Leibniz Universität Hannover
Fakultät für Mathematik und Physik
Institut für Angewandte Mathematik

Diplomarbeit

Ein hierarchischer Fehlerschätzer für Hindernisprobleme

von Cornelius Rüther
Matr.-Nr.: 2517350

26. Oktober 2014

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

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

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)

wichtig: Vorgehen einer adaptiven Verfeinerungsstrategie mit „solve → estimate → ....“ umschreiben

- 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.

Dieses Kapitel basiert auf [Bra13], [Sta08], [Ste12b], [Wal11], [Alt12].

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, $\|v\|_H^2 \coloneqq (v,v)_H$ für alle $v \in H$, ist, d.h. in dem jede Cauchy-Folge konvergiert.
- Es sei in diesem Kapitel H ein reeller Hlbertraum 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*
$$\|v - m_v\| = \operatorname{dist}(v, M) \coloneqq \inf_{w \in M} \|v - w\| \, .$$
Wir nennen $P_M : H \to M$ mit $v \mapsto m_v$ die Projektionen auf M.

Beweis. Der Beweis ist in [Wal11] Kapitel 7.1 Satz 7.2 zu finden. □

•

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

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

für alle $m \in M$.

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

„⇒“ 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 $2(v,m)_H \leq 0$.

„⇐“ 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 $\operatorname{dist}(v, M) = \|v\|_H^2$ und damit $0 = P_M(v)$. □

•

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

$$(P_M(w) - P_M(v), v - P_M(v))_H \leq 0 , \tag{2.1}$$
$$(P_M(v) - P_M(w), w - P_M(w))_H \leq 0 . \tag{2.2}$$

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

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

Nach Umstellen der Ungleichung folgt die Behauptung. □

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

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

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

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

d.h. jedes $v \in M$ 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. □

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

Beweis. Man kann zeigen, dass $\overline{\operatorname{span} M} = (M^{\perp})^{\perp} =: M^{\perp\perp}$ 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. □

2.2 Variationsformulierung

Stichpunkte für die Formulierung:

- Betrachte als Modellproblem Ausrenkung $u : \Omega \to \mathbb{R}$ 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 Γ.

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

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

  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 Γ 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 Ω. Durch Anwenden der 1. Green'schen Formel (bzw. Satz von Gauß) ergibt sich

$$\int_\Omega \nabla u \cdot \nabla v \, dx - \underbrace{\int_\Gamma v \partial_\nu u \, ds}_{=0,\ \text{da } v|_\Gamma = 0} = \int_\Omega f v \, dx$$

$$\iff \qquad \int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \tag{2.5}$$

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

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

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

  wenn die Gleichung

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

  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 $t \in [0,1]$

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

Daraus folgt die Behauptung. □

•

Lemma 2.11. *Sei H ein Hilbertraum. Das Funktional $J : H \to \mathbb{R}, J(v) = \frac{1}{2}a(v,v) - F(v)$ 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

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

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

•

Theorem 2.12. (Lax-Milgram) *Es sei H ein Hilbertraum und $a : H \times H \to \mathbb{R}$ eine symmetrische, in H stetige, koerzive Bilinearform. Weiter sei $F : H \to \mathbb{R}$ 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 $J : H \to \mathbb{R}$ nach Lemma 2.10 ein konvexes Funktional ist und J nach Lemma 2.11 Gâteaux-differenzierbar, gilt mit Satz A.10 für alle $v \in H$

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

(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 $v \in H$

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

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

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

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

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

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

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 $J(v_n) \to \eta$ für $n \to \infty$. Dann folgt mit der Koerzivität von a

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

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

- **Satz 2.13.** (Poincaré-Friedrich-Ungleichung) *Es sei Ω 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. □

Bemerkung 2.14. Für die Gültigkeit der Poincaré-Friedrich-Ungleichung, muss v nicht auf ganz Γ gleich Null sein, sondern es reicht aus, dass

$$v \in H^1_{\Gamma_u}(\Omega) \coloneqq \{v \in H^1(\Omega) \mid v = 0 \text{ auf } \Gamma_u\}$$

ist mit $\Gamma_u \subset \Gamma$ und einem Maß $\mu(\Gamma_u) \neq 0$ (vgl. [Bra13] Kapitel II, §1, Bemerkung 1.6).

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

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

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

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

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

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

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 Ω 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 2.15 (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 $(\cdot,\cdot)_H : H \times H \to \mathbb{R}$ 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 $F : H \to \mathbb{R}$ und damit folgt mit $a(\cdot,\cdot) := (\cdot,\cdot)_H$ aus dem Theorem von Lax-Milgram die Behauptung. □

•

Korollar 2.16. *Es sei H ein Hilbertraum mit Skalarprodukt $(\cdot,\cdot)_H$ und $a : H \times H \to \mathbb{R}$ eine stetige koerzive Bilinearform. Dann existiert genau ein linearer Operator $A : H \to H$, so dass gilt:*

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

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

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

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 $A : H \to H$ 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

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

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

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

denn

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

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

$$0 = Av_1 - Av_2 = A(v_1 - v_2) \overset{(2.8)}{\Longrightarrow} v_1 = v_2\,.$$

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

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

Sei $(v_n)_{n\in\mathbb{N}}$ eine Folge mit $v_k \in \operatorname{im} A$ 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 $Au_k = v_k \to v \in H$ geht, dann folgt

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

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 $u_n \to u$. Mit der Stetigkeit von A folgt dann

$$v_n = Au_n \xrightarrow[n\to\infty]{} Au = v\,,$$

d.h. $v \in \operatorname{im} A$ und damit ist $\operatorname{im} A$ abgeschlossen. Wir betrachten nun ein $v \in H$ mit $v \perp \operatorname{im} A \subset H$, 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 $\operatorname{im} A$ orthogonale Raum nur aus dem Nullelement und mit Korollar 2.7 gilt dann $\operatorname{im} A = \overline{\operatorname{im} A} = H$. 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. □

2.3 Finite Elemente Methode

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

$$a(u, v) = F(v) \quad \forall\, v \in H \tag{2.9}$$

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

$$a(u_h, v_h) = F(v_h) \quad \forall\, v_h \in V_h\,. \tag{2.10}$$

- **Satz 2.17.** *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.

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

$$\exists!\ \boldsymbol{\mu} \in \mathbb{R}^N : u_h(x) = \sum_{i=1}^{N} \mu_i\, \phi_i(x)\,. \tag{2.11}$$

- da $F(\cdot), a(u, \cdot)$ linear sind und alle $v_h \in V_h$ analog zu oben darstellbar sind, ist (2.10) 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

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

  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

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

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

- **Bemerkung 2.18.** 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

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

  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 Ω in endliche Elemente zerlegt werden. V_h wird dann bzgl. einer Zerlegung $\mathcal{T}_h$ beschrieben.
- betrachte im weiteren (wie oben schon angesprochen) Ω 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 2.19 (Triangulierung)**.** Es sei $\Omega \subset \mathbb{R}^2$ ein durch einen Polygonzug berandetes Gebiet. Dann heißt eine Zerlegung aus Dreiecken

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

  Triangulierung, wenn gilt:

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

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

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

  Der Radius des Umkreis eines Dreieckes 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 2.20.** natürlich auch im $\mathbb{R}^3$ analog mit Tetraedern definierbar

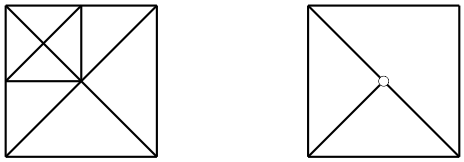

Abbildung 2.1: Zulässige und unzulässige Triangulierung (mit hängendem Knoten)

- vgl. wieder [Bra13] Seite 58

  Definition 2.21 ((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

Abbildung 2.2: Beispiele quasiuniformer Zerlegungen

-

  Bemerkung 2.22. 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 Ω bei gegebener Zerlegung eine Eigenschaft stückweise hat, wenn sie auf jedem Element diese Eigenschaft besitzt.

  Satz 2.23. *Sei $k \geq 1$ und $\Omega \subset \mathbb{R}^2$ ein polygonales Gebiet. Eine stückweise beliebig oft differenzierbare Funktion $v : \bar{\Omega} \to \mathbb{R}$ 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. □

- 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

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

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

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

  Definition 2.24 (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 $\mathcal{B}_h = \{\phi_1, \ldots, \phi_N\}$ mit $\phi_i \in \mathcal{P}_m, i = 1, \ldots, N$, 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 2.25. Sei $m \geq 0$. In einem Dreieck T seien auf $m+1$ Linien $l = 1+2+\ldots+(m+1)$ Punkte $z_1, \ldots, z_l$ angeordnet (s. Skizze). Dann gibt es zu jedem $C^0(T)$ genau ein Polynom p vom Grad m mit der Eigenschaft

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

  Beweis. Der Beweis steht in [Bra13] Kapitel II, §5, Bemerkung 5.4. □

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

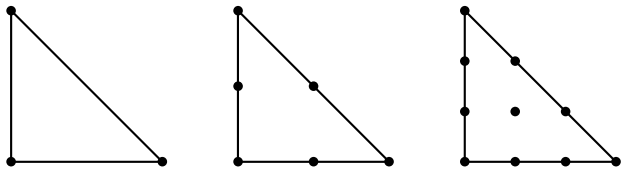

Abbildung 2.3: Dreiecke für nodale Basen (linear, quadratisch, kubisch)

- 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 $\mathcal{S}_h$ mit

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

- das Galerkin-Verfahren mit dem Ansatzraum $\mathcal{S}_h$ wird Finite-Elemente-Methode genannt
- Beispiel, um zu sehen, dass auch für kleine Gitter der Rechenaufwand sehr hoch werden kann.

  Beispiel 2.26. Wir betrachten unser Variationsproblem (2.10) auf $\Omega = [-1,1]^2$ mit $\mathcal{S}_h$ 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

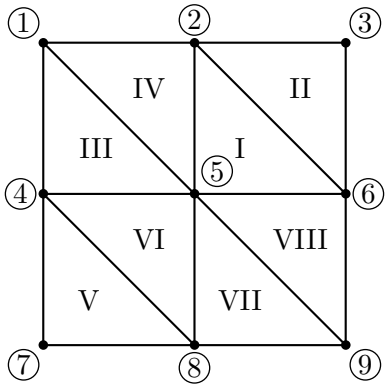

Abbildung 2.4: Triangulierung von $\Omega = [-1,1]^2$ in 8 *Courant-Elemente*

Wir stellen für die nodale Basisfunktion ϕ_5 die Einträge in der Steifig-

keitsmatrix 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 ϕ_5.

Damit rechnen wir nach, dass gilt

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

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 ϕ_5 an der Steifigkeitsmatrix A von der Form

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

Hierbei müssen die Einträge aus $\widetilde{A}$ 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 $\widetilde{A}$ 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. → 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)

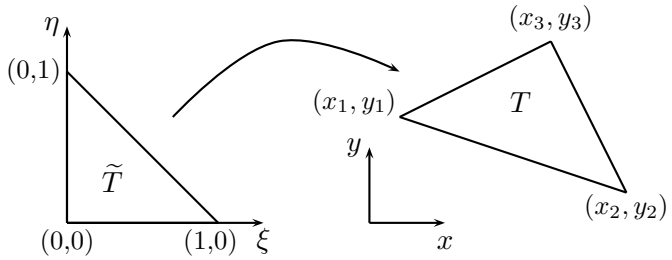

Abbildung 2.5: Referenzelement $\widetilde{T}$ für ein allgemeines Dreieck $T \in \mathcal{T}_h$

- 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 ⇒ Netzverfeinerung führt zur Konvergenz
-

Lemma 2.27. *Durch* $\|\cdot\|_E : H_0^1(\Omega) \to \mathbb{R}, \|v\|_E := (a(v,v))^{\frac{1}{2}}$ *mit einer stetigen koerziven Bilinearform a wird eine Norm auf* $H_0^1(\Omega)$ *definiert.*

Beweis. Aus der Stetigkeit und Koerzivität von a folgt direkt

$$\alpha \|v\|_1^2 \leq \underbrace{a(v,v)}_{=\|v\|_E^2} \leq c \|v\|_1^2 . \tag{2.12}$$

Damit ist $\|\cdot\|_E$ nach oben und unten durch die Norm auf $H^1(\Omega)$ beschränkt und somit eine zu dieser äquivalente Norm. □

•

Bemerkung. (a) Die Norm $\|\cdot\|_E$ bezeichnen wir als *Energie-Norm*. Sie gibt für die von uns später in der Strukturmechanik betrachtete Bilinearform die Verzerrungsenergie eines Kontinuums an.

(b) Für die Bilinearform

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

mit $u, v \in H_0^1(\Omega)$ gilt dann $\|\cdot\|_E = |\cdot|_1$ (s. Bemerkung A.6).

- Galerkin ist Bestapproximation

Satz 2.28. *Die Galerkin-Approximation u_h ist die beste Approximation von u bzgl. der Energie-Norm, also*

$$\|u - u_h\|_E = \inf_{v \in V_h} \|u - v\|_E .$$

Beweis. Zunächst betrachten wir die exakte und approximierte Variationsgleichung (2.9) und (2.10), d.h.

$$a(u,v) = F(v) \quad \forall\, v \in H , \tag{2.13}$$

$$a(u_h, v_h) = F(v_h) \quad \forall\, v_h \in V_h . \tag{2.14}$$

Da $V_h \subset H$ ist, gilt (2.13) auch für alle $v_h \in V_h$. Ersetzen wir dies in (2.13) und subtrahieren (2.13) und (2.14), so erhalten wir

$$a(u - u_h, v_h) = 0 \quad \forall\, v_h \in V_h . \tag{2.15}$$

Damit rechnen wir für ein beliebiges $v \in V_h$ einfach nach:

$$\begin{aligned}
\|u - u_h\|_E^2 &= a(u - u_h, u - u_h) \\
&= a(u - u_h, u - v + v - u_h) \\
&= a(u - u_h, u - v) + \underbrace{a(u - u_h, \underbrace{v - u_h}_{\in V_h})}_{=0 \text{ wegen } (2.14)} \\
&= a(u - u_h, u - v) \\
&\overset{\text{CS}}{\leq} \|u - u_h\|_E \|u - v\|_E
\end{aligned}$$

und damit folgt nach Division $\|u - u_h\|_E \leq \|u - v\|_E$, was zu zeigen war. □

•

Bemerkung. Die Gleichung (2.15) drückt aus, dass die Verbindung $u - u_h$ orthogonal zum Raum V_h steht und wird daher auch *Galerkin-Orthogonalität* genannt.

•

Satz 2.29 (Céa)**.** *Der Fehler der Galerkin-Approximation u_h hat in der H^1-Norm die Eigenschaft*

$$\|u - u_h\|_1 \leq \tilde{c} \inf_{v \in V_h} \|u - v\|_1 .$$

Beweis. Aus (2.12) und Satz 2.28 folgt

$$\|u - u_h\|_1 \leq \left(\frac{1}{\alpha}\right)^{\frac{1}{2}} \|u - u_h\|_E \leq \left(\frac{1}{\alpha}\right)^{\frac{1}{2}} \|u - v\|_E \leq \left(\frac{c}{\alpha}\right)^{\frac{1}{2}} \|u - v\|_1 .$$

Damit folgt die Behauptung mit $\tilde{c} := \sqrt{\frac{c}{\alpha}}$. □

• vgl. [Bra13] Satz 6.4

Theorem 2.30 (Approximationssatz)**.** *Es sei $k \geq 2$ und $\mathcal{T}_h$ eine quasi-uniforme Triangulierung von Ω. Dann gilt für die Interpolation I_h auf die stetigen, stückweise durch Polynome vom Grad $k-1$ gegebenen Funktionen mit einer von Ω, κ und k abhängigen Kontanten c die a priori Fehlerabschätzung*

$$\|u - I_h u\|_m \leq c h^{k-m} |u|_k$$

für $u \in H^k(\Omega)$ und $0 \leq m \leq k$.

Beweis. Für den Beweis würden wir noch weitere Ausführungen über affine Transformationen benötigen, die wir hier nicht weiter aufführen wollen. Der komplette Beweis ist in [Bra13] auf Seite 75ff einzusehen. □

• für $k = 2$ (lineare Polynome) und $m = 1$ (Norm in H^1) gilt dann

$$\|u - I_h u\|_1 \leq c h |u|_2$$

für $u \in H^2(\Omega)$.

•

Korollar 2.31. *Für lineare C^0-Elemente gilt bzgl. der Galerkin-Approximation u_h die a priori Fehlerschätzung für unser Modellproblem* (DP)

$$\|u - u_h\|_1 \leq \tilde{c}h|u|_2 \,.$$

Beweis. Mit Theorem 2.30 und Satz 2.29 folgt

$$\begin{aligned}\|u - u_h\|_1 &\leq \left(\frac{c_1}{\alpha}\right)^{\frac{1}{2}} \inf_{v \in V_h} \|u - v\|_1 \\ &\leq \left(\frac{c_1}{\alpha}\right)^{\frac{1}{2}} \|u - I_h u\|_1 \\ &\leq \left(\frac{c_1}{\alpha}\right)^{\frac{1}{2}} c_2 h|u|_2 \,.\end{aligned}$$

Mit $u \in H^2(\Omega)$ und $\tilde{c} := \left(\frac{c_1}{\alpha}\right)^{\frac{1}{2}} c_2$ folgt dann die Behauptung. □

$\Rightarrow$ Überleitung zu adaptiven Verfahren

- durch Verfeinerung des Netzes (Verkleinerung von h) wird der Fehler zwischen der exakten Lösung u und der Galerkin-Approximation u_h linear kleiner $\Rightarrow$ verfeinere das Netz hinreichend weit, um möglichst genaues Ergebnis zu erhalten

2.4 Adaptive Verfeinerungsstrategien

- durch Netzverfeinerung erhält das numerische Problem mehr Informationen (mehr Punkte werden betrachtet)

$\Rightarrow$ größeres LGS, also langsamer zu Lösen; dies ist nicht vorteilhaft

- Abhilfe: wir verfeinern nur Dreiecke bzw. Knoten, die einen großen lokalen Anteil am Gesamtfehler haben, d.h. an Stellen, wo der Fehler groß ist, erhöhen wir den Informationsgrad (mehr Punkte), um den Fehler zu Verfeinern

- die zu verfeinernden Knoten werden abhängig von der aktuellen Verfeinerung berechnet, d.h. adaptiv

2.4.1 A posteriori Fehlerschätzer

- für die zu verfeinernden Knoten muss der Fehler zwischen der Approximation u_h und der (nicht unbedingt bekannten) exakten Lösung u abschätzen

- hierfür a posteriori Schätzer zu benutzen $\rightarrow$ ein Fehlerschätzer, der den Fehler im Schritt $n+1$ durch den Fehler im Schritt n abschätzt.

- es gibt verschiedene Arten von Fehlerschätzern (vgl. [Bra13] Kapitel III, §8, Seite 176)
  (a) Residuale Schätzer
  (b) Schätzung über ein lokales Neumann-Problem
  (c) Schätzung über ein lokales Dirichlet-Problem
  (d) Schätzung durch Mitteilung
  (e) Hierarchische Schätzer
- wir wollen in dieser Arbeit nur hierarchische Schätzer betrachten
- daher: Idee von hierarchischem Schätzer erklären

2.4.2 Verfeinerung des Netzes

- wie kann ein Element verfeinert werden → Algorithmus, damit auch auch die Triangulierung konform bzw. zulässig bleibt

⇒ Regeln für die Zulässigkeit nötig: vgl. [Bra13] Seite 96 unten Punkt 8.1 oder [Sta08]

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):
$$\boldsymbol{\varepsilon} = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla^T \boldsymbol{u})$$
- Kinetik: Kräftegleichgewicht und äußere Kontaktlast
- Konzepte für ebene Spannungs- bzw. Verzerrungszustände (siehe hierfür auch FEM 1 Skript von Wriggers → [Wri09])
- Konstitutive Modelle (vor allem Materialgesetze) ⇒ Hier vor allem Hooke:
$$\boldsymbol{\sigma} = \mathcal{C}\boldsymbol{\varepsilon} = 2\mu\boldsymbol{\varepsilon} + \lambda(\operatorname{tr}\boldsymbol{\varepsilon})\boldsymbol{I}\,,$$
wobei λ, μ die Lamé-Konstanten sind (Materialabhängige Parameter). ⇒ Hier noch mal den Zusammengang von Konstanten zu E, ν aufzeigen.
- falls Tensorrechnungen konkret benötigt werden, können diese im Anhang dargelegt werden

Kapitel 3

Variationsungleichungen

Dieses Kapitel basiert auf [KO88], [Sta11], [Ste12b], [Ste12a], [Wri01], [Wri06], [HHNL80], [Glo08], [Fal74].

3.1 Ein Hindernisproblem

- Hindernisproblem: Auslenkung u einer Membran Ω unter Krafteinwirkung f, wobei die Membran durch ein Hindernis ψ behindert wird. Mathematische modelliert bedeutet dies:

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

  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 $K = \{v \in H_0^1(\Omega) \mid v \geq \psi$ 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 Ω. Dann gilt für $t \in [0,1]$

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

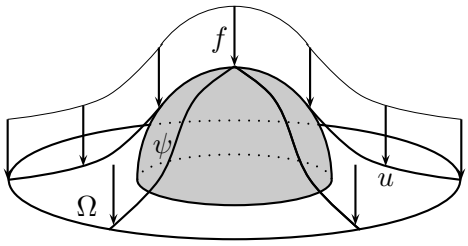

Abbildung 3.1: Ein Hindernisproblem mit Hindernis ψ, konstanter Streckenlast f und Lösung u

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

(ii) Es sei $(v_n)_{n\in\mathbb{N}} \subset K$ eine konvergente Folge mit $v_n \to v$ für $n \to \infty$. Da $H_0^1(\Omega)$ ein abgeschlossener Unterraum von $H^1(\Omega)$ (vgl. auch [Wal11] Bemerkung 6.7) ist, folgt direkt $v \in H_0^1(\Omega)$. Da weiter $v_n \geq \psi$ für alle $n \in \mathbb{N}$ gilt, folgt aus dem Spursatz (vgl. [Bra13] Kapitel II, §3, Satz 3.1), dass auch $v \geq \psi$ fast überall in Ω gilt und damit ist $v \in K$, d.h. K ist abgeschlossen. □

•

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

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

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

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

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

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

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. □

•

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

- auch für das Hindernisproblem gibt es analog zum homogenen Dirichlet-Problem (2.2) 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*

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

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)*.*

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

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

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

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

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

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

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. □

3.1.2 Existenz und Eindeutigkeit der Lösung

- für die Existenz und Eindeutigkeit der Lösung des Problems betrachten wir zunächst wieder das allgemeine reelle quadratische Funktional $J : H \to \mathbb{R}, J(v) = \frac{1}{2}a(v, v) - F(v)$.

- **Voraussetzung 3.5.** 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) $a : H \times H \to \mathbb{R}$ ist eine stetige koerzive Bilinearform,

(b) $F : H \to \mathbb{R}$ ist ein stetiges lineares Funktional,

(c) $K \neq \emptyset$ ist eine abgeschlossene konvexe Teilmenge von H.

•

Theorem 3.6 (Existenz und Eindeutigkeit)**.** *Unter den obigen Voraussetzungen hat die Variationsungleichung, finde $u \in K$, so dass*

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

ist, genau eine Lösung.

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

$$a(u_1, v-u_1) \geq F(v-u_1) \quad \forall\, v \in K\,, \tag{3.8}$$
$$a(u_2, v-u_2) \geq F(v-u_2) \quad \forall\, v \in K\,. \tag{3.9}$$

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 2.16 folgt, dass ein $A \in \mathcal{L}(H,H), l \in H$ existiert, so dass

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

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

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

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

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

mit $\|A\| := \sup_{v \in H} \frac{\|Av\|_H}{\|v\|_H}$. Also ist die Abbildung W_ϱ 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 ϱ 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. □

•

Korollar 3.7. *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 3.6 erfüllt. Damit hat das Problem, finde $u \in K$, so dass

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

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

•

Bemerkung 3.8. Insbesondere hat auch das Problem (3.5) nach Satz 3.4 und Theorem 3.6 eine eindeutige Lösung, wenn $u \in H^2(\Omega) \cap H_0^1(\Omega)$ ist.

3.1.3 Lösung des Hindernisproblems mittels FEM

- zur Lösung mittels FEM betrachten wir die Variationsungleichung (3.11) bzgl. eines endlich dimensionalen Unterraum

$$K_h \coloneqq \{v_h \in \mathcal{S}_h \mid v_h(p) \geq \psi(p) \, \forall \, p \in \mathcal{N} \cap \Omega\},$$

wobei $\mathcal{N}$ die Knotenmenge bzgl. der Triangulierung $\mathcal{T}_h$ bezeichne.

- damit ist (3.11) in diskreter Form: Finde $u_h \in K_h$, so dass

$$a(u_h, v_h - u_h) \geq (f, v_h - u_h) \quad \forall \, v_h \in K_h \,. \tag{3.12}$$

- vgl. [Wer11] Kapitel 4 Satz 7.15.

Satz 3.9 (Fixpunktsatz von Brouwer)**.** *Es sei $K \neq \emptyset$ eine kompakte konvexe Teilmenge eines endlich dimensionalen normierten Raumes H und $F : K \to K$ sei stetig. Dann besitzt F einen Fixpunkt $v \in K$.*

Beweis. Der Beweis ist in [Wer11] Kapitel 4 Satz 7.15 zu finden. □

•

Theorem 3.10 (Existenz und Eindeutigkeit)**.** *Das Problem* (3.12) *hat eine eindeutige Lösung $u_h \in K_h$.*

Beweis. Der Beweis ist analog zu Theorem 3.6 zu führen. Wir ersetzen lediglich H durch V_h und K durch K_h und verwenden im endlich dimensionalen Raum V_h den Fixpunktsatz von Brouwer. □

•

Bemerkung. In Kapitel 2.2 von [Sta08] sind die Argumente bzgl. der Existenz und Eindeutigkeit einer Lösung von (3.11) für den endlich dimensionalen Fall K_h auch noch einmal im Einzelnen aufgeführt.

- Es sei $\mathcal{B}_h = \{\phi_1, \ldots, \phi_N\}$ eine modale Basis von $\mathcal{S}_h$, d.h. analog zu (2.11) können wir u_h und v_h mit Koordinaten $\mu_i, \nu_i, i = 1, \ldots, N$ bzgl. $\mathcal{B}_h$ ausdrücken. Dann schreiben wir (3.12) als

$$\begin{aligned} & \sum_{i=1}^{N} \sum_{j=1}^{N} \mu_i \, a(\phi_i, \phi_j) \, (\nu_j - \mu_j) \geq \sum_{j=1}^{N} (f, \phi_j) \, (\nu_j - \mu_j) \\ \iff \quad & \boldsymbol{\mu}^T A (\boldsymbol{\nu} - \boldsymbol{\mu}) \geq \boldsymbol{f}^T (\boldsymbol{\nu} - \boldsymbol{\mu}) \end{aligned}$$

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

- Die Menge K_h ist bzgl. $\mathcal{S}_h$ äquivalent zu

$$K_{\mathcal{S}} \coloneqq \{\boldsymbol{\nu} \in \mathbb{R}^N \mid \nu_i \geq \psi(p_i), p_i \in \mathcal{N} \cap \Omega, i = 1, \ldots, N\}\,. \tag{3.13}$$

  Im Folgenden schreiben wir $\boldsymbol{\psi} \coloneqq [\psi(p_i)]_{i=1}^N$ mit $p_i \in \mathcal{N} \cap \Omega$.

- *Bemerkung.* $K_{\mathcal{S}}$ ist analog zu K konvex und abgeschlossen.

- Damit erhalten wir aus (3.12) die diskrete Variationsungleichung: Finde $\boldsymbol{\mu} \in K_{\mathcal{S}}$, so dass

$$(A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\nu} - \boldsymbol{\mu}) \geq 0 \quad \forall\, \boldsymbol{\nu} \in K_{\mathcal{S}}\,. \tag{3.14}$$

- **Satz 3.11.** *Das Problem* (3.14) *ist äquivalent zum linearen Komplementaritätsproblem: Bestimme* $\boldsymbol{\mu} \in K_{\mathcal{S}}$*, so dass*

$$A\boldsymbol{\mu} - \boldsymbol{f} \geq \mathbf{0} \quad \textit{und} \quad (A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\mu} - \boldsymbol{\psi}) = 0 \tag{3.15}$$

  gilt.

  Beweis. „⇒“ Sei $\boldsymbol{\mu} \in K_{\mathcal{S}}$ Lösung von (3.14). Wir setzen $\boldsymbol{\nu} = \boldsymbol{\mu} + \boldsymbol{e}_i \geq \boldsymbol{\psi}$ mit einem beliebigen $i \in \{1, \ldots, N\}$, wobei $\boldsymbol{e}_i$ den i-te Einheitsvektor bezeichne. Dann gilt

$$0 \leq (A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\nu} - \boldsymbol{\mu}) = (A\boldsymbol{\mu} - \boldsymbol{f})^T\boldsymbol{e}_i = (A\boldsymbol{\mu} - \boldsymbol{f})_i\,.$$

  Da i beliebig war, folgt $A\boldsymbol{\mu} - \boldsymbol{f} \geq \mathbf{0}$.

  Wir nun nehmen an, dass ein $i \in \{1, \ldots, N\}$ existiert, so dass $(A\boldsymbol{\mu} - \boldsymbol{f})_i(\boldsymbol{\mu} - \boldsymbol{\psi})_i > 0$ ist. Weiter wählen wir

$$\boldsymbol{\nu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{i-1} \\ 0 \\ \mu_{i+1} \\ \vdots \\ \mu_N \end{pmatrix} + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \psi_i \\ 0 \\ \vdots \\ 0 \end{pmatrix} \geq \boldsymbol{\psi}$$

  und damit folgt

$$0 > (A\boldsymbol{\mu} - \boldsymbol{f})_i(\boldsymbol{\psi} - \boldsymbol{\mu})_i = (A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\nu} - \boldsymbol{\mu}) \geq 0\,,$$

  was im Widerspruch zu (3.14) steht, daraus folgt die Behauptung.

„⇐“ Es sei $\boldsymbol{\mu} \in K_{\mathcal{S}}$ Lösung von (3.15). Dann rechnen wir nach, dass für ein beliebiges $\boldsymbol{\nu} \in K_{\mathcal{S}}$ gilt

$$\begin{aligned}(A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\nu} - \boldsymbol{\mu}) &= (A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\nu} - \boldsymbol{\psi} + \boldsymbol{\psi} - \boldsymbol{\mu}) \\ &= \underbrace{(A\boldsymbol{\mu} - \boldsymbol{f})^T}_{\geq \mathbf{0}} \underbrace{(\boldsymbol{\nu} - \boldsymbol{\psi})}_{\geq \mathbf{0}} - \underbrace{(A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\mu} - \boldsymbol{\psi})}_{=0} \\ &\geq 0\,.\end{aligned}$$

□

- das Problem (3.14) ist äquivalent zu einem quadratischen Optimierungsproblem

  Satz 3.12 (Äquivalenz zu quadratischem Programm)**.** *Das Problem* (3.14) *ist äquivalent zum quadratischen Programm*

  $$\min_{\boldsymbol{\nu} \in \mathbb{R}^N} J(\boldsymbol{\nu}) = \frac{1}{2}\boldsymbol{\nu}^T A\boldsymbol{\nu} - \boldsymbol{f}^T\boldsymbol{\nu} \quad s.t. \quad \boldsymbol{\nu} \geq \boldsymbol{\psi}\,. \tag{3.16}$$

  Beweis. Wir zeigen zunächst die Äquivalenz von (3.15) zu (3.16) und dann folgt mit Satz 3.11 die Behauptung.

  „⇒“ Es sei $\boldsymbol{\mu} \in \mathbb{R}^N$ Lösung vom Problem (3.15). Dann folgt mit einem beliebigen $\boldsymbol{\nu} \in K_{\mathcal{S}}$

  $$\begin{aligned}J(\boldsymbol{\nu}) - J(\boldsymbol{\mu}) &= \frac{1}{2}\boldsymbol{\nu}^T A\boldsymbol{\nu} - \boldsymbol{f}^T\boldsymbol{\nu} - \frac{1}{2}\boldsymbol{\mu}^T A\boldsymbol{\mu} + \boldsymbol{f}^T\boldsymbol{\mu} \\ &= \frac{1}{2}\underbrace{(\boldsymbol{\nu} - \boldsymbol{\mu})^T A(\boldsymbol{\nu} - \boldsymbol{\mu})}_{\geq 0 \text{ wegen Bem. 2.18}} + \boldsymbol{\mu}^T A\boldsymbol{\nu} - \boldsymbol{\mu}^T A\boldsymbol{\mu} - \boldsymbol{f}^T(\boldsymbol{\nu} - \boldsymbol{\mu}) \\ &\geq (A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\nu} - \boldsymbol{\psi} + \boldsymbol{\psi} - \boldsymbol{\mu}) \\ &= \underbrace{(A\boldsymbol{\mu} - \boldsymbol{f})^T}_{\geq 0} \underbrace{(\boldsymbol{\nu} - \boldsymbol{\psi})}_{\geq 0} - \underbrace{(A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\mu} - \boldsymbol{\psi})}_{=0} \\ &\geq 0\,.\end{aligned}$$

  Somit ist $\boldsymbol{\mu} \in K_{\mathcal{S}}$ auch Lösung des quadratischen Programms (3.16).

  „⇐“ Sei $\boldsymbol{\mu} \in K_{\mathcal{S}}$ Lösung von (3.16), dann gelten nach [NW06] Kapitel 12, Theorem 12.1 für die Lagrange-Funktion

  $$\mathcal{L}(\boldsymbol{\nu}, \boldsymbol{\lambda}) = J(\boldsymbol{\nu}) - \boldsymbol{\lambda}^T(\boldsymbol{\nu} - \boldsymbol{\psi})$$

  die Karush-Kuhn-Tucker Bedingungen für den Optimalpunkt $(\boldsymbol{\mu}, \boldsymbol{\lambda}^*)$

  $$\nabla_{\nu}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\lambda}^*) = \nabla J(\boldsymbol{\mu}) - \lambda^* = A\boldsymbol{\mu} - \boldsymbol{f} - \boldsymbol{\lambda}^* \stackrel{!}{=} \mathbf{0}\,, \tag{3.17}$$

  $$\boldsymbol{\mu} - \boldsymbol{\psi} \geq \mathbf{0}\,, \tag{3.18}$$

  $$\boldsymbol{\lambda}^* \geq \mathbf{0}\,, \tag{3.19}$$

  $$\lambda_i^*(\mu_i - \psi_i) = 0 \qquad \forall\, i = 1, \ldots, N\,. \tag{3.20}$$

Mit (3.17) gilt also $\boldsymbol{\lambda}^* = A\boldsymbol{\mu} - \boldsymbol{f}$ und daher folgt aus (3.19)

$$A\boldsymbol{\mu} - \boldsymbol{f} \geq \mathbf{0}\,.$$

Aus (3.20) folgt wegen $(A\boldsymbol{\mu} - \boldsymbol{f})_i(\mu_i - \psi_i) = 0$ für alle $i = 1, \ldots, N$ direkt

$$(A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\mu} - \boldsymbol{\psi}) = 0\,.$$

Also ist $\mu \in K_{\mathcal{S}}$ auch Lösung von (3.15). □

•

Bemerkung 3.13. Analog zu Satz 3.11 und 3.12 können wir auch zeigen, dass das quadratische Programm

$$\min_{\boldsymbol{\nu} \in \mathbb{R}^N} J(\boldsymbol{\nu}) = \frac{1}{2}\boldsymbol{\nu}^T A\boldsymbol{\nu} - \boldsymbol{f}^T\boldsymbol{\nu} \quad \text{s.t.} \quad B\boldsymbol{\nu} \geq \boldsymbol{\psi} \tag{3.21}$$

mit $B \in \mathbb{R}^{M \times N}$ äquivalent ist zur Variationsungleichung: Finde $\boldsymbol{\mu} \in \mathbb{R}^N$ mit $B\boldsymbol{\mu} \geq \boldsymbol{\psi}$, so dass

$$(A\boldsymbol{\mu} - \boldsymbol{f})^T(\boldsymbol{\nu} - \boldsymbol{\mu}) \geq 0 \quad \forall\, \boldsymbol{\nu} \in \mathbb{R}^N \text{ mit } B\boldsymbol{\nu} \geq \boldsymbol{\psi}\,. \tag{3.22}$$

•

Bemerkung 3.14. Die quadratischen Programme (3.16) und (3.21) hat mit den Voraussetzungen aus Bemerkung 2.18 eine globale Lösung $\boldsymbol{\mu}$, wenn diese die KKT-Bedingungen (B.3) erfüllt (vgl. Theorem B.1).

- mit dem im Anhang B.2 vorgestellten Active-Set-Algorithmus kann ein solches quadratisches Programm gelöst werden
- es bleibt noch zu prüfen, ob sich die Lösung der Variationsungleichung für Netzverfeinerung an die exakte Lösung konvergiert

•

Voraussetzung 3.15. Gegeben sei ein Parameter $h \to 0$. Weiter seien $(V_h)_h$ eine Familie aus abgeschlossenen Teilmengen von einem Hilbertraum $H, \emptyset \neq K \subset H$ eine konvexe abgeschlossene Teilmenge und $(K_h)_h$ eine Familie von abgeschlossenen konvexen nichtleeren Teilmengen von H, so dass $K_h \subset V_h$ für alle h.

Dabei sei K_h eine Approximation von K im folgenden Sinne:

(i) wenn $(v_h)_h$ eine in V beschränkte Folge mit $v_h \in K_h$ ist, dann folgt $v_h \to v \in K$,

(ii) es existiert ein $W \subset V$ mit $\overline{W} = K$ and ein $I_h : W \to K_h$, so dass

$$\lim_{h\to 0} I_h v = v$$

stark in V für alle $v \in W$ konvergiert.

- Konvergenz von u_h gegen die exakte Lösung u.

Theorem 3.16 (a priori Konvergenz)**.** *Mit den obigen Voraussetzungen für K und $(K_h)_h$ gilt für die Lösungen u von (3.7) und u_h vom approximierten Problem: Finde $u_h \in K_h$, so dass*

$$a(u_h, v_h - u_h) \geq F(v_h - u_h) \quad \forall\, v_h \in K_h\,, \tag{3.23}$$

der Zusammenhang

$$\lim_{h\to 0} \|u_h - u\|_H = 0\,.$$

Beweis. (i) Abschätzung von u_h: Es sei u_h Lösung von (3.23), dann gilt nach einer Umformung für alle $v_h \in K_h$

$$\begin{aligned} a(u_h, u_h) &\leq a(u_h, v_h) - F(v_h - u_h) \\ &= (Au_h, v_h)_H - (f, v_h - u_h)_H \\ &\overset{\text{CS}}{\leq} \underbrace{\|Au_h\|_H}_{\leq \|A\|\|u_h\|_H}\, \|v_h\|_H + \|f\|_H\, \underbrace{\|v_h - u_h\|_H}_{\leq \|v_h\|_H + \|u_h\|_H} \\ &\leq \|A\|\, \|u_h\|_H \|v_h\|_H + \|f\|_H\, (\|v_h\|_H + \|u_h\|_H)\,. \end{aligned}$$

Zusammen mit der Koerzivität folgt dann

$$\alpha\, \|u_h\|_H^2 \leq \|A\|\, \|u_h\|_H \|v_h\|_H + \|f\|_H\, (\|v_h\|_H + \|u_h\|_H)\,. \tag{3.24}$$

Wähle ein festes $v_0 \in W$, sodass $I_h v_0 = v_h \in K_h$ gilt. Aus Voraussetzungen (ii) folgt dann

$$\lim_{h\to 0} I_h v_0 = v_0$$

und daher muss v_h beschränkt sein, d.h. es existiert ein $m \in \mathbb{R}$: $\|v_h\|_H \leq m$ für alle h. Zusammen mit (3.24) gilt dann

$$\begin{aligned} \|u_h\|_H^2 &\leq \frac{1}{\alpha}(m\, \|A\|\, \|u_h\|_H + \|f\|_H (m + \|u_h\|_H)) \\ &= \underbrace{\Big(\frac{m}{\alpha}\, \|A\| + \|f\|_H\Big)}_{=:c_1} \|u_h\|_H + \underbrace{\frac{m}{\alpha}\|f\|_H}_{=:c_2} \\ &= c_1\, \|u_h\|_H + c_2 \end{aligned}$$

und damit können wir durch quadratischer Ergänzung folgern, dass es ein $c \in \mathbb{R}$ gibt mit $\|u_h\| \leq c$ für alle h, d.h. $(u_h)_h$ ist gleichmäßig beschränkt.

(ii) schwache Konvergenz: Da $(u_h)_h$ in H gleichmäßig beschränkt ist, folgt mit Bemerkung A.13 (b), dass es eine schwach konvergente Teilfolge $(u_{h_j})_{h_j} \in K_{h_j}$ mit einem Grenzwert u^* in H gibt, d.h.

$$u_{h_j} \rightharpoonup u^* \in H\,.$$

Mit den Voraussetzungen (i) für $(K_h)_h$ folgt direkt $u^* \in K$, außerdem ist u^* nach Bemerkung A.13 (e) eindeutig.

Wir zeigen nun, dass u^* eine Lösung von (3.7) ist. Für die oben betrachtete Teilfolge gilt

$$a(u_{h_j}, v_{h_j} - u_{h_j}) \geq F(v_{h_j} - u_{h_j}) \quad \forall\, v_{h_j} \in K_{h_j}\,. \tag{3.25}$$

Sei $v \in W$ mit $v_{h_j} = I_{h_j} v$. Dann gilt $v_{h_j} = I_{h_j} v \to v \in W$ für $h_j \to 0$. Mit (3.25) folgt

$$\begin{aligned} a(u_{h_j}, u_{h_j}) &\leq a(u_{h_j}, v_{h_j}) - F(v_{h_j} - u_{h_j}) \\ &= a(u_{h_j}, I_{h_j} v) - F(v_{h_j} - u_{h_j}) \\ \Longrightarrow \liminf_{h_j \to 0} a(u_{h_j}, u_{h_j}) &\leq a(u^*, v) - F(v - u^*)\,. \end{aligned} \tag{3.26}$$

Weiter schätzen wir durch Bemerkung A.13 (f) nach unten ab

$$\liminf_{h_j \to 0} a(u_{h_j}, u_{h_j}) = \liminf_{h_j \to 0} \|u_{h_j}\|_H^2 \geq \|u^*\|_H^2 = a(u^*, u^*)\,. \tag{3.27}$$

Insgesamt folgt also mit (3.26) und (3.27)

$$a(u^*, u^*) \leq \liminf_{h_j \to 0} a(u_{h_j}, u_{h_j}) \leq a(u^*, v) - F(v - u^*)$$

und damit nach Umformung

$$a(u^*, v - u^*) \geq F(v - u^*) \quad \forall\, v \in W\,.$$

Da W dicht in K liegt, d.h. $\overline{W} = K$, und a, F stetig sind, erhalten wir

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

mit $u^* \in K$, also ist $u^* =: u$ Lösung von (3.7). Da u ein Häufungspunkt von $(u_h)_h$ bzgl. der schwachen Topologie von H ist, konvergiert auch die Folge $(u_h)_h$ schwach gegen u.

(iii) starke Konvergenz: Aus der Koerzivität von a folgt

$$\begin{aligned} 0 \leq \alpha \|u_h - u\|_H^2 &\leq a(u_h - u, u_h - u) \\ &\leq a(u_h, u_h) - a(u_h, u) - a(u, u_h) + a(u, u)\,, \end{aligned} \tag{3.28}$$

wobei u_h Lösung vom approximierten Problem (3.23) und u Lösung vom exakten Problem (3.7) ist. Es sei $v \in W$ mit $I_h v = v_h \in K_h$, dann folgt aus (3.23)

$$a(u_h, u_h) \leq a(u_h, I_h v) - F(I_h v - u_h) \quad \forall\, v \in W\,. \tag{3.29}$$

Da $u_h \rightharpoonup u$ in H und $I_h v \to v$ in H für $h \to 0$, folgt aus (3.28) und (3.29) unter Verwendung von Voraussetzungen (ii)

$$0 \leq \alpha \lim_{h\to 0} \|u_h - u\|_H^2 \leq a(u, v-u) - F(v-u) \quad \forall\, v \in W\,. \tag{3.30}$$

Da a und F stetig sind und W dicht in K liegt, gilt (3.30) auch für alle $v \in K$. Setzen wir dann also $v = u$ in (3.30), dann folgt die Behauptung $\lim_{h\to 0}\|u_h - u\|_H^2 = 0$. □

- Überlegung: Inwiefern hält unser K_h diese Voraussetzungen ein.
- Es lässt sich folglich auch eine a priori Abschätzung für den Fehler von u und u_h machen. Die Herleitung ist detailliert in [Fal74] wiederzufinden.

  Theorem 3.17 (a priori Fehlerabschätzung)**.** *Es seien u und u_h die Lösungen von* (3.7) *und* (3.23)*. Dann existiert eine Konstante $C := C(\Omega, f, \psi)$ unabhängig von u, so dass*

  $$\|u_h - u\|_1 \leq Ch\,.$$

  Beweis. Vgl. [Fal74]. □

- damit führt die Netzverfeinerung also zur exakten Lösung der Variationsungleichung
- inwiefern adaptive Netzverfeinerung hier sinnvoll ist, wollen wir in Kapitel 4 betrachten

3.2 Kontaktprobleme

3.2.1 Mathematische Modellierung eines Kontaktproblems

-

  Voraussetzung 3.18. Wir treffen folgende Annahmen für unser Kontaktmodell:

  (a) Die in Kontakt stehenden Körper sind beschränkt.

  (b) Es liegen kleine Deformationen und linear elastische Materialien vor.

(c) Wir betrachten ein konstantes Temperaturfeld, d.h. thermodynamische Prozesse werden ausgeschlossen.

(d) Zu Beginn, also in der Ausgangskonfiguration, gilt für die Spannung und Verzerrung:

$$\boldsymbol{\sigma} = \mathbf{0}\,, \quad \boldsymbol{\varepsilon} = \mathbf{0}\,.$$

(e) Wir gehen von einem reibungslosen Kontakt aus. Dieses Kontaktproblem wird als *Signorini-Kontakt-Problem* bezeichnet.

(f) Wir gehen von ebenen Problemen aus, d.h. $\Omega \subset \mathbb{R}^2$. Im $\mathbb{R}^3$ sind alle Resultate analog.

- zur Herleitung der starken Kontaktformulierung wollen wir zwei Körper $\mathcal{B}^1, \mathcal{B}^2$ betrachten, welche durch zwei beschränkte Gebiete Ω^1, Ω^2 mathematisch beschrieben werden können
- diese Voraussetzung lässt sich noch weiter verallgemeinern (s. [CSW99])
- die Ränder Γ^i von $\Omega^i, i = 1, 2$, lassen sich in drei disjunkte Teile unterteilen

  Γ_u^i: Der *Dirichlet-Rand*, oder auch *Verschiebungsrand*, auf dem die Werte von der Verschiebung $\boldsymbol{u}$ vorgegeben sind.

  Γ_σ^i: Der *Neumann-Rand*, oder auch *Spannungsrand*, auf dem die Oberflächenlast bzw. -spannung $\bar{\boldsymbol{t}}$ vorgegeben ist.

  Γ_c^i: Der *Kontaktrand*, auf dem die Kontaktbedingungen definiert sind.

Skizze: zwei Körper, deren Randunterteilung zu erkennen ist.

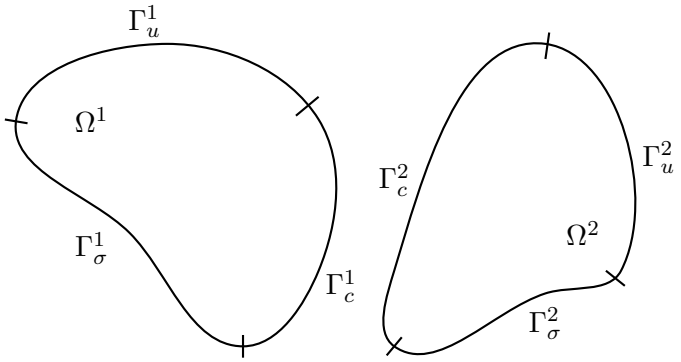

Abbildung 3.2: Körper $\mathcal{B}^1$ und $\mathcal{B}^2$ mit Randbezeichnungen

- zur Kontaktkinematik:

- für die Formulierung der Kontaktbedingungen werden den Körpern die „Werte“ *master* und *slave* zugeordnet.
- slave ist dabei die Menge an Punkten, die überprüft werden, ob sie in die master-Fläche eindringen.
- Zuordnung ist irrelevant → kein Unterschied für das Ergebnis
- o.B.d.A. sei $\mathcal{B}^1$ slave
- wir wollen zunächst die Kontaktkinematik etwas allgemeiner als in [Wri01] oder [Wri06] beschrieben einführen

Skizze: Skizze mit zwei Körpern (nichtglatter Rand!!) und den Bezeichnungen $\chi(X)$ und $n_c(X)$.

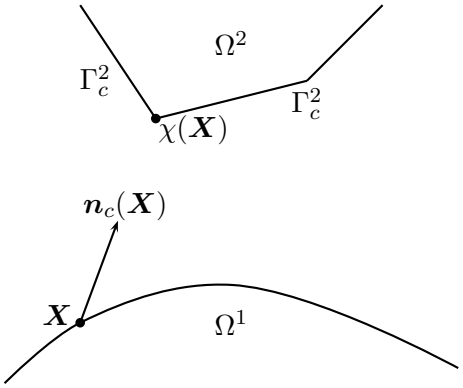

Abbildung 3.3: Kontaktformulierung zwischen zwei Körpern

- für gegebenen Punkt $\boldsymbol{X} \in \Omega^1$, bzw. $\boldsymbol{X} \in \Gamma_c^1$, in der Ausgangskonfiguration ist $\bar{\boldsymbol{X}} \coloneqq \chi(\boldsymbol{X})$ derjenige Punkt aus Γ^2, der minimalsten Abstand zu $\boldsymbol{X}$ hat, d.h.

$$\|\boldsymbol{X} - \bar{\boldsymbol{X}}\| = \min\{\|\boldsymbol{X} - \boldsymbol{Y}\| \mid \boldsymbol{Y} \in \Gamma^2\},$$

  also ist $\chi : \Gamma_c^1 \cup \Gamma_c^2 \to \Gamma^1 \cup \Gamma^2$ eine Abbildung der kleinsten Distanz

- damit definieren wir entsprechend die kritische Richtung mit Länge 1 als

$$\boldsymbol{n}_c(\boldsymbol{X}) \coloneqq \frac{\chi(\boldsymbol{X}) - \boldsymbol{X}}{\|\chi(\boldsymbol{X}) - \boldsymbol{X}\|}, \tag{3.31}$$

  wobei im Falle $\boldsymbol{X} = \chi(\boldsymbol{X})$, d.h. im Falle des Kontaktes, eine beliebige normierte Richtung gesetzt wird.

- Bem.: $\bar{\boldsymbol{X}}$ ist daher kritischer Punkt, da er wegen des kleinsten Abstandes zu $\boldsymbol{X}$ der wohlmöglich nächste Punkt ist, der in Kontakt tritt

Vorteil: diese Formulierung bzgl. der kritischen Richtung kann auch verwendet werden, wenn der Rand der Körper nicht hinreichend glatt ist.

- in den Koordinaten der Momentankonfiguration gilt $\boldsymbol{x} = \boldsymbol{X} + \boldsymbol{u}$ für das Verschiebungsfeld $\boldsymbol{u}$ und damit ergibt sich die *Nichtdurchdringungsbedingung*

$$(\bar{\boldsymbol{x}} - \boldsymbol{x})\, \boldsymbol{n}_c(\boldsymbol{X}) \geq 0\,, \tag{3.32}$$

  wobei $\bar{\boldsymbol{x}} \coloneqq \bar{\boldsymbol{X}} + u(\bar{\boldsymbol{X}})$ ist.

- das bedeutet, dass die Verbindung der Punkte in der Momantankonfiguration mit der kritischen Richtung einen Winkel $\alpha \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ einschließen muss, ansonsten läge $\bar{\boldsymbol{x}}$ „hinter“ $\boldsymbol{x}$, d.h. $\mathcal{B}^1$ wäre in $\mathcal{B}^2$ eingedrungen.

- aus (3.32) folgt

$$\begin{aligned} 0 \leq (\bar{\boldsymbol{x}} - \boldsymbol{x})\, \boldsymbol{n}_c(\boldsymbol{X}) &= (\bar{\boldsymbol{X}} + u(\bar{\boldsymbol{X}}) - \boldsymbol{X} - u(\boldsymbol{X}))\, \boldsymbol{n}_c(\boldsymbol{X}) \\ &= \underbrace{(\bar{\boldsymbol{X}} - \boldsymbol{X}) \frac{\bar{\boldsymbol{X}} - \boldsymbol{X}}{\|\bar{\boldsymbol{X}} - \boldsymbol{X}\|}}_{=\|\bar{\boldsymbol{X}} - \boldsymbol{X}\| =: g} + (u(\bar{\boldsymbol{X}}) - u(\boldsymbol{X}))\, \boldsymbol{n}_c(\boldsymbol{X}) \\ &= g + (u(\chi(\boldsymbol{X})) - u(\boldsymbol{X}))\, \boldsymbol{n}_c(\boldsymbol{X})\,. \end{aligned}$$

  damit erhalten wir die Nichtdurchdringungsbedingung bzgl. der Ausgangskonfiguration

- da wir kleine Deformationen vorausgesetzt haben, gilt unter anderem $\boldsymbol{X} \approx \boldsymbol{x}, \nabla_{\boldsymbol{X}} \approx \nabla_{\boldsymbol{x}}$ (vgl. [Alt12] S. 122f). Daher schreiben wir im folgenden immer $\boldsymbol{x}$ statt $\boldsymbol{X}$.

- damit schreibt sich die Nichtdurchdringungsbedingung als

$$(\boldsymbol{u} \circ \chi - \boldsymbol{u}) \cdot \boldsymbol{n}_c + g \geq 0 \quad \forall\, \boldsymbol{x} \in \Gamma_c\,, \tag{3.33}$$

  wobei $\Gamma_c \coloneqq \Gamma_c^1 \cup \Gamma_c^2$ ist.

- im Folgenden wollen wir der Einfachheit halber wir davon ausgehen, dass die Ränder Γ^i hinreichend glatt sind $\Rightarrow$ (3.33) gilt auch für $\boldsymbol{n}_c$ als Einheitsnormale von $\mathcal{B}^1$

- weiter wollen wir $\boldsymbol{u}(\bar{\boldsymbol{x}}) \equiv \boldsymbol{0}$ annehmen, d.h. falls $\mathcal{B}^2$ ein Verschiebungsfeld ungleich Null hat, können wir $\boldsymbol{u}$ bzgl. $\Omega = \Omega^1 \cup \Omega^2$ auch als Relativverschiebung interpretieren

Hinweis: Wähle später für die feste Ebene im Beispiel $\mathcal{B}^1$ zur Vereinfachung der Darstellung

- damit reduziert sich (3.33) auf

$$\boldsymbol{u} \cdot \boldsymbol{n} - g \leq 0 \quad \forall\, x \in \Gamma_c\,, \tag{3.34}$$

  wobei wir auch $u_n \coloneqq \boldsymbol{u} \cdot \boldsymbol{n}$ im Folgenden schreiben werden.

- weiter muss auf dem Kontaktrand Γ_c die Normalkraft eine Druckkraft sein oder es herrscht Kräftegleichgewicht, d.h. für die Spannung in Normalenrichtung $\sigma_n \coloneqq \boldsymbol{n} \cdot (\boldsymbol{\sigma} \cdot \boldsymbol{n})$ gilt

$$\sigma_n \leq 0 \;\text{ auf } \Gamma_c\,. \tag{3.35}$$

- wie oben schon angedeutet gilt: wenn die Kontaktbedingung nicht aktiv ist (d.h. nicht die Gleichheit gilt), so muss Kräftegleichgewicht herrschen, d.h. in (3.35) gilt die Gleichheit. Zusammen erhält man die Komplementaritätsbedingung

$$(u_n - g)\,\sigma_n = 0 \;\text{ auf } \Gamma_c \tag{3.36}$$

- laut Voraussetzung (e) betrachten wir Signorini-Kontakt und damit muss die Tangentialkraft auf dem Kontaktrand gleich Null sein, d.h.

$$\boldsymbol{\sigma}_t \coloneqq \boldsymbol{\sigma} \cdot \boldsymbol{n} - \sigma_n \boldsymbol{n} = 0 \;\text{ auf } \Gamma_c\,. \tag{3.37}$$

- materialunabhängige Gleichungen:

- wie in Kapitel 2.5 eingeführt, gelten auch hier die Gleichung des Kräftegleichgewichts. Da wir laut Voraussetzung (b) von kleinen Deformationen ausgehen, gilt

$$\operatorname{div} \boldsymbol{\sigma} + \bar{\boldsymbol{b}} = \boldsymbol{0}\,. \tag{3.38}$$

- weiter gilt nach dem *Cauchy-Theorem* (vgl. auch [Wri06] S. 38), dass die Spannung in Normalenrichtung auf der Oberfläche Γ von Ω gleich der von außen angebrachten Spannung $\bar{\boldsymbol{t}}$ ist, d.h.

$$\boldsymbol{\sigma} \cdot \boldsymbol{n} = \bar{\boldsymbol{t}} \;\text{ auf } \Gamma_\sigma\,, \tag{3.39}$$

  also auf dem Neumann-Rand.

- konstitutive Gleichungen:

- da wir laut Voraussetzung (b) von einem linear elastischen Material und kleinen Deformationen ausgehen, gilt ein linearer Zusammenhang bzgl. der Spannung $\boldsymbol{\sigma}$ und Verzerrung $\boldsymbol{\varepsilon}$, d.h. das Hooke'sche-Gesetz ($\boldsymbol{\sigma} = \lambda \operatorname{tr} \boldsymbol{\varepsilon} \cdot I + 2\mu\boldsymbol{\varepsilon}$, wobei λ, μ die Lamé Konstanten sind) und wir können wir den linearisierten Verzerrungstensor $\boldsymbol{\varepsilon}$ verwenden, d.h. mit einem 4 stufigem Materialtensor $\mathcal{C} = (c_{ijkl})$ gilt

$$\boldsymbol{\sigma} - \mathcal{C} : \boldsymbol{\varepsilon} = \boldsymbol{0} \;\text{ in } \Omega\,, \tag{3.40}$$

  wobei $\mathcal{C} : \boldsymbol{\varepsilon}$ das doppelt verjüngende Skalarprodukt beschreibt.

- zusammengefasst mit (3.34) bis (3.40) beschreiben wir also das Signorini-Kontakt-Problem in der starken Formulierung:

$$\begin{aligned}
\operatorname{div}\boldsymbol{\sigma} + \bar{\boldsymbol{b}} &= \mathbf{0} && \text{in } \Omega && \text{(3.41a)}\\
\boldsymbol{\sigma} - \mathcal{C} : \boldsymbol{\varepsilon} &= \mathbf{0} && \text{in } \Omega && \text{(3.41b)}\\
\boldsymbol{\sigma}\cdot\boldsymbol{n} &= \bar{\boldsymbol{t}} && \text{auf } \Gamma_\sigma && \text{(3.41c)}\\
\boldsymbol{u} &= \mathbf{0} && \text{auf } \Gamma_u && \text{(3.41d)}\\
\left.\begin{aligned} u_n - g &\le 0\\ \sigma_n &\le 0\\ (u_n - g)\,\sigma_n &= 0 \end{aligned}\right\} && \text{auf } \Gamma_c && \text{(3.41e)}\\
\boldsymbol{\sigma}_t &= \mathbf{0}
\end{aligned}$$

- was ändert sich, wenn wir Reibung betrachten

  Bemerkung 3.19. Ein Kontaktmodell mit $\boldsymbol{\sigma}_t \neq \mathbf{0}$ ist das Modell mit *Tresca-Reibung*. Für dieses Problem wird die letzte Bedingung aus (3.41e) durch die Bedingungen

$$\|\boldsymbol{\sigma}_t\| \le \mathcal{F}, \quad \boldsymbol{\sigma}_t\boldsymbol{u}_t + \mathcal{F}\,\|\boldsymbol{u}_t\| = 0 \tag{3.42}$$

  mit $\boldsymbol{u}_t \coloneqq \boldsymbol{u} - u_n\boldsymbol{n}$, dem tangentialen Anteil des Verschiebungsfeldes $\boldsymbol{u}$, ersetzt. Hierbei ist $\mathcal{F} \ge 0$ eine Schranke für die Reibung. Gilt $\|\boldsymbol{\sigma}_t\| < \mathcal{F}$, so folgt aus der zweiten Gleichung von (3.42), dass $\boldsymbol{u}_t = \mathbf{0}$ ist. Also kann $\boldsymbol{u}_t \neq \mathbf{0}$ nur gelten, wenn $\|\boldsymbol{\sigma}_t\| = \mathcal{F}$ ist.

  Mit $\mathcal{F} \coloneqq \mu\sigma_n$ erhalten wir das *Reibungsgesetz von Coulomb*, wobei μ den aus der Mechanik bekannten Reibungskoeffizienten darstellt.

  Da die Herleitung der zu diesem Problem äquivalenten Variationsungleichung zusätzliche mathematische Resultate erfordert, werden wir uns in der weiteren Herleitung auf das Signorini-Kontakt-Problem beziehen.

3.2.2 Variationsformulierung des Signorini-Kontaktproblems

- Sei $\Omega \subset \mathbb{R}^2$ (Voraussetzung (f))
- wir betrachten das Signorini-Kontakt-Problem (3.41)
- Es seien $H^1_{\Gamma_u}(\Omega) \coloneqq \{\boldsymbol{v} \in (H^1(\Omega))^2 \mid \boldsymbol{v} = \mathbf{0} \text{ auf } \Gamma_u\}$ und $\mathcal{K} \coloneqq \{\boldsymbol{v} \in H^1_{\Gamma_u}(\Omega) \mid v_n - g \le 0 \text{ auf } \Gamma_c\} \Rightarrow \mathcal{K}$ ist analog zu Lemma 3.1 konvex
- weiter seien $\boldsymbol{u}, \boldsymbol{v} \in K$, wobei $\boldsymbol{u}$ die Lösung des Signorini-Kontaktproblems darstellt und $\boldsymbol{v}$ (häufig in den Ingenieurswissenschaften als *virtuelle Verschiebung* bezeichnet) eine beliebige Testfunktion ist. Dann gilt

$\boldsymbol{w} = \boldsymbol{v} - \boldsymbol{u} \in (H^1_{\Gamma_u}(\Omega))^2$ eine Testfunktion, die wir mit (3.41a) multiplizieren und über Ω integrieren.

$$\begin{aligned}
0 &= \int_\Omega (\operatorname{div}\boldsymbol{\sigma} + \bar{\boldsymbol{b}}) \cdot \boldsymbol{w}\, d\Omega = \int_\Omega \operatorname{div}\boldsymbol{\sigma} \cdot \boldsymbol{w} + \bar{\boldsymbol{b}} \cdot \boldsymbol{w}\, d\Omega \\
&= \int_\Omega \operatorname{div}(\boldsymbol{w} \cdot \boldsymbol{\sigma}) - \operatorname{grad}\boldsymbol{w} : \boldsymbol{\sigma} + \bar{\boldsymbol{b}} \cdot \boldsymbol{w}\, d\Omega \\
&= \int_\Gamma \underbrace{\boldsymbol{w} \cdot \boldsymbol{\sigma} \cdot \boldsymbol{n}}_{=\boldsymbol{w}\cdot\|\boldsymbol{n}\|^2\cdot(\boldsymbol{\sigma}\cdot\boldsymbol{n})}\, d\Gamma - \int_\Omega \underbrace{\frac{1}{2}(\operatorname{grad}\boldsymbol{w} + \operatorname{grad}^T \boldsymbol{w})}_{=\boldsymbol{\varepsilon}(\boldsymbol{w})} : \boldsymbol{\sigma}\, d\Omega + \int_\Omega \bar{\boldsymbol{b}} \cdot \boldsymbol{w}\, d\Omega \\
&= \int_{\Gamma_c} w_n\, \sigma_n\, d\Gamma + \int_{\Gamma_\sigma} \bar{\boldsymbol{t}} \cdot \boldsymbol{w}\, d\Gamma - \int_\Omega \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\boldsymbol{w})\, d\Omega + \int_\Omega \bar{\boldsymbol{b}} \cdot \boldsymbol{w}\, d\Omega
\end{aligned} \tag{3.43}$$

(zweite Zeile ist Produktregel für die Divergenz, dritte Zeile Gauß, vierte Zeile die Aufteilung von $\Gamma = \Gamma_u \cup \Gamma_\sigma \cup \Gamma_c$, wobei $\boldsymbol{w} = \boldsymbol{0}$ auf Γ_u und $\boldsymbol{\sigma}\boldsymbol{n} = \bar{\boldsymbol{t}}$ auf Γ_σ).

- betrachte das Integral über Γ_c, dann gilt für den Integranden

$$\begin{aligned}
w_n\, \sigma_n &= (v_n - u_n)\, \sigma_n \overset{\text{„}+0\text{“}}{=} (v_n - g + g - u_n)\, \sigma_n \\
&= (v_n - g)\, \sigma_n - \underbrace{(u_n - g)\, \sigma_n}_{=0 \text{ auf } \Gamma_c} \\
&= \underbrace{(v_n - g)}_{\leq 0}\, \underbrace{\sigma_n}_{\leq 0} \geq 0
\end{aligned}$$

- damit ist das Integral ≥ 0 und aus (3.43) folgt

$$\begin{aligned}
0 &\geq \int_{\Gamma_\sigma} \bar{\boldsymbol{t}} \cdot \boldsymbol{w}\, d\Gamma - \int_\Omega \boldsymbol{\sigma} : \boldsymbol{\varepsilon}(\boldsymbol{w})\, d\Omega + \int_\Omega \bar{\boldsymbol{b}} \cdot \boldsymbol{w}\, d\Omega \\
\iff \int_\Omega \boldsymbol{\sigma}(\boldsymbol{u}) : \boldsymbol{\varepsilon}(\boldsymbol{v} - \boldsymbol{u})\, d\Omega &\geq \int_\Omega \bar{\boldsymbol{b}} \cdot (\boldsymbol{v} - \boldsymbol{u})\, d\Omega + \int_{\Gamma_\sigma} \bar{\boldsymbol{t}} \cdot (\boldsymbol{v} - \boldsymbol{u})\, d\Gamma
\end{aligned} \tag{3.44}$$

- mit (3.41b) kann die Spannung aus (3.44) bzgl. der Verzerrung mit $\boldsymbol{\sigma} = \mathcal{C} : \boldsymbol{\varepsilon}$ ausgedrückt werden

- mit der Bilinearform $a : H^1_{\Gamma_u} \times H^1_{\Gamma_u} \to \mathbb{R}$ und Linearform $F : H^1_{\Gamma_u} \to \mathbb{R}$ mit

$$\begin{aligned}
a(\boldsymbol{u}, \boldsymbol{v}) &\coloneqq \int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{u}) : \mathcal{C} : \boldsymbol{\varepsilon}(\boldsymbol{v})\, d\Omega\,, \\
F(\boldsymbol{v}) &\coloneqq \int_\Omega \bar{\boldsymbol{b}} \cdot \boldsymbol{v}\, d\Omega + \int_{\Gamma_\sigma} \bar{\boldsymbol{t}} \cdot \boldsymbol{v}\, d\Gamma
\end{aligned}$$

lässt sich (3.44) in der altbekannten Form: Finde $\boldsymbol{u} \in \mathcal{K}$, so dass gilt

$$a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) \geq F(\boldsymbol{v} - \boldsymbol{u}) \quad \forall\, \boldsymbol{v} \in \mathcal{K}\,. \tag{3.45}$$

noch zu zeigen: Kornsche Ungleichung (vgl. [Bra13]) auch für $a(\cdot,\cdot)$ unter der Bedingung, dass $\max\|C_{ijkl}\| \leq M$ für ein $M \geq 0$ ist. $\Rightarrow$ damit folgt die Koerzivität von a. Und noch Stetigkeit zeigen

- vgl. [KO88] S. 112, zeige: $|F(\boldsymbol{v})| \leq C(\|\bar{\boldsymbol{b}}\|_0 + \|\bar{\boldsymbol{t}}\|_{0,\Gamma_c})\,\|\boldsymbol{v}\|_1$, also mit $\bar{\boldsymbol{b}} \in (L^2(\Omega))^2$ und $\bar{\boldsymbol{t}} \in (L^2(\Gamma_c))^2$.
- als letztes: K ist analog zu Lemma 3.1 abgeschlossen und konvex.
- also hat das Problem (3.45) eine eindeutige Lösung. als Theorem

  Theorem 3.20.

  Beweis. □

- **Theorem 3.21.** *Es sei $\boldsymbol{v} : \Omega \to \mathbb{R}^2$ und K wie oben definiert. Die Variationsungleichung* (3.45) *ist äquivalent zum Minimierungsproblem:*

$$\min_{\boldsymbol{v} \in K} J(\boldsymbol{v}) = \frac{1}{2} a(\boldsymbol{v}, \boldsymbol{v}) - F(\boldsymbol{v}) \tag{3.46}$$

  Beweis. zeige die Voraussetzungen von Satz A.10, also Konvexität von J z.B. □

- **Bemerkung 3.22.** für das Kontaktproblem mit Tresca-Reibung gibt es ein analoges Funktional. dieses aufführen!! dieses ist nicht G-differenzierbar $\Rightarrow$ anders herleiten. dies führt auf eine sogenannte Variationsungleichung 2. Art (vgl. [Ste12a])

$$a(\boldsymbol{u}, \boldsymbol{v} - \boldsymbol{u}) + j(\boldsymbol{v}) - j(\boldsymbol{u}) \geq F(\boldsymbol{v} - \boldsymbol{u})$$

  mit dem Reibungsfunktional j.

3.2.3 Lösung des Kontaktproblems mittels FEM

- betrachte analog zu $\mathcal{S}_h$ den Raum der linearen mehrdimensionalen Ansatzfunktionen bzgl. einer quasi-uniformen Triangulierung $\mathcal{T}_h$

$$\mathcal{S}_h \coloneqq \{\boldsymbol{v} \in (C^0(\Omega))^2 \mid \boldsymbol{v}|_T \in \mathcal{P}_1^2 \text{ für } T \in \mathcal{T}_h, \boldsymbol{v}|_{\Gamma_u} = \boldsymbol{0}\} \subset H^1_{\Gamma_u}(\Omega) \tag{3.47}$$

- mit einer Basis $\mathcal{B}_h \coloneqq \{\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_N\}$ von $\mathcal{S}_h$ lässt sich jedes Element $\boldsymbol{v}_h \in \mathcal{S}_h$ als Linearkombination schreiben

$$\boldsymbol{v}_h(\bar{x}) = \sum_{i=1}^{N} x_i \, \boldsymbol{\psi}_i(\bar{x}) \quad \forall\, \bar{x} \in \Omega \tag{3.48}$$

  für genau ein $(x_1, \ldots, x_N)^T =: \boldsymbol{x} \in \mathbb{R}^N$.

- betrachten wir analog zu oben (3.45) diskret, so ergibt sich: Finde $\boldsymbol{u}_h \in \mathcal{S}_h$, so dass

$$a(\boldsymbol{u}_h, \boldsymbol{v}_h - \boldsymbol{u}_h) \geq F(\boldsymbol{v}_h - \boldsymbol{u}_h) \quad \forall\, \boldsymbol{v}_h \in K_h\,. \tag{3.49}$$

  mit $\mathcal{K}_h \coloneqq \{\boldsymbol{v}_h \in \mathcal{S}_h \mid \boldsymbol{v}_h(\bar{x}_i) \cdot \boldsymbol{n} - g(\bar{x}_i) \leq 0 \text{ mit } \bar{x}_i \in \mathcal{N} \cap \Gamma_c\}$, d.h. die punktuelle Form (der Nebenbedingung) von $\mathcal{K}$.

- analog zu $K_{\mathcal{S}}$ können wir auch hier bzgl. einer Basis $\mathcal{B}_h$ die Menge $\mathcal{K}_h$ äquivalent durch den Koordinatenvektor $\boldsymbol{x} \in \mathbb{R}^N$ ausdrücken, d.h.

$$\begin{aligned}\mathcal{K}_{\mathcal{S}} \coloneqq &\left\{ \boldsymbol{x} \in \mathbb{R}^N \;\middle|\; \sum_{j=1}^{N} x_j \, \boldsymbol{\psi}_j(\bar{x}_i) \cdot \boldsymbol{n} - g(x_i) \geq 0 \text{ für } \bar{x}_i \in \mathcal{N} \cap \Gamma_c \right\} \\ = &\{\boldsymbol{x} \in \mathbb{R}^N \mid B\boldsymbol{x} \geq \boldsymbol{c}, B = [-\boldsymbol{\psi}_j(\bar{x}_i) \cdot \boldsymbol{n}(\bar{x}_i)]_{\bar{x}_i \in \mathcal{N} \cap \Gamma_c, 1 \leq j \leq N}, \boldsymbol{c} = [-g(\bar{x}_i)]_{\bar{x}_i \in \mathcal{N} \cap \Gamma_c}\}\end{aligned}$$

- damit ist das diskrete Problem: Finde $\boldsymbol{x}^* \in \mathcal{K}_{\mathcal{S}}$, so dass

$$(A\boldsymbol{x}^* - \boldsymbol{b})^T(\boldsymbol{x} - \boldsymbol{x}^*) \geq 0 \quad \forall \boldsymbol{x} \in \mathcal{K}_{\mathcal{S}}\,, \tag{3.50}$$

  wobei

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

- Aus Bemerkung 3.13 folgt, dass die Variationsungleichung (3.50) äquivalent zu folgendem quadratischen Programm ist:

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

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

Kapitel 4

Ein hierarchischer Fehlerschätzer für Hindernisprobleme

- Vergleich Hindernisprobleme zu Kontaktproblemen → warum gerade dieser Fehlerschätzer bei Hindernis- bzw. Kontaktproblemen

Dieses Kapitel basiert größtenteils auf [ZVKG11].

4.1 Herleitung eines a posteriori hierarchischen Fehlerschätzers

- der Einfachheit halber gehen wir von folgendem Sachverhalt aus

  Voraussetzung 4.1. Das Hindernis wird durch eine stückweise lineare stetige Funktion ψ beschrieben.

- nicht nichtstetige oder auch glatte Hindernisse sind analoge Aussagen, aber schwerer, beweisbar

4.1.1 Diskretisierung

- $\mathcal{B}_h$ sei eine nodale Basis bzgl. einer quasi-uniformen Triangulierung $\mathcal{T}_h$ für $\mathcal{S}_h$ (s. auch Kapitel 2), K_h wie in Kapitel 3.1.3

  $$K_h = \{v_h \in \mathcal{S}_h \mid v_h(p) \geq \psi(p) \,\forall\, p \in \mathcal{N} \cap \Omega\}\,,$$

  wobei $\mathcal{N}$ wieder die Menge der Knoten von $\mathcal{T}_h$ darstellt.

- betrachte wieder die diskrete Variationsungleichung (3.12): Finde $u_h \in K_h$ mit

  $$a(u_h, v_h - u_h) \geq (f, v_h - u_h) \quad \forall\, v_h \in K_h\,.$$

- oder äquivalent die Minimierung des Funktionals $J(v) = \frac{1}{2}a(v,v) - (f,v)$ über K_h, d.h.

$$u_h \in K_h : \quad J(u_h) \leq J(v_h) \quad \forall\, v_h \in K_h \tag{4.1}$$

- wegen der Voraussetzung, dass ψ stückweise linear ist, gilt $K_h \subset K$, da die linearen Ansatzfunktionen nicht nur punktuell, sondern auch kontinuierlich die Nebenbedingung erfüllen

- damit ist (3.12) eine konforme FEM $\rightarrow$ nichtkonforme wollen wir hier nicht betrachten (s. bel. stetige Hindernisse)

- wir wollen einen a posteriori Fehlerschätzer für den Fehler bzgl. der Funktionswerte der Funktionale $J(u), J(u_h)$ herleiten. Hierbei gilt $J(u_h) - J(u) \geq 0$, denn aus den beiden Minimierungsproblemen über K und K_h folgt

$$J(u) \leq J(v) \,\forall\, v \in K\,, \quad J(u_h) \leq J(v_h) \,\forall\, v_h \in K_h\,.$$

  Da $K_h \subset K$ gilt, folgt auch $J(u) \leq J(v_h)$ für alle $v_h \in K_h$. Setze $v_h = u_h$, so gilt

$$J(u) \leq J(u_h) \Longleftrightarrow J(u_h) - J(u) \geq 0$$

-

  Bemerkung 4.2. Gilt $\psi = -\infty$, d.h. ist kein Hindernis vorhanden, so folgt

$$\begin{aligned}
J(u_h) - J(u) =& \frac{1}{2}a(u_h,u_h) - (f,u_h) - \left(\frac{1}{2}a(u,u) - (f,u)\right) \\
=& \frac{1}{2}a(u_h,u_h) - (f,u_h) - \frac{1}{2}a(u,u) + (f,u) \\
&+ \overbrace{(a(u,u-u_h) - \underbrace{(f,u-u_h)}_{=(f,u)-(f,u_h)})}^{=0} \\
=& \frac{1}{2}a(u_h,u_h) - \frac{1}{2}a(u,u) + a(u,u-u_h) \\
=& \frac{1}{2}a(u_h,u_h) - \frac{1}{2}a(u,u) + a(u,u) - a(u,u_h) \\
=& \frac{1}{2}(a(u_h,u_h) + a(u,u) - 2a(u,u_h)) \\
=& \frac{1}{2}a(u_h-u,u_h-u) = \frac{1}{2}\|u_h-u\|_E^2\,.
\end{aligned}$$

Ist nun ein $\psi > -\infty$ gegeben, dann addieren wir im zweiten Schritt nicht mehr Null, sondern es gilt für den Term

$$a(u, u - u_h) - (f, u - u_h) \leq 0$$

und damit gilt $J(u_h) - J(u) \geq \frac{1}{2}\|u_h - u\|_E^2$, d.h. eine obere Schranke des Fehlers im Funktional schätzt auch den Fehler zwischen exakter und approximierter Lösung in der Energienorm ab.

- Herleitung eines hierarchischen a posteriori Fehlerschätzers:
-

Notation. Um im Folgenden den hierarchischen Split leichter beschreiben zu können, schreiben wir für die Galerkin-Lösung u_h die Notation $u_{\mathcal{S}}$, um auszudrücken, dass diese im linearen Ansatzraum $\mathcal{S}_h$ liegt. Analog sind die im Weiteren übrigen verwendeten Indizes zu verstehen.

- wir führen Fehlerfunktion $e = u - u_{\mathcal{S}}$ ein
- weiter sei $\mathcal{I}(v) = \frac{1}{2}a(v, v) - \rho_{\mathcal{S}}(v)$ mit $\rho_{\mathcal{S}}(v) = (f, v) - a(u_{\mathcal{S}}, v), v \in H_0^1(\Omega)$.
-

Bemerkung 4.3. (a) Die Linearform $\rho_{\mathcal{S}}$ stellt das Residuum der Variationsgleichung (d.h. ohne Hindernis) dar.

(b) Nach dem Darstellungssatz von Riesz existiert ein $v^* \in H_0^1(\Omega)$, so dass

$$(v^*, v)_1 = \rho_{\mathcal{S}}(v) \quad \forall\, v \in H_0^1(\Omega)$$

ist. Wir können also v^* als Lagrange-Multiplikator bzgl. der Nebenbedingung $v \geq \psi$ interpretieren.

- neues Minimierungsproblem, jetzt für den Fehler e.

Satz 4.4 (Lösung des Defektproblems)**.** *Mit den obigen Bezeichnungen löst die Fehlerfunktion e folgendes Defektproblem:*

$$e \in \mathcal{A}: \quad \mathcal{I}(e) \leq \mathcal{I}(v) \quad \forall\, v \in \mathcal{A}\,, \tag{4.2}$$

wobei $\mathcal{A} \coloneqq \{v \in H_0^1(\Omega) \,|\, v \geq \psi - u_{\mathcal{S}}\} = -u_{\mathcal{S}} + K$.

Beweis. Es sei u die Lösung von (3.2) und $u_{\mathcal{S}}$ die Lösung von (4.1). Dann gilt

$$\begin{aligned} u \in K: \quad & J(u) \leq J(\tilde{v}) && \forall\, \tilde{v} \in K \qquad (*) \\ \Longleftrightarrow\, u \in K: \quad & J(u) - J(u_{\mathcal{S}}) \leq J(\tilde{v}) - J(u_{\mathcal{S}}) && \forall\, \tilde{v} \in K\,. \end{aligned}$$

Wir rechnen für die linke Seite nach, dass gilt

$$\begin{aligned} J(u) - J(u_{\mathcal{S}}) &= \frac{1}{2}a(u,u) - (f,u) - \left(\frac{1}{2}a(u_{\mathcal{S}},u_{\mathcal{S}}) - (f,u_{\mathcal{S}})\right) \\ &= \frac{1}{2}a(u,u) + \frac{1}{2}a(u_{\mathcal{S}},u_{\mathcal{S}}) - a(u_{\mathcal{S}},u_{\mathcal{S}}) - (f,u-u_{\mathcal{S}}) \\ &= \frac{1}{2}a(u,u) + \frac{1}{2}a(u_{\mathcal{S}},u_{\mathcal{S}}) - a(u_{\mathcal{S}},u) - ((f,u-u_{\mathcal{S}}) - a(u_{\mathcal{S}},u-u_{\mathcal{S}})) \\ &= \frac{1}{2}a(u-u_{\mathcal{S}},u-u_{\mathcal{S}}) - \rho_{\mathcal{S}}(u-u_{\mathcal{S}}) \\ &= \frac{1}{2}a(e,e) - \rho_{\mathcal{S}}(e) = \mathcal{I}(e)\,. \end{aligned}$$

Analog gilt für die rechte Seite $J(\tilde{v}) - J(u_{\mathcal{S}}) = \mathcal{I}(\tilde{v} - u_{\mathcal{S}})$. Mit $v \coloneqq \tilde{v} - u_{\mathcal{S}}$ gilt $v \in \mathcal{A}$ und damit ist $(*)$ äquivalent zu: Finde $e \in \mathcal{A}$, so dass

$$\mathcal{I}(e) \leq \mathcal{I}(v) \quad \forall\, v \in \mathcal{A}\,. \qquad \square$$

•

Korollar 4.5. *Das Problem* (4.2) *ist äquivalent zur Variationsungleichung: Finde $e \in \mathcal{A}$ mit*

$$a(e, v-e) \geq \rho_{\mathcal{S}}(v-e) \quad \forall\, v \in \mathcal{A}\,. \tag{4.3}$$

Beweis. Analog zu Lemma 3.1 lässt sich zeigen, dass $\mathcal{A}$ abgeschlossen und konvex ist. Mit Satz A.10 folgt dann die Behauptung. $\square$

- da ψ stückweise linear ist, liegt $0 \in \mathcal{A}$, d.h. das „gewünschte“ Ergebnis für e liegt im betrachteten Raum

•

Bemerkung. Wir werden noch zeigen, dass $\rho_{\mathcal{S}}$ eine Schlüsselgröße für die a posteriori Abschätzung darstellt.

- a posteriori Schätzer in 2 Schritten
  - (i) diskreditiere (4.3) bzgl. einer Erweiterung von $\mathcal{S}_h$ (hier quadratische Funktionen), so dass e hinreichend genau approximiert wird.
  - (ii) Aufteilung des neuen Raumes, sodass (4.3) lokal in der Erweiterung exakt gelöst werden kann
- als Erweiterung von $\mathcal{S}_h$ betrachten wir einen Raum $\mathcal{Q}_h$ mit $\mathcal{S}_h \subset \mathcal{Q}_h$.
- hier bietet sich an: $\mathcal{Q}_h \coloneqq \{v \in C^0(\Omega) \mid v|_T \in \mathcal{P}_2 \text{ für } T \in \mathcal{T}_h, v|_{\partial\Omega} = 0\}$, also der Raum der quadratischen Spline über einer quasi-uniformen Zerlegung $\mathcal{T}_h$.

- damit definiere $\mathcal{N}_\mathcal{Q} \coloneqq \mathcal{N} \cup \{x_E \mid E \in \mathcal{E}\}$, wobei x_E den Mittelpunkt der Kante E darstellt und $\mathcal{E}$ somit die Menge aller Kanten ist.
- damit ergibt sich $\mathcal{A}$ über $\mathcal{Q}$ diskret als

$$\mathcal{A}_\mathcal{Q} \coloneqq \{v \in \mathcal{Q}_h \mid v(p) \geq \psi(p) - u_\mathcal{S}(p) \,\forall\, p \in \mathcal{N}_\mathcal{Q} \cap \Omega\} \tag{4.4}$$

- im Bezug zu (4.4) ergibt sich dann das diskrete Defektproblem

$$e_\mathcal{Q} \in \mathcal{A}_\mathcal{Q} : \quad a(e_\mathcal{Q}, v - e_\mathcal{Q}) \geq \rho_\mathcal{S}(v - e_\mathcal{Q}) \quad \forall\, v \in \mathcal{A}_\mathcal{Q} \tag{4.5}$$

- **Bemerkung 4.6.** Im Allgemeinen gilt hierbei nicht $\mathcal{A}_\mathcal{Q} \subset \mathcal{A}$. So kann man sich anschaulich eine quadratische Funktion $v_\mathcal{Q} \in \mathcal{A}_\mathcal{Q}$ vorstellen, die allerdings zwischen den übereinstimmenden Werten aufgrund ihrer Krümmung das lineare Hindernis aus $\mathcal{A}$ durchdringt.

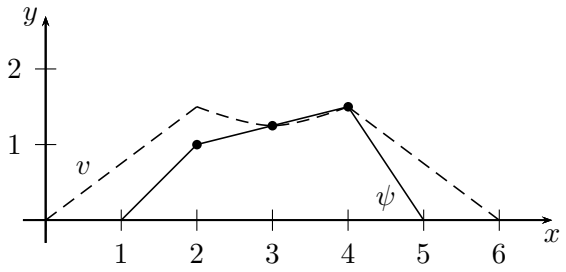

Abbildung 4.1: Beispiel eines affinen Hindernisses ψ mit $v \in \mathcal{A}_\mathcal{Q}$ in $\mathbb{R}$

- hierarchische Aufteilung von $\mathcal{Q}_h$ durch $\mathcal{Q}_h = \mathcal{S}_h \oplus \mathcal{V}_h$, wobei $\mathcal{V}_h \coloneqq \{\phi_E \mid E \in \mathcal{E}\}$ ist und ϕ_E die *Bubble-Funktion* mit

$$\phi_E(p) = \delta_{x_E,p} = \begin{cases} 1, & p = x_E \\ 0, & \text{sonst} \end{cases}$$

ist

- **Beispiel 4.7.** allgemeine Skizze und die drei bubble Funktionen auf einem Referenzdreieck

- **Satz 4.8.** *Mit den oben verwendeten Notationen gilt* $\mathcal{Q}_h = \mathcal{S}_h \oplus \mathcal{V}_h$.

Beweis. Wir zeigen, dass $\mathcal{Q}_h = \mathcal{S}_h \oplus \mathcal{V}_h$ auf dem Referenzdreieck gilt und damit gilt es auch für beliebige Dreiecke $T \in \mathcal{T}_h$, da ein allgemeines Dreieck T aus dem Referenzelement $\tilde{T}$ durch affine Transformation hervorgeht.

Auf dem Referenzelement $\tilde{T}$ ist $\{\phi_1, \phi_2, \phi_3\}$ eine Basis von $\mathcal{S}_h$ mit

$$\phi_1(\xi,\eta) = 1 - \xi - \eta\,, \quad \phi_2(\xi,\eta) = \xi\,, \quad \phi_3(\xi,\eta) = \eta$$

und $\{\phi_4, \phi_5, \phi_6\}$ eine Basis von $\mathcal{V}_h$ mit

$$\phi_1(\xi,\eta) = 4\xi(1-\xi-\eta)\,, \quad \phi_2(\xi,\eta) = 4\xi\eta\,, \quad \phi_3(\xi,\eta) = 4\eta(1-\xi-\eta)\,.$$

Damit ist $\{\phi_1, \ldots, \phi_6\}$ ein Erzeugendensystem von $\mathcal{Q}_h$, da jedes Element

$$a_0 + a_1\xi + a_2\eta + a_3\xi^2 + a_4\xi\eta + a_5\eta^2 \in \mathcal{Q}_h$$

als Linearkombination aus den Funktionen beschrieben werden kann (ϕ_1 bis ϕ_6 enthalten alle vorkommenden Summanden eines Polynom 2. Grades). Außerdem ist leicht nachzurechnen, dass die Funktionen $\phi_i, i = 1, \ldots, 6$, linear unabhängig sind und damit gilt

$$\mathcal{Q}_h = \operatorname{span}\{\phi_1, \ldots, \phi_6\}\,.$$

Aus der linearen Unabhängigkeit folgt damit auch $\mathcal{S}_h \cap \mathcal{V}_h = \{0\}$ gilt und damit die Behauptung. □

- daher kann jedes Element $v_\mathcal{Q} \in \mathcal{Q}_h$ als $v_\mathcal{Q} = v_\mathcal{S} + v_\mathcal{V}$ mit $v_\mathcal{S} \in \mathcal{S}_h, v_\mathcal{V} \in \mathcal{V}_h$ geschrieben werden
- aus diesem Grund führen wir folgende Bilinearform ein:

$$a_\mathcal{Q}(v,w) \coloneqq a(v_\mathcal{S}, w_\mathcal{S}) + \sum_{E\in\mathcal{E}} u_\mathcal{V}(x_E) w_\mathcal{V}(x_E) a(\phi_E, \phi_E) \quad \forall\, v, w \in \mathcal{Q}_h\,,$$

  welche aufgrund der Eigenschaften der direkten Summe von $\mathcal{S}_h$ und $\mathcal{V}_h$ wohldefiniert ist.
- dabei ergibt sich $a_\mathcal{Q}$ durch Entkopplung von $\mathcal{S}_h$ und $\mathcal{V}_h$ und anschließender „Diagonalisierung“ auf $\mathcal{V}$
- sinnvoll $a_\mathcal{Q}$ so einzuführen, denn:

  Satz 4.9. *Die zu $a_\mathcal{Q}$ assoziierte Energienorm*

$$\|v\|_\mathcal{Q} \coloneqq a_\mathcal{Q}(v,v)^{\frac{1}{2}}\,, \quad v \in \mathcal{Q}_h$$

  ist äquivalent zur Energienorm $\|\cdot\|_E$, d.h. es gibt Konstanten c_1, c_2 (die insbesondere nur von der Quasi-Uniformität von $\mathcal{T}_h$ abhängen), so dass

$$c_1\|v\|_E \le \|v\|_\mathcal{Q} \le c_2\|v\|_E\,, \quad \forall\, v \in \mathcal{Q}_h\,.$$

Beweis. Die Aussage folgt aus Theorem 4.1 bzw. Bemerkung 4.3 in [HK92] zusammen mit dem Lemma auf Seite 14 in [DLY89]. □

- daher führen wir die approximierte Energie

$$\mathcal{I}_{\mathcal{Q}}(v) \coloneqq \frac{1}{2} a_{\mathcal{Q}}(v,v) - \rho_{\mathcal{S}}(v)\,, \quad v \in \mathcal{Q}_h \tag{4.6}$$

ein.

- das damit verbundene Defektproblem ist allerdings noch durch die Nebenbedingung aus $\mathcal{A}_{\mathcal{Q}}$ mit $\mathcal{S}_h$ gekoppelt und daher noch nicht alleine auf die Raumerweiterung $\mathcal{V}_h$ bezogen.

- Als Abhilfe ignorieren wir einfach die linearen Beiträge in $\mathcal{A}_{\mathcal{Q}}$ und führen eine echte Teilmenge

$$\mathcal{A}_{\mathcal{V}} \coloneqq \{v \in \mathcal{V} \mid v(x_E) \geq \psi(x_E) - u_{\mathcal{S}}(x_E) \,\forall\, E \in \mathcal{E}\} \tag{4.7}$$

von $\mathcal{A}_{\mathcal{Q}}$ ein.

- zusammen mit (4.6) und (4.7) erhalten wir das lokale diskrete Defektproblem

$$\varepsilon_{\mathcal{V}} \in \mathcal{A}_{\mathcal{V}} : \quad \mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}}) \leq \mathcal{I}_{\mathcal{Q}}(v) \quad \forall\, v \in \mathcal{A}_{\mathcal{V}} \tag{4.8}$$

bzw. die dazu äquivalente Variationsungleichung

$$\varepsilon_{\mathcal{V}} \in \mathcal{A}_{\mathcal{V}} : \quad a_{\mathcal{Q}}(\varepsilon_{\mathcal{V}}, v - \varepsilon_{\mathcal{V}}) \geq \rho_{\mathcal{S}}(v - \varepsilon_{\mathcal{V}}) \quad \forall\, v \in \mathcal{A}_{\mathcal{V}}\,. \tag{4.9}$$

-

Bemerkung 4.10. (a) Da ψ stetig stückweise linear ist und somit $u_{\mathcal{S}} \geq \psi$ gilt, folgt $0 \in \mathcal{A}_{\mathcal{V}}$. Damit ist auch hier die gewünschte Lösung für $\varepsilon_{\mathcal{V}}$ in $\mathcal{A}_{\mathcal{V}}$ enthalten

(b) Auch für $\mathcal{A}_{\mathcal{V}}$ lässt sich mit analogem Vorgehen zu Lemma 3.1 die Konvexität zeigen.

-

Lemma 4.11. *Das Energiefunktional $\mathcal{I}_{\mathcal{Q}}$ ist konvex.*

Beweis. Da a eine stetige koerzive Bilinearform, werden aufgrund der Konstruktion von $a_{\mathcal{Q}}$ diese Eigenschaften auch auf $a_{\mathcal{Q}}$ übertragen. Weiterhin ist leicht zu überprüfen, dass $\rho_{\mathcal{S}}$ eine stetige Linearform ist. Dann folgt aus Lemma 2.10 direkt die Behauptung. □

- Lösung des lokalen Defektproblems

Satz 4.12. *Die Lösung von* (4.8) *bzw.* (4.9) *ist explizit gegeben durch*

$$\varepsilon_{\mathcal{V}}(x_E) = \frac{\max\{-d_E, \rho_E\}}{\|\phi_E\|} \tag{4.10}$$

wobei

$$d_E = (u_{\mathcal{S}}(x_E) - \psi(x_E))\|\phi_E\| \geq 0\,, \quad \rho_E = \frac{\rho_{\mathcal{S}}(\phi_E)}{\|\phi_E\|}\,. \tag{4.11}$$

Beweis. Es sei $M = |\mathcal{E}|$ die Anzahl der Kanten. Zunächst berechnen wir zur besseren Übersicht $\varepsilon_{\mathcal{V}}(x_E)$ konkret, d.h.

$$\begin{aligned}
\varepsilon_{\mathcal{V}}(x_E) &= \frac{\max\{-d_E, \rho_E\}}{\|\phi_E\|} \\
&= \frac{\max\left\{(\psi(x_E) - u_{\mathcal{S}}(x_E))\|\phi_E\|, \frac{\rho_{\mathcal{S}}(\phi_E)}{\|\phi_E\|}\right\}}{\|\phi_E\|} \\
&= \max\left\{\psi(x_E) - u_{\mathcal{S}}(x_E), \frac{\rho_{\mathcal{S}}(\phi_E)}{\|\phi_E\|^2}\right\} \\
&= \max\left\{\psi(x_E) - u_{\mathcal{S}}(x_E), \frac{1}{\|\phi_E\|^2}((f,\phi_E) - a(u_{\mathcal{S}},\phi_E))\right\}\,.
\end{aligned} \tag{4.12}$$

Da $\varepsilon_{\mathcal{V}} = \sum_{E\in\mathcal{E}} \varepsilon_{\mathcal{V}}(x_E)\phi_E$ ist, können wir (4.8) bzgl. der Basis $\{\phi_E \mid E \in \mathcal{E}\}$ von $\mathcal{V}_h$ diskret schreiben als

$$\min \frac{1}{2}\boldsymbol{v}^T D\boldsymbol{v} - \boldsymbol{g}^T\boldsymbol{v} \quad \text{s.t.} \quad \boldsymbol{v} \geq \boldsymbol{\psi} - \boldsymbol{u}_{\mathcal{S}}\,,$$

wobei $\boldsymbol{v} = [\varepsilon_{\mathcal{V}}(x_{E_i})]_{1\leq i\leq M}, D = \operatorname{diag}(a(\phi_{E_1},\phi_{E_1}),\ldots,a(\phi_{E_M},\phi_{E_M})), \boldsymbol{g} = [(f,\phi_{E_i}) - a(u_{\mathcal{S}},\phi_{E_i})]_{1\leq i\leq M}, \boldsymbol{\psi} = [\psi(x_{E_i})]_{1\leq i\leq M}$ und $\boldsymbol{u}_{\mathcal{S}} = [u_{\mathcal{S}}(x_{E_i})]_{1\leq i\leq M}$. Da $\mathcal{A}_{\mathcal{V}}$ und $\mathcal{I}_{\mathcal{Q}}$ konvex sind, existiert ein Minimum $\boldsymbol{v}^* \in \mathcal{A}_{\mathcal{V}}$ von $\mathcal{I}_{\mathcal{Q}}$, das die KKT-Bedingungen erfüllt. Damit gilt

$$D\boldsymbol{v} - \boldsymbol{g} - \boldsymbol{\lambda} = \boldsymbol{0}\,, \tag{4.13a}$$

$$\boldsymbol{\lambda} \geq 0\,, \tag{4.13b}$$

$$\boldsymbol{v} \geq \boldsymbol{\psi} - \boldsymbol{u}_{\mathcal{S}}\,, \tag{4.13c}$$

$$\lambda_i\,(\boldsymbol{v} - \boldsymbol{\psi} + \boldsymbol{u}_{\mathcal{S}})_i = 0 \quad \forall\, i = 1,\ldots,M\,. \tag{4.13d}$$

Es sei $k \in \{1,\ldots,M\}$ beliebig.

Fall 1: Gilt $\lambda_k = 0$, so folgt aus (4.13a)

$$\varepsilon_{\mathcal{V}}(x_{E_k}) = v_k = \frac{g_k}{a(\phi_{E_k},\phi_{E_k})} = \frac{1}{\|\phi_{E_k}\|^2}((f,\phi_{E_k}) - a(u_{\mathcal{S}},\phi_{E_k}))\,.$$

Fall 2: Gilt $\lambda_k \neq 0$, dann folgt wegen (4.13d)

$$\varepsilon_{\mathcal{V}}(x_{E_k}) = v_k = (\boldsymbol{\psi} - \boldsymbol{u}_{\mathcal{S}})_k = \psi(x_{E_k}) - u_{\mathcal{S}}(x_{E_k})\,.$$

Insgesamt folgt mit (4.13c) und (4.12) die Behauptung. □

- wir wollen im weiteren den a posteriori Fehlerschätzer

$$-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}}) = -\frac{1}{2}a_{\mathcal{Q}}(\varepsilon_{\mathcal{V}}, \varepsilon_{\mathcal{V}}) + \rho_{\mathcal{S}}(\varepsilon_{\mathcal{V}})$$

  betrachten und werden zeigen, dass er äquivalent zu $J(u_{\mathcal{S}}) - J(u)$ ist (vgl. Kapitel 4.1.4 und 4.1.5)

- zunächst aber Einführung des lokalen Anteils des Fehlerschätzers $-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$

4.1.2 Lokaler Anteil des Fehlerschätzers

- *Notation.* (a) Wir schreiben im Folgenden „$\lesssim$“ statt „$\leq C$“, wenn die Konstante C nur von der Quasi-Uniformität von $\mathcal{T}_h$ abhängt.
  (b) Weiter schreiben wir „$A \approx B$“ für „$A \lesssim B$“ und „$B \lesssim A$“.

- zunächst zeigen wir ein paar Eigenschaften von der Fehlerfunktion $e = u - u_{\mathcal{S}}$

- **Lemma 4.13.** *Die Fehlerfunktion $e = u - u_{\mathcal{S}}$ erfüllt die Ungleichungen*

$$\frac{1}{2}\|e\|^2 \leq \frac{1}{2}\rho_{\mathcal{S}}(e) \leq -\mathcal{I}(e) \leq \rho_{\mathcal{S}}(e)\,. \tag{4.14}$$

  Beweis. Wir erinnern uns, dass

$$-\mathcal{I}(e) \coloneqq -\frac{1}{2}\underbrace{a(e,e)}_{\geq 0} + \rho_{\mathcal{S}}(e) \leq \rho_{\mathcal{S}}(e)\,,$$

  da a koerziv ist. Dann gilt weiter

$$\begin{aligned}
-\mathcal{I}(e) &= -\frac{1}{2}a(e,e) + \rho_{\mathcal{S}}(e) \\
&= -\frac{1}{2}a(u - u_{\mathcal{S}}, e) + \rho_{\mathcal{S}}(e) \\
&= -\frac{1}{2}a(u,e)\underbrace{\frac{1}{2}a(u_{\mathcal{S}}, e) - \frac{1}{2}(f,e)}_{=-\frac{1}{2}\rho_{\mathcal{S}}(e)} + \frac{1}{2}(f,e) + \rho_{\mathcal{S}}(e) \\
&= -\frac{1}{2}(\underbrace{a(u, u - u_{\mathcal{S}}) - (f, u - u_{\mathcal{S}})}_{\leq 0}) + \frac{1}{2}\rho_{\mathcal{S}}(e) \geq \frac{1}{2}\rho_{\mathcal{S}}(e)\,.
\end{aligned}$$

Es bleibt also die erste Ungleichung von (4.14) zu zeigen. Wir rechnen nach, dass

$$\begin{aligned}\frac{1}{2}\rho_{\mathcal{S}}(e) &= \frac{1}{2}(f,e) - \frac{1}{2}a(u_{\mathcal{S}},e)\\ &= \frac{1}{2}(\underbrace{(f,u-u_{\mathcal{S}}) - a(u,u-u_{\mathcal{S}})}_{\geq 0} + a(u-u_{\mathcal{S}},e))\\ &\geq \frac{1}{2}a(u-u_{\mathcal{S}},e) = \frac{1}{2}a(e,e) = \frac{1}{2}\|e\|^2\end{aligned}$$

gilt, womit (4.14) insgesamt bewiesen ist. □

•

Korollar 4.14. *Für die Lösungen $e_{\mathcal{Q}}, \varepsilon_{\mathcal{V}}$ von (4.5) und (4.9) gilt*

$$\frac{1}{2}\|e_{\mathcal{Q}}\|^2 \leq \frac{1}{2}\rho_{\mathcal{S}}(e_{\mathcal{Q}}) \leq -\mathcal{I}(e_{\mathcal{Q}}) \leq \rho_{\mathcal{S}}(e_{\mathcal{Q}})\,, \tag{4.15}$$

$$\frac{1}{2}\|\varepsilon_{\mathcal{V}}\|_{\mathcal{Q}}^2 \leq \frac{1}{2}\rho_{\mathcal{S}}(\varepsilon_{\mathcal{V}}) \leq -\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}}) \leq \rho_{\mathcal{S}}(\varepsilon_{\mathcal{V}})\,. \tag{4.16}$$

Beweis. Da $e_{\mathcal{Q}}$ und $\varepsilon_{\mathcal{V}}$ Lösungen der Variationsungleichungen (4.5) und (4.9) sind, folgt die Behauptung analog zum Beweis von Lemma 4.13. □

- wegen (4.16) ist $\rho_{\mathcal{S}}(\varepsilon_{\mathcal{V}})$ äquivalent zum Fehlerschätzer $-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$ und kann daher als Indikator für $-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$ verwendet werden (verkleinern wir $\rho_{\mathcal{S}}$, so wird auch $-\mathcal{I}_{\mathcal{Q}}$ kleiner)
- in Kapitel 4.1.4 und 4.1.5 werden wir die Äquivalenz von $-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$ zum exakten Fehler in den Funktionalen $J(u_{\mathcal{S}}) - J(u) = -\mathcal{I}(e)$ zeigen
- damit folgt auch aus Lemma 4.13, dass der Fehler $J(u_{\mathcal{S}}) - J(u)$ äquivalent zu $\rho_{\mathcal{S}}(e)$ ist ⇒ daher betrachten wir ein paar weitere Eigenschaften von $\rho_{\mathcal{S}}$.
- nun zu den lokalen Anteilen von $\rho_{\mathcal{S}}(\varepsilon_{\mathcal{V}})$:
- es sei $u_{\mathcal{S}}$ die Lösung von (3.12), dann auf jedem $T \in \mathcal{T}_h$ die Gleichung $\Delta u_{\mathcal{S}} = 0$, da $u_{\mathcal{S}}$ auf jedem T linear ist.

- dann gilt mit $\Omega = \bigcup_{T\in\mathcal{T}_h} T$ für alle $v \in H^1(\Omega)$

$$\begin{aligned}\rho_{\mathcal{S}}(v) = (f,v) - a(u_{\mathcal{S}},v) &= \int_\Omega fv\,d\Omega - \int_\Omega \nabla u_{\mathcal{S}}\nabla v\,d\Omega \\ &= \int_\Omega fv\,d\Omega - \sum_{T\in\mathcal{T}_h}\int_T \nabla u_{\mathcal{S}}\nabla v\,dT \\ &= \int_\Omega fv\,d\Omega - \sum_{T\in\mathcal{T}_h}\left(\int_{\partial T} v\partial_{\boldsymbol{n}} u_{\mathcal{S}}\,d\Gamma - \int_T \underbrace{\Delta u_{\mathcal{S}}}_{=0} v\,dT\right) \\ &= \int_\Omega fv\,d\Omega - \sum_{T\in\mathcal{T}_h}\int_{\partial T} v\partial_{\boldsymbol{n}} u_{\mathcal{S}}\,d\Gamma\,, \end{aligned} \tag{4.17}$$

  wobei im vorletzten Schritt die 1. Green'sche Formel angewendet wurde und d$\boldsymbol{n}$ die äußere Einheitsnormale von T ist.

- Betrachten wir zwei beliebige Dreiecke T_1, T_2 wie in Abbildung **??**, wobei $\boldsymbol{n}$ hierbei die Einheitsnormale, die von T_1 nach T_2 zeigt, bezeichnet, so können wir die Summe aus (4.17) bzgl. der Menge der Kanten $\mathcal{E}$ darstellen, da der Rand $\partial T = E_1 \cup E_2 \cup E_3$ für jedes T disjunkt in seine Kantenstücke aufgeteilt werden kann.

  Dabei sei E nun die Kante, die T_1 und T_2 zugleich enthalten, d.h. $\boldsymbol{n}$ steht rechtwinklig auf E. Dann gilt, dass die Richtungsableitung $\partial_{\boldsymbol{n}} u_{\mathcal{S}}|_{T_2}$ negativ ist bzgl. (4.17) wegen der negativen Orientierung von $\boldsymbol{n}$ bzgl. T_2.

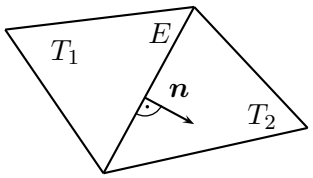

Abbildung 4.2: Dreiecke T_1 und T_2 mit Einheitsnormalen $\boldsymbol{n}$

Hiermit ergibt sich aus (4.17)

$$\begin{aligned}\rho_{\mathcal{S}}(v) &= \int_\Omega fv\,d\Omega - \sum_{T\in\mathcal{T}_h}\int_{\partial T} v\partial_{\boldsymbol{n}} u_{\mathcal{S}}\,d\Gamma \\ &= \int_\Omega fv\,d\Omega - \sum_{E\in\mathcal{E}}\int_E v\,(\underbrace{\partial_{\boldsymbol{n}} u_{\mathcal{S}}|_{T_1} - \partial_{\boldsymbol{n}} u_{\mathcal{S}}|_{T_2}}_{=:-j_E})\,d\Gamma \\ &= \int_\Omega fv\,d\Omega + \sum_{E\in\mathcal{E}}\int_E j_E v\,d\Gamma\,. \end{aligned} \tag{4.18}$$

- da für die nodalen Basisfunktionen $\{\phi_p \,|\, p \in \mathcal{N} \cap \Omega\}$ gilt

$$\sum_{p \in \mathcal{N}} \phi_p = 1 \text{ auf ganz } \Omega\,,$$

sodass wir $\rho_{\mathcal{S}}$ wie folgt in lokale Anteile aufteilen können:

$$\rho_p(v) \coloneqq \rho_{\mathcal{S}}(v\phi_p)\,, \quad v \in H^1(\Omega)\,. \tag{4.19}$$

-

Lemma 4.15. *Für ρ_p gilt*

$$\rho_p(v) = \int_{\omega_p} f v\phi_p \, d\Omega + \sum_{E \in \mathcal{E}_p} \int_E j_E v\phi_p \, d\Gamma\,, \quad v \in H^1(\Omega)$$

mit $\omega_p \coloneqq \operatorname{supp}\phi_p$ und $\mathcal{E}_p \coloneqq \{E \in \mathcal{E} \,|\, E \ni p\}$, d.h. die Menge der Kanten, in denen p enthalten ist.

Beweis. Wir rechnen einfach mit der Definition (4.19) und (4.18) nach, dass für ein beliebiges $v \in H^1(\Omega)$ gilt

$$\begin{aligned}\rho_p(v) = \rho_{\mathcal{S}}(v\phi_p) &= \int_\Omega f v\phi_p \, d\Omega + \sum_{E \in \mathcal{E}} \int_E j_E v\phi_p \, d\Gamma \\ &= \int_{\omega_p} f v\phi_p \, d\Omega + \sum_{E \in \mathcal{E}_p} \int_E j_E v\phi_p \, d\Gamma\,,\end{aligned}$$

da $\phi_p \equiv 0$ auf $\mathcal{O} \coloneqq \overline{\Omega \setminus \omega_p}$ und damit auch auf $\mathcal{F} \coloneqq \mathcal{E} \setminus \mathcal{E}_p$, da $\mathcal{F} \subset \mathcal{O}$.

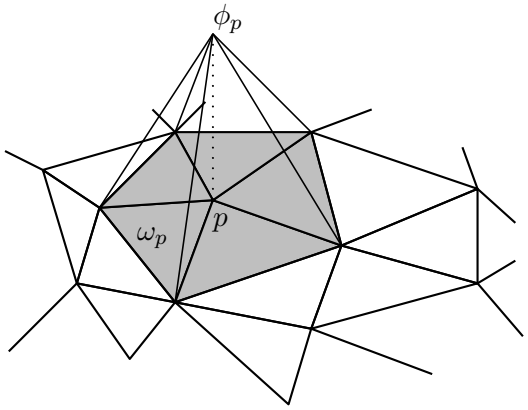

Abbildung 4.3: Darstellung von ω_p (grau) und $\mathcal{E}_p$ (abgehende Kanten von p) für ein beliebiges ϕ_p

□

- **Korollar 4.16.** *Der Indikator $\rho_{\mathcal{S}}$ lässt sich schreiben als*

$$\rho_{\mathcal{S}} = \sum_{p \in \mathcal{N}} \rho_p \,. \tag{4.20}$$

Beweis. Die Behauptung folgt direkt aus Lemma 4.15 zusammen mit

$$\Omega = \bigcup_{p \in \mathcal{N}} \omega_p \,, \quad \mathcal{E} = \bigcup_{p \in \mathcal{N}} \mathcal{E}_p \quad \text{und} \quad \sum_{p \in \mathcal{N}} \phi_p = 1$$

durch einfaches Nachrechnen. □

- im unbeschränkten Fall gilt $\rho_{\mathcal{S}} = 0 \Leftrightarrow e = 0$, denn zu $\rho_{\mathcal{S}} = 0$ ist äquivalent

$$a(e, v) = \rho_{\mathcal{S}}(v) = 0 \quad \forall\, v \in \mathcal{V} \,.$$

Da $e \in \mathcal{V}$ ist, folgt wegen der Galerkin-Orthogonalität, dass $e = 0$ sein muss. Die Umkehrung gilt analog.

- bei Variationsungleichungen gilt dies im allgemeinen nicht.

- aber: aus Lemma 4.13 folgt allgemeiner, falls $\rho_{\mathcal{S}}(v) \leq 0$ für alle $v \in \mathcal{A}$ gilt

$$\frac{1}{2}\|e\|^2 \leq \rho_{\mathcal{S}}(e) \leq 0 \implies \|e\| = 0 \implies e = 0 \,,$$

wodurch $\rho_{\mathcal{S}} = 0$ folgt, dass $e = 0$ ist.

- es gilt, ist $u_{\mathcal{S}}$ die Lösung von (3.12), so gilt für alle $p \in \mathcal{N} \cap \Omega$, dass $v = u_{\mathcal{S}} + \phi_p \geq \psi$, d.h. $v \in K_h$.

Damit folgt mit Einsetzen von v in (3.12)

$$\begin{aligned} & a(u_{\mathcal{S}}, u_{\mathcal{S}} + \phi_p - u_{\mathcal{S}}) \geq (f, u_{\mathcal{S}} + \phi_p - u_{\mathcal{S}}) \\ \Longleftrightarrow\, & 0 \geq (f, \phi_p) - a(u_{\mathcal{S}}, \phi_p) = \rho_{\mathcal{S}}(\phi_p) \end{aligned} \tag{4.21}$$

dies bedeutet, dass die lineare Approximation des Fehlers e gleich Null ist.

- falls an einem Punkt p kein Kontakt zwischen $u_{\mathcal{S}}$ und ψ vorliegt, also $u_{\mathcal{S}}(p) > \psi(p)$ ist, dann können wir ein $\alpha > 0$ hinreichend klein wählen, sodass $v = u_{\mathcal{S}} - \alpha\phi_p \in K_h$ liegt. Dann folgt analog durch Einsetzen von v in (3.12)

$$\begin{aligned} & 0 \geq (f, -\alpha\phi_p) - a(u_{\mathcal{S}}, -\alpha\phi_p) \\ \Longleftrightarrow\, & 0 \leq (f, \phi_p) - a(u_{\mathcal{S}}, \phi_p) = \rho_{\mathcal{S}}(\phi_p) \underset{(4.21)}{\leq} 0 \end{aligned}$$

und damit gilt $\rho_{\mathcal{S}}(\phi_p) = 0$

- zusammen ergeben sich die Bedingungen

$$\rho_{\mathcal{S}}(\phi_p) \leq 0\,, \quad \psi(p) - u_{\mathcal{S}}(p) \leq 0\,, \quad \rho_{\mathcal{S}}(\phi_p)(\psi(p) - u_{\mathcal{S}}(p)) = 0 \tag{4.22}$$

- dies berechtigt zur Definition von Kontakt- und Nichtkontaktpunkten

  Definition 4.17. Wir definieren die Mengen von *Kontaktpunkten* $\mathcal{N}^0$ und *Nichtkontaktpunkten* $\mathcal{N}^+$ durch

$$\mathcal{N}^0 \coloneqq \{p \in \mathcal{N} \cap \Omega \mid u_{\mathcal{S}}(p) = \psi(p)\}\,, \quad \mathcal{N}^+ \coloneqq \{p \in \mathcal{N} \cap \Omega \mid u_{\mathcal{S}}(p) > \psi(p)\}\,.$$

-

  Bemerkung 4.18. Die Bedingungen (4.22) können wir auch auf den lokalen Anteil ρ_p übertragen, damit ergibt sich für alle $p \in \mathcal{N} \cap \Omega$

$$\rho_p(1) \leq 0\,, \tag{4.23a}$$
$$u_{\mathcal{S}}(p) > \psi(p) \implies \rho_p(1) = 0\,, \tag{4.23b}$$

  denn $\rho_p(1) = \rho_{\mathcal{S}}(\phi_p)$.

- damit ist also die Approximation von e über $\mathcal{S}_h$ gleich Null, wenn die lokalen Anteile (im Vektor später) kleiner gleich Null sind

4.1.3 Oszillationsterme

- in Kapitel 4.1.4 werden wir zeigen, dass $-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$ eine obere Schranke von $-\mathcal{I}(e)$ bis auf Terme höherer Ordnung bereitstellen, d.h. Terme, die nicht in $\mathcal{V}$ enthalten sind $\rightarrow$ Oszillationsterme (hier eingeführt)

- man kann in den numerischen Beispielen später sehen, dass (wie auch in der Theorie) eine Verkleinerung der Oszillation auch eine Verringerung des Fehlers mit sich bringt $\Rightarrow$ wir führen die Oszillationsterme ein (auch ohne präzise Beweise)

- die Oszillation ist in zwei Teile kaufteilbar

$$\operatorname{osc}(u_{\mathcal{S}}, \psi, f) \coloneqq \left(\operatorname{osc}_1(u_{\mathcal{S}}, \psi)^2 + \operatorname{osc}_2(u_{\mathcal{S}}, \psi, f)^2\right)^{\frac{1}{2}} \tag{4.24}$$

-

  Bemerkung. In [ZVKG11] werden die Oszillationsterme (4.24) durch

$$\operatorname{osc}(u_{\mathcal{S}}, \psi, f) = \operatorname{osc}_1(u_{\mathcal{S}}, \psi) + \operatorname{osc}_2(u_{\mathcal{S}}, \psi, f)$$

  eingeführt. Wir wählen hier absichtlich eine leicht veränderte Darstellung, da diese für die spätere Implementierung bzgl. der lokalen Anteile der Oszillationen von Vorteil ist.

- im unbeschränkten Fall ist die Oszillation nur von f abhängig (s. [MNS00]) und dort wird daher von „Daten-Oszillation“ gesprochen
- osc_1 ist ein Maß für die Oszillation zwischen Hindernis ψ und der Galerkin-Lösung $u_{\mathcal{S}}$, d.h.

$$\operatorname{osc}_1(u_{\mathcal{S}}, \psi) \coloneqq \left(\sum_{p \in \mathcal{N}^{0+}} \|\nabla(\psi - u_{\mathcal{S}})\|^2_{0,\omega_p} \right)^{\frac{1}{2}}, \tag{4.25}$$

  wobei $\mathcal{N}^{0+} \coloneqq \{p \in \mathcal{N}^0 \mid u_{\mathcal{S}} > \psi \text{ in } \omega_p \setminus \{p\}\}$, also die Megne der *isolierten Kontaktknoten*, d.h. $u_{\mathcal{S}}$ ist in ω_p nur mit p in Kontakt
- anschaulich: da $\psi, u_{\mathcal{S}}$ linear, gilt: je größer die Differenz zwischen ψ und $u_{\mathcal{S}}$, umso größer die Differenz $\nabla(\psi - u_{\mathcal{S}})$, d.h. auch osc_1
- das kontinuierliche Gegenstück zu $\mathcal{N}^{0+}$ ist die Menge der *isolierten Kontaktpunkte* x_c, die aufgrund von $u - \psi > 0$ für alle $x \in \mathcal{U}(x_c, \varepsilon) \subset \Omega$ mit $u(x_c) = \psi(x_c)$ alle strikten Minima $x_c \in \Omega$ enthält $\Rightarrow (\nabla u - \nabla\psi) = 0$ für alle isolierten Kontaktpunkte, wenn u, ψ hinreichend glatt sind
- da laut Theorem **??** $u_h \to u$ für $h \to 0$ geht, folgt: wenn ein isolierter Kontaktknoten $p \in \mathcal{N}^{0+}$ bei Verfeinerung bestehen bleibt, so hat die exakte Lösung u einen korrespondierenden Kontaktpunkt $\tilde{p}$, dann gilt

$$\bigcup_{p \in \mathcal{N}^{0+}} \omega_p \xrightarrow[h \to 0]{} \tilde{p}$$

- damit gilt osc_1 hat wenigstens den Grad vom Fehler e (warum?)
- wegen oben (mit dem hinreichend glatten u, ψ) verschwindet osc_1 für $h \to 0$
- osc_2 ist über zwei Mengen definiert:

$$\mathcal{N}^{++} \coloneqq \{p \in \mathcal{N}^+ \mid \rho_E \geq -d_E \,\forall\, E \in \mathcal{E}_p\} \tag{4.26}$$

  d.h. alle Punkte ohne Kontakt, in denen der Fehler $\varepsilon_{\mathcal{V}}$ nicht in Kontakt mit $\mathcal{A}_{\mathcal{V}}$ steht (wie in Beweis von Satz **??** ersichtlich)

$$\mathcal{N}^{0-} \coloneqq \{p \in \mathcal{N}^0 \mid u_{\mathcal{S}} = \psi, f \leq 0 \text{ auf } \omega_p, j_E \leq 0 \,\forall\, E \in \mathcal{E}_p\} \tag{4.27}$$

  d.h. voller Kontakt (s. auch [SV07] Gleichung (2.11)) mit Last f auf Druck und negativem Normalenfluss j_E
- aus der Nebenbedingung von $\mathcal{N}^{0-}$ folgt

$$0 \geq f + \sum_{E \in \mathcal{E}_p} j_E$$

durch Multiplikation mit geeigneten Testfunktionen v und multiplizieren über ω_p ergibt

$$\begin{aligned} 0 &\geq \int_{\omega_p} fv\, d\Omega + \sum_{E\in\mathcal{E}_p} \int_E j_E v\, d\Gamma \\ &= \int_{\omega_p} fv\, d\Omega - \int_{\omega_p} \underbrace{\nabla u_{\mathcal{S}}}_{=\nabla\psi} \nabla v\, d\Omega \end{aligned}$$

und damit gilt

$$\int_{\omega_p} \nabla\psi\nabla v\, d\Omega \geq \int_{\omega_p} fv\, d\Omega \tag{4.28}$$

es gilt also laut Satz 3.4, dass $-\Delta\psi - f \geq 0$ auf ω_p im distributionellem Sinne (vgl. auch [Wal11] Kapitel 3)

dies ist laut Satz 3.4 auch notwendig, damit $u = \psi$ auf ω_p ist

- damit ergibt sich osc_2 als

$$\mathrm{osc}_2(u_{\mathcal{S}}, \psi, f) \coloneqq \left(\sum_{p\in\mathcal{N}^{++}} h_p^2 \|f - \bar{f}_p\|_{0,\omega_p}^2 + \sum_{p\in\mathcal{N}\setminus(\mathcal{N}^{0-}\cup\mathcal{N}^{++})} h_p^2 \|f\|_{0,\omega_p}^2 \right)^{\frac{1}{2}} \tag{4.29}$$

wobei $h_p \coloneqq \max_{E\in\mathcal{E}_p}|E|$ für jedes $p \in \mathcal{N}$ (h_p ist ein Maß für den Durchmesser von ω_p) und $\bar{f}_p$ den Mittelwert von f über ω_p bezeichne, d.h.

$$\bar{f}_p = \frac{1}{|\omega_p|} \int_{\omega_p} f\, d\Omega \tag{4.30}$$

- anschaulich: damit kann man die Summanden der ersten Summe als Varianz der Last f auf ω_p interpretieren

  $\mathcal{N}\setminus(\mathcal{N}^{0-}\cup\mathcal{N}^{++})$ ist die Menge von Punkten, die keinen vollen Kontakt und in der $\varepsilon\mathcal{V}$ keinen Kontakt mit $\mathcal{A}_{\mathcal{V}}$ hat

- Beachte: Im Term osc_2 fehlen nur die Punkte, die vollen Kontakt haben, d.h. wir betrachten also wirklich nur die Punkte außerhalb des Hindernisses!! (genauer noch: $\mathcal{N} \setminus (\mathcal{N}^{0-} \cup \mathcal{N}^{++})$ sind nur die Randpunkte!)

- damit enthält osc_2 nur Anteile aus Knoten, ohne vollen Kontakt

- Bem.: die Oszillationsterme können leicht berechnet werden (siehe hierfür auch Kapitel 5)

- im unbeschränkten Fall, also $\psi = -\infty$, ist $\varepsilon_{\mathcal{V}}$ nicht im Kontakt mit dem Hindernis für alle Punkte aus $\mathcal{N}$, also gilt $\mathcal{N}^{++} = \mathcal{N}$.
- damit wird (**??**) (osc_2) zu

$$\mathrm{osc}_2(u_{\mathcal{S}}, \psi, f) = \left(\sum_{p \in \mathcal{N} \cap \Omega} h_p^2 \|f - \bar{f}_p\|_{0,\omega_p}^2 + \sum_{p \in \mathcal{N} \cap \partial\Omega} h_p^2 \|f\|_{0,\omega_p}^2 \right)^{\frac{1}{2}} \tag{4.31}$$

- damit ist (4.28) eine Verallgemeinerung von (4.30): wenn der Teil ohne Kontakt also bekannt wäre, dann wäre der beschränkte Fall auf dieser Menge äquivalent zu einem unbeschränkten Dirichtlet-Problem
- WICHTIG: Noch einmal in [Zha07] schauen, ob dies in Verbindung des letzten Absatzes im Mainpaper verwendet werden kann!!!

4.1.4 Zuverlässigkeit des Fehlerschätzers

- wir wollen in diesem Kapitel eine obere Schranke des Fehlers im Energiefunktional, die vom hierarchischen Fehlerschätzer abhängt, herleiten.
- die Reduktion des Fehlers $e = u - u_{\mathcal{S}} \in H_0^1(\Omega)$ auf den approximierten Fehler $\varepsilon_{\mathcal{V}} \in \mathcal{V}$ erhalten wir durch lokale Projektionen für jedes $p \in \mathcal{N}$ mit

$$\pi_p : H^1(\Omega) \to \mathcal{Q}_p = \mathrm{span}\{\phi_p\} \cup \mathcal{V}_p\,, \quad \mathcal{V}_p = \mathrm{span}\{\phi_E \mid E \in \mathcal{E}_p\}\,. \tag{4.32}$$

- π_p ist für jedes $v \in H^1(\Omega)$ aus Dimensionsgründen ($\dim(\mathcal{Q}_p) = p + 1$) eindeutig bestimmt durch

$$\int_E \pi_p v \, d\Gamma = \int_E v \, d\Gamma \quad \forall\, E \in \mathcal{E}_p \text{ und } \begin{cases} \int_{\omega_p} \pi_p v \, d\Omega = \int_{\omega_p} v \, d\Omega & , p \in \mathcal{N}^{++} \\ 0 & , \text{sonst} \end{cases}. \tag{4.33}$$

- **Lemma 4.19.** *Es sei π_p die oben beschriebene Projektion. Dann gelten für die Koordinaten bzgl. der Basis $\{\phi_p\} \cup \{\phi_E \mid E \in \mathcal{E}_p\}$ von*

$$\pi_p v = \alpha_p(v)\phi_p + \sum_{E \in \mathcal{E}_p} \alpha_E(v)\phi_E$$

die Beziehungen

$$\alpha_p(v) = \begin{cases} \frac{c_p(v)}{c_p(\phi_p)} & , p \in \mathcal{N}^{++} \\ 0 & , sonst \end{cases}, \tag{4.34a}$$

$$\alpha_E(v) = \frac{\int_E v \, d\Gamma - \alpha_p(v) \int_E \phi_p \, d\Gamma}{\int_E \phi_E \, d\Gamma}, \tag{4.34b}$$

wobei

$$c_p(v) = \int_{\omega_p} v \, d\Omega - \sum_{E \in \mathcal{E}_p} \left(\int_E v \, d\Gamma \right) \left(\int_{\omega_p} \phi_E \, d\Omega \right) \left(\int_E \phi_E \, d\Gamma \right)^{-1}$$

Insbesondere gilt $c_p(\phi_p) = -\frac{1}{6} |\omega_p|$.

Beweis. Für eine bessere Übersicht im Beweis werden wir die Differentialformen $d\Omega$ und $d\Gamma$ weg. Es sei $v \in H^1(\Omega)$ beliebig. Dann gilt für jede Kante $E \in \mathcal{E}_p$ mit

$$\pi_p v = \alpha_p(v)\phi_p + \alpha_E(v)\phi_E \in \mathcal{Q}_p$$

nach (4.33), dass

$$\begin{aligned} \int_E v = \int_E \pi_p v &= \int_E \alpha_p(v)\phi_p + \alpha_E(v)\phi_E \\ \Longrightarrow \alpha_E(v) &= \left(\int_E v - \alpha_p(v) \int_E \phi_p \right) \left(\int_E \phi_E \right)^{-1}. \end{aligned} \tag{4.35}$$

Wenn $p \notin \mathcal{N}^{++}$ ist, so gilt $\pi_p v \in \mathcal{V}_p = \operatorname{span}\{\phi_E \mid E \in \mathcal{E}_p\}$, d.h. $\alpha_p(v) = 0$.

Es sei nun also $p \in \mathcal{N}^{++}$. Dann folgt aus der zweiten Eigenschaft von (4.33) und (4.35) für $\pi_p v = \alpha_p(v)\phi_p + \sum_{E \in \mathcal{E}_p} \alpha_E(v)\phi_E \in \mathcal{Q}_p$

$$\begin{aligned} \int_{\omega_p} v = \int_{\omega_p} \pi_p v &= \int_{\omega_p} \alpha_p(v)\phi_p + \sum_{E \in \mathcal{E}_p} \alpha_E(v)\phi_E \\ &= \alpha_p(v) \int_{\omega_p} \phi_p + \sum_{E \in \mathcal{E}_p} \alpha_E(v) \int_{\omega_p} \phi_E \\ &= \alpha_p(v) \int_{\omega_p} \phi_p + \sum_{E \in \mathcal{E}_p} \left(\int_E v - \alpha_p(v) \int_E \phi_p \right) \left(\int_E \phi_E \right)^{-1} \left(\int_{\omega_p} \phi_E \right) \\ &= \alpha_p(v) \Bigg(\underbrace{\int_{\omega_p} \phi_p - \sum_{E \in \mathcal{E}_p} \left(\int_E \phi_p \right) \left(\int_{\omega_p} \phi_E \right) \left(\int_E \phi_E \right)^{-1}}_{= c_p(\phi_p)} \Bigg) \\ &\quad + \sum_{E \in \mathcal{E}_p} \left(\int_E v \right) \left(\int_{\omega_p} \phi_E \right) \left(\int_E \phi_E \right)^{-1}. \end{aligned}$$

Nach dem Umformen nach $\alpha_p(v)$ ergibt sich dann

$$\alpha_p(v) = \frac{c_p(v)}{c_p(\phi_p)}$$

mit dem oben definierten $c_p(\cdot)$.

Es bleibt also zu zeigen, dass $c_p(\phi_p) = -\frac{1}{6}|\omega_p|$. Hierfür betrachten wir die einzelnen Summanden von $c_p(\phi_p)$. Zunächst berechnet das Integral von ϕ_p über ω_p das Volumen der von ϕ_p erzeugten Pyramide mit Grundfläche $|\omega_p|$, d.h.

$$\int_{\omega_p} \phi_p = \frac{1}{3}|\omega_p|. \tag{4.36}$$

Weiter ist ϕ_p auf jeder Kante $E \in \mathcal{E}_p$ eine von 1 zu 0 abfallende Gerade und damit ist das Integral über E gerade der Flächeninhalt vom darüber liegenden Dreieck, also

$$\int_E \phi_p = \frac{1}{2}|E|. \tag{4.37}$$

Die letzten beiden Integrale berechnen wir über die Referenzelemente in $\mathbb{R}$ oder $\mathbb{R}^2$ für das Kurven- bzw. Flächenintegral. Die Funktion ϕ_E über eine Kante E ist eine nach unten geöffnete Parabel. Auf dem Referenzelement $[-1,1] \subset \mathbb{R}$ ist dies die Funktion

$$\hat{\phi}_E = 1 - \xi^2$$

und mit einer affinen Transformation $s : [-1,1] \to [a,b] = E, s(\xi) = \frac{b-a}{2}\xi + \frac{b+a}{2}$ lässt sich das Referenzelement auf das Element E abbilden. Damit ergibt sich mit dem Transformationssatz der Integration

$$\int_E \phi_E = \frac{b-a}{2}\int_{-1}^{1} \hat{\phi}_E = \frac{1}{2}|E| \cdot \frac{4}{3} = \frac{2}{3}|E|. \tag{4.38}$$

Der letzte Fall ist komplizierter zu beschreiben. Zunächst sei erwähnt, dass $\operatorname{supp}(\phi_E) = T_i \cup T_j, T_i, T_j \subset \omega_p, i \neq j$ gilt, ϕ_E also nur auf zwei Dreiecken, die in ω_p enthalten sind, ungleich Null ist. Damit wird für jede Kante $E \in \mathcal{E}_p$ über jedes Dreieck $T \subset \omega_p$ genau zweimal integriert.

Auf dem Referenzelement

$$\hat{T} \coloneqq \{(\xi,\eta) \in \mathbb{R}^2 \mid 0 \leq \xi \leq 1, 0 \leq \eta \leq 1 - \xi\}$$

haben wir die drei Bubble-Funktionen

$$\hat{\phi}_{E_1} = 4\xi(1-\xi-\eta), \quad \hat{\phi}_{E_2} = 4\xi\eta, \quad \hat{\phi}_{E_3} = 4\eta(1-\xi-\eta),$$

für die man leicht nachrechnen kann, dass

$$\int_{\hat{T}} \hat{\phi}_{E_1} = \int_{\hat{T}} \hat{\phi}_{E_2} = \int_{\hat{T}} \hat{\phi}_{E_3} = \frac{1}{6}$$

gilt. Es sei nun J_T die Jacobi-Determinante bzgl. einer affinen Transformation $r : \hat{T} \to T$, dann gilt nach Transformationssatz mit einem $T \subset \operatorname{supp}(\phi_E)$

$$\int_T \phi_E = |J_T| \int_{\hat{T}} \hat{\phi}_E = \frac{1}{6} |J_T| .$$

Weiter rechnen wir nach, dass

$$|T| = \int_T d\Omega = |J_T| \int_{\hat{T}} d\Omega = \frac{1}{2} |J_T| \implies |J_T| = 2 |T|$$

gilt und damit folgt insgesamt zusammen mit (4.36) bis (4.38)

$$\begin{aligned} c_p(\phi_p) &= \int_{\omega_p} \phi_p - \sum_{E \in \mathcal{E}_p} \left(\int_E \phi_p \right) \left(\int_{\omega_p} \phi_E \right) \left(\int_E \phi_E \right)^{-1} \\ &= \frac{1}{3} |\omega_p| - 2 \sum_{T \subset \omega_p} \frac{1}{2} |E| \cdot \frac{1}{6} |J_T| \cdot \frac{3}{2} |E|^{-1} \\ &= \frac{1}{3} |\omega_p| - \sum_{T \subset \omega_p} \frac{1}{2} |T| = \left(\frac{1}{3} - \frac{1}{2} \right) |\omega_p| \\ &= -\frac{1}{6} |\omega_p| . \end{aligned}$$

□

- für den wichtigsten Satz dieser Arbeit benötigen wir noch eine Eigenschaft der lokalen Projektionen

Lemma 4.20. *Die Koeffizienten in* (4.34) *erfüllen die Eigenschaft*

$$\max_{\mathcal{Q} \in \{p\} \cup \mathcal{E}_p} |\alpha_{\mathcal{Q}}(v)| \lesssim h_p^{-1} (\|v\|_{0,\omega_p} + h_p \|\nabla v\|_{0,\omega_p}) \tag{4.39}$$

und π_p ist stabil im Sinne von

$$\|\pi_p v\|_{0,\omega_p} \lesssim \|v\|_{0,\omega_p} + h_p \|\nabla v\|_{0,\omega_p} . \tag{4.40}$$

Insbesondere gilt, wenn $p \notin \mathcal{N}^{++}$ ist, dass für $\alpha_E(v) = \left(\int_E v \, d\Gamma \right) \left(\int_E \phi_E \right)^{-1}$ die Eigenschaft gilt:

$$\int_E v \, d\Gamma \geq \int_E (\psi - u_{\mathcal{S}}) \, d\Gamma \Longrightarrow \alpha_E(v) \gtrsim \psi(x_E) - u_{\mathcal{S}}(x_E) \quad \forall \, E \in \mathcal{E}_p . \tag{4.41}$$

Beweis. Beweis machen??? □

- folgendes Lemma ist zentral, um die obere Schranke von $J(u_{\mathcal{S}}) - J(u)$ bzgl. $-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$ zu zeigen

  Lemma 4.21. *Es sei Voraussetzung 4.1 erfüllt. Dann gilt*

  $$\rho_{\mathcal{S}}(e) \lesssim \sum_{E\in\mathcal{E}} \eta_E \, |\rho_E| + \operatorname{osc}(u_{\mathcal{S}}, \psi, f)^2 \tag{4.42}$$

  mit ρ_E wie in (4.11), $\operatorname{osc}(u_{\mathcal{S}}, \psi, f)$ *wie in* (4.24) *und* $\eta_E = |\varepsilon_{\mathcal{V}}(x_E)| \, \|\phi_E\|$.

  Beweis. Die Idee des Beweises beruht auf Gleichung (4.20); damit können wir den Indikator in die lokalen Anteile bzgl. der Punkte $p \in \mathcal{N}$ aufteilen, d.h.

  $$\rho_{\mathcal{S}}(e) = \sum_{p\in\mathcal{N}} \rho_p(e) \, . \tag{4.43}$$

  Hierbei ist die Abschätzung der lokalen Anteile $\rho_p(e)$ abhängig von einer Anwendung der Poincaré-Friedrichungleichung (Satz 2.13), die auf die *verallgemeinerte Poincaré-Friedrich-Ungleichung* (vgl. auch [Rud91])

  $$\|v - c\|_{0,\omega_p} \lesssim h_p \|\nabla v\|_{0,\omega_p} \tag{4.44}$$

  mit einer Konstanten c und $v \in H^1(\Omega)$, so dass $v = c$ auf einer Menge $\Gamma \subset \partial\omega_p$ mit einem Maß $\mu(\Gamma) \neq 0$ gilt, führt. Da (4.44) von $p \in \mathcal{N}$ abhängt, werden wir sehen, dass die Anwendung von der Poincaré-Friedrich-Ungleichung vom Typ des Knotens p abhängt. Deshalb betrachten wir die disjunkte Vereinigung

  $$\begin{aligned} &\overbrace{\mathcal{N}^{++} \cup (\mathcal{N}^{+} \setminus \mathcal{N}^{++})}^{=\mathcal{N}^{+}} \cup (\mathcal{N} \cap \partial\Omega) \cup \overbrace{(\mathcal{N}^{0} \setminus (\mathcal{N}^{0+} \cup \mathcal{N}^{0-})) \cup \mathcal{N}^{0+} \cup \mathcal{N}^{0-}}^{=\mathcal{N}^{0}} \\ &= \mathcal{N}^{+} \cup \mathcal{N}^{0} \cup (\mathcal{N} \cap \partial\Omega) = (\mathcal{N} \cap \Omega) \cup (\mathcal{N} \cap \partial\Omega) = \mathcal{N} \, . \end{aligned} \tag{4.45}$$

  Wir wollen im Folgenden die in (4.45) aufgeführten Fälle chronologisch abarbeiten.

  Fall 1: Es sei $p \in \mathcal{N}^{++}$. Wir behaupten, dass

  $$\rho_p(e) \lesssim \left(\sum_{E\in\mathcal{E}_p^+} |\rho_E| + h_p \|f - \bar{f}_p\|_{0,\omega_p} \right) \|\nabla e\|_{0,\omega_p} \tag{4.46}$$

  gilt, wobei $\mathcal{E}_p^+ = \{E \in \mathcal{E}_p \mid \rho_E \geq -d_E\}$. Da $p \in \mathcal{N}^{++}$ ist, gilt für alle $E \in \mathcal{E}_p$ die Ungleichung $\rho_E \geq -d_E$ ist, d.h. $\mathcal{E}_P^+ = \mathcal{E}_p$. Wir setzen

  $$w = (e - c)\phi_p \, , \quad c = \frac{1}{|\omega_p|} \int_{\omega_p} e \, d\Omega \, .$$

Da $\mathcal{N}^{++} \subset \mathcal{N}^{+} \cap \Omega$ ist, d.h. $u_{\mathcal{S}}(p) > \psi(p)$ ist, gilt

$$\rho_p(c) = c\,\rho_p(1) \overset{(4.23b)}{=} c \cdot 0 = 0\,.$$

Damit erhalten wir

$$\begin{aligned}
\rho_p(e) = \rho_p(e) - \rho_p(c) = \rho_p(e-c) &\overset{\text{Lem. 4.15}}{=} \int_{\omega_p} f w \,d\Omega + \sum_{E\in\mathcal{E}_p} \int_E j_E w \,d\Gamma \\
&\overset{\text{„}+0\text{“}}{=} \int_{\omega_p} f \pi_p w \,d\Omega + \sum_{E\in\mathcal{E}_p} \underbrace{\int_E j_E \pi_p w \,d\Gamma}_{\underset{(4.33)}{=}\int_E j_E w\,d\Gamma} + \int_{\omega_p} f(w-\pi_p w)\,d\Omega \\
&= \rho_{\mathcal{S}}(\pi_p w) + \int_{\omega_p} f(w-\pi_p w)\,d\Omega - \bar{f}_p \underbrace{\int_{\omega_p} (w-\pi_p w)\,d\Omega}_{=0 \text{ wegen } (4.33)} \\
&= \rho_{\mathcal{S}}(\pi_p w) + \int_{\omega_p} (f-\bar{f}_p)(w-\pi_p w)\,d\Omega \\
&\overset{\text{C.S.}}{\leq} \sum_{E\in\mathcal{E}_p} \alpha_E(w)\rho_E\|\phi_E\| + \|f-\bar{f}_p\|_{0,\omega_p}\|w-\pi_p w\|_{0,\omega_p}\,, \qquad (4.47)
\end{aligned}$$

wobei im letzten Schritt zusätzlich zur Cauchy-Schwarz-Ungleichung im zweiten Summanden noch angewendet wurde, dass

$$\begin{aligned}
\rho_{\mathcal{S}}(\pi_p w) &= \rho_{\mathcal{S}}\left(\alpha_p(w)\phi_p + \sum_{E\in\mathcal{E}_p} \alpha_E(w)\phi_E\right) \\
&= \alpha_p(w)\underbrace{\rho_{\mathcal{S}}(\phi_p)}_{=\rho_p(1)=0} \sum_{E\in\mathcal{E}_p} \alpha_E(w) \underbrace{\rho_{\mathcal{S}}(\phi_E)}_{\underset{(4.11)}{=}\rho_E\|\phi_E\|} \\
&= \sum_{E\in\mathcal{E}_p} \alpha_E(w)\rho_E\|\phi_E\|
\end{aligned}$$

ist. Da $\|\phi_p\|_{\infty,\omega_p} \leq 1$, gilt mit der Poincaré-Friedrich-Ungleichung (4.44)

$$\|w\|_{0,\omega_p} = \|(e-c)\phi_p\|_{0,\omega_p} \leq \|e-c\|_{0,\omega_p} \lesssim h_p\|\nabla e\|_{0,\omega_p}\,, \qquad (4.48)$$

wobei die erste Ungleichung aus dem Mittelwertsatz der Integralrechnung folgt. Es sei weiterhin darauf hingewiesen, dass $\|\nabla\phi_p\|_{\infty,\omega_p} \lesssim h_p^{-1}$, da die Steigung der Hutfunktion nur von der Form von ω_p abhängt.

Damit erhalten wir dann durch Anwenden von (4.39) für alle $E \in \mathcal{E}_p$

$$\begin{aligned}
|\alpha_E(w)| &\lesssim h_p^{-1}(\|w\|_{0,\omega_p} + h_p\|\nabla w\|_{0,\omega_p}) \\
&= h_p^{-1}(\|(e-c)\phi_p\|_{0,\omega_p} + h_p\|\nabla((e-c)\phi_p)\|_{0,\omega_p}) \\
&\leq h_p^{-1}(h_p\|\nabla e\|_{0,\omega_p} + h_p\|\underbrace{\nabla(e-c)}_{=\nabla e}\phi_p + (e-c)\nabla\phi_p\|_{0,\omega_p}) \\
&\overset{\triangle\neq}{\leq} \|\nabla e\|_{0,\omega_p} + \|\nabla e\,\phi_p\|_{0,\omega_p} + \|(e-c)\nabla\phi_p\|_{0,\omega_p} \\
&\lesssim 2\,\|\nabla e\|_{0,\omega_p} + h_p^{-1}\|e-c\|_{0,\omega_p} \overset{(4.48)}{\lesssim} \|\nabla e\|_{0,\omega_p}\,.
\end{aligned} \tag{4.49}$$

Über das Referenzelement $\hat{T}$ kann man zeigen, dass

$$\|\phi_E\| = a(\phi_E,\phi_E)^{\frac{1}{2}} = \left(\int_\Omega \nabla\phi_E\nabla\phi_E\,d\Omega\right)^{\frac{1}{2}} = \left(\frac{8}{3}(|J_{T_1}| + |J_{T_2}|)\right)^{\frac{1}{2}} \lesssim \tilde{c}\,,$$

da die Jacobi-Determinanten J_{T_1}, J_{T_2} (wobei $E \subset T_i, i = 1,2$ gilt) endlich sind, jedoch von der Form von ω_p abhängen. Damit gilt

$$\|\nabla e\|_{0,\omega_p} \approx \|\phi_E\|^{-1}\|\nabla e\|_{0,\omega_p}\,. \tag{4.50}$$

Analog folgt mit Anwendung von (4.40) und (4.48), dass gilt:

$$\begin{aligned}
\|w - \pi_p w\|_{0,\omega_p} &\overset{\triangle\neq}{\leq} \|w\|_{0,\omega_p} + \|\pi_p w\|_{0,\omega_p} \\
&\lesssim h_p\|\nabla e\|_{0,\omega_p} + \|w\|_{0,\omega_p} + h_p\|\nabla e\|_{0,\omega_p} \lesssim h_p\|\nabla e\|_{0,\omega_p}
\end{aligned} \tag{4.51}$$

Setzen wir (4.49) bis (4.51) in (4.47), so folgt nach Ausklammern von $\|\nabla e\|_{0,\omega_p}$ die Behauptung (4.46) mit $\mathcal{E}_p = \mathcal{E}_p^+$.

Fall 2: Es sei $p \in \mathcal{N}^+ \setminus \mathcal{N}^{++}$. Wir behaupten, dass gilt:

$$\rho_p(e) \lesssim \left(\sum_{E\in\mathcal{E}_p^+} |\rho_E| + h_p\|f\|_{0,\omega_p}\right)\|\nabla e\|_{0,\omega_p}\,. \tag{4.52}$$

Auch hier können wir analog zu (4.47) eine Ungleichung herleiten, wobei die zweite Addition der Null nicht gilt, da $p \notin \mathcal{N}^{++}$. Damit erhalten wir die Aussage

$$\rho_p(e) \leq \sum_{E\in\mathcal{E}_p} \alpha_E(w)\rho_E\|\phi_E\| + \|f\|_{0,\omega_p}\|w - \pi_p w\|_{0,\omega_p}\,, \tag{4.53}$$

wobei wir hier

$$w = (e-c)\phi_p\,, \quad c = \min\left\{\left(\int_E e\phi_p\,d\Gamma\right)\left(\int_E \phi_p\,d\Gamma\right)^{-1} \mid E \in \mathcal{E}_p\right\} \tag{4.54}$$

setzen. Damit gilt

$$\begin{aligned}\alpha_E(w) &= \left(\int_E w\,d\Gamma\right)\left(\int_E \phi_E\,d\Gamma\right)^{-1} = \left(\int_E (e-c)\phi_p\,d\Gamma\right)\left(\int_E \phi_E\,d\Gamma\right)^{-1}\\ &= \left(\int_E e\phi_p\,d\Gamma - c\int_E \phi_p\,d\Gamma\right)\left(\int_E \phi_E\,d\Gamma\right)^{-1}\\ &\overset{(4.54)}{\geq} \left(\int_E e\phi_p\,d\Gamma - \left(\int_E e\phi_p\,d\Gamma\right)\left(\int_E \phi_p\,d\Gamma\right)^{-1}\left(\int_E \phi_p\,d\Gamma\right)\right)\left(\int_E \phi_E\,d\Gamma\right)^{-1}\\ &= 0\,.\end{aligned}$$

Daraus können wir folgern, dass für die Kanten $E \in \mathcal{E}_p$ mit $\rho_E < -d_E \leq 0$

$$\alpha_E(w)\rho_E \leq 0$$

gilt. Ersetzen wir dies in (4.53), so folgt (4.52) insgesamt mit denselben Abschätzungen aus (4.49), (4.50) und (4.51).

Fall 3: Sei $p \in \mathcal{N} \cap \partial\Omega$. Wir behaupten für diesen Fall, dass

$$\rho_p(e) \lesssim \sum_{E\in\mathcal{E}_p^0} d_E\,|\rho_E| + \left(\sum_{E\in\mathcal{E}_p^+} |\rho_E| + h_p\|f\|_{0,\omega_p}\right)\|\nabla e\|_{0,\omega_p} \tag{4.55}$$

mit $\mathcal{E}_p^0 = \mathcal{E}_p \setminus \mathcal{E}_p^+ = \{E \in \mathcal{E}_p \mid \rho_E < -d_E\}$. Auch hier betrachten wir die Ungleichung (4.53), wobei wir dieses Mal $w = e\phi_p$ setzen, d.h. mit der obigen Wahl $c = 0$ setzen. In diesem Fall kann kein anderes c gewählt werden, da wir wegen $p \notin \Omega$ nicht direkt $\rho_p(c) = 0$ aus (4.23b) folgern können. Da $p \in \partial\Omega$ ist, gilt mindestens auf einer Kante $E \in \mathcal{E}_p$ von $\partial\omega_p$

$$e = u - u_{\mathcal{S}} = 0\,.$$

Damit ist e auf einer Teilmenge, vom Maße ungleich Null, des Randes gleich Null und wir können daher die allgemeine Poincaré-Friedrich-Ungleichung (4.44) anwenden. Damit erhalten wir die Anteile von $E \in \mathcal{E}_p^+$ durch die Abschätzungen (4.49), (4.50) und (4.51). Um die Anteile für die Kanten $E \in \mathcal{E}_p^0$ zu erhalten, betrachten wir

$$\begin{aligned}u_{\mathcal{S}} + w = u_{\mathcal{S}} + e\phi_p &= u_{\mathcal{S}} + (u - u_{\mathcal{S}})\phi_p\\ &= \underbrace{(1-\phi_p)}_{\in[0,1]}u_{\mathcal{S}} + \underbrace{\phi_p}_{\in[0,1]}u\\ &\geq (1-\phi_p)\psi + \phi_p\psi = \psi\,.\end{aligned}$$

Da die Ungleichung punktweise gilt, folgt auch

$$\int_E w\,d\Gamma \geq \int_E \psi - u_{\mathcal{S}}\,d\Gamma \overset{(4.41)}{\Longrightarrow} \alpha_E(w) \gtrsim \psi(x_E) - u_{\mathcal{S}}(x_E) \overset{(4.11)}{=} -d_E\,\|\phi_E\|^{-1}$$

und daher gilt $\alpha_E(w) \lesssim d_E \|\phi_E\|^{-1}$. Alle Aussagen ersetzt in (4.53) ergeben dann die Behauptung.

Fall 4: Sei $p \in \mathcal{N}^0 \setminus (\mathcal{N}^{0-} \cup \mathcal{N}^{0+})$. Wir behaupten, dass

$$\rho_p(e) \lesssim \sum_{E \in \mathcal{E}_p^0} d_E \, |\rho_E| + \left(\sum_{E \in \mathcal{E}_p^+} |\rho_E| + h_p \|f\|_{0,\omega_p} \right) \|\nabla e\|_{0,\omega_p} \tag{4.56}$$

gilt. Um dies zu zeigen, spalten wir den Fehler $e = e^+ + e^-$ auf mit $e^+ := \max\{e, 0\}$ und $e^- := \min\{e, 0\}$. Damit lässt sich der lokale Anteil des Indekators $\rho_{\mathcal{S}}$ schreiben als

$$\rho_p(e) = \rho_p(e^+ + e^-) = \rho_p(e^+) + \rho_p(e^-) \, . \tag{4.57}$$

Wir betrachten zunächst $\rho_p(e^+)$ und setzen analog zum Fall 2

$$w = (e^+ - c)\phi_p \, , \quad c = \min \left\{ \left(\int_E e^+ \phi_p \, d\Gamma \right) \left(\int_E \phi_p \, d\Gamma \right)^{-1} \mid E \in \mathcal{E}_p \right\} . \tag{4.58}$$

Da $e^+ \geq 0$ ist, gilt auch $c \geq 0$. Wegen $\mathcal{N}^0 \setminus (\mathcal{N}^{0-} \cup \mathcal{N}^{0+}) \subset \mathcal{N} \cap \Omega$ gilt nach (4.23a) $\rho_p(1) \leq 0$, also analog zu Fall 2

$$\begin{aligned} \rho_p(e^+) &\leq \rho_p(e^+) - \underbrace{c \, \rho_p(1)}_{\leq 0} = \rho_p(e^+ - c) \\ &\leq \sum_{E \in \mathcal{E}_p} \alpha_E(w) \rho_E \|\phi_E\| + \|f\|_{0,\omega_p} \|w - \pi_p w\|_{0,\omega_p} \, . \end{aligned} \tag{4.59}$$

Die Aussagen (4.58), (4.59) sind identisch zu denen aus Fall 2, wenn wir $e = e^+$ setzen. Damit folgt mit $\|\nabla e^+\|_{0,\omega_p} \leq \|\nabla e\|_{0,\omega_p}$ und denselben Argumenten wie in Fall 2, dass

$$\rho_p(e^+) \lesssim \left(\sum_{E \in \mathcal{E}_p^+} |\rho_E| + h_p \|f\|_{0,\omega_p} \right) \|\nabla e\|_{0,\omega_p} \, . \tag{4.60}$$

Wir betrachten nun $\rho_p(e^-)$. Analog zu Fall 3 setzen wir $w = (e^- - c)\phi_p, c = 0$ und leiten damit wieder die obere Schranke

$$\rho_p(e^-) \leq \sum_{E \in \mathcal{E}_p} \alpha_E(w) \rho_E \|\phi_E\| + \|f\|_{0,\omega_p} \|w - \pi_p\|_{0,\omega_p} \tag{4.61}$$

her. Weiter gilt punktweise

$$\begin{aligned} w = e^- \phi_p &\geq e^- = \min\{e, 0\} = \min\{u - u_{\mathcal{S}}, 0\} \\ &\geq \min\{\underbrace{\psi - u_{\mathcal{S}}}_{\leq 0}, 0\} = \psi - u_{\mathcal{S}} \end{aligned}$$

und damit auch die Aussage über $E \in \mathcal{E}_p$ integriert; also folgt aus (4.41)

$$0 \geq \alpha_E(w) \gtrsim \psi(x_E) - u_{\mathcal{S}}(x_E) = -d_E \|\phi_E\|^{-1} \quad \forall\, E \in \mathcal{E}_p\,. \tag{4.62}$$

Damit folgt speziell für alle $E \in \mathcal{E}_p^0$ die Abschätzung

$$\begin{aligned} |\alpha_E(w)\rho_E \|\phi_E\|| &= |\alpha_E(w)|\,|\rho_E|\,\|\phi_E\| \\ &\overset{(4.62)}{\lesssim} d_E \|\phi_E\|^{-1}\,|\rho_E|\,\|\phi_E\| = d_E\,|\rho_E|\,. \end{aligned} \tag{4.63}$$

Nun bleiben noch die oberen Schranken von $|\alpha_E(w)|$ für $E \in \mathcal{E}_p^+$ und $\|w - \pi_p w\|_{0,\omega_p}$ zu zeigen. Da $p \notin \mathcal{N}^{0+}$, also kein isolierter Knoten ist, gibt es mindestens eine Kante $E \in \mathcal{E}_p$, so dass $u_{\mathcal{S}} = \psi$ gilt. Damit folgt

$$0 = \psi - u_{\mathcal{S}} \leq e^- = \min\{e, 0\} \leq 0 \implies e^- = 0 \text{ auf } E\,.$$

Wie in Fall 3 ist damit die allgemeine Poincaré-Friedrich-Ungleichung anwendbar und wir können die Aussagen (4.49), (4.50) und (4.51) mit Anwendung von $\|\nabla e^-\|_{0,\omega_p} \leq \|\nabla e\|_{0,\omega_p}$ zeigen. Insgesamt folgt dann mit (4.63)

$$\rho_p(e^-) \lesssim \sum_{E \in \mathcal{E}_p^0} d_E\,|\rho_E| + \left(\sum_{E \in \mathcal{E}_p^+} |\rho_E| + h_p \|f\|_{0,\omega_p} \right) \|\nabla e\|_{0,\omega_p}\,. \tag{4.64}$$

Zusammen mit (4.57), (4.60) und (4.64) folgt dann die Behauptung (4.56).

Fall 5: Es sei nun $p \in \mathcal{N}^{0+}$. Wir behaupten, dass

$$\begin{aligned} \rho_p(e) \lesssim{}& \sum_{E \in \mathcal{E}_p^0} d_E\,|\rho_E| \\ &+ \left(\sum_{E \in \mathcal{E}_p^+} |\rho_E| + h_p \|f\|_{0,\omega_p} \right) \left(\|\nabla e\|_{0,\omega_p} + \|\nabla(\psi - u_{\mathcal{S}})\|_{0,\omega_p} \right) \end{aligned} \tag{4.65}$$

gilt. Wie in Fall 4 verwenden wir die Aufteilung des Indikators nach Gleichung (4.57). Mit genau demselben Vorgehen wie in Fall 4 können wir für $\rho_p(e^+)$ zeigen, dass die Abschätzung (4.60) gilt. Für $\rho_p(e^-)$ können die Aussagen (4.61) bis (4.63) wie in Fall 4 gezeigt werden. Es bleiben also auch hier noch die oberen Schranken von $|\alpha_E(w)|$ für $E \in \mathcal{E}_p^+$ und $\|w - \pi_p w\|_{0,\omega_p}$ zu zeigen.

Wir erinnern uns, dass $\psi - u_{\mathcal{S}} \leq e^- \leq w \leq 0$ gilt und damit folgt

$$\begin{aligned} 0 \geq \alpha_E(w) &= \left(\int_E w \, d\Gamma\right)\left(\int_E \phi_E \, d\Gamma\right)^{-1} \\ &\geq \left(\int_E \psi - u_{\mathcal{S}} \, d\Gamma\right)\left(\int_E \phi_E \, d\Gamma\right)^{-1} \\ &= \alpha_E(\psi - u_{\mathcal{S}}) \,. \end{aligned} \tag{4.66}$$

Aus (4.66) folgt

$$\begin{aligned} |\alpha_E(w)| \, \|\phi_E\| &\leq |\alpha_E(\psi - u_{\mathcal{S}})| \, \|\phi_E\| \\ &= \left| \left(\int_E \psi - u_{\mathcal{S}} \, d\Gamma\right) \underbrace{\left(\int_E \phi_E \, d\Gamma\right)^{-1}}_{\lesssim h_p^{-1}} \right| \cdot \underbrace{\left(\int_{\omega_p} \nabla\phi_E \nabla\phi_E \, d\Omega\right)^{\frac{1}{2}}}_{\lesssim h_p^{\frac{1}{2}}} \\ &\lesssim h_p^{-\frac{1}{2}} \left| \int_E \psi - u_{\mathcal{S}} \, d\Gamma \right| \overset{\text{C.S.}}{\lesssim} h_p^{-\frac{1}{2}} \|\psi - u_{\mathcal{S}}\|_{0,E} \\ &\lesssim \|\nabla(\psi - u_{\mathcal{S}})\|_{0,\omega_p} \,, \end{aligned} \tag{4.67}$$

wobei im letzten Schritt eine *skalierte* Version der Poincaré-Friedrich-Ungleichung [1] verwendet wurde, die darauf basiert, dass $(\psi - u_{\mathcal{S}})(p) = 0$ und $\psi - u_{\mathcal{S}}$ linear ist wegen Voraussetzung 4.1. Wegen $\psi - u_{\mathcal{S}} \leq w$ folgt auch, dass

$$\|w\|_{0,\omega_p} \leq \|\psi - u_{\mathcal{S}}\|_{0,\omega_p} \lesssim h_p \|\nabla(\psi - u_{\mathcal{S}})\|_{0,\omega_p} \,, \tag{4.68}$$

wobei in (4.68) im letzten Schritt wegen $(\psi - u_{\mathcal{S}})(p) = 0$ wieder die skalierte Version der Poincaré-Friedrich-Ungleichung verwendet wurde. Aus (4.67) folgern wir unter Verwendung der Dreiecksungleichung

$$\begin{aligned} \|\pi_p w\|_{0,\omega_p} &\leq \sum_{E \in \mathcal{E}_p} |\alpha_E(w)| \, \|\phi_E\|_{0,\omega_p} \\ &\lesssim \sum_{E \in \mathcal{E}_p} |\alpha_E(w)| \, h_p \underbrace{\|\nabla\phi_E\|_{0,\omega_p}}_{=\|\phi_E\|} \\ &\lesssim h_p \|\nabla(\psi - u_{\mathcal{S}})\|_{0,\omega_p} \,. \end{aligned} \tag{4.69}$$

Also folgt aus (4.68) und (4.69) insgesamt

$$(\mathbf{??})\|w - \pi_p w\|_{0,\omega_p} \leq \|w\|_{0,\omega_p} + \|\pi_p w\|_{0,\omega_p} \lesssim h_p \|\nabla(\psi - u_{\mathcal{S}})\|_{0,\omega_p} \,. \tag{4.70}$$

[1] Anschaulich können wir uns dies folgendermaßen vorstellen: Da $\psi - u_{\mathcal{S}}$ linear und am Punkt p gleich ist, ist der Gradient $\nabla(\psi - u_{\mathcal{S}}) = \text{const}$. Damit kann man mit dieser Konstanten als Höhe mit dem Mittelwertsatz der Integralrechnung die beiden Normen gegenseitig abschätzen, wobei diese Ungleichung abhängig von einer Konstanten c ist, die wiederum nur von der Form von ω_p bzw. E abhängt.

Setzen wir nun für die Kanten $E \in \mathcal{E}_p^0$ die Abschätzung (4.63) und für die Kanten $E \in \mathcal{E}_p^+$ die Ungleichungen (4.67) und (**??**) in die Bedingung (4.61) ein, so erhalten wir die Aussage

$$\rho_p(e^-) \lesssim \sum_{E\in\mathcal{E}_p^0} d_E\,|\rho_E| + \left(\sum_{E\in\mathcal{E}_p^+} |\rho_E| + h_p\|f\|_{0,\omega_p}\right)\|\nabla(\psi - u_{\mathcal{S}})\|_{0,\omega_p}\,. \tag{4.71}$$

Damit folgt mit der Aufteilung (4.57) und den Abschätzungen (4.60) und (4.71) die Behauptung.

Fall 6: Es sei $p \in \mathcal{N}^{0-}$. In diesem Fall haben wir vollen Kontakt, also $u_{\mathcal{S}} = \psi$ auf ω_p und daher gilt

$$e = u - u_{\mathcal{S}} = u - \psi \geq 0 \text{ auf ganz } \omega_p\,.$$

Weiter gelten für alle $p \in \mathcal{N}^{0-}$ die Eigenschaften $f \leq 0$ auf ω_p und $j_E \leq 0$ für alle $E \in \mathcal{E}_p$. Daher rechnen wir leicht nach, dass gilt:

$$\rho_p(e) = \rho_{\mathcal{S}}(e\phi_p) = \int_{\omega_p} \underbrace{f e\phi_p}_{\leq 0}\,d\Omega + \sum_{E\in\mathcal{E}_p}\int_E \underbrace{j_E e\phi_p}_{\leq 0}\,d\Gamma \leq 0\,. \tag{4.72}$$

Bevor wir die sechs Fälle zusammenführen, machen wir uns klar, dass gilt

$$\begin{aligned}&(\mathcal{N}^+ \setminus \mathcal{N}^{++}) \cup (\mathcal{N}\cap\partial\Omega) \cup \overbrace{(\mathcal{N}^0 \setminus (\mathcal{N}^{0-}\cup\mathcal{N}^{0+}))\cup\mathcal{N}^{0+}}^{=\mathcal{N}^0\setminus\mathcal{N}^{0-}}\\ &= (\mathcal{N}\cap\Omega)\setminus(\mathcal{N}^{0-}\cup\mathcal{N}^{++}) \cup (\mathcal{N}\cap\partial\Omega) = \mathcal{N}\setminus(\mathcal{N}^{0-}\cup\mathcal{N}^{++})\,.\end{aligned}$$

Verwenden wir nun die sechs gezeigten Fälle, so ergibt sich:

$$\begin{aligned}\rho_{\mathcal{S}}(e) &= \sum_{p\in\mathcal{N}} \rho_p(e)\\ &= \sum_{E\in\mathcal{E}^0} d_E\,|\rho_E| + \sum_{p\in\mathcal{N}\setminus\mathcal{N}^{0-}} \left(\sum_{E\in\mathcal{E}_p^+}|\rho_E|\right)\|\nabla e\|_{0,\omega_p}\\ &\quad + \sum_{p\in\mathcal{N}\setminus(\mathcal{N}^{0-}\cup\mathcal{N}^{++})} h_p\|f\|_{0,\omega_p}\|\nabla e\|_{0,\omega_p} + \sum_{p\in\mathcal{N}^{++}} h_p\|f-\bar{f}_p\|_{0,\omega_p}\|\nabla e\|_{0,\omega_p}\\ &\quad + \sum_{p\in\mathcal{N}^{0+}} \left(\sum_{E\in\mathcal{E}_p^+}|\rho_E| + h_p\|f\|_{0,\omega_p}\right)\|\nabla(\psi-u_{\mathcal{S}})\|_{0,\omega_p}\,.\end{aligned}$$

Damit folgt nach der Cauchy-Schwarz-Ungleichung, dass mit einer Konstante $C > 0$, die nur von der Quasi-Uniformität von $\mathcal{T}_h$ abhängt,

gilt:

$$\begin{aligned}
C\rho_{\mathcal{S}}(e) &\leq \sum_{E\in\mathcal{E}^0} d_E\,|\rho_E| + \Bigg(\sum_{E\in\mathcal{E}^+}|\rho_E|^2 + \sum_{p\in\mathcal{N}\setminus(\mathcal{N}^{0-}\cup\mathcal{N}^{++})} h_p^2\|f\|_{0,\omega_p}^2 \\
&\quad + \sum_{p\in\mathcal{N}^{++}} h_p^2\|f-\bar{f}_p\|_{0,\omega_p}^2\Bigg)^{\frac{1}{2}}\Bigg(\|\nabla e\|_{0,\Omega}^2 + \sum_{p\in\mathcal{N}^{0+}}\|\nabla(\psi-u_{\mathcal{S}})\|_{0,\omega_p}^2\Bigg)^{\frac{1}{2}} \\
&= \sum_{E\in\mathcal{E}^0} d_E\,|\rho_E| + \Bigg(\sum_{E\in\mathcal{E}^+}|\rho_E|^2 + \operatorname{osc}_2(u_{\mathcal{S}},\psi,f)^2\Bigg)^{\frac{1}{2}}\Big(\|\nabla e\|_{0,\Omega}^2 + \operatorname{osc}_1(u_{\mathcal{S}},\psi)^2\Big)^{\frac{1}{2}} \\
&\leq \sum_{E\in\mathcal{E}^0} d_E\,|\rho_E| + \frac{\varepsilon}{2}\Bigg(\sum_{E\in\mathcal{E}^+}|\rho_E|^2 + \operatorname{osc}_2(u_{\mathcal{S}},\psi,f)^2\Bigg) + \frac{1}{2\varepsilon}\Big(\|\nabla e\|_{0,\Omega}^2 + \operatorname{osc}_1(u_{\mathcal{S}},\psi)^2\Big),
\end{aligned}$$

wobei wir als letztes die Ungleichung von Young mit einem $\varepsilon > 0$ verwendet haben und

$$\mathcal{E}^0 = \bigcup_{p\in\mathcal{N}}\mathcal{E}_p^0, \quad \mathcal{E}^+ = \bigcup_{p\in\mathcal{N}}\mathcal{E}_p^+.$$

Wählen wir $\varepsilon \leq C$, so ergibt sich nach leichtem Umstellen der Ungleichung

$$\begin{aligned}
c(\varepsilon)\rho_{\mathcal{S}}(e) &\leq \frac{\varepsilon}{2}\left(\sum_{E\in\mathcal{E}^+}|\rho_E|^2 + \operatorname{osc}_2(u_{\mathcal{S}},\psi,f)^2\right) + \frac{1}{2\varepsilon}\operatorname{osc}_1(u_{\mathcal{S}},f)^2 + \sum_{E\in\mathcal{E}^0} d_E\,|\rho_E| \\
&\leq \left(1+\frac{1}{2\varepsilon}\right)\sum_{E\in\mathcal{E}_p}\eta_E\,|\rho_E| + \left(\frac{\varepsilon}{2}+\frac{1}{2\varepsilon}\right)\left(\operatorname{osc}_2(u_{\mathcal{S}},\psi,f)^2 + \operatorname{osc}_1(u_{\mathcal{S}},f)^2\right) \\
&\lesssim \sum_{E\in\mathcal{E}_p}\eta_E\,|\rho_E| + \operatorname{osc}(u_{\mathcal{S}},\psi,f)^2
\end{aligned}$$

mit $c(\varepsilon) = \varepsilon - \frac{1}{2\varepsilon}$ und η_E wie in (4.11) definiert. Damit folgt die Behauptung. □

- Theorem für obere Schranke des Fehlerschätzers:

Theorem 4.22. *Es sei Voraussetzung 4.1 für ψ erfüllt. Dann ist der hierarchische Fehlerschätzer $-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$ eine obere Schranke für den Fehler im Energiefunktional bis auf Addition von Oszillationstermen und einer Konstante C, die nur von der Quasi-Uniformität von $\mathcal{T}_h$ abhängt, d.h.*

$$J(u_{\mathcal{S}}) - J(u) \lesssim -\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}}) + \operatorname{osc}(u_{\mathcal{S}},\psi,f)^2. \tag{4.73}$$

Beweis. Die Aussage folgt direkt durch Lemma 4.13 und 4.21, denn

$$\begin{aligned} J(u_{\mathcal{S}}) - J(u) &= -\mathcal{I}(e) \leq \rho_{\mathcal{S}}(e) \\ &\lesssim \underbrace{\sum_{E\in\mathcal{E}} \eta_E \left|\rho_E\right|}_{=\rho_{\mathcal{S}}(\varepsilon_{\mathcal{V}})} + \operatorname{osc}(u_{\mathcal{S}}, \psi, f)^2 \\ &\leq 2 \cdot (-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})) + \operatorname{osc}(u_{\mathcal{S}}, \psi, f)^2 \\ &\leq 2 \cdot (-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}}) + \operatorname{osc}(u_{\mathcal{S}}, \psi, f)^2) \end{aligned}$$

und damit folgt die Behauptung. □

- an dieser Abschätzung können wir sehen, dass es sinnvoll ist, nicht nur den hierarchischen Fehlerschätzer $-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$ zum Abschätzen des Fehlers zu verwenden, sondern auch die Oszillationsterme (diese sollten von Verfeinerungsschritt zum Verfeinerungsschritt kleiner werden, sonst ist die Verringerung im exakten Fehler nicht gesichert)
- damit nachher eine Abschätzung analog zu [MNS00] Lemma 3.8 (wäre schön, wenn diese noch gezeigt werden würde....)

  Lemma 4.23. *Es sei $0 < \gamma < 1$ ein Parameter, der die Reduktion der Größe des Dreiecks bei Verfeinerung wiedergibt. Weiter sei $0 < \hat{\theta} < 1$ gegeben und eine Menge an Punkten $\hat{\mathcal{N}} \subset \mathcal{N}$, die die zu verfeinernden Dreiecke anzeigen, gegeben, so dass*

$$\operatorname{osc}(u_{\mathcal{S}}, \psi, f, \hat{\mathcal{N}}) \geq \hat{\theta} \operatorname{osc}(u_{\mathcal{S}}, \psi, f, \mathcal{N}) .$$

  Dann existiert ein $\hat{\alpha} \in (0,1)$, so dass

$$\operatorname{osc}(u_{\mathcal{S}}, \psi, f, \tilde{\mathcal{N}}) \leq \hat{\alpha} \operatorname{osc}(u_{\mathcal{S}}, \psi, f, \mathcal{N}) , \tag{4.74}$$

  wobei $\tilde{\mathcal{N}}$ die Menge an Punkten nach Verfeinerung der Triangulierung $\mathcal{T}_h$ bzgl. der Punkte $\hat{\mathcal{N}}$ ist.

  Hierfür vllt eine äquivalente Darstellung von osc_2 bzgl. der Dreiecke und dann das Lemma nur auf osc_2 beziehen.

4.1.5 Effektivität des Fehlerschätzers

- wir zeigen, dass der hierarchische Fehlerschätzer $-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$ ist auch eine untere Schranke für $-\mathcal{I}(e) = J(u_{\mathcal{S}}) - J(u)$
-

Theorem 4.24. *Das Hindernis ψ sei stückweise linear und stetig. Dann ist der hierarchische a posteriori Fehlerschätzer $\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}})$ auch eine untere Schranke für den Fehler im Energiefunktional im Sinne von*

$$-\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}}) \leq 6(J(u_{\mathcal{S}}) - J(u))\,. \tag{4.75}$$

Beweis. Zunächst folgt mit (4.16)

$$\begin{aligned} -\mathcal{I}_{\mathcal{Q}}(\varepsilon_{\mathcal{V}}) \leq \rho_{\mathcal{S}}(\varepsilon_{\mathcal{V}}) &= \rho_{\mathcal{S}}\left(\sum_{E\in\mathcal{E}} \varepsilon_{\mathcal{V}}(x_E)\phi_E\right) \\ &= \sum_{E\in\mathcal{E}} \varepsilon_{\mathcal{V}}(x_E)\rho_{\mathcal{S}}(\phi_E) \\ &= \sum_{E\in\mathcal{E}} \eta_E\,|\rho_E| \end{aligned} \tag{4.76}$$

mit $\eta_E = |\varepsilon_{\mathcal{V}}(x_E)| \cdot \|\phi_E\|$ und $\rho_E = \frac{\rho_{\mathcal{S}}(\phi_E)}{\|\phi_E\|}$, wobei man zeigen kann, dass $\operatorname{sign}(\varepsilon_{\mathcal{V}}(x_E)) = \operatorname{sign}(\rho_{\mathcal{S}}(\phi_E))$ gilt. Weiter sollte man erwähnen, dass (4.76) äquivalent ist zu [SV07] Gleichung (2.16).

Das weitere Vorgehen ist ähnlich zum Beweis von Theorem 3.2 aus [SV07]. Es sei

$$\varphi = \frac{1}{3}\sum_{E\in\mathcal{E}} \beta_E\phi_E$$

eine Linearkombination aus Bubble-Funktionen. Dann lässt sich $u_{\mathcal{S}}+\varphi$ auf jedem $T \in \mathcal{T}_h$ durch eine Konvexkombination aus $v_E \coloneqq u_{\mathcal{S}} + \beta_E\phi_E, E \in \mathcal{E}$ schreiben, d.h.

$$(u_{\mathcal{S}} + \varphi)\Big|_T = \frac{1}{3}\sum_{E\in\mathcal{E}, E\subset T} v_E\Big|_T\,.$$

Da $\mathbb{R}^2 \ni x \mapsto \frac{1}{2}|x|^2$ konvex ist, rechnen wir mit den obigen Bezeichnungen schnell nach, dass gilt

$$\begin{aligned} J(u_{\mathcal{S}} + \varphi) &= \int_\Omega \frac{1}{2}|\nabla(u_{\mathcal{S}} + \varphi)|^2 - f(u_{\mathcal{S}} + \varphi)\,d\Omega \\ &= \sum_{T\in\mathcal{T}_h} \int_T \frac{1}{2}\left|\nabla(u_{\mathcal{S}} + \varphi)\Big|_T\right|^2 - f(u_{\mathcal{S}} + \varphi)\Big|_T\,d\Omega \\ &= \sum_{T\in\mathcal{T}_h} \int_T \frac{1}{2}\left|\left(\frac{1}{3}\sum_{E\in\mathcal{E}, E\subset T} \nabla v_E\Big|_T\right)\right|^2 - f\left(\frac{1}{3}\sum_{E\in\mathcal{E}, E\subset T} v_E\Big|_T\right)\,d\Omega \\ &\leq \frac{1}{3}\sum_{E\in\mathcal{E}, E\subset T}\sum_{T\in\mathcal{T}_h} \int_T \frac{1}{2}\left|\nabla v_E\Big|_T\right|^2 - f v_E\Big|_T\,d\Omega\,. \end{aligned}$$

Da wir drei Kanten pro Dreieck T haben, gilt analog die Gleichung

$$J(u_{\mathcal{S}}) = \frac{1}{3} \sum_{E\in\mathcal{E}, E\subset T} \sum_{T\in\mathcal{T}_h} \int_T \frac{1}{2} |\nabla u_{\mathcal{S}}|^2 - f u_{\mathcal{S}} \, d\Omega \,.$$

Durch Subtraktion der letzten beiden Terme und einigen Umformungen ergibt sich dann

$$J(u_{\mathcal{S}}) - J(u_{\mathcal{S}} + \varphi) \geq \frac{1}{3} \sum_{E\in\mathcal{E}} (J(u_{\mathcal{S}}) - J(u_{\mathcal{S}} + \beta_E \phi_E)) \,. \tag{4.77}$$

Wir rechnen nach, dass für alle $E \in \mathcal{E}$

$$\begin{aligned} J(u_{\mathcal{S}} + \beta_E\phi_E) &= \frac{1}{2} a(u_{\mathcal{S}} + \beta_E\phi_E, u_{\mathcal{S}} + \beta_E\phi_E) - (f, u_{\mathcal{S}} + \beta_E\phi_E) \\ &= J(u_{\mathcal{S}}) + \frac{1}{2} a(\beta_E\phi_E, \beta_E\phi_E) - ((f, \beta_E\phi_E) - a(u_{\mathcal{S}}, \beta_E\phi_E)) \\ &= J(u_{\mathcal{S}}) + \mathcal{I}(\beta_E\phi_E) \end{aligned}$$

gilt. Damit ist das Minimieren von $J(u_{\mathcal{S}} + \beta_E\phi_E)$, so dass $\beta_E \geq -d_E$, mit d_E wie oben definiert, äquivalent ist zum Problem:

$$\min_{\beta_E \geq -d_E} \mathcal{I}(\beta_E\phi_E) \,,$$

was den nächsten Schritt legitimiert. Wir setzen nun $\beta_E = \varepsilon\mathcal{V}(x_E)$, dann gilt, dass $u_{\mathcal{S}} + \beta_E\phi_E \in K$ ist für alle $E \in \mathcal{E}$ und damit aufgrund der Konvexität von K auch $u_{\mathcal{S}} + \varphi \in \mathcal{K}$. Damit folgt insgesamt

$$\begin{aligned} J(u_{\mathcal{S}}) - J(u) &\geq J(u_{\mathcal{S}}) - J(u_{\mathcal{S}} + \varphi) \\ &\geq \frac{1}{3} \sum_{E\in\mathcal{E}} (J(u_{\mathcal{S}}) - J(u_{\mathcal{S}} + \beta_E\phi_E)) = \frac{1}{3} \sum_{E\in\mathcal{E}} -\mathcal{I}(\beta_E\phi_E) \\ &= \frac{1}{3} \sum_{E\in\mathcal{E}} \left(\rho_{\mathcal{S}}(\beta_E\phi_E) - \frac{1}{2} a(\beta_E\phi_E, \beta_E\phi_E) \right) \\ &= \frac{1}{3} \sum_{E\in\mathcal{E}} \left(\beta_E \, \rho_{\mathcal{S}}(\phi_E) - \frac{1}{2} \beta_E^2 a(\phi_E, \phi_E) \right) \\ &\geq \frac{1}{3} \sum_{E\in\mathcal{E}} \left(\frac{\max\{-d_E, \rho_E\}}{\|\phi_E\|} \, \rho_{\mathcal{S}}(\phi_E) - \frac{1}{2} \frac{\max\{-d_E, \rho_E\}^2}{\|\phi_E\|^2} \|\phi_E\|^2 \right) \\ &= \frac{1}{3} \sum_{E\in\mathcal{E}} \left(\underbrace{\max\{-d_E, \rho_E\}\rho_E}_{=\eta_E |\rho_E|} - \frac{1}{2} \underbrace{\max\{-d_E, \rho_E\}^2}_{\geq \eta_E |\rho_E|} \right) \\ &\geq \frac{1}{3} \sum_{E\in\mathcal{E}} \frac{1}{2} \eta_E \, |\rho_E| = \frac{1}{6} \sum_{E\in\mathcal{E}} \eta_E \, |\rho_E| \,. \end{aligned}$$

Zusammen mit (4.76) folgt dann die Behauptung. □

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 ρ_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

- numerisches Beispiel aus [SV07] oder auch [BCH07]

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

[DLY89] Deuflhard, P. ; Leinen, P. ; Yserentant, H.: Concepts of an Adaptive Hierarchical Finite Element Code. In: *Impact of Computing in Science and Engineering* 1 (1989), S. 3–35

[Fal74] Falk, Richard S.: Error estimates for the approximation of a class of variational inequalities. In: *Math. Comp.* 28 (1974), S. 963–971

[Glo08] Glowinski, Roland: *Numerical methods for nonlinear variational problems.* Reprint. Springer, 2008

[GRT09] Göpfert, A. ; Riedrich, T. ; Tammer, C.: *Angewandte Funktionalanalysis.* Vieweg und Teubner, 2009

[HH80] Haslinger, J. ; Hlavácek, I.: Contact between elastic bodies. I. Continuous problems. In: *Apl. Mat.* 25 (1980), S. 324–347

[HHNL80] Hlavácek, I. ; Haslinger, J. ; Necas, J. ; Lovísek, J.: *Solution of Variational Inequalities in Mechanics*. 9. Auflage. Springer, 1980

[HK92] Hoppe, R. ; Kornhuber, R.: Adaptive Multilevel-Methods for Obstacle Problems. In: *Preprint SC 91-16* (1992), April

[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

[MNS00] Morin, P. ; Nochetto, R.H. ; Siebert, K.G.: Data Oscillation and convergence of adaptive FEM. In: *SIAM J. Numer. Anal.* 38 (2000), Nr. 2, S. 466–488

[NW06] Nocedal, Jorge ; Wright, Stephen J.: *Numerical Optimization*. 2. ed. New York, NY : Springer, 2006

[QSS02] Quarteroni, A. ; Sacco, R. ; Saleri, F.: *Numerische Mathematik 2*. 1. Auflage. Springer, 2002

[Rud91] Rudin, Walter: *Functional Analysis*. 2. Auflage. McGraw-Hill, 1991

[Sta08] Starke, Gerhard: Numerik partieller Differentialgleichungen / IFAM - Universität Hannover. 2008. – Vorlesungsskript

[Sta11] Starke, Gerhard: Variationsungleichungen / IFAM - Universität Hannover. 2011. – Vorlesungsskript

[Ste12a] Stephan, Ernst P.: Contact Problems – Numerical Analysis and Implementation / IFAM - Universität Hannover. 2012. – Vorlesungsskript

[Ste12b] Stephan, Ernst P.: Numerik partieller Differentialgleichungen I / IFAM - Universität Hannover. 2012. – Vorlesungsskript

[Sto99] Stoer, Josef: *Numerische Mathematik I.* 8. Auflage. Springer, 1999

[SV07] Siebert, K.G. ; Veeser, A.: A Unilaterally Constrained Quadratic Minimization with Adaptive Finite Elements. In: *SIAM J. Optim.* 18 (2007), Nr. 1, S. 260–289

[Wal11] Walker, Christoph: Partielle Differentialgleichungen / IFAM - Universität Hannover. 2011. – Vorlesungsskript

[Wer11] Werner, Dirk: *Funktionalanalysis.* 7. Auflage. Springer, 2011

[Wri01] Wriggers, Peter: *Nichtlineare Finite-Element-Methoden.* 5. Auflage. Springer, 2001

[Wri06] Wriggers, Peter: *Computional Contact Mechanics.* 2. Auflage. Springer, 2006

[Wri09] Wriggers, Peter: Finite-Elemente-Methode / IKM - Universität Hannover. 2009. – Vorlesungsskript

[Zha07] Zhang, Yongmin: Convergence of free boundaries in discrete obstacle problems. In: *Numerische Mathematik* (2007), Nr. 106, S. 157–164

[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] Kaptitel II, §2 und [Wal11] Kapitel 6).

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

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

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 Ω ist, d.h.

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

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

Definition A.2. Der Raum

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

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 $v(x) = \operatorname{sign}(x)$, 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 $1 \le p \le \infty$ 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*

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

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 dimensionalen Vektorräumen existiert.

Definition A.9. Es seien V ein Vektorraum, $M \subset V$ und W ein normierter Raum, sowie $F : M \to W$ 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

$$\mathcal{D}_v F(x_0) := \frac{d}{dt} F(x_0 + tv)\Big|_{t=0} := \lim_{t \to 0} \frac{F(x_0 + tv) - F(x_0)}{t} \tag{A.1}$$

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) $\lim_{t\to 0}$ durch $\lim_{t\to +0}$ 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 $F : M \to \mathbb{R}$ ein konvexes Funktional. Dann gilt für $x_0, x \in M$:*

x_0 ist Lösung von $\min\limits_{x \in M} F(x)$ 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 $F : U \to \mathbb{R}$ eine Gâteaux-differenzierbare konvexe Funktion. Dann ist $x_0 \in U$ genau dann Lösung von $\min\limits_{x \in U} F(x)$, wenn für alle $u \in U$ gilt*

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

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

A.3 Konvergenzbegriffe

Definition A.12. Es sei $m \in \mathbb{N}, 1 \leq p < \infty, 1 = \frac{1}{p} + \frac{1}{p'}$.

(a) Eine Folge (u_j) in L_p konvergiert schwach gegen $u \in L_p(\Omega)$

$$\begin{aligned} &:\Longleftrightarrow u_j \rightharpoonup u \text{ in } L_p(\Omega) \\ &:\Longleftrightarrow \forall\, v \in L_{p'}(\Omega) : \int_\Omega u_j v \,\mathrm{d}x \longrightarrow \int_\Omega uv \,\mathrm{d}x \text{ in } \mathbb{K}\,. \end{aligned}$$

(b) Eine Folge $(u_j) \in W_p^m(\Omega)$ konvergiert schwach gegen $u \in W_p^m(\Omega)$

$$:\Longleftrightarrow u_j \rightharpoonup u \text{ in } W_p^m(\Omega)$$
$$:\Longleftrightarrow \partial^\alpha u_j \rightharpoonup \partial^\alpha u \text{ in } L_p(\Omega) \,\forall\, |\alpha| \leq m\,.$$

Bemerkung A.13. Sei $1 \leq p < \infty, m \in \mathbb{N}$, dann ist:

(a) Ist $u_j \to u$ in $W_p^m(\Omega)$, dann folgt $u_j \rightharpoonup u$ in $W_p^m(\Omega)$, d.h. „starke Konvergenz ist stärker als schwache Konvergenz".

Beweis. $\forall\, v \in L_{p'}(\Omega), |\alpha| \leq m$ gilt

$$\left| \int_\Omega (\partial^\alpha u_j - \partial^\alpha u) v \,\mathrm{d}x \right| \underset{\text{Hölder}}{\leq} \|v\|_{L_{p'}(\Omega)} \|\partial^\alpha u_j - \partial^\alpha u\|_{L_p(\Omega)} \longrightarrow 0\,. \qquad \square$$

(b) Sei $1 < p < \infty, (u_j) \subset W_p^m(\Omega)$ beschränkt (bzgl. $\|\cdot\|_{W_p^m}$), dann folgt, dass eine Teilfolge $(u_{j'})$ und ein $u \in W_p^m(\Omega)$ existiert, so dass $u_{j'} \rightharpoonup u$ in $W_p^m(\Omega)$, d.h. „beschränkte Folgen sind relativ schwach kompakt".

Beweis. Vgl. [Rud91]. $\square$

(c) Es sei $M \subset W_p^m(\Omega)$ konvex und abgeschlossen (bzgl. $\|\cdot\|_{W_p^m}$), sowie $(u_j) \subset M$ mit $u_j \rightharpoonup u$ in $W_p^m(\Omega)$, dann ist $u \in M$, d.h. „abgeschlossene konvexe Mengen sind schwach abgeschlossen" (Theorem von Mazun; ohne Beweis, vgl. [Rud91]).

(d) Es sei $u_j \rightharpoonup u$ in $W_m^p(\Omega)$, dann folgt (u_j) ist beschränkt in $W_p^m(\Omega)$ (bzgl. $\|\cdot\|_{W_p^m}$), d.h. „schwach konvergente Folgen sind beschränkt".

Beweis. Theorem von Mackey, vgl. [Rud91]. $\square$

(e) $u_j \rightharpoonup u$ in $W_p^m(\Omega), u_j \rightharpoonup v$ in $W_p^m(\Omega)$, dann gilt $u = v$, d.h. „Grenzwerte von schwach konvergenten Folgen sind eindeutig".

Beweis. Aus dem Hauptsatz der Variationsrechnung folgt die Behauptung. $\square$

(f) Sei $u_j \rightharpoonup u$ in $W_p^m(\Omega)$, dann folgt $\|u\|_{W_p^m(\Omega)} \leq \liminf \|u_j\|_{W_p^m(\Omega)}$.

Theorem A.14. *In einem reflexiven Raum V, d.h. der Bidualraum V'' ist isomorph zu V, besitzt jede beschränkte Folge $(v_n)_{n\in\mathbb{N}}$ eine schwach konvergente Teilfolge (v_{n_j}).*

Beweis. Der Beweis befindet sich in [Wer11] Kapitel III, Theorem 3.7. $\square$

Bemerkung A.15. Jeder Hilbertraum H ist reflexiv.

Beweis. Dies folgt aus dem Darstellungssatz von Riesz (Satz 2.15). $\square$

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

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

wobei $\mathcal{E}$ und $\mathcal{I}$ die Indexmengen der Gleichungs- und Ungleichungsnebenbedingungen darstellen und $\boldsymbol{c}, \boldsymbol{x}, \boldsymbol{a}_i \in \mathbb{R}^n, b_i \in \mathbb{R}, i \in \mathcal{E} \cup \mathcal{I}$, 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 $\boldsymbol{x}^*$, die die Bedingungen 1. Ordnung erfüllt, auch globale Lösung des Problems ist (s. Theorem **??**). 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

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

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

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

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

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

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

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

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

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

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

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 $\boldsymbol{x}_k$ 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 $\boldsymbol{x}_k$ im k-ten Schritt heißt *working set* und kann wie folgt beschrieben werden

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

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 x_k nicht erfüllt wird.

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

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

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

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

Zusammengefasst müssen wir also im k-ten Schritt das Teilproblem

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

lösen. Die Lösung im k-ten Schritt von (B.5) bezeichnen wir mit $\boldsymbol{p}_k$. Umgekehrt gilt damit, analog zur obigen Rechnung, natürlich auch, dass für alle $i \in \mathcal{W}_k$ die Restriktion aktiv bleibt für $\boldsymbol{x}_k + \alpha \boldsymbol{p}_k$ mit beliebigem α. 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 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{p}_k$ 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 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{p}_k$, ansonsten suchen wir das größtmögliche $\alpha_k \in [0,1]$, so dass

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

zulässig bleibt. Hierfür betrachten wir zwei Fälle.

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

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

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

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

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

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

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

gilt. Eine Restriktion $i \notin \mathcal{W}_k$, für die das Minimum für α_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 $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{p}_k$ 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

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

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

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

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

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

d.h. die blocking constraint ist für die neue Iterierte $\boldsymbol{x}_{k+1}$ 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 $\hat{\boldsymbol{x}}$ gefunden haben. Dies ist leicht zu erkennen, da wir (B.1) auf $\mathcal{W}_k$ nicht weiter minimieren können, sobald es keinen Schritt p gibt, in dessen Richtung wir q verringern können, d.h. wenn $\boldsymbol{p} = \boldsymbol{0}$ die Lösung für das Teilproblem (B.5) ist. Dann ist der optimale Punkt $\hat{\boldsymbol{x}}$ bzgl. des working sets $\hat{\mathcal{W}} \subset \mathcal{A}(\hat{\boldsymbol{x}})$ gefunden.

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

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

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

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

wobei A die Gradienten $\boldsymbol{a}_i^T$ der aktiven Restriktionen $\hat{\mathcal{W}}$ zeilenweise enthält. Damit werden die ersten beiden KKT-Bedingungen aus (B.3) erfüllt. Da die Schrittlänge α_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 $\boldsymbol{x}^* = \hat{\boldsymbol{x}}$. 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 wegfallen lassen (vlg. [NW06], Kapitel 12.3). Dies zeigt das folgende Theorem.

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

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

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

Dann ist p eine zulässige Richtung für die Nebenbedingung j, d.h. $\boldsymbol{a}_j^T \boldsymbol{p} \geq 0$. Weiterhin gilt sogar $\boldsymbol{a}_j^T \boldsymbol{p} > 0$ und p ist eine Abstiegsrichtung für q, wenn $\boldsymbol{p}$ 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 λ_i proportional zu $|\lambda_i|$ ist, eliminieren wir gerade die Restriktion mit kleinstem Langrange-Multiplikator. Es kann allerdings sein, dass der folgende zu berechnende Schritt $\boldsymbol{p}$ 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 $\boldsymbol{x}_0$ für (B.1) und definiere $\mathcal{W}_0$ z.B. mit allen aktiven Restriktionen bzgl. $\boldsymbol{x}_0$.

for k = 0,1,2,... **do**
  Löse (B.5) zur Berechnung von $\boldsymbol{p}_k$;
  if $\boldsymbol{p}_k = \mathbf{0}$ **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 $\boldsymbol{x}^* = \hat{\boldsymbol{x}}$;
    else
      $j \leftarrow \arg\min_{j \in \mathcal{W}_k \cap \mathcal{I}} \hat{\lambda}_j$;
      $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k, \mathcal{W}_{k+1} \leftarrow W_k \setminus \{j\}$;
    end if
  else ($\boldsymbol{p}_k \neq \mathbf{0}$)
    Berechne α_k mit (B.7);
    $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + \alpha_k \boldsymbol{p}_k$;
    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

Tensorrechnung

hier auch ein paar Integralsätzer???

Anhang D

Quellcode

D.1 Implementierung des Fehlerschätzers für das Hindernisproblem

Index