u D_j\varphi &= \int_{\Omega_1} u|_{\Omega_1} D_j\varphi + \int_{\Omega_2} u|_{\Omega_2} D_j\varphi \\
&= \int_{\Omega_1} u_1 D_j\varphi + \int_{\Omega_2} u_2 D_j\varphi \\
&= -\int_{\Omega_1} D_j u_1\varphi - \int_{\Omega_2} D_j u_2\varphi + \int_\Gamma u_1\varphi(n_1)_j + \int_\Gamma u_2\varphi(n_2)_j \\
&= -\int_{\Omega_1} D_j u_1\varphi - \int_{\Omega_2} D_j u_2\varphi + \int_\Gamma \underbrace{(u_1 - u_2)}_{=0}\varphi(n_1)_j \\
&= -\int_\Omega D_j u\varphi, \ \varphi \in C_0^\infty(\Omega).
\end{aligned}$$

Seien nun $u_i, i = 1, 2$ Lösungen von (7.15). Wir setzen

$$u := \begin{cases} u_1 & \text{in } \Omega_1, \\ u_2 & \text{in } \Omega_2, \end{cases}$$

und folgern aus $u_i \in X_i$ und $(7.15)_3$, dass $u \in X$. Für beliebiges $v \in X$ gilt: $\eta := \mathcal{T}_\Gamma v \in H^{\frac{1}{2}}(\Gamma)$. Sei $\mathcal{F}_\eta$ wie in (7.16). Da $(v|_{\Omega_i} - \mathcal{F}_i\eta) \in X_i^0$ folgt aus $(7.15)_1$, $(7.15)_2$ und $(7.15)_3$, dass

$$\begin{aligned}
b(u, v) &= \sum_{i=1}^2 b_i(u_i, v|_{\Omega_i}) = \sum_{i=1}^2 b_i(u_i, v|_{\Omega_i} - \mathcal{F}_i\eta) + b_i(u_i, \mathcal{F}_i\eta) \\
&= \sum_{i=1}^2 f_i(v|_{\Omega_i} - \mathcal{F}_i\eta) + f_i(\mathcal{F}_i\eta) \\
&= \sum_{i=1}^2 f_i(v|_{\Omega_i}) = f(v) \ \forall v \in X.
\end{aligned}$$

□

7.7 Bemerkung

$(7.15)_4$ ist die schwache formulierung der Neumann-Bedingung $(7.2)_8$.
Nächstes Ziel: Herleitung der Steklov-Poincaré Interface Gleichung im schwachen Sinne.

Dazu starten wir von der schwachen Neumann-Bedingung $(7.15)_4$, wobei wir wie oben die u_i zerlegen: $u_i = H_i\lambda + R_iq$, jetzt mit $H_i\lambda \in X_i$ und $R_iq \in X_i^0$, wobei $H_i\lambda, R_iq$ Lösungen von

$$b_i(H_i\lambda, v_i) = 0 \ \forall v_i \in X_i^0; \ H_i\lambda = \lambda \text{ auf } \Gamma, \tag{7.17}$$

$$b_i(R_iq, v_i) = f_i(v_i) \ \forall v_i \in X_i^0. \tag{7.18}$$

Einsetzen in $(7.15)_4$ liefert

$$\sum_{i=1}^{2} b_i(H_i\lambda, \mathcal{F}_i\eta) = \sum_{i=1}^{2} \left[f_i(\mathcal{F}_i\eta) - b_i(R_iq, \mathcal{F}_i\eta) \right] \ \forall \eta \in H^{\frac{1}{2}}(\Gamma). \tag{7.19}$$

Da $H_i\lambda$ Lösung von (7.17) gilt

$$\|H_i\lambda\|_{H^1(\Omega_i)} \leq c \cdot \|\lambda\|_{H^{\frac{1}{2}}(\Gamma)}, \tag{7.20}$$

(siehe zum Beispiel [Lions, Magenes 1972]) und wir können $\mathcal{F}_i\eta = H_i\eta$ in (7.19) wählen. Wir erhalten die Steklov-Poincaré Interface Ungleichung im schwachen Sinne:
Finde $\lambda \in H^{\frac{1}{2}}(\Gamma)$, so dass

$$\sum_{i=1}^{2} b_i(H_i\lambda, H_i\eta) = \sum_{i=1}^{2} \left[f_i(H_i\eta) - b_i(R_iq, H_i\eta) \right] \ \forall \eta \in H^{\frac{1}{2}}(\Gamma).1 \tag{7.21}$$