



Fakultät für Mathematik und Physik  
Institut für Angewandte Mathematik

Diplomarbeit

# Ein hierarchischer Fehlerschätzer für Hindernisprobleme

von Cornelius Rüther  
Matr.-Nr.: 2517350

1. Oktober 2014

Erstprüfer: Prof. Dr. Gerhard Starke  
Zweitprüfer: Prof. Dr. Peter Wriggers

# Inhaltsverzeichnis

Abbildungsverzeichnis	iii
Tabellenverzeichnis	iv
<b>1 Einleitung</b>	<b>6</b>
<b>2 Grundlagen</b>	<b>7</b>
2.1 Hilberträume . . . . .	7
2.2 Variationsformulierung . . . . .	9
2.3 Finite Elemente Methode . . . . .	18
2.3.1 A priori Fehlerabschätzung . . . . .	24
2.4 Adaptive Verfeinerungsstrategien . . . . .	25
2.4.1 A posteriori Fehlerschätzer . . . . .	25
2.5 Einführung in die Strukturmechanik . . . . .	25
<b>3 Variationsungleichungen</b>	<b>26</b>
3.1 Ein Hindernisproblem . . . . .	26
3.1.1 Variationsformulierung für das Hindernisproblem . . .	26
3.1.2 Existenz und Eindeutigkeit der Lösung . . . . .	29
3.1.3 Lösung des Hindernisproblems mittels FEM . . . . .	31
3.2 Kontaktprobleme . . . . .	32
3.2.1 Mathematische Modellierung von Kontaktproblemen .	32
3.2.2 Variationsformulierung für Kontaktprobleme . . . . .	32
3.2.3 Lösung des Kontaktproblems mittels FEM . . . . .	33
<b>4 Ein hierarchischer Fehlerschätzer für Hindernisprobleme</b>	<b>34</b>
4.1 Herleitung eines a posteriori hierarchischen Fehlerschätzers .	34
4.1.1 Diskretisierung . . . . .	34
4.1.2 Lokaler Anteil des Fehlerschätzers . . . . .	34
4.1.3 Oszillationsterme . . . . .	34
4.1.4 Zuverlässigkeit des Fehlerschätzers . . . . .	34
4.1.5 Effektivität des Fehlerschätzers . . . . .	34
4.2 Ein adaptiver Algorithmus . . . . .	34
4.3 Erfüllung einer Saturationseigenschaft . . . . .	34

4.4	Übertragung des Fehlerschätzers auf Kontaktprobleme . . . .	34
<b>5</b>	<b>Implementierung des Fehlerschätzers in Matlab</b>	<b>35</b>
<b>6</b>	<b>Validierung</b>	<b>36</b>
6.1	Numerisches Beispiel zum Hindernisproblem . . . . .	36
6.2	Numerisches Beispiel zum Kontaktproblem . . . . .	36
<b>7</b>	<b>Zusammenfassung und Ausblick</b>	<b>37</b>
	<b>Literaturverzeichnis</b>	<b>38</b>
<b>A</b>	<b>Funktionalanalysis</b>	<b>40</b>
A.1	Sobolev-Räume . . . . .	40
A.2	Optimalitätskriterien . . . . .	42
<b>B</b>	<b>Optimierung</b>	<b>43</b>
B.1	Quadratische Programmierung . . . . .	43
B.2	Active Set-Methode für konvexe QPs . . . . .	44
B.3	Algorithmus . . . . .	48
<b>C</b>	<b>Quellcode</b>	<b>49</b>
C.1	Implementierung des Fehlerschätzers für das Hindernisproblem	49
	<b>Index</b>	<b>49</b>

# Abbildungsverzeichnis

# Tabellenverzeichnis

2.1	Ableitungen der nodalen Basisfunktion $\phi_5$ . . . . .	23
-----	--	----

# Kapitel 1

## Einleitung

- Thema (worum geht es?) → Fehlerabschätzung → analytische Lösung oftmals nicht bekannt und damit Fehlerschätzer interessant
- in FEM soll Lösung genauer mit weniger Rechenzeit sein, daraus folgt Anwendung adaptiver Verfahren mit verschiedenen Fehlerschätzern
- Lücke zum neuen (Kontaktproblematik) füllen in dieser Arbeit
- Übertragung unseres Fehlerschätzers auf Kontaktprobleme, wie und warum?! → möglicher Grund: Hindernisprobleme beinhalten Kontaktbereiche (später für Kapitel 4 interessant)
- Struktur der Arbeit

## Kapitel 2

# Grundlagen

In diesem Kapitel wollen wir uns mit grundlegender Theorie beschäftigen, die nicht im Anhang aufgeführt ist, zum Verständnis von den darauffolgenden Kapiteln jedoch notwendig ist.

### 2.1 Hilberträume

- benötigen in den Variationsformulierungen immer wieder Hilberträume, daher werden Eigenschaften dieser hier nochmal eingeführt

- 

**Definition 2.1.** Ein *Hilbertraum* ist ein reeller oder komplexer Vektorraum  $H$  mit Skalarprodukt  $(\cdot, \cdot)_H$ , der vollständig bzgl. der durch das Skalarprodukt induzierten Norm,  $\|v\|_H^2 := (v, v)_H$  für alle  $v \in H$ , ist, d.h. in dem jede Cauchy-Folge konvergiert.

- Es sei in diesem Kapitel  $H$  ein reeller Hilbertraum mit Skalarprodukt  $(\cdot, \cdot)_H$  und der dazu induzierten Norm  $\|v\|_H^2 = (v, v)_H$  für alle  $v \in H$ .

- 

*Bemerkung.* Für alle  $v, w \in H$  gilt die Cauchy-Schwarz'sche Ungleichung

$$(v, w)_H \leq \|v\|_H \|w\|_H.$$

- 

**Satz 2.2** (Approximationssatz). *Es sei  $\emptyset \neq M \subset H$  konvex und abgeschlossen. Dann existiert für alle  $v \in H$  ein  $m_v \in M$  mit*

$$\|v - m_v\| = \text{dist}(v, M) := \inf_{w \in M} \|v - w\|.$$

*Wir nennen  $P_M : H \rightarrow M$  mit  $v \mapsto m_v$  die Projektionen auf  $M$ .*

## 2. Grundlagen

---

*Beweis.* Der Beweis ist in [Wal11] Kapitel 7.1 Satz 7.2 zu finden.  $\square$

•

**Satz 2.3** (Charakterisierung der Projektionen).  $\emptyset \neq M \subset H$  sei abgeschlossen und konvex und  $v \in H$ . Dann gilt:

$$m_0 = P_M(v) \iff (m - m_0, v - m_0)_H \leq 0$$

für alle  $m \in M$ .

*Beweis.* Es sei o.B.d.A.  $0 \in M$  und  $m_0 = 0$ .

„ $\Rightarrow$ “ Wegen  $0 = P_M(x)$  muss  $\|v - tm\|_H \geq \|v\|_H$  für  $m \in M$  und  $0 \leq t \leq 1$  sein. Dann ist

$$\|v\|_H^2 \leq \|v\|_H^2 - 2t(v, m)_H + t^2\|m\|_H^2 \implies 0 \leq -2t(v, m)_H + \underbrace{t^2\|m\|_H^2}_{\geq 0}.$$

Damit ist  $2(v, m)_H \leq 0$ .

„ $\Leftarrow$ “ Für alle  $m \in M$  ist  $(v, m)_H \leq 0$ . Es folgt

$$\|v\|_H^2 \leq \|v\|_H^2 + \|m\|^2 - 2(v, m)_H = \|v - m\|_H^2.$$

Wegen  $0 \in M$  ist  $\text{dist}(v, M) = \|v\|_H^2$  und damit  $0 = P_M(v)$ .  $\square$

•

**Satz 2.4.** Es sei  $\emptyset \neq M \subset H$  konvex und abgeschlossen. Dann gilt:

$$\|P_M(v) - P_M(w)\|_H \leq \|v - w\|_H \quad \forall v, w \in H.$$

*Beweis.* Da  $P_M(v), P_M(w) \in M$  für alle  $v, w \in H$  ist, folgt aus Satz 2.3

$$(P_M(w) - P_M(v), v - P_M(v))_H \leq 0, \quad (2.1)$$

$$(P_M(v) - P_M(w), w - P_M(w))_H \leq 0. \quad (2.2)$$

Addieren wir (2.1) und (2.2), so erhalten wir

$$\begin{aligned} 0 &\geq (P_M(w) - P_M(v), v - P_M(v))_H + (P_M(v) - P_M(w), w - P_M(w))_H \\ &= (P_M(w) - P_M(v), v - w + P_M(w) - P_M(v))_H \\ &= \|P_M(w) - P_M(v)\|_H^2 - (P_M(w) - P_M(v), w - v)_H \\ &\stackrel{\text{CS}}{\geq} \|P_M(w) - P_M(v)\|_H^2 - \|P_M(w) - P_M(v)\|_H \|w - v\|_H. \end{aligned}$$

Nach Umstellen der Ungleichung folgt die Behauptung.  $\square$



•

**Definition 2.5.** Es sei  $\emptyset \neq M \subset H$  und wir definieren das *orthogonale Komplement* von  $M$  durch

$$M^\perp := \{v \in H \mid v \perp M\} := \{v \in H \mid (v, m)_H = 0 \forall m \in M\}.$$

•

**Satz 2.6.** Es sei  $M$  ein abgeschlossener Untervektorraum von  $H$ . Dann ist

$$H = M \oplus M^\perp,$$

d.h. jedes  $v \in H$  hat eine eindeutige Zerlegung  $v = v_M + v_{M^\perp}$  mit  $v_M \in M$  und  $v_{M^\perp} \in M^\perp$ .

*Beweis.* Der Beweis findet sich in [Wal11] Kapitel 7.1 Theorem 7.6.  $\square$

•

**Korollar 2.7.** Es sei  $\emptyset \neq M \subset H$  ein Untervektorraum. Dann ist  $\overline{M} = H$  genau dann, wenn  $M^\perp = \{0\}$  ist.

*Beweis.* Man kann zeigen, dass  $\overline{\text{span } M} = (M^\perp)^\perp =: M^{\perp\perp}$  ist und dann unter Verwendung von Satz 2.6 die Behauptung folgern. Den kompletten Beweis können wir in [Wal11] Kapitel 7.1 Korollar 7.7 (iii) einsehen.  $\square$

## 2.2 Variationsformulierung

Stichpunkte für die Formulierung:

- Betrachte als Modellproblem Auslenkung  $u : \Omega \rightarrow \mathbb{R}$  einer in  $\Omega \subset \mathbb{R}^d$  eingespannten Membran unter Kraft  $f$
- mathematisch beschrieben wird dies durch das *Dirichlet-Problem*

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u &= g \text{ auf } \partial\Omega, \end{aligned} \tag{2.3}$$

- in der Praxis  $d = 2, 3$  übliche Dimensionen
- Richtiger Punkt:

*Notation.* der Einfachheit halber sei im Folgenden  $d = 2$  und  $\Omega \subset \mathbb{R}^2$  ein durch ein Polygonzug berandetes Gebiet, den Rand  $\partial\Omega$  bezeichnen wir mit  $\Gamma$ .

## 2. Grundlagen

---

- allgemeiner berandete Gebiete können durch polygonale beliebig genau approximiert werden
- Transformation: Sei  $u_0 : \Omega \rightarrow \mathbb{R}$  eine zulässige Funktion, d.h. deren Regularität für (2.3) ausreichend ist, und für die  $u_0 = g$  auf  $\Gamma$  gilt. Dann gilt für  $\tilde{u} = u - u_0$

$$\begin{aligned} -\Delta \tilde{u} &= \tilde{f} \text{ in } \Omega, \\ \tilde{u} &= 0 \text{ auf } \Gamma \end{aligned} \quad (2.4)$$

mit  $\tilde{f} = f - \Delta u_0$ .

- $\Rightarrow$  wir beschränken uns auf das *homogene Dirichlet-Problem* (2.4), d.h. sei  $g \equiv 0$  in (2.3)
- Sei im Folgenden  $H_0^1(\Omega)$  wie in Bemerkung A.8 der Raum der schwach differenzierbaren Funktionen, die am Rand  $\Gamma$  verschwinden im Sinne der Spur.
- für  $v \in H_0^1(\Omega)$  gilt dann mit (2.3)

$$\int_{\Omega} -\Delta u \cdot v \, dx = \int_{\Omega} f v \, dx.$$

Betrachte also (2.3) im Mittel über das ganze Gebiet  $\Omega$ . Durch Anwenden der 1. Green'schen Formel (bzw. Satz von Gauß) ergibt sich

$$\begin{aligned} \int_{\Omega} \nabla u \cdot \nabla v \, dx - \underbrace{\int_{\Gamma} v \partial_{\nu} u \, ds}_{=0, \text{ da } v|_{\Gamma}=0} &= \int_{\Omega} f v \, dx \\ \iff \int_{\Omega} \nabla u \cdot \nabla v \, dx &= \int_{\Omega} f v \, dx \end{aligned} \quad (2.5)$$

- kurz geschrieben ist (2.5) mit der Notation aus Satz A.5 (b)

$$(\nabla u, \nabla v)_0 = (f, v)_0.$$

- wir definieren die Bilinearform  $a : (H_0^1(\Omega))^2 \rightarrow \mathbb{R}$ ,  $a(u, v) := (\nabla u, \nabla v)_0$  und  $(f, v) := (f, v)_0$ .

**Definition 2.8.** Eine Funktion  $u \in H_0^1(\Omega)$  heißt *schwache Lösung* vom homogenen Dirichlet-Problem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega, \\ u &= 0 \text{ auf } \Gamma, \end{aligned} \quad (\text{DP})$$

wenn die Gleichung

$$a(u, v) = (f, v) \quad \forall v \in H_0^1(\Omega) \quad (2.6)$$

gilt.

- Wir betrachten im folgenden alle Hilberträume über  $\mathbb{R}$ .
- Frage nach der Existenz und Eindeutigkeit einer schwachen Lösung für (DP)  $\Rightarrow$  hierfür wird ein Hilbertraum benötigt (nachher im Beweis ersichtlich)  $\rightarrow$  Lösung liefert der Satz von Lax-Milgram.
- zuvor noch eine Definition.

**Definition 2.9.** Sei  $H$  ein Hilbertraum. Die Bilinearform  $a : H \times H \rightarrow \mathbb{R}$  heißt *stetig*, falls mit einem  $c > 0$

$$|a(u, v)| \leq c \|u\|_H \|v\|_H \quad \forall u, v \in H$$

gilt. Sie heißt *H-elliptisch* (oder kurz *elliptisch* oder *koerziv*), falls es ein  $\alpha > 0$  gibt, so dass

$$a(v, v) \geq \alpha \|v\|_H^2 \quad \forall v \in H$$

gilt.

- Bevor Existenz der Lösung gezeigt, betrachte Funktional  $J(v) = \frac{1}{2}a(v, v) - F(v)$  genauer
- 

**Lemma 2.10.** *Es sei  $H$  ein Hilbertraum. Das Funktional*

$$J : H \rightarrow \mathbb{R}, \quad J(v) := \frac{1}{2}a(v, v) - F(v),$$

*wobei  $a : H \times H \rightarrow \mathbb{R}$  eine stetige bilineare koerzive und  $F : H \rightarrow \mathbb{R}$  eine lineare Abbildung ist, ist konvex.*

*Beweis.* Es seien  $u, v \in H$ , dann gilt  $u + t(v - u) = (1 - t)u + tv \in H$  (dies gilt auch, wenn wir den Satz auf eine konvexe Teilmenge  $M \subset H$

beschränken). Damit folgt mit  $t \in [0, 1]$

$$\begin{aligned}
 J((1-t)u + tv) &= \frac{1}{2}a((1-t)u + tv, (1-t)u + tv) - F((1-t)u + tv) \\
 &= (1-t)J(u) + tJ(v) + \frac{1}{2}a((1-t)u + tv, (1-t)u + tv) \\
 &\quad - \frac{1}{2}(1-t)a(u, u) - \frac{1}{2}ta(v, v) \\
 &= (1-t)J(u) + tJ(v) + \frac{1}{2}a(u, u) + ta(u, v-u) \\
 &\quad + \frac{t^2}{2}a(v-u, v-u) - \frac{1}{2}(1-t)a(u, u) - \frac{1}{2}ta(v, v) \\
 &= (1-t)J(u) + tJ(v) + \frac{t^2}{2}a(v-u, v-u) \\
 &\quad + \underbrace{ta(u, v) - \frac{1}{2}ta(u, u) - \frac{1}{2}ta(v, v)}_{= -\frac{1}{2}ta(v-u, v-u)} \\
 &= (1-t)J(u) + tJ(v) - \frac{1}{2}\underbrace{t(1-t)}_{\geq 0} \underbrace{a(v-u, v-u)}_{\geq \alpha\|v-u\|_H^2 \geq 0} \\
 &\leq (1-t)J(u) + tJ(v).
 \end{aligned}$$

Daraus folgt die Behauptung.  $\square$

•

**Lemma 2.11.** *Sei  $H$  ein Hilbertraum. Das Funktional  $J : H \rightarrow \mathbb{R}$ ,  $J(v) = \frac{1}{2}a(v, v) - F(v)$  aus Lemma 2.10 ist Gâteaux-differenzierbar (s. Definition A.9).*

*Beweis.* Wir rechnen einfach nach, dass der Grenzwert des Differenzenquotienten existiert und verwenden dabei die Bilinearität von  $a$  und Linearität von  $F$ . Seien  $u, v \in H$ , dann gilt

$$\begin{aligned}
 \mathcal{D}_v J(u) &= \lim_{t \rightarrow 0} \frac{J(u + tv) - J(u)}{t} \\
 &= \lim_{t \rightarrow 0} \frac{J(u) + t(a(u, v) - F(v)) + \frac{t^2}{2}a(v, v) - J(u)}{t} \\
 &= \lim_{t \rightarrow 0} (a(u, v) - F(v)) + \frac{t}{2}a(v, v) \\
 &= a(u, v) - F(v) < \infty,
 \end{aligned}$$

da  $a$  und  $F$  jeweils stetig sind und daher durch  $\|u\|_H, \|v\|_H$  beschränkt sind. Damit folgt die Behauptung.  $\square$

•

**Theorem 2.12.** (Lax-Milgram) *Es sei  $H$  ein Hilbertraum und  $a : H \times H \rightarrow \mathbb{R}$  eine symmetrische, in  $H$  stetige, koerzive Bilinearform. Weiter sei  $F : H \rightarrow \mathbb{R}$  ein stetiges lineares Funktional, d.h.*

$$|F(v)| \leq c \|v\|_H \quad \forall v \in H$$

*mit einer Konstante  $c > 0$ . Dann gibt es eine eindeutige Lösung  $u \in H$ , für die*

$$a(u, v) = F(v) \quad \forall v \in H.$$

*gilt. Diese minimiert den Ausdruck*

$$J(v) = \frac{1}{2}a(v, v) - F(v)$$

*unter allen  $v \in H$ .*

*Beweis.* (i) Zunächst zeigen wir die Äquivalenz der beiden oberen Probleme.

„ $\Rightarrow$ “ Es sei  $u \in H$ , so dass  $a(u, v) = F(v) \forall v \in H$ . Für  $t > 0$  und  $v \in H$  gilt dann

$$\begin{aligned} J(u + tv) &= \frac{1}{2}a(u + tv, u + tv) - F(u + tv) \\ &= \frac{1}{2}a(u, u) + t a(u, v) + \frac{t^2}{2}a(v, v) - F(u) - t F(v) \\ &= \frac{1}{2}a(u, u) - F(u) + t \underbrace{(a(u, v) - F(v))}_{=0} + \frac{t^2}{2} \underbrace{a(v, v)}_{\substack{\geq 0, \text{ da } a \\ \text{koerziv}}} \\ &> \frac{1}{2}a(u, u) - F(u) = J(u), \end{aligned}$$

also ist  $u = \arg \min_{v \in H} J(v)$ .

„ $\Leftarrow$ “ Es sei  $u \in H$  das Minimum von dem Problem

$$\min_{v \in H} J(v) = \frac{1}{2}a(v, v) - F(v).$$

Da  $J : H \rightarrow \mathbb{R}$  nach Lemma 2.10 ein konvexes Funktional ist und  $J$  nach Lemma 2.11 Gâteaux-differenzierbar, gilt mit Satz A.10 für alle  $v \in H$

$$\begin{aligned} 0 = \mathcal{D}_v J(u) &= \left. \frac{d}{dt} J(u + tv) \right|_{t=0} \\ &= \left. \frac{d}{dt} (J(u) + t(a(u, v) - F(v)) + \frac{t^2}{2}a(v, v)) \right|_{t=0} \\ &= a(u, v) - F(v) + t a(v, v) \Big|_{t=0} = a(u, v) - F(v) \end{aligned}$$

## 2. Grundlagen

---

(ii) Eindeutigkeit: Es seien  $u, \tilde{u} \in H$  Lösungen der Variationsungleichung, d.h.

$$a(u, v) = F(v) \wedge a(\tilde{u}, v) = F(v) \quad \forall v \in H.$$

Damit folgt durch Subtraktion der beiden Gleichungen für alle  $v \in H$

$$a(u, v) = a(\tilde{u}, v) \iff a(u - \tilde{u}, v) = 0. \quad (2.7)$$

Da  $H$  ein Vektorraum ist, gilt auch  $u - \tilde{u} \in H$ . Ersetzen wir also in (2.7)  $v = u - \tilde{u}$ , dann ergibt sich

$$0 = a(u - \tilde{u}, u - \tilde{u}) \stackrel{a \text{ koerziv}}{\geq} \underbrace{\alpha}_{>0} \|u - \tilde{u}\|_H^2 \geq 0 \implies \|u - \tilde{u}\|_H^2 = 0,$$

also folgt  $u = \tilde{u}$ .

(iii) Existenz: Die Existenz einer Lösung weisen wir über das Funktional nach.

$$\begin{aligned} J(v) &= \frac{1}{2}a(v, v) - F(v) \stackrel{a \text{ koerziv}}{\underset{F \text{ linear}}{\geq}} \frac{1}{2}\alpha\|v\|_H^2 - c\|v\|_H \\ &= \frac{1}{2}\alpha \left( \|v\|_H^2 - \frac{2c}{\alpha}\|v\|_H \right) = \frac{1}{2}\alpha \left( \|v\|_H - \frac{c}{\alpha} \right)^2 - \frac{c^2}{2\alpha} \\ &\geq -\frac{c^2}{2\alpha} \end{aligned}$$

Folglich ist  $J$  nach unten beschränkt. Sei  $\eta := \inf\{J(v) \mid v \in H\}$  und  $(v_n)_{n \in \mathbb{N}}$  eine Folge mit  $J(v_n) \rightarrow \eta$  für  $n \rightarrow \infty$ . Dann folgt mit der Koerzivität von  $a$

$$\begin{aligned} \alpha\|v_n - v_m\|_H^2 &\leq a(v_n - v_m, v_n - v_m) \\ &= a(v_n, v_n) + a(v_m, v_m) - a(v_n, v_m) - a(v_m, v_n) \\ &= 2a(v_n, v_n) + 2a(v_m, v_m) - \underbrace{a(v_n, v_n + v_m) - a(v_m, v_n + v_m)}_{=-a(v_n + v_m, v_n + v_m)} \\ &= 2a(v_n, v_n) - 4F(v_n) + 2a(v_m, v_m) - 4F(v_m) \\ &\quad - a(v_n + v_m, v_n + v_m) + 4F(v_n + v_m) \\ &= 4J(v_n) + 4J(v_m) - 4a\left(\frac{v_n + v_m}{2}, \frac{v_n + v_m}{2}\right) + 8F\left(\frac{v_n + v_m}{2}\right) \\ &= 4J(v_n) + 4J(v_m) - 8J\left(\frac{v_n + v_m}{2}\right) \\ &\leq 4J(v_n) + 4J(v_m) - 8\eta \xrightarrow{n, m \rightarrow \infty} 4\eta + 4\eta - 8\eta = 0, \end{aligned}$$

d.h.  $(v_n)_{n \in \mathbb{N}}$  ist eine Cauchy-Folge. Da  $H$  ein Hilbertraum ist, gilt somit:  $\exists u \in H : v_n \xrightarrow{n \rightarrow \infty} u$  mit  $J(u) = \eta$ .  $\square$

•

**Satz 2.13.** (Poincaré-Friedrich-Ungleichung) *Es sei  $\Omega$  in einem  $d$ -dimensionalen Würfel der Kantenlänge  $s > 0$  enthalten. Dann gilt*

$$\|v\|_0 \leq s \|\nabla v\|_0 \quad \forall v \in H_0^1(\Omega),$$

wobei  $\|\cdot\|_0$  die durch das Skalarprodukt  $(\cdot, \cdot)_0$  induzierte Norm ist.

*Beweis.* Der Beweis ist in [Bra13] Kapitel II, §1 Sobolev-Räume, Satz 1.5 oder [Sta08] Satz 1.5 zu finden.  $\square$

- Greifen wieder die Frage auf, ob das Problem (2.6) mit  $a : (H_0^1(\Omega))^2 \rightarrow \mathbb{R}, a(u, v) = (\nabla u, \nabla v)_0$  und  $F : H_0^1(\Omega) \rightarrow \mathbb{R}, F(v) := (f, v)$  eine eindeutige Lösung hat.
- Kann nun mit Theorem 2.12 beantwortet werden. Es seien  $u, v \in H_0^1(\Omega)$ , dann gilt

$$\begin{aligned} a(v, v) &= \int_{\Omega} \nabla v \nabla v \, dx = \|\nabla v\|_0^2 \\ &\geq \frac{s^2 + 1}{(1 + s)^2} \|\nabla v\|_0^2 \stackrel{\text{Satz 2.13}}{\geq} \frac{1}{(1 + s)^2} (\|v\|_0^2 + \|\nabla v\|_0^2) \\ &= \frac{1}{(1 + s)^2} \|v\|_1^2. \end{aligned}$$

Damit ist  $a$  mit  $\alpha := \frac{1}{(1+s)^2}$  koerziv. Weiter rechnen wir nach:

$$\begin{aligned} |a(u, v)| &= \left| \int_{\Omega} \nabla u \nabla v \, dx \right| \leq \sum_{i=1}^d \int_{\Omega} |\partial_i u| |\partial_i v| \, dx \\ &\stackrel{\text{CS}}{\leq} \sum_{i=1}^d \left( \int_{\Omega} |\partial_i u|^2 \, dx \right)^{\frac{1}{2}} \left( \int_{\Omega} |\partial_i v|^2 \, dx \right)^{\frac{1}{2}} \\ &\leq \left( \sum_{i=1}^d \int_{\Omega} |\partial_i u|^2 \, dx \right)^{\frac{1}{2}} \left( \sum_{i=1}^d \int_{\Omega} |\partial_i v|^2 \, dx \right)^{\frac{1}{2}} \\ &\leq \left( \int_{\Omega} |\nabla u|^2 \, dx + \int_{\Omega} u^2 \, dx \right)^{\frac{1}{2}} \left( \int_{\Omega} |\nabla v|^2 \, dx + \int_{\Omega} v^2 \, dx \right)^{\frac{1}{2}} \\ &= \|u\|_1 \|v\|_1, \end{aligned}$$

d.h.  $a$  ist stetig mit  $c := 1$ . Die Symmetrie von  $a$  ist trivial, also bleibt nur noch die Stetigkeit von  $F$  zu zeigen. Es sei  $v \in H_0^1(\Omega)$ , dann gilt

$$\begin{aligned} |F(v)| &= |(f, v)| = \left| \int_{\Omega} f v \, dx \right| \stackrel{\text{CS}}{\leq} \left( \int_{\Omega} |f|^2 \, dx \right)^{\frac{1}{2}} \left( \int_{\Omega} |v|^2 \, dx \right)^{\frac{1}{2}} \\ &\leq c \left( \int_{\Omega} |\nabla v|^2 + |v|^2 \, dx \right)^{\frac{1}{2}} = c \|v\|_1 \end{aligned}$$

## 2. Grundlagen

---

mit  $0 < c := \int_{\Omega} |f|^2 dx < \infty$ , wenn  $f \in L_2(\Omega)$  ist. Damit ist  $F$  ein stetiges lineares Funktional und somit existiert nach Theorem 2.12 eine eindeutige Lösung  $u \in H_0^1(\Omega)$  für die schwache Formulierung des homogenen Dirichlet-Problems.

- Weiter minimiert die Lösung  $u \in H_0^1(\Omega)$  das Funktional

$$J(v) = \frac{1}{2} \int_{\Omega} \nabla v \nabla v dx - \int_{\Omega} f v dx ,$$

welches die gespeicherte Energie der durch die Kraft  $f$  belasteten Membran  $\Omega$  beschreibt.

•

*Bemerkung.* Die Stetigkeit vom Funktional  $F$  zeigt, welche Eigenschaft die Kraft  $f$  aus dem Dirichlet-Problem wenigstens quadratisch integrierbar sein muss, damit es eine schwache Lösung geben kann.

•

*Bemerkung.* (a) Mit  $H'$  bezeichnen wir den Dualraum zu einem Hilbertraum  $H$ .

(b) Den Dualraum zu  $H^1(\Omega)$  bezeichnen wir mit  $H^{-1}(\Omega)$ .

- Hier noch eine Folgerung aus dem Satz von Lax-Milgram:

**Satz 2.14.** (Riesz'scher Darstellungssatz) *Es sei  $H$  ein Hilbertraum mit einem Skalarprodukt  $(\cdot, \cdot)_H$ . Es sei  $F \in H'$ , dann existiert genau ein  $u \in H$ , so dass*

$$(u, v)_H = F(v) \quad \forall v \in H .$$

*Beweis.* Dies ist eine direkte Folgerung aus dem Theorem 2.12. Die Abbildung  $(\cdot, \cdot)_H : H \times H \rightarrow \mathbb{R}$  ist als Skalarprodukt bilinear, symmetrisch und positiv definit, damit auch bzgl. der auf  $H$  durch das Skalarprodukt induzierten Norm  $\|v\|_H := \sqrt{(v, v)_H}$ , koerziv.  $F$  ist als Element des Dualraumes  $H'$  eine lineare stetige Abbildung  $F : H \rightarrow \mathbb{R}$  und damit folgt mit  $a(\cdot, \cdot) := (\cdot, \cdot)_H$  aus dem Theorem von Lax-Milgram die Behauptung.  $\square$

•

**Korollar 2.15.** *Es sei  $H$  ein Hilbertraum mit Skalarprodukt  $(\cdot, \cdot)_H$  und  $a : H \times H \rightarrow \mathbb{R}$  eine stetige koerzive Bilinearform. Dann existiert genau ein linearer Operator  $A : H \rightarrow H$ , so dass gilt:*

$$a(u, v) = (Au, v)_H \quad \forall u, v \in H .$$



## 2. Grundlagen

---

*Beweis.* Es sei  $u \in H$  fest, dann ist  $L : H \rightarrow \mathbb{R}, L(v) := a(u, v)$  eine lineare Abbildung, die stetig ist, da

$$|L(v)| = |a(u, v)| \stackrel{\text{stetig}}{\leq} c \|u\|_H \|v\|_H = \tilde{c} \|v\|_H$$

mit  $0 < \tilde{c} := c \|u\|_H$  gilt. Damit folgt nach dem Darstellungssatz von Riesz, dass es ein eindeutiges  $l \in H$  gibt, so dass

$$a(u, v) = L(v) = (l, v)_H \quad \forall v \in H$$

gilt. Da  $u \in H$  jedoch beliebig ist, bleibt zu zeigen, dass es ein eindeutiges  $A : H \rightarrow H$  gibt, so dass  $Au = l$  ist.

Wir zeigen zunächst mithilfe der Bilinearform  $a$ , dass  $A$  linear ist. Es gilt für  $\lambda, \mu \in \mathbb{R}$  und  $u, v \in H$

$$\begin{aligned} (A(\lambda u + \mu v), w)_H &= a(\lambda u + \mu v, w) = \lambda a(u, w) + \mu a(v, w) \\ &= \lambda (Au, w)_H + \mu (Av, w)_H \\ &= (\lambda Au + \mu Av, w)_H \end{aligned}$$

für alle  $w \in H$ . Weiter gilt

$$\|Au\|_H^2 = (Au, Au)_H = a(u, Au) \stackrel{\text{stetig}}{\leq} c \|u\|_H \|Au\|_H,$$

d.h.  $\|Au\|_H \leq c \|u\|_H$  und damit ist nach [Wer11] Satz II.1.2 der Operator  $A$  stetig.

Betrachten wir den Kern von  $A$ , so ergibt sich

$$\ker A := \{v \in H \mid Av = 0\} = \{0\}, \quad (2.8)$$

denn

$$\alpha \|v\|_H^2 \stackrel{\text{koerziv}}{\leq} a(v, v) = (Av, v)_H \stackrel{\text{CS}}{\leq} \|Av\|_H \|v\|_H$$

und damit gilt  $\|Av\|_H \geq \alpha \|v\|_H$ , d.h.  $Av = 0 \Leftrightarrow v = 0$ . Dies impliziert, dass  $A$  injektiv ist, denn mit  $v_1, v_2 \in H, Av_1 = Av_2$  folgt

$$0 = Av_1 - Av_2 = A(v_1 - v_2) \stackrel{(2.8)}{\implies} v_1 = v_2.$$

Weiter betrachten wir das Bild von  $A$ , d.h.

$$\text{im } A := \{v \in H \mid \exists u \in H : Au = v\} \subset H.$$

Sei  $(v_n)_{n \in \mathbb{N}}$  eine Folge mit  $v_k \in \text{im } A$  für alle  $k \in \mathbb{N}$ . Dann folgt, dass für jedes  $v_k$  ein  $u_k \in H$  existiert mit  $Au_k = v_k$ . Es gelte, dass  $Au_k = v_k \rightarrow v \in H$  geht, dann folgt

$$\begin{aligned} \alpha \|u_n - u_m\|_H &\leq \|A(u_n - u_m)\|_H = \|Au_n - Au_m\|_H \\ &= \|v_n - v_m\|_H \xrightarrow{n, m \rightarrow \infty} 0, \end{aligned}$$

d.h.  $(u_n)_{n \in \mathbb{N}} \subset H$  ist eine Cauchy-Folge und konvergiert daher in  $H$ . Also existiert ein  $u \in H$  mit  $u_n \rightarrow u$ . Mit der Stetigkeit von  $A$  folgt dann

$$v_n = Au_n \xrightarrow{n \rightarrow \infty} Au = v,$$

d.h.  $v \in \text{im } A$  und damit ist  $\text{im } A$  abgeschlossen. Wir betrachten nun ein  $v \in H$  mit  $v \perp \text{im } A \subset H$ , dann gilt

$$(Au, v)_H = 0 \quad \forall u \in H.$$

Damit folgt mit  $u = v \in H$  oben eingesetzt

$$0 = (Av, v)_H = a(v, v) \geq \alpha \|v\|_H^2 \implies v = 0.$$

Also besteht der zu  $\text{im } A$  orthogonale Raum nur aus dem Nullelement und mit Korollar 2.7 gilt dann  $\text{im } A = \overline{\text{im } A} = H$ . Damit ist  $A$  bijektiv.

Es seien nun  $0 \neq l \in H$  sowie  $A_1, A_2 \in \mathcal{L}(H, H)$  zwei lineare Operatoren mit  $A_1 u = l$  und  $A_2 u = l$ , die nach der obigen Weise konstruiert sind. Dann gilt

$$0 = A_1 u - A_2 u = (A_1 - A_2)u \implies A_1 = A_2,$$

da  $u \neq 0$  und die Summe zweier bijektiver linearer Operatoren wieder bijektiv ist, also ist ein so konstruierter Operator eindeutig.  $\square$

## 2.3 Finite Elemente Methode

- FEM  $\rightarrow$  einleitend ansprechen, dass analytische nicht immer lösbar
- Unter FEM verstehen wir das Galerkin-Verfahren
- Galerkin-Verfahren bedeutet, wir wollen die Variationsgleichung

$$a(u, v) = F(v) \quad \forall v \in H$$

auf einem endlich dimensionalen Unterraum  $V_h \subset H$  lösen, d.h. finde  $u_h \in V_h$ , so dass

$$a(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h. \tag{2.9}$$

•

**Satz 2.16.** *Das „Galerkin-Problem“ hat eine eindeutige Lösung.*

da  $V_h$  als Unterraum von  $H$  auch ein Hilbertraum ist und die Eigenschaften von  $a, F$  weiterhin erfüllt sind, gilt auch hier der Satz von Lax-Milgram, was die Eindeutigkeit und Existenz einer Lösung sichert.

## 2. Grundlagen

---

- ist  $\mathcal{B}_h := \{\phi_1, \dots, \phi_N\}$  eine Basis von  $V_h$ , dann gilt für  $u_h \in V_h$ :

$$\exists! \boldsymbol{\mu} \in \mathbb{R}^N : u_h(x) = \sum_{i=1}^N \mu_i \phi_i(x). \quad (2.10)$$

- da  $F(\cdot), a(u, \cdot)$  linear sind und alle  $v_h \in V_h$  analog zu oben darstellbar sind, ist (2.9) äquivalent zum Problem

$$a(u_h, \phi_i) = F(\phi_i) \quad \forall i = 1, \dots, N,$$

mit  $u_h = \sum \mu_i \phi_i$  eingesetzt ergibt sich

$$a(u_h, \phi_i) = a\left(\sum_{j=1}^N \mu_j \phi_j, \phi_i\right) = \sum_{j=1}^N \mu_j a(\phi_j, \phi_i),$$

also

$$\sum_{j=1}^N \mu_j a(\phi_j, \phi_i) = F(\phi_i) \quad \forall i = 1, \dots, N.$$

Damit ergibt sich das LGS

$$A\boldsymbol{\mu} = \mathbf{f}$$

mit  $A = [a(\phi_j, \phi_i)]_{i,j=1}^N$ ,  $\boldsymbol{\mu} = [\mu_i]_{i=1}^N$  und  $\mathbf{f} = [F(\phi_i)]_{i=1}^N$ .

•

*Bemerkung.* Ist die Bilinearform  $a$  symmetrisch, so ist es auch die Matrix  $A$ , denn

$$a_{ij} = a(\phi_i, \phi_j) = a(\phi_j, \phi_i) = a_{ji}.$$

Außerdem folgt aus der Koerzivität von  $a$ , dass mit  $0 \neq v \in \mathbb{R}^N$  gilt

$$\begin{aligned} v^T A v &= \sum_{i,j=1}^N v_i a_{ij} v_j = \sum_{i=1}^N v_i \sum_{j=1}^N a(\phi_i, \phi_j) v_j \\ &= \sum_{i=1}^N v_i a\left(\phi_i, \sum_{j=1}^N v_j \phi_j\right) = a\left(\sum_{i=1}^N v_i \phi_i, \sum_{j=1}^N v_j \phi_j\right) \\ &= a(v_h, v_h) \geq \alpha \|v_h\|_H^2 > 0, \end{aligned}$$

da  $v_h \neq 0$  wegen  $v \neq 0$ . Damit ist  $A$  also positiv definit.

- in Ingenieurwissenschaften, insbesondere bei kontinuumsmechanischen Problemen, wird  $A$  als Steifigkeitsmatrix bezeichnet.

- um eine Basis  $\mathcal{B}_h$  bzgl.  $V_h$  beschreiben zu können, muss das Gebiet  $\Omega$  in endliche Elemente zerlegt werden.  $V_h$  wird dann bzgl. einer Zerlegung  $\mathcal{T}_h$  beschrieben.
- betrachte im weiteren (wie oben schon angesprochen)  $\Omega$  als zweidimensionales Gebiet, das mit einem Polygonzug berandet ist.
- eine gebräuchliche Zerlegung  $\mathcal{T}_h$  kann durch Dreiecke oder auch Vierecke geschehen. Wir wollen hier nur Zerlegungen durch Dreiecke betrachten
- hierfür führen wir den Begriff der Triangulierung ein (vgl. [Bra13] Seite 58 oder [Sta08] Seite 19)

**Definition 2.17** (Triangulierung). Es sei  $\Omega \subset \mathbb{R}^2$  ein durch einen Polygonzug berandetes Gebiet. Dann heißt eine Zerlegung aus Dreiecken

$$\mathcal{T} = \{T_1, T_2, \dots, T_M\}$$

*Triangulierung*, wenn gilt:

- Für alle Dreiecke  $T \in \mathcal{T}$  gilt:  $T$  ist abgeschlossen.
- Ganz  $\Omega$  wird durch alle Dreiecke aus  $\mathcal{T}$  überdeckt, d.h.  $\bar{\Omega} = \bigcup_{T \in \mathcal{T}} T$ .
- Der Schnitt zweier Dreiecke  $T_i \cap T_j$  mit  $i \neq j$  überlappt sich nicht, d.h.  $\text{int}(T_i) \cap \text{int}(T_j) = \emptyset$ .

Wir nennen eine Triangulierung *konform* oder *zulässig*, wenn zusätzlich gilt:

- Für jede Kante  $k$  eines Dreiecks  $T \in \mathcal{T}$  gilt entweder  $k \subset \partial\Omega$  oder  $k \subset \tilde{T}$  für ein weiteres Dreieck  $\tilde{T} \in \mathcal{T}$ .

Der Radius des Umkreises eines Dreiecks  $T$  wird mit  $h$  bezeichnet und beschreibt die Größe eines Dreiecks. Wenn jedes Dreieck  $T \in \mathcal{T}$  höchstens einen Radius von  $h$  hat, so schreiben wir  $\mathcal{T}_h$  statt  $\mathcal{T}$ .

Skizze von einer zulässigen und einer nicht zulässigen Triangulierung (hierbei eine Skizze mit hängenden Knoten machen)

•

**Bemerkung 2.18.** natürlich auch im  $\mathbb{R}^3$  analog mit Tetraedern definierbar

- vgl. wieder [Bra13] Seite 58

**Definition 2.19** ((quasi-) uniforme Zerlegung). Eine Familie von Zerlegungen  $\{\mathcal{T}_h\}$  heißt *quasi-uniform*, wenn es eine Zahl  $\kappa > 0$  gibt, so dass jedes  $T \in \mathcal{T}_h$  einen Kreis vom Radius

$$\rho_T \geq \frac{h_T}{\kappa}$$

enthält, wobei  $h_T$  der Radius des Dreiecks  $T$  ist.

Eine Familie von Zerlegungen  $\{\mathcal{T}_h\}$  heißt *uniform*, wenn es eine Zahl  $\kappa > 0$  gibt, so dass jedes  $T \in \mathcal{T}_h$  einen Kreis vom Radius

$$\rho_T \geq \frac{h}{\kappa}$$

enthält, wobei  $h := \max_{T \in \mathcal{T}_h} h_T$  ist.

Skizze Beispiele für eine quasiuniforme Zerlegung (vgl. [Bra13] Seite 59) → hier noch was zur Erklärung schreiben, wie im Braess

•

**Bemerkung 2.20.** Wie man leicht sehen kann, ist jede uniforme Zerlegung auch quasi-uniform. Umgekehrt gilt dies nicht (s. Abbildung oben).

Allerdings lassen uniforme Zerlegungen keine lokalen Verfeinerungen zu. Da dies für adaptive Verfeinerungsstrategien allerdings ausschlaggebend ist, gehen wir im Folgenden immer von einer quasi-uniformen Zerlegung  $\mathcal{T}_h$  aus.

- nun wollen wir uns Gedanken über unseren Ansatzraum  $V_h$  machen.
- hierfür gibt es viele Möglichkeiten, vergleiche hierzu auch [Bra13] Kapitel II, §5, Tabelle 2, durch Konstruktion der Elemente
- wir wollen uns weitestgehend aber nur auf ein Element konzentrieren
- zuvor hierfür ein wichtiges Resultat; sei noch bemerkt, dass eine Fkt.  $u$  auf  $\Omega$  bei gegebener Zerlegung eine Eigenschaft stückweise hat, wenn sie auf jedem Element diese Eigenschaft besitzt.

**Satz 2.21.** Sei  $k \geq 1$  und  $\Omega \subset \mathbb{R}^2$  ein polygonales Gebiet. Eine stückweise beliebig oft differenzierbare Funktion  $v : \bar{\Omega} \rightarrow \mathbb{R}$  liegt in  $H^k(\Omega)$  genau dann, wenn  $v \in C^{k-1}(\bar{\Omega})$  ist.

*Beweis.* Der Beweis ist in [Bra13] Kapitel II, §5, Satz 5.2 zu finden.  $\square$

## 2. Grundlagen

---

- dies rechtfertigt, dass für unser Modellproblem (2.6), welches für  $u, v \in H_0^1(\Omega)$  gestellt ist, auf einer Triangulierung  $\mathcal{T}_h$  ein Ansatzraum mit stetigen Funktionen  $v \in C^0(\Omega)$  verwendet werden kann, also

$$V_h := \{v \in C^0(\Omega) \mid v|_T \in \mathcal{P}_m \text{ für } T \in \mathcal{T}_h, v|_{\partial\Omega} = 0\},$$

wobei  $\mathcal{P}_m$  der Raum der Polynome vom Grad  $m$  ist.

- wie diesen Raum  $V_h$  aufspannen?  $\rightarrow$  die einfachste Möglichkeit solch einen Raum aufzuspannen sind nodale Basisfunktionen

**Definition 2.22** (nodale Basisfunktion). Zu einem Finiten Element Raum  $V_h$  und einer gegebenen Zerlegung  $\mathcal{T}_h$  sei eine Menge von Punkten  $P$  bekannt mit  $|P| = N$ . Die Menge  $\mathcal{B}_h = \{\phi_1, \dots, \phi_N\}$  mit  $\phi_i \in \mathcal{P}_m, i = 1, \dots, N$ , heißt *nodale Basis* (oder *Lagrange-Basis*), wenn

$$\phi_i(x_j) = \delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

für alle  $\phi_i \in \mathcal{B}_h$  und  $x_j \in P$  gilt.

- folgende Bemerkung erklärt die Anordnung der in der letzten Definition beschriebenen Punkte  $P$

**Bemerkung 2.23.** Sei  $m \geq 0$ . In einem Dreieck  $T$  seien auf  $m+1$  Linien  $l = 1+2+\dots+(m+1)$  Punkte  $z_1, \dots, z_l$  angeordnet (s. Skizze). Dann gibt es zu jedem  $C^0(T)$  genau ein Polynom  $p$  vom Grad  $m$  mit der Eigenschaft

$$p(z_i) = f(z_i) \quad \forall i = 1, \dots, m.$$

*Beweis.* Der Beweis steht in [Bra13] Kapitel II, §5, Bemerkung 5.4.  $\square$

Skizze Dreiecke für die nodalen Basen (linear, quadratisch, kubisch).

- damit lässt sich für  $V_h$  mit einem beliebigen Polynomgrad  $m$  eine eindeutige nodale Basis finden, die den Raum aufspannt.
- im weiteren wollen wir lineare Ansatzfunktionen verwenden. Wir bezeichnen, sofern nicht anders beschrieben, also im Folgenden  $V_h$  mit

$$V_h := \{v \in C^0(\Omega) \mid v|_T \in \mathcal{P}_1 \text{ für } T \in \mathcal{T}_h, v|_{\partial\Omega} = 0\}.$$

- Beispiel, um zu sehen, dass auch für kleine Gitter der Rechenaufwand sehr hoch werden kann.

**Beispiel 2.24.** Wir betrachten unser Variationsproblem (2.9) auf  $\Omega = [-1, 1]^2$  mit  $V_h$  wie oben eingeführt als den Raum der linearen Ansatzfunktionen auf einer Zerlegung  $\mathcal{T}_h$  wie unten aufgeführt.

Skizze: mit 8 Courant-Elementen, wie auf in <http://www.math.uni-hamburg.de/home/struckmeier/numpde06/Kap2.pdf> auf Seite 95

Wir stellen für die nodale Basisfunktion  $\phi_5$  die Einträge in der Steifigkeitsmatrix auf. Man rechnet leicht nach, dass

$$\phi_5(x, y) = \begin{cases} 1 - x - y, & \text{auf I} \\ 1 + x, & \text{auf III} \\ 1 - y, & \text{auf IV} \\ 1 + x + y, & \text{auf VI} \\ 1 + y, & \text{auf VII} \\ 1 - x, & \text{auf VIII} \\ 0, & \text{sonst} \end{cases}$$

ist und damit ergeben sich folgende Ableitungen.

	I	II	III	IV	V	VI	VII	VIII
$\partial_x \phi_5$	-1	0	1	0	0	1	0	-1
$\partial_y \phi_5$	-1	0	0	-1	0	1	1	0

Tabelle 2.1: Ableitungen der nodalen Basisfunktion  $\phi_5$ .

Damit rechnen wir nach, dass gilt

$$\begin{aligned} a(\phi_5, \phi_5) &= \int_{\Omega} \nabla \phi_5 \nabla \phi_5 \, dxdy \\ &= \int_{\text{I} \cup \dots \cup \text{VIII}} \underbrace{(\partial_x \phi_5)^2}_{\geq 0} + \underbrace{(\partial_y \phi_5)^2}_{\geq 0} \, dxdy \\ &= 2 \int_{\text{I} \cup \text{III} \cup \text{IV}} (\partial_x \phi_5)^2 + (\partial_y \phi_5)^2 \, dxdy \\ &= 2 \left( \int_{\text{I} \cup \text{III}} \underbrace{(\partial_x \phi_5)^2}_{=1} \, dxdy + \int_{\text{I} \cup \text{IV}} \underbrace{(\partial_y \phi_5)^2}_{=1} \, dxdy \right) \\ &= 2(A(\text{I}) + A(\text{III}) + A(\text{I}) + A(\text{IV})) \\ &= 8 \cdot A(\text{I}) = 8 \cdot \frac{1}{2} = 4, \end{aligned}$$

wobei verwendet wurde, dass die Dreiecke kongruent zueinander sind. Analog können wir auch die übrigen acht nodalen Basisfunktionen

aufstellen und damit die Einträge der Steifigkeitsmatrix

$$\begin{aligned} a(\phi_5, \phi_2) &= a(\phi_5, \phi_4) = a(\phi_5, \phi_6) = a(\phi_5, \phi_8) = -1, \\ a(\phi_5, \phi_1) &= a(\phi_5, \phi_3) = a(\phi_5, \phi_7) = a(\phi_5, \phi_9) = 0 \end{aligned}$$

berechnen. Damit ist der Einteil der Basisfunktion  $\phi_5$  an der Steifigkeitsmatrix  $A$  von der Form

$$\tilde{A} = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{pmatrix}.$$

Hierbei müssen die Einträge aus  $\tilde{A}$  in die Matrix  $A \in \mathbb{R}^{9 \times 9}$  an die richtige Stelle zugeordnet werden, wie durch die Formel  $a_{ij} = a(\phi_i, \phi_j)$  beschrieben wird. Dabei nennen wir  $\tilde{A}$  lokale Steifigkeitsmatrix bzgl. des Knoten 5.

Dieses Vorgehen müssten wir noch für die übrigen Basisfunktion analog durchführen, um die vollständige Steifigkeitsmatrix  $A$  zu erhalten. Dies soll hier aber nicht weiter ausgeführt werden.

- wie man schön erkennt, ist das Vorgehen aus dem obigen Beispiel sehr aufwendig.  $\rightarrow$  außerdem ist es schwer dieses zu verallgemeinern, damit man es gut implementieren kann, da die Ansatzfunktionen auf das Gitter bezogen von individueller Form sind.
- Abhilfe durch local-global node ordering zur Effizienzsteigerung (Erklärung):
- hierbei ist die Idee die lokale Steifigkeitsmatrix für ein Element durch Transformation auf ein Referenzelement zu berechnen und somit die Berechnung von lokalen Steifigkeitsmatrizen zu verallgemeinern

Skizze vom Referenzelement (Stephan NPDE Seite 13)

- Dann die Herleitung von der lokalen Steifigkeitsmatrix (Stephan NPDE Seite 13+14)

### 2.3.1 A priori Fehlerabschätzung

- man kann zeigen, dass der Fehler von  $h$  zwischen  $u_h$  und  $u$  (exakte Lösung) abhängt  $\Rightarrow$  Netzverfeinerung führt zur Konvergenz

$\Rightarrow$  Überleitung zu adaptiven Verfahren



## 2.4 Adaptive Verfeinerungsstrategien

### 2.4.1 A posteriori Fehlerschätzer

- Fehlerschätzer → alle aufführen (s. Braess) → damit verbundene adaptive Verfeinerungsstrategien (wie arbeitet Matlab mit Verfeinerung und welche Verfeinerungen gibt es?)

## 2.5 Einführung in die Strukturmechanik

- Beschreibung der Kinematik: Referenz- bzw. Ausgangskonfiguration, Deformationsgradient, Verzerrungsmaße (Konti-Buch)
- Linearisierung der Verzerrungsmaße für unseren Fall (kleine Deformationen) mittels "Taylor" (siehe auch Gateaux-Ableitung - Seite 24 Konti Skript):

$$\boldsymbol{\varepsilon} = \frac{1}{2}(\nabla \mathbf{u} + \nabla^T \mathbf{u})$$

- Kinetik: Kräftegleichgewicht und äußere Kontaktlast
- Konzepte für ebene Spannungs- bzw. Verzerrungszustände
- Konstitutive Modelle (vor allem Materialgesetze) ⇒ Hier vor allem Hooke:

$$\boldsymbol{\sigma} = \mathcal{C}\boldsymbol{\varepsilon} = 2\mu\boldsymbol{\varepsilon} + \lambda(\text{tr } \boldsymbol{\varepsilon})\mathbf{I},$$

wobei  $\lambda, \mu$  die Lamé-Konstanten sind (Materialabhängige Parameter).  
⇒ Hier noch mal den Zusammengang von Konstanten zu  $E, \nu$  aufzeigen.

- falls Tensorrechnungen konkret benötigt werden, können diese im Anhang dargelegt werden

## Kapitel 3

# Variationsungleichungen

### 3.1 Ein Hindernisproblem

- Hindernisproblem: Auslenkung  $u$  einer Membran  $\Omega$  unter Krafteinwirkung  $f$ , wobei die Membran durch ein Hindernis  $\psi$  behindert wird. Mathematisch modelliert bedeutet dies:

$$\min_{v \in K} J(v) = \frac{1}{2}a(v, v) - (f, v) \quad (3.1)$$

mit  $K := \{v \in H_0^1(\Omega) \mid v \geq \psi \text{ fast überall in } \Omega\}$ .

- $J$  gibt wieder die in der Membran gespeicherte Energie an.
- wobei jetzt die Lösung nicht auf ganz  $H_0^1(\Omega)$  gesucht ist, sondern in einer Teilmenge  $K \subset H_0^1(\Omega)$ .
- wir können auch hier eine Variationsformulierung, die äquivalent zu (3.1) ist, herleiten
- zu Beginn noch eine Skizze von einem Hindernisproblem

#### 3.1.1 Variationsformulierung für das Hindernisproblem

- zeigen zunächst, dass  $K$  konvex und abgeschlossen ist.

**Lemma 3.1.** *Die Menge  $K = \{v \in H_0^1(\Omega) \mid v \geq \psi \text{ fast überall in } \Omega\}$  ist eine konvexe abgeschlossene Teilmenge von  $H_0^1(\Omega)$ .*

*Beweis.* (i) Es seien  $u, v \in K$ , d.h.  $u \geq \psi$  und  $v \geq \psi$  fast überall in  $\Omega$ . Dann gilt für  $t \in [0, 1]$

$$(1-t)u + tv \geq (1-t)\psi + t\psi = \psi,$$

somit ist  $(1-t)u + tv \in K$ , also  $K$  konvex.

(ii) abgeschlossen fehlt noch

□

•

**Satz 3.2.** Es sei  $K = \{v \in H_0^1(\Omega) \mid v \geq \psi \text{ fast überall in } \Omega\}$ . Das Minimierungsproblem

$$\min_{v \in K} J(v) = \frac{1}{2}a(v, v) - (f, v) \quad (3.2)$$

ist äquivalent zur Variationsungleichung: Finde  $u \in K$ , so dass

$$a(u, v - u) \geq (f, v - u) \quad \forall v \in K. \quad (3.3)$$

*Beweis.* Aus Lemma 2.10 folgt, dass  $J$  konvex ist und damit gilt mit Satz A.10, dass  $u \in K$  genau dann eine Lösung von (3.2) ist, wenn

$$\mathcal{D}_{v-u}J(u) \geq 0 \quad \forall v \in K \quad (3.4)$$

gilt. Analog zu der berechneten Gâteaux-Ableitung von  $J$  in Lemma 2.11, gilt

$$\mathcal{D}_{v-u}J(u) = \frac{d}{dt}J(u + t(v - u)) \Big|_{t=0} = a(u, v - u) - (f, v - u)$$

und damit folgt mit (3.4) die Behauptung.  $\square$

•

**Bemerkung 3.3.** Wie man mit Satz A.10 sehen kann, gilt analog zu Satz 3.2 auch allgemeiner: Es sei  $K \subset H$  eine konvexe Teilmenge eines Hilbertraumes  $H$ . Dann ist

$$\min_{v \in K} J(v) = \frac{1}{2}a(v, v) - F(v)$$

äquivalent zur Variationsungleichung: Finde  $u \in K$ , so dass

$$a(u, v - u) \geq F(v - u) \quad \forall v \in K,$$

wobei  $F : H \rightarrow \mathbb{R}$  eine lineare stetige Abbildung ist.

- auch für das Hindernisproblem gibt es analog zum homogenen Dirichlet-Problem (2.4) eine äquivalente starke Formulierung

•

**Satz 3.4** (Starke Formulierung des Hindernisproblems). Jede Lösung  $u \in H^2(\Omega) \cap H_0^1(\Omega)$  des Problems

$$\begin{aligned} -\Delta u - f &\geq 0 \\ u - \psi &\geq 0 \\ (u - \psi)(-\Delta u - f) &= 0 \end{aligned} \quad (3.5)$$

mit  $\psi \in H^1(\Omega)$  erfüllt die Variationsungleichung (3.3). Umgekehrt ist jede Lösung  $u \in H^2(\Omega) \cap K$  von (3.3) auch eine Lösung von (3.5).

### 3. Variationsungleichungen

---

*Beweis.* „ $\Rightarrow$ “ Sei  $u \in H^2(\Omega) \cap H_0^1(\Omega)$  eine Lösung von (3.5), dann gilt für ein beliebiges  $v \in K$

$$\begin{aligned} \int_{\Omega} (-\Delta u - f)(v - u) dx &= \underbrace{- \int_{\Omega} \Delta u (v - u) dx}_{\stackrel{\text{Green}}{=} \int_{\Omega} \nabla u \nabla (v - u) dx - \underbrace{\int_{\Gamma} (v - u) \partial_{\nu} u ds}_{=0}} - \int_{\Omega} f(v - u) dx \\ &= \int_{\Omega} \nabla u \nabla (v - u) dx - \int_{\Omega} f(v - u) dx \\ &= a(u, v - u) - (f, v - u). \end{aligned}$$

Mit  $\Omega_0 := \{x \in \Omega \mid u = \psi\}$  folgt, dass  $-\Delta u = f$  auf  $\Omega_1 := \Omega \setminus \bar{\Omega}_0$  gelten muss.

$$\Rightarrow \int_{\Omega = \Omega_0 \cup \Omega_1} \underbrace{(-\Delta u - f)(v - u)}_{=0 \text{ auf } \Omega_1} dx = \int_{\Omega_0} \underbrace{(-\Delta u - f)}_{\geq 0} \underbrace{(v - \psi)}_{\geq 0} dx \geq 0$$

Damit ist  $u$  eine Lösung von (3.3)

$$a(u, v - u) \geq (f, v - u) \quad \forall v \in K.$$

„ $\Leftarrow$ “ Es sei  $u \in H^2(\Omega) \cap K$  Lösung von (3.3). Weiter sei  $v \in K$  beliebig, dann gilt

$$\begin{aligned} 0 &\leq a(u, v - u) - (f, v - u) \\ &= \int_{\Omega} \nabla u \nabla (v - u) dx - \int_{\Omega} f(v - u) dx \\ &\stackrel{\text{Green}}{=} \int_{\Omega} -\Delta u (v - u) dx - \int_{\Omega} f(v - u) dx \quad (3.6) \\ &= \int_{\Omega} (-\Delta u - f)(v - u) dx. \end{aligned}$$

Wir nehmen an, dass  $-\Delta u - f < 0$  in einem Ball  $B_{r_0} := B_{r_0}(x_0) \subset \Omega$  mit Radius  $r_0$  um  $x_0 \in \Omega$  gilt. Sei weiter  $\chi \in C^\infty(\Omega)$  mit  $\chi = 0$  auf  $\Omega \setminus \bar{B}_{r_0}$ ,  $\rho(r) := \left(1 - \frac{r}{r_0}\right)^2 \chi > 0$  und  $v := u + \rho(r) \in K$ , da  $u \in K$  und  $\rho(r) > 0$ . Dann gilt

$$\int_{\Omega} (-\Delta u - f)(v - u) dx = \int_{B_{r_0}} \underbrace{(-\Delta u - f)}_{<0} \underbrace{\rho(r)}_{>0} dx < 0,$$

was im Widerspruch zu (3.6) steht. Also muss  $-\Delta u - f \geq 0$  gelten.

Nun nehmen wir an, dass  $-\Delta u - f > 0$  und  $u > \psi$  fast überall in einem Ball  $B_{r_0}$  gilt. Wir betrachten  $v := u + \varepsilon \rho(r)(\psi - u) \in K$  mit

$0 < \varepsilon \leq 1$ , dann folgt

$$\int_{\Omega} (-\Delta u - f)(v - u) dx = \varepsilon \int_{B_{r_0}} \underbrace{(-\Delta u - f)}_{>0} \underbrace{\rho(r)}_{>0} \underbrace{(\psi - u)}_{<0} dx < 0,$$

was wiederum im Widerspruch zu (3.6) steht. Damit muss  $u = \psi$  gelten, wenn  $-\Delta u = f$  ist. Es folgt, dass  $u \in H^2(\Omega) \cap K$  eine Lösung von (3.5) ist.  $\square$

### 3.1.2 Existenz und Eindeutigkeit der Lösung

- Kapitel 3 in [KO88] mit Theorem 3.1-3.4 (**Beweis vgl. NPDE I von Stephan Seite 39**, auch in Solution of Variational Inequalities in Mechanics (Theorem 1.1 Seite 4))
- für die Existenz und Eindeutigkeit der Lösung des Problems betrachten wir zunächst wieder das allgemeine reelle quadratische Funktional  $J : H \rightarrow \mathbb{R}, J(v) = \frac{1}{2}a(v, v) - F(v)$ .

•

*Voraussetzungen.* Sei  $H$  ein reeller Hilbertraum mit Skalarprodukt  $(\cdot, \cdot)_H$  und der damit induzierten Norm  $\|\cdot\|_H$ . Mit  $H'$  bezeichnen wir den Dualraum zu  $H$ . Weiter sei vorausgesetzt:

- (a)  $a : H \times H \rightarrow \mathbb{R}$  ist eine stetige koerzive Bilinearform,
- (b)  $F : H \rightarrow \mathbb{R}$  ist ein stetiges lineares Funktional,
- (c)  $K \neq \emptyset$  ist eine abgeschlossene konvexe Teilmenge von  $H$ .

•

**Theorem 3.5** (Existenz und Eindeutigkeit). *Unter den obigen Voraussetzungen hat die Variationsungleichung, finde  $u \in K$ , so dass*

$$a(u, v - u) \geq F(v - u) \quad \forall v \in K \quad (3.7)$$

*ist, genau eine Lösung.*

*Beweis.* (i) Eindeutigkeit: Es seien  $u_1, u_2 \in K$  zwei Lösungen der Variationsungleichung (3.7), d.h.

$$a(u_1, v - u_1) \geq F(v - u_1) \quad \forall v \in K, \quad (3.8)$$

$$a(u_2, v - u_2) \geq F(v - u_2) \quad \forall v \in K. \quad (3.9)$$

### 3. Variationsungleichungen

---

Addieren wir (3.8) und (3.9) miteinander und setzen zuvor  $v = u_2$  in (3.8) und  $v = u_1$  in (3.9), so erhalten wir

$$\begin{aligned} 0 &\leq a(u_1, u_2 - u_1) - F(u_2 - u_1) + a(u_2, u_1 - u_2) - \underbrace{F(u_1 - u_2)}_{=F(u_2-u_1)} \\ &= a(u_1, u_2 - u_1) - a(u_2, u_2 - u_1) = -a(u_2 - u_1, u_2 - u_1) \\ &\leq -\alpha \|u_2 - u_1\|_H^2. \end{aligned}$$

Also gilt  $\|u_2 - u_1\|_H^2 \leq 0 \Rightarrow \|u_2 - u_1\|_H^2 = 0$  und damit folgt  $u_1 = u_2$ .

(ii) Existenz: Aus dem Darstellungssatz von Riesz bzw. das Korollar 2.15 folgt, dass ein  $A \in \mathcal{L}(H, H), l \in H$  existiert, so dass

$$\begin{aligned} a(u, v) &= (Au, v)_H \quad \forall u, v \in H, \\ F(v) &= (l, v)_H \quad \forall v \in H. \end{aligned}$$

Damit gilt

$$\begin{aligned} F(v - u) - a(u, v - u) &= (l, v - u)_H - (Au, v - u)_H \\ &= (l - Au, v - u)_H \leq 0. \end{aligned}$$

Durch Multiplikation mit  $\varrho > 0$  und Addition der Null erhalten wir das äquivalente Problem: Finde  $u \in K$ , so dass

$$(u - \varrho(Au - l) - u, v - u)_H \leq 0 \quad \forall v \in K. \quad (3.10)$$

Nach Satz 2.3 ist  $u$  damit das Bild der Projektion von  $u - \varrho(Au - l)$  auf  $K$ , d.h.

$$u = P_K(u - \varrho(Au - l)).$$

Es bleibt zu zeigen, dass  $W_\varrho : H \rightarrow K, W_\varrho(v) := P_K(v - \varrho(Av - l))$  einen Fixpunkt besitzt. Mit Anwendung von Satz 2.4 und der Koerzivität von  $a$  rechnen wir nach, dass

$$\begin{aligned} \|W_\varrho(v_1) - W_\varrho(v_2)\|_H^2 &= \|P_K(v_1 - \varrho(Av_1 - l)) - P_K(v_2 - \varrho(Av_2 - l))\|_H^2 \\ &\leq \|v_1 - \varrho(Av_1 - l) - (v_2 - \varrho(Av_2 - l))\|_H^2 \\ &= \|(v_1 - v_2) - \varrho A(v_1 - v_2)\|_H^2 \\ &= \|v_1 - v_2\|_H^2 + \varrho^2 \|A(v_1 - v_2)\|_H^2 \\ &\quad - \underbrace{\varrho (A(v_1 - v_2), v_1 - v_2)_H - \varrho (v_1 - v_2, A(v_1 - v_2))_H}_{=2\varrho (A(v_1 - v_2), v_1 - v_2)_H = 2\varrho a(v_1 - v_2, v_1 - v_2)} \\ &\leq \|v_1 - v_2\|_H^2 + \varrho^2 \|A\|^2 \|v_1 - v_2\|_H^2 - 2\varrho\alpha \|v_1 - v_2\|_H^2 \\ &= (1 - 2\varrho\alpha + \varrho^2 \|A\|^2) \|v_1 - v_2\|_H^2 \end{aligned}$$

### 3. Variationsungleichungen

---

mit  $\|A\| := \sup_{v \in H} \frac{\|Av\|_H}{\|v\|_H}$ . Also ist die Abbildung  $W_\varrho$  eine Kontraktion, wenn gilt

$$1 - 2\varrho\alpha + \varrho^2 \|A\|^2 < 1 \implies 0 < \varrho < \frac{2\alpha}{\|A\|^2}.$$

Nach dem Banach'scher Fixpunktsatz (vgl. [Sto99] Satz 5.2.3) existiert für solch ein  $\varrho$  ein  $u \in H$  mit  $u = W_\varrho(u) = P_K(u - \varrho(Au - l))$ .

Insgesamt gibt es also für das Problem (3.7) genau eine Lösung.  $\square$

•

**Korollar 3.6.** *Das Problem (3.1) hat eine eindeutige Lösung.*

*Beweis.* Da laut Lemma 3.1 die Menge

$$K = \{v \in H_0^1(\Omega) \mid v \geq \psi \text{ fast überall in } \Omega\}$$

abgeschlossen und konvex ist,  $F(v) = (f, v)$  ein stetiges lineares Funktional und

$$a(u, v) = \int_{\Omega} \nabla u \nabla v \, dx$$

stetig bilinear und koerziv, sind die Voraussetzungen für Theorem 3.5 erfüllt. Damit hat das Problem, finde  $u \in K$ , so dass

$$a(u, v - u) \geq (f, v - u) \quad \forall v \in K, \quad (3.11)$$

genau eine Lösung. Nach Satz 3.2 ist (3.1) äquivalent zu (3.11) und damit folgt die Behauptung.  $\square$

•

**Bemerkung 3.7.** Insbesondere hat auch das Problem (3.5) nach Satz 3.4 und Theorem 3.5 eine eindeutige Lösung, wenn  $u \in H^2(\Omega) \cap H_0^1(\Omega)$  ist.

#### 3.1.3 Lösung des Hindernisproblems mittels FEM

- Analog zum vorherigen Kapitel kann man auch im  $\mathbb{R}^n$  Existenz und Eindeutigkeit der Lösung unter bestimmten Voraussetzungen zeigen. (vgl. Vug Skript Kapitel 2)  $\Rightarrow$  Beachte hierfür auch den Fixpunktsatz von Brouwer.

## 3.2 Kontaktprobleme

### 3.2.1 Mathematische Modellierung von Kontaktproblemen

- Starke Formulierung (s. Wriggers Paper) für Kontaktproblem mit Signorini-Kontakt (ohne Reibung).

$$\operatorname{div} \boldsymbol{\sigma} + \mathbf{b} = \mathbf{0} \text{ in } \Omega \quad (3.12)$$

$$\boldsymbol{\sigma} - \mathcal{C}\boldsymbol{\varepsilon} = \mathbf{0} \text{ in } \Omega \quad (3.13)$$

$$\boldsymbol{\sigma} \cdot \mathbf{n} = \mathbf{t} \text{ auf } \Gamma_N \quad (3.14)$$

$$\mathbf{u} = \mathbf{0} \text{ auf } \Gamma_D \quad (3.15)$$

$$(\mathbf{u} \circ \chi - \mathbf{u}) \cdot \mathbf{n}_c + g \geq 0 \text{ auf } \Gamma_C \quad (3.16)$$

sowie auf  $\Gamma_C$  muss  $\sigma_n \leq 0$  (Normalenkraft  $\sigma_n = \mathbf{n} \cdot (\boldsymbol{\sigma} \cdot \mathbf{n})$ ),  $\boldsymbol{\sigma}_t = \mathbf{0}$  (keine Tangentialkraft, da keine Reibung –  $\boldsymbol{\sigma}_t = \boldsymbol{\sigma} \cdot \mathbf{n} - \sigma_n \mathbf{n}$ ) und  $((\mathbf{u} \circ \chi - \mathbf{u}) \cdot \mathbf{n}_c + g)\sigma_n = 0$ , d.h. wenn kein Kontakt ist, ist die Normalkraft in den Punkten Null, also herrscht Kräftegleichgewicht.

- Anreißen von Kontaktproblem mit Tresca-Reibung (vgl. Numerik für Kontaktmechanik von Stephan und Vug von Starke)  $\Rightarrow$  Herleitung der Variationsungleichung durch Ableitung nicht mehr möglich, da Reibungspotential nicht mehr differenzierbar.

### 3.2.2 Variationsformulierung für Kontaktprobleme

- Minimierung von Energiefunktional (vgl. [KO88] Seite 112 unten) mit  $\mathbf{u} : \Omega \rightarrow \mathbb{R}^3$ :

$$E(u) = \frac{1}{2}a(u, u) - f(u) \text{ mit}$$

$$a(u, u) = \int_{\Omega} \mathcal{C}\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{u}) \, d\Omega, \quad f(u) = \int_{\Omega} \mathbf{b} \cdot \mathbf{u} \, d\Omega + \int_{\Gamma_N} \mathbf{t} \cdot \mathbf{u} \, d\Gamma$$

unter der Nebenbedingung  $\mathbf{n} \cdot \mathbf{u} - g \leq 0$  auf  $\Gamma_C$  (siehe Vug Skript), bzw.  $(\mathbf{u} \circ \chi - \mathbf{u}) \cdot \mathbf{n}_c + g \geq 0$  auf  $\Gamma_C$  (etwas allgemeiner, vgl. Wriggers Paper).

- Herleitung auch über starke Formulierung möglich, vgl. Stephan – Kontaktprobleme.
- Herleitung der Variationsformulierung: Finde  $\mathbf{u} \in K$ :  $a(\mathbf{u}, \mathbf{v} - \mathbf{u}) \geq f(\mathbf{v} - \mathbf{u}) \, \forall \mathbf{v} \in K$  (s. auch Wriggers Paper) analog zum Hindernisproblem (nicht mehr ausführlich, wenn oben schon ausführlich).
- [KO88] Seite 113 für Bedingung für die Eindeutigkeit und Existenz der Lösung des Problems (hierfür wird Korn's Ungleichung benötigt  $\Rightarrow$  vielleicht Anhang?).



### 3.2.3 Lösung des Kontaktproblems mittels FEM

- Beschreibe das diskrete Problem, was man bekommt mit: Finde  $\mathbf{x}^* \in \mathbb{R}^N$  mit  $B\mathbf{x}^* \geq \mathbf{c}$ , so dass

$$(\mathbf{A}\mathbf{x}^* - \mathbf{b})^T(\mathbf{x} - \mathbf{x}^*) \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^N \text{ mit } B\mathbf{x} \geq \mathbf{c},$$

wobei

$$A = \left[ \int_{\Omega} \mathcal{C}\boldsymbol{\varepsilon}(\boldsymbol{\Psi}_j) : \boldsymbol{\varepsilon}(\boldsymbol{\Psi}_i) d\Omega \right]_{1 \leq i, j \leq N}, \quad \mathbf{b} = \left[ \int_{\Omega} \mathbf{b} \cdot \boldsymbol{\Psi}_i d\Omega + \int_{\Gamma_N} \mathbf{t} \cdot \boldsymbol{\Psi}_i ds \right]_{1 \leq i \leq N}$$

$$B = [(\boldsymbol{\Psi}_j(\chi(\mathbf{x}_i)) - \boldsymbol{\Psi}_j(\mathbf{x}_i)) \cdot \mathbf{n}_c(\mathbf{x}_i)]_{\mathbf{x}_i \in \Gamma_c, 1 \leq j \leq N}, \quad \mathbf{c} = [-g(\mathbf{x}_i)]_{\mathbf{x}_i \in \Gamma_c}$$

Dieses Problem ist (wie vorher schon gezeigt) äquivalent zu einem quadratischen Problem

$$\min_{\mathbf{x} \in \mathbb{R}^N} \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{x} - \mathbf{b}^T \mathbf{x} \text{ s.t. } B\mathbf{x} \geq \mathbf{c},$$

d.h. Lösbarkeit des quadratischen Programms sollte auch gezeigt sein (vgl. Vug Skript oder auch nichtlineare Optimierung).

## Kapitel 4

# Ein hierarchischer Fehlerschätzer für Hindernisprobleme

- Herleitung des Fehlerschätzers bei Hindernisproblemen (s. Mainpaper)
- Vergleich Hindernisprobleme zu Kontaktproblemen → warum gerade dieser Fehlerschätzer bei Hindernis- bzw. Kontaktproblemen

### 4.1 Herleitung eines a posteriori hierarchischen Fehlerschätzers

#### 4.1.1 Diskretisierung

#### 4.1.2 Lokaler Anteil des Fehlerschätzers

#### 4.1.3 Oszillationsterme

#### 4.1.4 Zuverlässigkeit des Fehlerschätzers

#### 4.1.5 Effektivität des Fehlerschätzers

### 4.2 Ein adaptiver Algorithmus

### 4.3 Erfüllung einer Saturationseigenschaft

### 4.4 Übertragung des Fehlerschätzers auf Kontaktprobleme

## Kapitel 5

# Implementierung des Fehlerschätzers in Matlab

- Grundlegender Aufbau des Programms
- Gründe warum wo was.
- Warum Verwendung von Sparse, IPM und large scale?
- Berechnung der einzelnen lokalen Element-Steifigkeitsmatrizen bzw. Element-Vektoren (siehe hierfür auch die Berechnung für den Vektor  $\rho_S$  – hier ist die Berechnung durch lokalen Vektoren auch schneller gemacht worden).
- Anmerkungen: Wie verfeinert refinemesh in Matlab eigentlich? (siehe auch Bachelorarbeit von Christina)
- dokumentierter Quellcode ist im Anhang zu finden

# Kapitel 6

## Validierung

- numerisches Beispiel (Problemstellung) → vielleicht mit Kontakt und nur Hindernis
- Vergleich mit Analytischer Lösung?! (Tabelle mit Ergebnissen) → Ergebnisse diskutieren

### 6.1 Numerisches Beispiel zum Hindernisproblem

### 6.2 Numerisches Beispiel zum Kontaktproblem

## Kapitel 7

# Zusammenfassung und Ausblick

- kurz einleiten, worum es ging (Einleitung in einem Absatz zusammenfassen)
- Was ist rausgekommen?!
- Ausblick: Was ist noch offen geblieben, was kann man noch machen...  
In dieser Arbeit linearisierte Verzerrung verwendet; kann verallgemeinert werden durch allgemeine Verzerrungstensoren (bzgl. der jeweiligen Konfiguration).

# Literaturverzeichnis

- [Alt12] ALTENBACH, Holm: *Kontinuumsmechanik*. 2. Auflage. Springer, 2012
- [BCH05] BARTELS, S. ; CARSTENSEN, C. ; HECHT, A.: 2D isoparametric FEM in MATLAB / Humboldt-Universität, Berlin. 2005. – Forschungsbericht
- [BCH07] BRAESS, D. ; CARSTENSEN, C. ; HOPPE, R.: Convergence analysis of a conforming adaptive finite element method for an obstacle problem. In: *Numerische Mathematik* 107 (2007), S. 455–471
- [Bra05] BRAESS, Dietrich: A Posteriori Error Estimators for Obstacle Problems – Another Look / Faculty of Mathematics, Ruhr-University. 2005. – Forschungsbericht
- [Bra13] BRAESS, Dietrich: *Finite Elemente – Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. 5. Auflage. Springer-Verlag, 2013
- [CSW99] CARSTENSEN, C. ; SCHERF, O. ; WRIGGERS, P.: Adaptive finite elements for elastic bodies in contact. In: *SIAM J. Sci. Comput.* 20 (1999), Nr. 5, S. 1605–1626
- [GRT09] GÖPFERT, A. ; RIEDRICH, T. ; TAMMER, C.: *Angewandte Funktionalanalysis*. Vieweg und Teubner, 2009
- [Joh92] JOHNSON, Claes: Adaptive finite element methods for the obstacle problem. In: *Math. Models Methods Appl. Sci.* 2 (1992), Nr. 4, S. 483–487
- [KO88] KIKUCHI, N. ; ODEN, J.T.: *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*. SIAM, 1988
- [KZ11] KORNHUBER, Ralf ; ZOU, Qingsong: Efficient and reliable hierarchical error estimates for the discretization error of elliptic obstacle problems. In: *Mathematics of Computation* 80 (2011), Nr. 273, S. 69–88

- [NW06] NOCEDAL, Jorge ; WRIGHT, Stephen J.: *Numerical Optimization*. 2. ed. New York, NY : Springer, 2006
- [Sta08] STARKE, Gerhard: Numerik partieller Differentialgleichungen / IFAM - Universität Hannover. 2008. – Vorlesungsskript
- [Sta11] STARKE, Gerhard: Variationsungleichungen / IFAM - Universität Hannover. 2011. – Vorlesungsskript
- [Ste12] STEPHAN, Ernst P.: Numerik partieller Differentialgleichungen I / IFAM - Universität Hannover. 2012. – Vorlesungsskript
- [Sto99] STOER, Josef: *Numerische Mathematik I*. 8. Auflage. Springer, 1999
- [Wal11] WALKER, Christoph: Partielle Differentialgleichungen / IFAM - Universität Hannover. 2011. – Vorlesungsskript
- [Wer11] WERNER, Dirk: *Funktionalanalysis*. 7. Auflage. Springer, 2011
- [Zou11] ZOU, Qingsong: Efficient and reliable hierarchical error estimates for an elliptic obstacle problem. In: *Applied Numerical Mathematics* 61 (2011), S. 344–355
- [ZVKG11] ZOU, Q. ; VEESER, A. ; KORNHUBER, R. ; GRÄSER, C.: Hierarchical error estimates for the energy functional in obstacle problems. In: *Numerische Mathematik* (2011), Nr. 117, S. 653–677

## Anhang A

# Funktionalanalysis

### A.1 Sobolev-Räume

Sei im Weiteren  $\emptyset \neq \Omega \subset \mathbb{R}^n$ . Wir definieren den Sobolev-Raum allgemein wie folgt (vgl. [Bra13] Kapitel II, §2 und [Wal11] Kapitel 6).

**Definition A.1.** Seien  $1 \leq p \leq \infty$  und  $m \in \mathbb{N}$ . Die Menge

$$W_p^m(\Omega) := \left( \{u \in L_p(\Omega) \mid \partial^\alpha u \in L_p(\Omega) \forall |\alpha| \leq m\}, \|\cdot\|_{W_p^m} \right)$$

heißt *Sobolev-Raum* der Ordnung  $m$ . Dabei ist

$$\|u\|_{W_p^m} := \|u\|_{W_p^m(\Omega)} := \left( \sum_{|\alpha| \leq m} \|\partial^\alpha u\|_{L_p}^p \right)^{\frac{1}{p}},$$

wenn  $1 \leq p < \infty$ . Im Fall  $p = \infty$  ist  $\|u\|_{W_p^m} := \max_{|\alpha| \leq m} \|\partial^\alpha u\|_\infty$ .

Weiterhin bezeichne  $L_p(\Omega)$  den *Lebesgue-Raum*, d.h. den Raum der messbaren Funktionen, deren  $p$ -te Potenz Lebesgue-integrierbar über  $\Omega$  ist, d.h.

$$L_p(\Omega) := \left( \{u : \Omega \rightarrow \mathbb{R} \mid u \text{ messbar}, \|u\|_{L_p} < \infty\}, \|\cdot\|_{L_p} \right),$$

wobei  $\|u\|_{L_p} := \|u\|_{L_p(\Omega)} = \|u\|_{W_p^0}$ .

**Definition A.2.** Der Raum

$$\mathcal{D}(\Omega) := C_c^\infty(\Omega) = \{\varphi \in C^\infty(\Omega) \mid \text{supp}(\varphi) \subset\subset \Omega\}$$

heißt der *Raum der Testfunktionen*, wobei  $K \subset\subset \Omega : \Leftrightarrow \bar{K} \subset \Omega$  kompakt.

**Bemerkung A.3.** Seien  $u \in W_p^m(\Omega)$ ,  $\varphi \in \mathcal{D}(\Omega)$  und  $\alpha \in \mathbb{N}^n$  mit  $|\alpha| \leq m$ . Dann bezeichnen wir  $v = \partial^\alpha u$  als *schwache Ableitung* von  $u$ , wenn gilt

$$\int_{\Omega} v \cdot \varphi \, dx = (-1)^{|\alpha|} \int_{\Omega} u \cdot \partial^\alpha \varphi \, dx.$$



**Beispiel A.4.** Es sei  $\Omega = (-1, 1) \subset \mathbb{R}$  und  $u(x) = |x| \in L_2(\Omega)$ . Betrachten wir  $v(x) = \text{sign}(x)$ , so ergibt sich für  $\varphi \in \mathcal{D}(\Omega)$

$$\begin{aligned} \int_{\Omega} v \cdot \varphi \, dx &= \int_{-1}^0 -1 \cdot \varphi(x) \, dx + \int_0^1 1 \cdot \varphi(x) \, dx \\ &= -x\varphi(x) \Big|_{-1}^0 - \int_{-1}^0 -x\varphi'(x) \, dx + x\varphi(x) \Big|_0^1 - \int_0^1 x\varphi'(x) \, dx \\ &= - \int_{-1}^1 |x| \varphi'(x) \, dx = (-1)^1 \int_{\Omega} u \cdot \varphi' \, dx, \end{aligned}$$

da  $\varphi(-1) = \varphi(1) = 0$ . Also ist  $v = \partial u$  und somit  $u \in W_2^1(\Omega)$ . Analog kann man nachrechnen, dass

$$\int_{\Omega} v \cdot \varphi' \, dx = -2\varphi(0)$$

ist und somit  $u$  nicht zweimal schwach ableitbar ist, d.h.  $u \notin W_2^2(\Omega)$ .

Wir wollen in der Theorie der Finiten Elemente Methode vor allem Sobolev-Räume über dem Raum  $L_2(\Omega)$  betrachten, daher ist folgender Satz essentiell.

**Satz A.5.** Es seien  $1 \leq p \leq \infty$  und  $m \in \mathbb{N}$ . Dann gilt:

- (a)  $W_p^m(\Omega)$  ist ein Banachraum.
- (b)  $H^m(\Omega) := W_2^m(\Omega)$  ist ein Hilbertraum mit Skalarprodukt

$$(u, v)_m := (u, v)_{H^m(\Omega)} := \sum_{|\alpha| \leq m} (\partial^\alpha u, \partial^\alpha v)_0 \quad \forall u, v \in H^m(\Omega),$$

wobei

$$(u, v)_0 := (u, v)_{L_2(\Omega)} := \int_{\Omega} uv \, dx.$$

**Bemerkung A.6.** (a) Die Norm auf  $H^m(\Omega)$  ergibt sich analog zur Norm des allgemeinen Sobolev-Raumes durch das Skalarprodukt, d.h.  $\|u\|_m := \|u\|_{H^m(\Omega)} := \|u\|_{W_2^m}$ .

(b) Analog dazu definieren wir die Halbnorm  $|\cdot|_m$  auf  $H^m$  wie folgt:

$$|u|_m := |u|_{H^m(\Omega)} := \left( \sum_{|\alpha|=m} \|\partial^\alpha u\|_{L_2}^2 \right)^{\frac{1}{2}}.$$

**Definition A.7.** Der Raum  $H_0^m(\Omega)$  ist die Vervollständigung von  $\mathcal{D}(\Omega)$  bzgl. der Norm  $\|\cdot\|_m$ .

**Bemerkung A.8.** Die Funktionen  $u \in H_0^m(\Omega)$  können als die Funktionen  $u \in H^m(\Omega)$  mit  $u = 0$  auf  $\partial\Omega$  aufgefasst werden.

## A.2 Optimalitätskriterien

Zunächst definieren wir einen verallgemeinerten Begriff der Richtungsableitung, der auch auf unendlich dimensional Vektorräumen existiert.

**Definition A.9.** Es seien  $V$  ein Vektorraum,  $M \subset V$  und  $W$  ein normierter Raum, sowie  $F : M \rightarrow W$  eine Abbildung,  $x_0 \in M$  und  $v \in V$ . Dann heißt  $F$  *Gâteaux-differenzierbar* (bzw. in Richtung  $v$  an der Stelle  $x_0$  differenzierbar), falls es ein  $\varepsilon > 0$  mit  $[x_0 - \varepsilon v, x_0 + \varepsilon v] \subset M$  gibt und der Grenzwert

$$\mathcal{D}_v F(x_0) := \left. \frac{d}{dt} F(x_0 + tv) \right|_{t=0} := \lim_{t \rightarrow 0} \frac{F(x_0 + tv) - F(x_0)}{t} \quad (\text{A.1})$$

in  $W$  existiert.  $\mathcal{D}_v F(x_0)$  heißt dann *Gâteaux-Ableitung* von  $F$  an der Stelle  $x_0$  in Richtung  $v$ .

Falls wir nur  $[x_0, x_0 + \varepsilon v] \subset M$  voraussetzen, so können wir in (A.1)  $\lim_{t \rightarrow 0}$  durch  $\lim_{t \rightarrow +0}$  ersetzen. Dann nennen wir (A.1) die *rechtsseitige Gâteaux-Ableitung* und bezeichnen diese mit  $\mathcal{D}_v^+ F(x_0)$ .

Für die Variationsrechnung sind folgende zwei Sätze für uns von besonderer Bedeutung.

**Satz A.10.** (Charakterisierungssatz der konvexen Optimierung) *Es seien  $M \subset V$  eine konvexe Menge,  $V$  ein Vektorraum und  $F : M \rightarrow \mathbb{R}$  eine konvexe Funktional. Dann gilt für  $x_0, x \in M$ :*

*$x_0$  ist Lösung von  $\min_{x \in M} F(x)$  genau dann, wenn für alle  $x \in M$  gilt*

$$\mathcal{D}_{x-x_0}^+ F(x_0) \geq 0.$$

*Beweis.* Siehe [GRT09], Kapitel 3.3.3, Satz 3.34. □

**Satz A.11.** *Es sei  $U \subset V$  ein (Unter-)Vektorraum,  $V$  ein Vektorraum und  $F : U \rightarrow \mathbb{R}$  eine Gâteaux-differenzierbare konvexe Funktion. Dann ist  $x_0 \in U$  genau dann Lösung von  $\min_{x \in U} F(x)$ , wenn für alle  $u \in U$  gilt*

$$\mathcal{D}_u F(x_0) = 0.$$

*Beweis.* Siehe [GRT09], Kapitel 3.3.3, Satz 3.35. □

## Anhang B

# Optimierung

### B.1 Quadratische Programmierung

Um im folgenden die Idee des Algorithmus zu verstehen, führen wir zunächst grundlegende Begriffe ein. Ein quadratisches Problem mit Gleichungs- und Ungleichungsnebenbedingungen ist von der Form

$$\begin{aligned} \min_{\mathbf{x}} \quad & q(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T G \mathbf{x} + \mathbf{x}^T \mathbf{c} \\ \text{s.t.} \quad & \mathbf{a}_i^T \mathbf{x} = b_i, \quad i \in \mathcal{E}, \\ & \mathbf{a}_i^T \mathbf{x} \geq b_i, \quad i \in \mathcal{I}, \end{aligned} \tag{B.1}$$

wobei  $\mathcal{E}$  und  $\mathcal{I}$  die Indexmengen der Gleichungs- und Ungleichungsnebenbedingungen darstellen und  $\mathbf{c}, \mathbf{x}, \mathbf{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}, i \in \mathcal{E} \cup \mathcal{I}$ , sowie  $G$  eine symmetrische  $(n \times n)$ -Matrix ist, welche die Hesse-Matrix des Problems darstellt. Damit ist die Hesse-Matrix konstant und daher das Problem konvex, wenn  $G$  positiv semidefinit ist. (Ist  $G$  positiv definit, so nennen wir das Problem strikt konvex. Wenn  $G$  indefinit ist, ist (B.1) „nicht konvex“.)

Da sonst das quadratische Problem (und damit der Active-Set Algorithmus) zu kompliziert wird, betrachten wir hier nur den konvexen Fall. Für diesen Fall können wir leicht zeigen, dass eine Lösung  $\mathbf{x}^*$ , die die Bedingungen 1. Ordnung erfüllt, auch globale Lösung des Problems ist (s. Theorem B.1). Anschaulich kann es im indefiniten Fall mehrere optimale Punkte geben, die voneinander getrennt liegen, d.h. die Menge der optimalen Punkte ist nicht zusammenhängend, wodurch das Auffinden des globalen Minimums erschwert wird.

Die notwendigen Bedingungen 1. Ordnung sind die KKT-Bedingungen und können hier angewendet werden, da die Restriktionen und die Zielfunktion stetig differenzierbar sind. Die Lagrangefunktion  $\mathcal{L}$  für das quadratische Problem ist

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2} \mathbf{x}^T G \mathbf{x} + \mathbf{x}^T \mathbf{c} - \sum_{i \in \mathcal{I} \cup \mathcal{E}} \lambda_i (\mathbf{a}_i^T \mathbf{x} - b_i). \tag{B.2}$$

Damit ergeben sich – vgl. [NW06], Theorem 12.1 – mit der Menge der aktiven Nebenbedingungen  $\mathcal{A}(\mathbf{x}^*) = \{i \in \mathcal{E} \cup \mathcal{I} : \mathbf{a}_i^T \mathbf{x}^* = b_i\}$  die KKT-Bedingungen

$$\begin{aligned} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) &= G\mathbf{x}^* + \mathbf{c} - \sum_{i \in \mathcal{A}(\mathbf{x}^*)} \lambda_i^* \mathbf{a}_i = 0, \\ \mathbf{a}_i^T \mathbf{x}^* &= b_i, \quad \forall i \in \mathcal{A}(\mathbf{x}^*), \\ \mathbf{a}_i^T \mathbf{x}^* &\geq b_i, \quad \forall i \in \mathcal{I} \setminus \mathcal{A}(\mathbf{x}^*), \\ \lambda_i^* &\geq 0, \quad \forall i \in \mathcal{I} \cap \mathcal{A}(\mathbf{x}^*). \end{aligned} \tag{B.3}$$

Hierbei ist  $\mathbf{x}^*$  Lösung von (B.1) und erfüllt die LICQ-Bedingung;  $\boldsymbol{\lambda}^*$  ist dazugehöriger optimaler Lagrange-Multiplikator. In (B.3) wird die Komplementaritätsbedingung  $\lambda_i^* c_i(\mathbf{x}^*) = 0$  impliziert durch  $\lambda_i^* = 0 \forall i \notin \mathcal{A}(\mathbf{x}^*)$ .

**Theorem B.1.** *Wenn  $\mathbf{x}^*$  die Bedingungen (B.3) erfüllt mit  $\lambda_i^*, i \in \mathcal{A}(\mathbf{x}^*)$  und  $G$  ist positiv semidefinit, dann ist  $\mathbf{x}^*$  eine globale Lösung von (B.1).*

*Beweis.* Wenn  $\mathbf{x}$  ein beliebiger weiterer zulässiger Punkt für (1.1) ist, gelten die Restriktionen  $\mathbf{a}_i^T \mathbf{x} = b_i, i \in \mathcal{E}$ , sowie  $\mathbf{a}_i^T \mathbf{x} \geq b_i, i \in \mathcal{I} \cap \mathcal{A}(\mathbf{x}^*)$  für  $\mathbf{x}$  und damit gilt zusammen mit der ersten Bedingung von (B.3), dass

$$(\mathbf{x} - \mathbf{x}^*)^T (G\mathbf{x}^* + \mathbf{c}) = \sum_{i \in \mathcal{E}} \underbrace{\lambda_i^* \mathbf{a}_i^T (\mathbf{x} - \mathbf{x}^*)}_{\geq 0} + \sum_{i \in \mathcal{A}(\mathbf{x}^*) \cap \mathcal{I}} \underbrace{\lambda_i^* \mathbf{a}_i^T (\mathbf{x} - \mathbf{x}^*)}_{\geq 0} \geq 0.$$

Dann drücken wir  $q(\mathbf{x})$  durch  $q(\mathbf{x}^*)$  aus und wenden die obere Ungleichung sowie die positive Semidefinitheit für  $G$  an, um zu zeigen, dass  $q(\mathbf{x}) \geq q(\mathbf{x}^*)$  ist. Damit ist  $\mathbf{x}^*$  globale Lösung des quadratischen Problems.  $\square$

Daher ist im positiv semidefiniten Fall gesichert, dass ein optimaler Punkt auch gleichzeitig globale Lösung ist.

## B.2 Active Set-Methode für konvexe QPs

Wenn wir eine Lösung  $\mathbf{x}^*$  für das Problem (B.1) kennen, so ist auch die Menge der aktiven Nebenbedingungen  $\mathcal{A}(\mathbf{x}^*)$  bekannt und wir können (B.1) vereinfachen zum Optimierungsproblem

$$\min_{\mathbf{x}} \quad q(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T G \mathbf{x} + \mathbf{x}^T \mathbf{c}, \quad \text{s.t.} \quad \mathbf{a}_i^T \mathbf{x} = b_i, \quad i \in \mathcal{A}(\mathbf{x}^*). \tag{B.4}$$

Dieses könnten wir dann beispielsweise mit direkten Verfahren wie der Schur-Komplement-Methode oder der Nullraum-Methode lösen. Natürlich ist die optimale Lösung zu Beginn noch nicht bekannt und damit auch nicht die aktiven Restriktionen. Jedoch können wir diese Idee für die Active-Set-Methode verwenden.

Das Hauptziel der Active-Set-Methode ist, die Menge der aktiven Restriktionen bzgl. der optimalen Lösung zu finden, wobei wir hier die primale

Variante betrachten wollen, in der die Approximierte  $\mathbf{x}_k$  zulässig bzgl. des primalen Problems ist.

Die Grundidee ist, ein quadratisches Teilproblem zu lösen, bei dem wir bestimmte Nebenbedingungen aus Problem (B.1) bzgl.  $\mathcal{I}$  als aktiv annehmen. Die dadurch beschriebene Indexmenge der aktiven Restriktionen für  $\mathbf{x}_k$  im  $k$ -ten Schritt heißt *working set* und kann wie folgt beschrieben werden

$$\mathcal{W}_k = \{i \mid \mathbf{a}_i^T \mathbf{x}_k = b_i, i \in \mathcal{E} \cup \mathcal{J}, \mathcal{J} \subset \mathcal{I}\}.$$

Hierbei muss vorausgesetzt werden, dass die Nebenbedingungen in  $\mathcal{W}_k$  die LICQ-Bedingung erfüllen, selbst wenn diese bezogen auf alle Nebenbedingungen an der Stelle  $\mathbf{x}_k$  nicht erfüllt wird.

Wir betrachten nun den  $k$ -ten Schritt mit der Approximierten  $\mathbf{x}_k$  und dem working set  $\mathcal{W}_k$ . Wir berechnen die neue Iterierte  $\mathbf{x}_{k+1}$ , indem wir eine Richtung  $\mathbf{p}$  finden, in der wir unter den Nebenbedingungen  $\mathcal{W}_k$  die Funktion  $q$  minimieren. Hierfür betrachten wir  $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}$  und setzen  $\mathbf{x}_{k+1}$  in  $q$  ein:

$$\begin{aligned} q(\mathbf{x}_{k+1}) &= q(\mathbf{x}_k + \mathbf{p}) = \frac{1}{2}(\mathbf{x}_k + \mathbf{p})^T G(\mathbf{x}_k + \mathbf{p}) + (\mathbf{x}_k + \mathbf{p})^T \mathbf{c} \\ &= \frac{1}{2}\mathbf{x}_k^T G \mathbf{x}_k + \underbrace{\mathbf{x}_k^T G \mathbf{p}}_{\text{da } G \text{ symm.}} + \frac{1}{2}\mathbf{p}^T G \mathbf{p} + \mathbf{x}_k^T \mathbf{c} + \mathbf{p}^T \mathbf{c} \\ &= \frac{1}{2}\mathbf{p}^T G \mathbf{p} + \mathbf{g}_k^T \mathbf{p} + \rho_k, \end{aligned}$$

wobei  $\mathbf{g}_k = G\mathbf{x}_k + \mathbf{c}$  und  $\rho_k = \frac{1}{2}\mathbf{x}_k^T G \mathbf{x}_k + \mathbf{x}_k^T \mathbf{c}$ . Da wir den Parameter  $\mathbf{p}$  so wählen wollen, so dass  $q(\mathbf{x}_{k+1})$  minimal wird, ist der Term  $\rho_k$  bzgl. des Problems konstant und kann somit für die Lösung jenes weggelassen werden. Da weiterhin auch  $\mathbf{x}_{k+1}$  die aktiven Nebenbedingungen  $\mathcal{W}_k$  erfüllen soll, gilt

$$\mathbf{a}_i^T \mathbf{p} = \mathbf{a}_i^T (\mathbf{x}_{k+1} - \mathbf{x}_k) = \underbrace{\mathbf{a}_i^T \mathbf{x}_{k+1}}_{=b_i} - \underbrace{\mathbf{a}_i^T \mathbf{x}_k}_{=b_i} = 0 \quad \forall i \in \mathcal{W}_k.$$

Zusammengefasst müssen wir also im  $k$ -ten Schritt das Teilproblem

$$\begin{aligned} \min_{\mathbf{p}} \quad & \frac{1}{2}\mathbf{p}^T G \mathbf{p} + \mathbf{g}_k^T \mathbf{p}, \\ \text{s.t.} \quad & \mathbf{a}_i^T \mathbf{p} = 0, \quad \forall i \in \mathcal{W}_k \end{aligned} \tag{B.5}$$

lösen. Die Lösung im  $k$ -ten Schritt von (B.5) bezeichnen wir mit  $\mathbf{p}_k$ . Umgekehrt gilt damit, analog zur obigen Rechnung, natürlich auch, dass für alle  $i \in \mathcal{W}_k$  die Restriktion aktiv bleibt für  $\mathbf{x}_k + \alpha \mathbf{p}_k$  mit beliebigem  $\alpha$ . Da  $G$  positiv definit ist, kann (B.5) nun – wie schon bei (B.4) erwähnt – mit Schur-Komplement-Methode oder Nullraum-Methode gelöst werden.

Wie wir schon wissen, ist die neue Iterierte  $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$  bzgl.  $\mathcal{W}_k$  immer noch zulässig. Nun müssen wir jedoch feststellen, ob diese Iterierte

auch alle übrigen Restriktionen mit  $i \notin \mathcal{W}_k$  erfüllt. Ist dies der Fall, so setzen wir  $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$ , ansonsten suchen wir das größtmögliche  $\alpha_k \in [0, 1]$ , so dass

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$$

zulässig bleibt. Hierfür betrachten wir zwei Fälle.

Fall 1: Gilt für ein  $i \notin \mathcal{W}_k$ , dass  $\mathbf{a}_i^T \mathbf{p}_k \geq 0$  ist, so folgt

$$\mathbf{a}_i^T (\mathbf{x}_k + \alpha_k \mathbf{p}_k) = \mathbf{a}_i^T \mathbf{x}_k + \underbrace{\alpha_k \mathbf{a}_i^T \mathbf{p}_k}_{\geq 0} \geq \mathbf{a}_i^T \mathbf{x}_k \geq b_i,$$

da  $\alpha_k \geq 0$ , d.h. für diese Nebenbedingungen müssen wir für die Wahl von  $\alpha_k$  nichts beachten.

Fall 2: Existiert ein  $i \notin \mathcal{W}_k$ , für das  $\mathbf{a}_i^T \mathbf{p}_k < 0$  ist, so gilt

$$\begin{aligned} & \mathbf{a}_i^T (\mathbf{x}_k + \alpha_k \mathbf{p}_k) \geq b_i \\ \iff & \mathbf{a}_i^T \mathbf{x}_k + \alpha_k \mathbf{a}_i^T \mathbf{p}_k \geq b_i \\ \iff & \alpha_k \underbrace{\mathbf{a}_i^T \mathbf{p}_k}_{< 0} \geq b_i - \mathbf{a}_i^T \mathbf{x}_k \\ \iff & \alpha_k \leq \frac{b_i - \mathbf{a}_i^T \mathbf{x}_k}{\mathbf{a}_i^T \mathbf{p}_k}. \end{aligned} \tag{B.6}$$

Damit folgt mit (B.6) und den vorherigen Überlegungen, dass zusammengefasst

$$\alpha_k = \min \left\{ 1, \min_{i \notin \mathcal{W}_k, \mathbf{a}_i^T \mathbf{p}_k < 0} \frac{b_i - \mathbf{a}_i^T \mathbf{x}_k}{\mathbf{a}_i^T \mathbf{p}_k} \right\} \tag{B.7}$$

gilt. Eine Restriktion  $i \notin \mathcal{W}_k$ , für die das Minimum für  $\alpha_k$  angenommen wird, nennen wir *blocking constraint*; diese muss nicht eindeutig sein, da wir beispielsweise anschaulich auch von einer Ecke geblockt werden können. Ist  $\alpha_k = 1$ , so werden alle Restriktion außerhalb vom working set mit dem Schritt  $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$  erfüllt, d.h. es gibt keine blocking constraint. Gibt es eine Nebenbedingung  $j \notin \mathcal{W}_k$ , die aktiv ist, obwohl sie nicht zum working set gehört, so gilt

$$\begin{aligned} \alpha_k &= \min \left\{ 1, \min_{i \notin \mathcal{W}_k, \mathbf{a}_i^T \mathbf{p}_k < 0} \frac{b_i - \mathbf{a}_i^T \mathbf{x}_k}{\mathbf{a}_i^T \mathbf{p}_k} \right\} \\ &= \min \left\{ 1, \frac{b_j - \mathbf{a}_j^T \mathbf{x}_k}{\mathbf{a}_j^T \mathbf{p}_k} \right\} \\ &= \min \left\{ 1, \frac{b_j - b_j}{\mathbf{a}_j^T \mathbf{p}_k} \right\} = 0. \end{aligned}$$

Es sei  $j \notin \mathcal{W}_k$  nun ein Index einer blocking constraint. Dann ist

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k = \mathbf{x}_k + \frac{b_j - \mathbf{a}_j^T \mathbf{x}_k}{\mathbf{a}_j^T \mathbf{p}_k} \mathbf{p}_k.$$

Setzen wir  $\mathbf{x}_{k+1}$  in die  $j$ -te Restriktion ein, so erhalten wir

$$\begin{aligned} \mathbf{a}_j^T \mathbf{x}_{k+1} &= \mathbf{a}_j^T \left( \mathbf{x}_k + \frac{b_j - \mathbf{a}_j^T \mathbf{x}_k}{\mathbf{a}_j^T \mathbf{p}_k} \mathbf{p}_k \right) = \mathbf{a}_j^T \mathbf{x}_k + \frac{b_j - \mathbf{a}_j^T \mathbf{x}_k}{\mathbf{a}_j^T \mathbf{p}_k} \cdot \cancel{\mathbf{a}_j^T \mathbf{p}_k} \\ &= \mathbf{a}_j^T \mathbf{x}_k + b_j - \mathbf{a}_j^T \mathbf{x}_k = b_j, \end{aligned}$$

d.h. die blocking constraint ist für die neue Iterierte  $\mathbf{x}_{k+1}$  nach Konstruktion aktiv. Daher setzen wir als neues working set  $\mathcal{W}_{k+1} = \mathcal{W}_k \cup \{j\}$ .

Das oben beschriebene Vorgehen wiederholen wir so lange, bis wir das working set  $\hat{\mathcal{W}}$  mit dem Minimum des quadratischen Problems  $\hat{\mathbf{x}}$  gefunden haben. Dies ist leicht zu erkennen, da wir (B.1) auf  $\mathcal{W}_k$  nicht weiter minimieren können, sobald es keinen Schritt  $\mathbf{p}$  gibt, in dessen Richtung wir  $q$  verringern können, d.h. wenn  $\mathbf{p} = \mathbf{0}$  die Lösung für das Teilproblem (B.5) ist. Dann ist der optimale Punkt  $\hat{\mathbf{x}}$  bzgl. des working sets  $\hat{\mathcal{W}} \subset \mathcal{A}(\hat{\mathbf{x}})$  gefunden.

Wir müssen jetzt überprüfen, ob  $\hat{\mathbf{x}}$  die KKT-Bedingungen erfüllt. Wir wissen, dass für  $\mathbf{p} = \mathbf{0}$  die KKT-Bedingungen für (B.5)

$$\begin{pmatrix} G & A^T \\ A & 0 \end{pmatrix} \cdot \begin{pmatrix} -\mathbf{p} \\ \hat{\lambda} \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{g}} \\ \hat{\mathbf{h}} \end{pmatrix}$$

mit  $\hat{\mathbf{g}} = \mathbf{c} + G\hat{\mathbf{x}}$ ,  $\hat{\mathbf{h}} = A\hat{\mathbf{x}} + \mathbf{b}$  und  $\mathbf{p} = \mathbf{0}$  erfüllt. Daraus folgt

$$\begin{aligned} A^T \hat{\lambda} &= \hat{\mathbf{g}} \iff \sum_{i \in \hat{\mathcal{W}}} \mathbf{a}_i \hat{\lambda}_i = G\hat{\mathbf{x}} + \mathbf{c}, \\ \mathbf{0} &= \hat{\mathbf{h}} \iff A\hat{\mathbf{x}} = \mathbf{b}, \end{aligned}$$

wobei  $A$  die Gradienten  $\mathbf{a}_i^T$  der aktiven Restriktionen  $\hat{\mathcal{W}}$  zeilenweise enthält. Damit werden die ersten beiden KKT-Bedingungen aus (B.3) erfüllt. Da die Schrittweite  $\alpha_k$  mit (B.6) so gewählt ist, dass die übrigen Restriktionen erfüllt bleiben, gilt auch die dritte Bedingung aus (B.3). Es bleibt zu überprüfen, ob die Lagrange-Multiplikatoren  $\hat{\lambda}_i \geq 0$  sind.

Gilt  $\hat{\lambda}_i \geq 0$  für alle  $i \in \hat{\mathcal{W}} \cap \mathcal{I}$ , so sind alle KKT-Bedingungen erfüllt und damit  $\mathbf{x}^* = \hat{\mathbf{x}}$ . Existiert allerdings ein  $j \in \hat{\mathcal{W}} \cap \mathcal{I}$ , so dass  $\hat{\lambda}_j < 0$  ist, so können wir den Wert von  $q$  noch weiter verringern, indem wir die  $j$ -te Restriktion wegfällen lassen (vgl. [NW06], Kapitel 12.3). Dies zeigt das folgende Theorem.

**Theorem B.2.** *Der Punkt  $\hat{\mathbf{x}}$  erfülle die notwendigen Bedingungen 1. Ordnung für das Teilproblem (B.5) auf  $\hat{\mathcal{W}}$ . Weiter seien die Gradienten  $\mathbf{a}_i, i \in$*

$\hat{\mathcal{W}}$ , linear unabhängig (LICQ) und es gebe einen Index  $j \in \mathcal{W}$  mit  $\hat{\lambda}_j < 0$ .  
Es sei  $\mathbf{p}$  die Lösung vom Teilproblem (B.5) ohne die Restriktion  $j$ , d.h.

$$\begin{aligned} \min_{\mathbf{p}} \quad & \frac{1}{2} \mathbf{p}^T G \mathbf{p} + (G\hat{\mathbf{x}} + \mathbf{c})^T \mathbf{p}, \\ \text{s.t.} \quad & \mathbf{a}_i^T \mathbf{p} = 0, \quad \forall i \in \hat{\mathcal{W}} \setminus \{j\}. \end{aligned}$$

Dann ist  $\mathbf{p}$  eine zulässige Richtung für die Nebenbedingung  $j$ , d.h.  $\mathbf{a}_j^T \mathbf{p} \geq 0$ .  
Weiterhin gilt sogar  $\mathbf{a}_j^T \mathbf{p} > 0$  und  $\mathbf{p}$  ist eine Abstiegsrichtung für  $q$ , wenn  $\mathbf{p}$  die hinreichenden Bedingungen 2. Ordnung erfüllt.

Da wir zeigen können, dass der erzielte Abstieg für  $q$  durch das Weglassen einer Nebenbedingung mit negativem Lagrange-Multiplikator  $\lambda_i$  proportional zu  $|\lambda_i|$  ist, eliminieren wir gerade die Restriktion mit kleinstem Lagrange-Multiplikator. Es kann allerdings sein, dass der folgende zu berechnende Schritt  $\mathbf{p}$  aufgrund einer blocking constraint kurz ist, wodurch nicht garantiert ist, dass  $q$  den größtmöglichen Abstieg erfährt.

### B.3 Algorithmus

---

**Algorithm B.3.1** Active-Set-Methode für konvexe quadratische Probleme

Gegeben sei ein zulässiger Startpunkt  $\mathbf{x}_0$  für (B.1) und definiere  $\mathcal{W}_0$  z.B. mit allen aktiven Restriktionen bzgl.  $\mathbf{x}_0$ .

```

for  $k = 0, 1, 2, \dots$  do
  Löse (B.5) zur Berechnung von  $\mathbf{p}_k$ ;
  if  $\mathbf{p}_k = \mathbf{0}$  then
    Berechne die Lagrange-Multiplikatoren mittels (2.5a)
    und setze  $\hat{\mathcal{W}} = \mathcal{W}_k$ ;
    if  $\hat{\lambda}_i \geq 0 \forall i \in \hat{\mathcal{W}} \cap \mathcal{I}$  then
      stop mit der Lösung  $\mathbf{x}^* = \hat{\mathbf{x}}$ ;
    else
       $j \leftarrow \arg \min_{j \in \mathcal{W}_k \cap \mathcal{I}} \hat{\lambda}_j$ ;
       $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k$ ,  $\mathcal{W}_{k+1} \leftarrow \mathcal{W}_k \setminus \{j\}$ ;
    end if
  else ( $\mathbf{p}_k \neq \mathbf{0}$ )
    Berechne  $\alpha_k$  mit (B.7);
     $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k + \alpha_k \mathbf{p}_k$ ;
    if  $\alpha_k < 1$  (blocking constraint existiert) then
      Bestimme blocking constraint  $j$  und setze  $\mathcal{W}_{k+1} \leftarrow \mathcal{W}_k \cup \{j\}$ 
    else
       $\mathcal{W}_{k+1} \leftarrow \mathcal{W}_k$ 
    end if
  end if
end for

```

---



## Anhang C

# Quellcode

### C.1 Implementierung des Fehlerschätzers für das Hindernisproblem

# Index

Bilinearform  
    elliptisch, 9  
    koerziv, 9  
    stetig, 9

Dirichlet-Problem, 7  
    homogenes, 8

Gâteaux-Ableitung, 22  
    rechtsseitig, 22

Gâteaux-differenzierbar, 22

homogenen Dirichlet-Problem, 8

Lebesgue-Raum, 20

Raum der Testfunktionen, 20

schwache Ableitung, 20  
schwache Lösung, 8  
Sobolev-Raum, 20