Vorlesungsmitschrift

Numerik III (FEM)

Vorlesung von F. T. Suttmeier

Universität Siegen WS 2005/2006

Frank Gimbel

23. Februar 2006

Literatur

[1] D. Braess, Finite Elemente, Springer 1997

[2] R. Glowinski, Numerical methods for nonlinear variational problems, Springer Series in Comp. Physics, Springer 1983

[3] C. Johnson, Numerical solution of partial differential equations by the finite element method, Studentlitteratur 1987

[4] D. Kinderlehrer and G. Stampacchia, An introduction to variational inequalities and their applications, Acadeic Press 1980

[5] C. Großmann und H.-G. Roos, Numerik partieller Differentialgleichungen, Teubner Studienbücher 1992

[6] H. R. Schwarz, Methode der finiten Elemente, Teubner 1984

[7] R. Seydel, Tools for Computational Finance, Springer 2002

Inhaltsverzeichnis

I Motivation

II Einleitung zur Finite Elemente-Methode (FEM)

II.1 Modellbeispiel

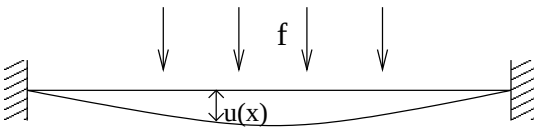

Abbildung II.1: Modellbeispiel: elastischer Draht

Beispiel: Verformung eines elastischen Drahtes im Kraftfeld f (Abbildung II.1)
Die Auslenkung des Drahtes werde durch $u(x)$ beschrieben. Aus der Physik weiß man, daß die elastische Energie proportional zur Längenänderung des Drahtes ist, also

$$\mathcal{U}_E \propto \Delta l = \int_0^1 \sqrt{1+(\partial_x u)^2}dx - \int_0^1 1dx$$
$$\approx \frac{1}{2}\int_0^1 (\partial_x u(x))^2 dx$$

Durch Einwirkung der Kraft f besitzt der Draht eine potentielle Energie

$$\mathcal{U}_f = -\int_0^1 f(x)u(x)dx$$

Die stabile Gleichgewichtslage ist dadurch charakterisiert, daß die Gesamtenergie

$$\mathcal{U}(u) = \mathcal{U}_E(u) + \mathcal{U}_f(u)$$

minimal wird. Gesucht ist also eine Funktion $u \in V$, so daß

$$\mathcal{U}(u) \leq \mathcal{U}(v) \qquad \forall \quad v \in V \qquad \text{Ⓜ}$$

V ist der Raum der Vergleichsfunktionen. V ist festgelegt durch

1. Alle stetigen Funktionen mit Nullrandwerten
2. stückweise stetige, beschränkte 1. Ableitungen („Integrale müssen Sinn ergeben“)

II.2 Klassische und variationelle Formulierung

Variationsrechnung: Wähle $\varphi \in V$ beliebig aber fest.
Betrachte $v = u + \varepsilon\varphi$ für $\varepsilon \in \mathbb{R}$.
Aufgrund der Minimaleigenschaft

$$\mathcal{U}(u) \leq \mathcal{U}(v) = \mathcal{U}(u+\varepsilon\varphi) \qquad \forall \quad \varepsilon$$

folgt die notwendige Bedingung

$$\begin{aligned}
& \frac{d}{d\varepsilon}\mathcal{U}(u+\varepsilon\varphi)\Big|_{\varepsilon=0} = 0 \\
\Leftrightarrow \quad & \frac{d}{d\varepsilon}\left[\frac{1}{2}\int_0^1 (\partial_x u + \varepsilon\partial_x\varphi)^2 dx - \int_0^1 f\cdot(u+\varepsilon\varphi)dx\right]_{\varepsilon=0} = 0 \\
\Leftrightarrow \quad & \left[\frac{1}{2}\int_0^1 2(\partial_x u + \varepsilon\partial_x\varphi)\cdot\partial_x\varphi dx - \int_0^1 f\cdot\varphi dx\right]_{\varepsilon=0} = 0 \\
\Leftrightarrow \quad & \int_0^1 \partial_x u\partial_x\varphi dx - \int_0^1 f\varphi dx = 0 \\
\Leftrightarrow \quad & \int_0^1 \partial_x u\partial_x\varphi dx = \int_0^1 f\varphi dx \qquad \forall \quad \varphi\in V \qquad \text{(V)}
\end{aligned}$$

(Variationelle Formulierung)

Bemerkung II.1 *: $\mathcal{U}(.)$ ist für das Modellbeispiel konvex. $\curvearrowright$ (V) ist auch hinreichend. Obige Rechnung zeigt*

Satz II.1

$$\text{(M)} \quad \Rightarrow \quad \text{(V)}$$

Umgekehrt gilt

Satz II.2

$$\text{(V)} \quad \Rightarrow \quad \text{(M)}$$

Beweis: Schreibweise: $(v,w) := \int_0^1 v(x)w(x)dx$ für stückweise stetige, beschränkte Funktionen

1. *u sei Lösung von (V)*
2. *wähle $v\in V$ und setze $w = v-u$*
3. *$v = w+u \in V$ und ebenfalls $w\in V$*

Rechnung:

$$\begin{aligned}
\mathcal{U}(v) = \mathcal{U}(u+w) &= \frac{1}{2}(\partial_x u + \partial_x w, \partial_x u + \partial_x w) - (f, u+w) \\
&= \frac{1}{2}(\partial_x u, \partial_x u) - (f,u) + \underbrace{(\partial_x u, \partial_x w) - (f,w)}_{=0,\ \text{(V)}} + \frac{1}{2}\underbrace{(\partial_x w, \partial_x w)}_{\geq 0} \\
&\geq \frac{1}{2}(\partial_x u, \partial_x u) - (f,u) = \mathcal{U}(u) \\
\Rightarrow \quad \mathcal{U}(u) &\leq \mathcal{U}(v) \quad \forall \quad v\in V
\end{aligned}$$

■

Satz II.3 (Klassische Formulierung) *Sei u Lösung von Ⓥ. Zusätzlich existiere $\partial_x^2 u$ und $\partial_x^2 u$ sei stetig. Dann gilt*

$$-\partial_x^2 u(x) = f(x), \quad u(0) = u(1) = 0 \qquad \text{Ⓓ}$$

(für Dirichlet-Problem)

Beweis: $\int_0^1 \partial_x u \cdot \partial_x v dx - \int_0^1 f v dx = 0 \quad \forall \quad v \in V$

Partielle Integration liefert:

$$\underbrace{[\partial_x u \cdot v]_0^1}_{=0 \quad (v(0)=v(1)=0)} - \int_0^1 \partial_x^2 u \cdot v dx - \int_0^1 f \cdot v dx = 0$$

$$\curvearrowright \quad \int_0^1 (-\partial_x^2 u - f) \cdot v dx = 0 \quad \forall \quad v \in V$$

$$\Rightarrow \quad -\partial_x^2 u - f = 0 \ \textit{auf}\ (0,1)$$

■

Analog dazu:

Satz II.4 Ⓓ $\Rightarrow$ Ⓥ

Zusammengefasst: $\boxed{\text{Ⓓ} \quad \Rightarrow \quad \text{Ⓥ} \Leftrightarrow \text{Ⓜ}}$

II.3 Näherungsverfahren, Ritz-Galerkin-Verfahren

Idee: Approximation des Raumes V durch einen endlichdimensionalen Teilraum V^N mit $\dim V^N = N$.

Betrachte die schwache Formulierung von Ⓥ:

$$\int_0^1 \partial_x u^N \partial_x \varphi^N dx = \int_0^1 f \varphi^N dx \quad \forall \quad \varphi^N \in V^N$$

Dadurch ist die diskrete Lösung $u^N \in V^N$ charakterisiert.

Schreibweise: $a(v,w) = (\partial_x v, \partial_x w)$

Also kompakt: $\boxed{a(u^N, \varphi^N) = (f, \varphi^N) \quad \forall \varphi^N \in V^N}$

Wähle Basis für $V^N = \langle \varphi_1, \ldots, \varphi_N \rangle$

Darstellung für $\varphi \in V^N$: $\quad \varphi = \sum_{j=1}^N v_j \varphi_j, \quad v_j \in \mathbb{R}$

Es genügt $\boxed{a(u^N, \varphi_i) = (f, \varphi_i) \quad \forall \quad i = 1, \ldots, N}$
zu erfüllen.

Rechnung: $a(u^N, \sum_{i=1}^N v_i \varphi_i) - (f, \sum_{i=1}^N v_i \varphi_i) = \sum_{i=1}^N v_i \underbrace{\big(a(u^N, \varphi_i) - (f, \varphi_i)\big)}_{=0} = 0$

Frage: Wie berechnet man u^N ?

Ansatz für u^N: $u^N = \sum_{j=1}^{N} u_j \varphi_j, \quad u_j \in \mathbb{R}$ *(u_j ist noch zu bestimmen!)*

Einsetzen liefert:

$$a\left(\sum_{j=1}^{N} u_j \varphi_j, \varphi_i\right) = (f, \varphi_i), \quad i = 1, \ldots, N$$

$$\Leftrightarrow \quad \sum_{j=1}^{N} u_j a(\varphi_j, \varphi_i) = (f, \varphi_i), \quad i = 1, \ldots, N$$

$$\Leftrightarrow \quad \sum_{j=1}^{N} u_j \left(\int_0^1 \partial_x \varphi_j \partial_x \varphi_i dx\right) = \int_0^1 f \varphi_i dx, \quad i = 1, \ldots, N$$

Kompakt: $Ax = b$ *mit* $x^T = (u_1, \ldots, u_N)$, $b^T = \left(\ldots, \int_0^1 f\varphi_i dx, \ldots\right)$
und $A \in \mathbb{R}^{N \times N}$: $\quad A_{ij} = \int_0^1 \partial_x \varphi_j \partial_x \varphi_i dx$
Erste Fehlerabschätzung, Galerkin-Eigenschaft

$$(\partial_x u, \partial_x \varphi) = (f, \varphi) \quad \forall \quad \varphi \in V$$

$$(\partial_x u^N, \partial_x \varphi^N) = (f, \varphi^N) \quad \forall \quad \varphi^N \in V^N \overset{!}{\subset} V$$

(Differenz beider Gleichungen für $\varphi^N \in V^N$*)*

$$(\partial_x u - \partial_x u^N, \partial_x \varphi^N) = 0 \quad \forall \quad \varphi^N \in V^N \quad \text{„Galerkin-Orthogonalität“}$$

Abschätzung des Fehlers:

$$\begin{aligned} \left\|\partial_x u - \partial_x u^N\right\|^2 &= (\partial_x u - \partial_x u^N, \partial_x u - \partial_x u^N) \\ &= (\partial_x u - \partial_x u^N, \partial_x u - \partial_x \varphi^N + \partial_x \varphi^N - \partial_x u^N) \qquad , \varphi^N \in V^N \\ &= (\partial_x u - \partial_x u^N, \partial_x u - \partial_x \varphi^N) + \underbrace{(\partial_x u - \partial_x u^N, \partial_x \varphi^N - \partial_x u^N)}_{=0, wg.\ Galerkin\text{-}Orthogonalität} \end{aligned}$$

$$\Rightarrow \quad \|\partial_x u - \partial_x u^N\|^2 \leq \|\partial_x u - \partial_x u^N\| \cdot \|\partial_x u - \partial_x \varphi^N\|$$

$$\Rightarrow \quad \|\partial_x u - \partial_x u^N\| \leq \|\partial_x u - \partial_x \varphi^N\| \quad \forall \quad \varphi^N \in V^N$$

II.4 Einfache finite Elemente

allgemeine Überlegung:

- *Lösung u sollte gut approximierbar sein*
- *Leichte Berechnung der Einträge in der Matrix* A_{ij}
- *A sollte für die Numerik gute Eigenschaften haben („dünn besetzt“ , moderate Kondition)*

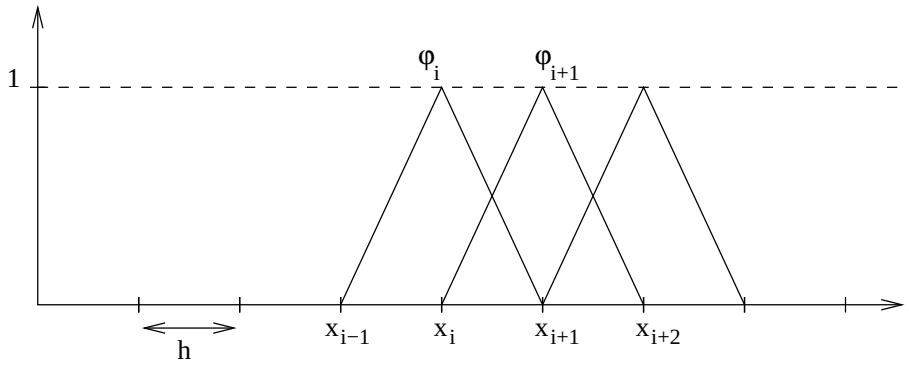

Abbildung II.2: Basisfunktionen bei linearen finiten Elementen

Lineare finite Elemente

Idee: V^N *besteht aus stückweise linearen Funktionen, die global stetig sind. Die Basisfunktionen* φ_i *(Abbildung II.2) sind definiert durch*

$$\varphi_i(x_i) = 1, \quad \varphi_i(x_{i+1}) = 0, \quad \varphi_i \text{ linear auf } (x_i, x_{i+1})$$

(auf (x_{i-1}, x_i) *analog)*
Damit ist $\partial_x \varphi_i = \pm \frac{1}{h}$ *und*

$$\begin{aligned}
\int_0^1 \partial_x\varphi_i \partial_x\varphi_i dx &= \int_{x_{i-1}}^{x_{i+1}} \partial_x\varphi_i \partial_x\varphi_i dx = 2 \int_{x_i}^{x_{i+1}} \partial_x\varphi_i \partial_x\varphi_i dx \\
&= 2\frac{1}{h^2} \cdot h = \frac{2}{h} = A_{ii} \\
\int_0^1 \partial_x\varphi_i \partial_x\varphi_{i+1} dx &= \int_{x_i}^{x_{i+1}} \partial_x\varphi_i \partial_x\varphi_{i+1} dx = -\frac{1}{h^2} h \\
&= -\frac{1}{h} = A_{i,i+1} = A_{i,i-1}
\end{aligned}$$

$$A = \frac{1}{h} \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}$$

Rechte Seite: $b_i = \int_0^1 f\varphi_i dx$

Falls f *konstant:* $b_i = f \int_{x_{i-1}}^{x_{i+1}} \varphi_i dx = hf$

Interpolationsfehler

Motivation: $\|\partial_x u - \partial_x u_h\| \leq \inf_{\varphi_h \in V_h} \|\partial_x u - \partial_x \varphi_h\| \leq \|\partial_x u - \partial_x I_h u\|$
Dabei bezeichnet $I_h u$ die Interpolierende von u.

Satz II.5 *Auf einem Teilintervall T, $T = (a_1, a_2)$, $h_T = a_2 - a_1$ der Zerlegung des Rechengebietes $(0,1)$ gilt:*

Teil 1: $\|v - I_h v\|_{L_\infty(T)} \leq c h_T^2 \|\partial_x^2 v\|_{L_\infty(T)}$
Teil 2: $\|\partial_x (v - I_h v)\|_{L_\infty(T)} \leq c h_T \|\partial_x^2 v\|_{L_\infty(T)}$

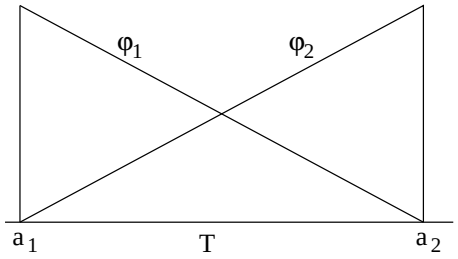

Abbildung II.3: zu Satz II.5

Beweis:

1. *Vorbereitende Bemerkungen: Die „Hutfunktionen“ φ_1, φ_2 bestimmen eine Basis für den Raum der linearen Funktionen auf T $P_1(T)$. Allgemein gilt für eine Funktion* $w \in P_1(T)$: $w(x) = \sum_{i=1}^{2} w(a_i)\varphi_i(x), \quad x \in T,$

   also $I_h v(x) = \sum_{i=1}^{2} v(a_i) \cdot \varphi_i(x), \quad x \in T$ *(I)*

   Betrachte die Taylorentwicklung von v um x in a_i:
   $$v(a_i) = v(x) + \partial_x v(x)(a_i - x) + \tfrac{1}{2}\partial_x^2 v(\xi_i)(a_i - x)^2 \qquad \text{(II)}$$
   Einsetzen von (II) in (I): $I_h v(x) = \sum_{i=1}^{2} \left(v(x) + \partial_x v(x)(a_i - x) + \tfrac{1}{2}\partial_x^2 v(\xi_i)(a_i - x)^2 \right) \varphi_i(x)$

   Umsortieren:
   $I_h v(x) = v(x) \sum_{i=1}^{2} \varphi_i(x) + \sum_{i=1}^{2} \partial_x v(x)(a_i - x)\varphi_i(x) + \sum_{i=1}^{2} \tfrac{1}{2}\partial_x^2 v(\xi_i)(a_i - x)^2 \varphi_i(x)$ *(III)*

   Wir zeigen später:

   - $\sum_{i=1}^{2} \varphi_i(x) = 1$
   - $\sum_{i=1}^{2} \partial_x v(x)(a_i - x)\varphi_i(x) = 0$

   Was bleibt zunächst?

$$|I_h v(x) - v(x)| = \left| \sum_{i=1}^{2} \frac{1}{2} \partial_x^2 v(\xi_i)(a_i - x)^2 \varphi_i(x) \right|$$

Wegen $\varphi_i(x) \leq 1$ *und* $|a_i - x| \leq h_T$ *gilt:*

$$|I_h v(x) - v(x)| \leq \max_{\xi} \left|\partial_x^2 v(\xi)\right| \cdot h_T^2$$

Bleibt zu zeigen: $\sum_{i=1}^{2} \varphi_i(x) = 1$

Rechnung: Betrachte $v(x) \equiv 1 \quad \Rightarrow \quad \partial_x v(x) = \partial_x^2 v(x) = 0 \quad \Rightarrow \quad I_h v = 1$

Einsetzen in (III): $I_h v = 1 = 1 \cdot \sum_{i=1}^{2} \varphi_i(x)$

Bleibt zu zeigen: $\sum_{i=1}^{2} \partial_x v(x)(a_i - x)\varphi_i(x) = 0$

Rechnung: Sei v *gegeben, für festes* x*:* $d = \partial_x v(x)$

Ansatz:

$$\begin{aligned} w(x') &= d \cdot x' \\ \curvearrowright \quad I_h w &= w \\ \curvearrowright \quad \partial_x w &= d \\ \curvearrowright \quad \partial_x^2 w &= 0 \end{aligned}$$

Einsetzen in (III): $I_h w = w \cdot 1 + \sum_{i=1}^{2} d \cdot (a_i - x)\varphi_i(x) + 0$

$$\curvearrowright \quad 0 = \sum_{i=1}^{2} d \cdot (a_i - x)\varphi_i(x) = \sum_{i=1}^{2} \partial_x v(x)(a_i - x)\varphi_i(x)$$

2. *zu zeigen:* $\|\partial_x(v - I_h v)\|_{L_\infty(T)} \leq c h_T \|\partial_x^2 v\|_{L_\infty(T)}$

Betrachte: $\partial_x(I_h v(x)) = \sum_{i=1}^{2} v(a_i)\partial_x \varphi(x)$

Einsetzen der Taylorentwicklung für $v(a_i)$*:*

$$\partial_x I_h v(x) = v(x) \sum_{i=1}^{2} \partial_x \varphi_i(x) + \sum_{i=1}^{2} \partial_x v(x)(a_i - x)\partial_x \varphi_i(x) + \sum_{i=1}^{2} \frac{1}{2} \partial_x^2 v(\xi_i)(a_i - x)^2 \partial_x \varphi_i(x)$$

Nun gilt $\partial_x \varphi_1 = -\frac{1}{h_T}, \quad \partial_x \varphi_2 = \frac{1}{h_T}$

Damit: $\sum_{i=1}^{2} \partial_x \varphi_i(x) = 0$

Weiterhin:

$$\begin{aligned} \sum_{i=1}^{2} \partial_x v(x)(a_i - x)\partial_x \varphi_i(x) &= \partial_x v(x)(a_1 - x)\left(-\frac{1}{h_T}\right) + \partial_x v(x)(a_2 - x)\frac{1}{h_T} \\ &= \partial_x v(x)\frac{a_2 - a_1}{h_T} = \partial_x v(x) \end{aligned}$$

Es bleibt also

$$\begin{aligned} |\partial x I_h v(x) - \partial_x v(x)| &= \left| \sum_{i=1}^{2} \frac{1}{2} \partial_x^2 v(\xi_i)(a_i - x)^2 \partial_x \varphi_i(x) \right| \\ &\leq \max_{\xi \in T} \left|\partial_x^2 v(\xi)\right| h_T^2 \cdot \frac{1}{h_T} \end{aligned}$$

■

Satz II.6 (Interpolationsfehler auf (0,1)) *Auf* $(0,1)$ *gilt bei gegebener Zerlegung in Teilintervalle* $T \subset (0,1)$ *mit maximaler Größe* h

$$\left\|\partial_x^i (v - I_h v)\right\| \leq c \cdot h^{2-i} \left\|\partial_x^2 v\right\|_{L_\infty((0,1))} \qquad \text{für } i = 0, 1$$

Beweis:

$$\begin{aligned}
\left\|\partial_x^i (v - I_h v)\right\|^2 &= \sum_T \int_T \left(\partial_x^i (v - I_h v)\right)^2 dx \\
&\leq \sum_T \int_T c h_T^{2\cdot(2-i)} \left(\left\|\partial_x^2 v\right\|_{L_\infty(T)}\right)^2 dx \qquad \textit{(Benutze Satz II.5)} \\
&\leq \sum_T c h_T^{2(2-i)} \left(\left\|\partial_x^2 v\right\|_{L_\infty(T)}\right)^2 \cdot \int_T 1 dx \\
&\leq c h^{2(2-i)} \left\|\partial_x^2 v\right\|_{L_\infty((0,1))}{}^2 \cdot 1
\end{aligned}$$

■

Satz II.7 (Energiefehler) *Auf* $(0,1)$ *gilt bei gegebener Zerlegung in Teilintervalle* T *mit maximaler Größe* h*:*

$$\left\|\partial_x (u - u_h)\right\| \leq c \cdot h \cdot \left\|\partial_x^2 u\right\|_{L_\infty((0,1))}$$

Beweis:

$$\begin{aligned}
\left\|\partial_x (u - u_h)\right\| &\leq \inf_{\varphi \in V_h} \left\|\partial_x (u - \varphi)\right\| \\
&\leq \left\|\partial_x (u - I_h u)\right\| \\
&\leq c h \left\|\partial_x^2 u\right\|_{L_\infty((0,1))}
\end{aligned}$$

■

II.5 Variationsungleichungen

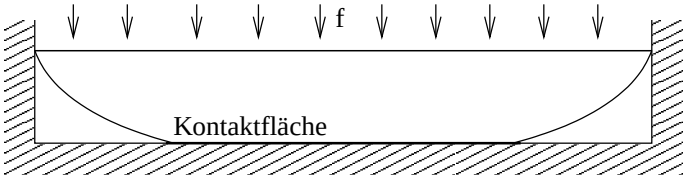

Abbildung II.4: freies Randwertproblem/ Hindernisproblem

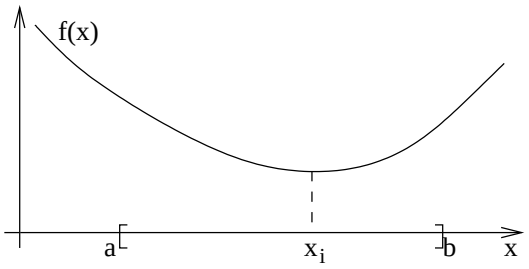

Abbildung II.5: zur eindimensionalen Minimumsuche

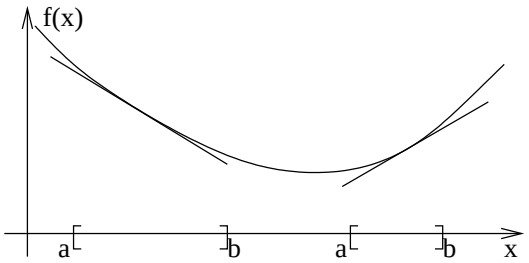

Abbildung II.6: zur Fallunterscheidung

II.5.1 Minimumsuche (eindimensional) unter Invervallrestriktion

f sei stetig differenzierbar.
Fallunterscheidung (auch für Minima am Rand, Abbildung II.6)

- $f(a) \leq f(x) \quad \forall \quad x \in [a,b] \quad \curvearrowright \quad f'(a) \geq 0$
- $f(b) \leq f(x) \quad \forall \quad x \in [a,b] \quad \curvearrowright \quad f'(b) \leq 0$
- $f(x_i) \leq f(x) \quad \forall \quad x \in [a,b] \quad \curvearrowright \quad f'(x_i) = 0$

kompakte Schreibweise: $f'(x_0) \cdot (x - x_0) \geq 0 \quad \forall x$
Das ist notwendige Bedingung für eine Minimalstelle x_0 auf einem Intervall.

II.5.2 Minimierung auf konvexer Menge $K \subset \mathbb{R}^n$

für $f : K \to \mathbb{R}$
gesucht: x_0 *mit* $f(x_0) \leq f(x) \quad \forall \quad x \in K$
Betrachte $F(\varepsilon) = f(x_0 + \varepsilon(x - x_0))$ *(Abbildung II.7). Aus dem eindimensionalen Fall ist bekannt:*

$$\begin{aligned} & \frac{d}{d\varepsilon} F(\varepsilon)\Big|_{\varepsilon=0} \cdot \varepsilon \geq 0 \quad \forall \quad \varepsilon \geq 0 \\ \Rightarrow \quad & \nabla f(x_0)(x - x_0) \cdot \varepsilon \geq 0 \quad \forall \quad \varepsilon \geq 0 \\ \Rightarrow \quad & \nabla f(x_0)(x - x_0) \geq 0 \end{aligned}$$

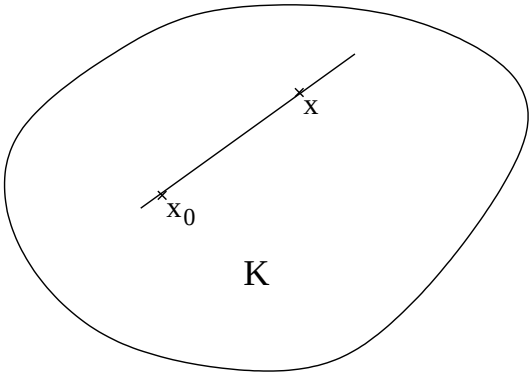

Abbildung II.7: Zur Minimierung auf einer konvexen Menge

II.5.3 Minimierung auf $K \subset V$

$K = \{v \in V | v \geq g\}$
Bemerkung: K ist konvex.
Rechnung: $v_1, v_2 \in K$, $\alpha \in (0,1)$

$$\alpha v_1 + (1-\alpha)v_2 \geq \alpha g + (1-\alpha)g = (1+\alpha-\alpha)g = g$$

Welches Funktional wird minimiert?

$$\mathcal{U}(v) = \frac{1}{2}\int_I (\partial_x v)^2 dx - \int_I f \cdot v dx$$

Betrachte: $F(\varepsilon) = \mathcal{U}\left(u + \varepsilon(v-u)\right)$
$\curvearrowright \frac{d}{d\varepsilon}F(\varepsilon)\big|_{\varepsilon=0} \cdot \varepsilon \geq 0$
Ausrechnen liefert

$$\left\{\int_I \partial_x(u+\varepsilon(v-u))\partial_x(v-u)dx - \int_I f\cdot(v-u)dx\right\}\Bigg|_{\varepsilon=0} \cdot \varepsilon \geq 0 \quad \forall \quad \varepsilon \geq 0$$

$$\Rightarrow \quad \left\{\int_I \partial_x u \partial_x(v-u)dx - \int_I f\cdot(v-u)dx\right\}\cdot\varepsilon \geq 0$$

Zusamengefasst:

$$\boxed{(\partial_x u, \partial_x(v-u)) \geq (f, v-u) \quad \forall \quad v \in K}$$

Bemerkung: $(\partial_x u, \partial_x(v-u)) \geq (f, v-u)$ ist der Prototyp einer sogenannten elliptischen Variationsungleichung.
Diskretisierung von II.5.3 mit linearen finiten Elementen

$I \quad a(u, \varphi - u) \geq (f, \varphi - u) \quad \forall \quad \varphi \in K$

$II \quad a\left(u_h, \varphi_h - u_h\right) \geq (f, \varphi_h - u_h) \quad \forall \quad \varphi_h \in K_h = V_h \cap K$

(insbesondere $K_h \subset K$)

Satz II.8 (Energiefehler) *Voraussetung wie in Satz II.7. Dann gilt*

$$\|\partial_x (u - u_h)\| \leq \mathcal{O}(h)$$

Beweis: Ausgangspunkt:

$$\begin{aligned} a(u-u_h, u-u_h) &= a(u-u_h, u-u_i+u_i-u_h) \qquad ; u_i = I_h u \\ &= a(u-u_h, u-u_i) + a(u-u_h, u_i-u_h) \qquad (*) \end{aligned}$$

Bei Variationsgleichungen war der zweite Term Null.
Hier: $a(u-u_h, u_i-u_h) = \underbrace{(f, u_i-u_h) - a(u_h, u_i-u_h)}_{\text{Term 1}} + a(u, u_i-u) - (f, u_i-u) +$

$$+ \underbrace{a(u, u-u_h) - (f, u-u_h)}_{\text{Term 2}}$$

Term 1 ≤ 0, *wegen Test mit* $\varphi_h = u_i$ *in II.*
Term 2 ≤ 0, *wegen Test mit* $\varphi = u_h$ *in I.*
Nun weiter in (*):

$$\begin{aligned} \ldots &\leq \|\partial_x(u-u_h)\| \cdot \|\partial_x(u-u_i)\| + a(u, u_i-u) - (f, u_i-u) \\ &\leq \frac{1}{2}\|\partial_x(u-u_h)\|^2 + \frac{1}{2}\|\partial_x(u-u_i)\|^2 - \int_I (\partial_x^2 u)(u_i-u)dx - (f, u_i-u) \\ &\leq \frac{1}{2}\|\partial_x(u-u_h)\|^2 + \frac{1}{2}\underbrace{\|\partial_x(u-u_i)\|^2}_{\mathcal{O}(h^2)} + \|\partial_x^2 u\| \underbrace{\|u-u_i\|}_{\mathcal{O}(h^2)} + \|f\| \underbrace{\|u_i-u\|}_{\mathcal{O}(h^2)} \end{aligned}$$

$$\Rightarrow \quad \frac{1}{2}\|\partial_x(u-u_h)\|^2 \leq \mathcal{O}(h^2) + \left(\|\partial_x^2 u\| + \|f\|\right)\mathcal{O}(h^2) = \mathcal{O}(h^2)$$

Jetzt Wurzel ziehen. ■

II.6 Aposteriori-Fehlerschätzer

Bisher: $\|\partial_x(u-u_h)\| \leq ch \cdot \|\partial_x^2 u\|_{L_\infty(I)}$ *mit unbekannter Lösung u.*
Ziel*:* $\|\partial_x(u-u_h)\| \leq \eta(u_h, f)$ *wobei* u_h, f ***bekannt***

Vorbereitungen:

Satz II.9 *Auf einem Intervall* $T = (x_i, x_{i+1})$ *gilt für* $v \in V$ *und* $v(x_i) = 0$

$$\|v\|_{L^2(T)} \leq h \|\partial_x v\|_{L^2(T)}$$

Beweis: Wähle $y \in (x_i, x_{i+1}]$

$$\begin{aligned} v(y) &= \int_{x_i}^{y} \partial_x v(x) dx \\ &\leq \left(\int_{x_i}^{y} 1^2 dx \right)^{1/2} \left(\int_{x_i}^{y} (\partial_x v)^2 dx \right)^{1/2} \qquad \textit{(Höldersche Ungleichung)} \\ &\leq \sqrt{h} \left\| \partial_x v \right\|_{L^2(T)} \\ \Rightarrow \qquad (v(y))^2 &\leq h \left\| \partial_x v \right\|^2_{L^2(T)} \\ \Rightarrow \qquad \int_T (v(y)^2) dy &\leq \int_T h \left\| \partial_x v \right\|^2_{L^2(T)} dy \\ &\leq h^2 \left\| \partial_x v \right\|^2_{L^2(T)} \end{aligned}$$

Wurzelziehen $\curvearrowright$ ■

Satz II.10 (Stabilität der Interpolation) *Auf* $I \subset \mathbb{R}$ *gilt bei gegebener Zerlegung in Zellen* T*:*

$$\left\| \partial_x (I_h v) \right\|_{L^2(T)} \leq \left\| \partial_x v \right\|_{L^2(T)}$$

Beweis: $y \in (x_i, x_{i+1})$

$$\begin{aligned} \partial_y I_h(y) &= \frac{v(x_{i+1}) - v(x_i)}{h} \\ &= \frac{1}{h} \int_{x_i}^{x_{i+1}} \partial_x v(x) dx \\ &\leq \frac{1}{h} \left(\int_{x_i}^{x_{i+1}} 1^2 dx \right)^{1/2} \left(\int_{x_i}^{x_{i+1}} (\partial_x v)^2 dx \right)^{1/2} \qquad \textit{Höldersche Ungleichung} \\ &= \frac{1}{\sqrt{h}} \left\| \partial_x v \right\|_{L^2(T)} \end{aligned}$$

Quadrieren, Integrieren:

$$\begin{aligned} \int_T \left(\partial_x I_h v(y) \right)^2 dy &\leq \int_T \frac{1}{h} \left\| \partial_x v \right\|^2_{L^2(T)} dy \\ \curvearrowright \quad \left\| \partial_x I_h v \right\|_{L^2(T)} &\leq \left\| \partial_x v \right\|_{L^2(T)} \end{aligned}$$

■

Satz II.11 (Energiefehlerschätzer) *Auf dem Intervall I mit einer Zerlegung in Zellen T gilt:*

$$\|\partial_x(u-u_h)\|_{L^2(I)} \le c\left(\sum_T h_T^2\rho_T^2\right)^{1/2} \qquad \text{mit } \rho_T = 2\left\|f+\partial_x^2 u_h\right\|_{L^2(T)}$$

(ρ_T heißt Residuum: $\partial_x^2 u + f = 0$)

Beweis: Schreibweise: $e = u - u_h$, $e_i = I_h e$

$$\begin{aligned}
\|\partial_x(u-u_h)\|^2 &= (\partial_x u - \partial_x u_h, \partial_x e - \partial_x e_i) \qquad \textit{(Galerkin-Eigenschaft)}\\
&= (f, e-e_i) - (\partial_x u_h, \partial_x e - \partial_x e_i) \qquad ((\partial_x u, \partial_x\varphi) = (f,\varphi))\\
&= (f, e-e_i) - \sum_T (\partial_x u_h, \partial_x e - \partial_x e_i)_T\\
&= (f, e-e_i) - \sum_T \left\{ \left(-\partial_x^2 u_h, e-e_i\right)_T + \underbrace{[\partial_x u_h\cdot(e-e_i)]_{x_i}^{x_{i+1}}}_{=0} \right\}\\
&= \sum_T \left(f+\partial_x^2 u_h, e-e_i\right)_T
\end{aligned}$$

Betrachte:

$$\begin{aligned}
\left(f+\partial_x^2 u_h, e-e_i\right)_T &\le \left\|f+\partial_x^2 u_h\right\|_{L^2(T)} \|e-e_i\|_{L^2(T)} \qquad \textit{(Höldersche Ungleichung)}\\
&\le \left\|f+\partial_x^2 u_h\right\|_{L^2(T)} h_T \|\partial_x(e-e_i)\|_{L^2(T)} \qquad \textit{(Satz II.9)}\\
&\le \left\|f+\partial_x^2 u_h\right\|_{L^2(T)} h_T \left\{\|\partial_x e\|_{L^2(T)} + \|\partial_x e\|_{L^2(T)}\right\} \qquad \textit{(Satz II.10)}\\
&\le \left\|f+\partial_x^2 u_h\right\|_{L^2(T)} \cdot 2h_T \|\partial_x e\|_{L^2(T)}
\end{aligned}$$

Einsammeln:

$$\begin{aligned}
\|\partial_x(u-u_h)\|_{L^2(I)}^2 &\le \sum_T h_T\rho_T \|\partial_x e\|_{L^2(T)}\\
&\le \left(\sum_T h_T^2\rho_T^2\right)^{1/2} \underbrace{\left(\sum_T \|\partial_x e\|_{L^2(T)}^2\right)^{1/2}}_{\|\partial_x e\|_{L^2(I)}}
\end{aligned}$$

$$\Rightarrow \quad \|\partial_x(u-u_h)\|_{L^2(I)} \le \left(\sum_T h_T^2\rho_T^2\right)^{1/2}$$

■

II.7 Referenzelement, Gebietstransformation

Ziel: alle Rechnungen (zum Basisaufbau) auf einem sogenannten Referenzelement (z. B. Einheitsintervall)

+ Basisfunktionen nur einmal ausrechnen
+ (numerische) Integrationsformeln werden nur auf dem Referenzelement benötigt.
Vorbereitungen für die Substitutionsregel:

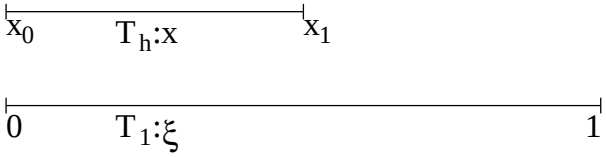

Abbildung II.8: Einheitsintervall als Referenzelement

$$
\begin{aligned}
F_h: &\quad T_1 \to T_h \\
&\quad x = x_0 + (x_1 - x_0)\xi \\
\frac{d}{dx}: &\quad 1 = (x_1 - x_0)\frac{d\xi}{dx} \quad \curvearrowright \quad dx = \underbrace{(x_1 - x_0)}_{J} d\xi \\
F_h^{-1}: &T_h \to T_1 \\
&\xi = \frac{x - x_0}{x_1 - x_0} \\
&\xi_x = \frac{d\xi}{dx} = \frac{1}{x_1 - x_0}
\end{aligned}
$$

Eine Basisfunktion φ_i^h auf T_h wird wie folgt angesetzt:

$$\varphi_i^h(x) := \varphi_i^1\left(F_h^{-1}(x)\right) = \varphi_i^1(\xi)$$

allgemein: $u(x) = v\left(F_h^{-1}(x)\right)$
$\curvearrowright \quad \partial_x u(x) = \partial_\xi v(\xi)\partial_x F_h^{-1}(x) = \partial_\xi v(\xi) \cdot \xi_x$
Beispiel: $\int\limits_{T_h} \partial_x\varphi_i^h(x)\partial_x\varphi_j^h(x)dx = \int\limits_{T_1} (\partial_\xi\varphi_i^1) \cdot \xi_x(\partial_\xi\varphi_j^1)\xi_x J d\xi$

$$\int\limits_{T_h} f(x)\varphi_i^h dx = \int\limits_{T_1} f(F_h(\xi))\varphi_i^1(\xi) \cdot J d\xi$$

Numerische Integration:
allgemein: $\int\limits_{T_1} g(\xi)d\xi \approx \sum\limits_{k=1}^{q} \omega_k g(\xi_k)$ *mit Integrationsgewichten* ω_k *und Stützstellen* ξ_k.
Beispiel: $q = 2$, $\xi_1 = 0, \xi_2 = 1$, $\omega_1 = \omega_2 = 1/2$ *(Trapezregel)*
Bei unserem Beispiel: $\int\limits_{T_h} f(x)\varphi_i^h dx \approx \sum\limits_{k=1}^{q} \omega_k f(F_h(\xi_k))\varphi_i^1(\xi_k) \cdot J$
Entsprechend:

$$A_{ij} += \int\limits_{T_h} \partial_x\varphi_j^h\partial_x\varphi_i^h(x)dx \approx \sum_{k=1}^{q} \omega_k \partial_\xi\varphi_j^1(\xi_k) \cdot \xi_x \cdot \partial_\xi\varphi_i^1(\xi_k) \cdot \xi_x \cdot J$$

Bemerkung: „Wähle q groß genug“
- *Integrationsfehler sollte von mindestens gleicher Ordnung sein, wie der Diskretisierungsfehler.*
- *Zu geringes q kann zu singulärer Gesamtmatrix A führen.*

II.8 Rechentechnische Betrachtung

$$A_{ij} = \int_I \partial_x \varphi_j \partial_x \varphi_i dx = \sum_T \int_T \partial_x \varphi_j \partial_x \varphi_i dx$$

for i=1 to n
  for j=1 to n
    $A_{ij} = \int_I \partial_x \varphi_j \partial_x \varphi_i dx$
Praxis: Summation vertauschen

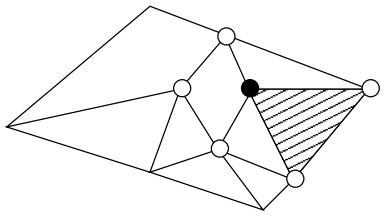

Abbildung II.9: dim> 1: Welche Zellen sind benachbart?

forall T
  for i=1 to n
    for j=1 to n
      $A_{ij} += \int_T \partial_x \varphi_j \partial_x \varphi_i dx$

Was passiert auf einer Zelle?
for k=1 to q
  berechne Basisfunktionen auf Referenzzelle T^1 *in* ξ_k
  for i=1 to local_n
    for j=1 to local_n
      $\boxed{A_{ij} += \omega_k \cdot \partial_\xi \varphi_j^1 \cdot \xi_x \cdot \partial_\xi \varphi_i^1 \xi_x \cdot J}$ *(DGL-spezifisch)*

III FEM für elliptische Probleme

III.1 Poisson-Problem

Rechengebiet $\Omega \subset \mathbb{R}^2$*, beschränkt (meistens* $(0,1)^2$*)*
Betrachte Aufgabe: $\min \left(\frac{1}{2} \int_\Omega (\nabla u)^2 dx - \int_\Omega f \cdot u dx \right)$ Ⓜ
mit $u = u(x)$*,* $f = f(x)$*,* $x = (x_1, x_2) \in \Omega$

$f, u : \Omega \to \mathbb{R} \qquad \nabla u = (\partial_{x_1} u, \partial_{x_2} u)$
Suche die Lösung von Ⓜ *in*

$V = \{\varphi \,|\varphi$*ist stetig auf* $\Omega, \partial_{x_1}\varphi, \partial_{x_2}\varphi$*sind stückweise stetig und beschränkt,* $\varphi = 0$ *auf* $\partial\Omega\}$

Satz III.1 (Greensche Formel) *Für hinreichend glatte Funktionen* v, w *gilt*

$$\int_\Omega \nabla v \nabla w dx = -\int_\Omega v \Delta w dx + \int_{\partial\Omega} v \partial_n w d\Gamma$$

(„Analogon zur partiellen Integration“)
mit $\Delta = (\partial^2_{x_1} + \partial^2_{x_2})$ *und* $\partial_n w = \nabla w \cdot n$ *(Mit dem Normalenvektor* n*, Abb. III.1)*

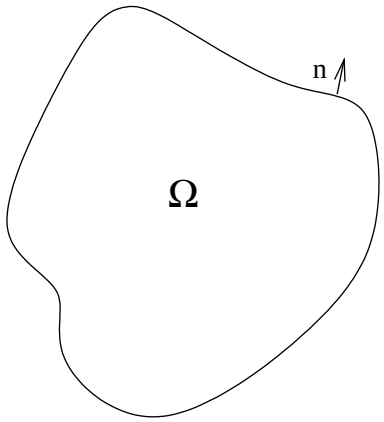

Abbildung III.1: Normalenvektor

<u>Beweis</u>: *Benutze Divergenzsatz in 2D für vektorwertige Funtkionen. (Übungsaufgabe)*
<u>*Klassisches Poisson-Problem*</u>

$$-\Delta u = f \text{ auf } \Omega \qquad \text{Ⓓ}$$
$$u = 0 \text{ auf } \partial\Omega$$

Satz III.2 *Analog zum eindimensionalen Fall gilt*

$$\text{Ⓓ} \Rightarrow \text{Ⓜ}$$

<u>Beweis</u>: *Benutze die Greensche Formel.* ■
Betrachte die variationelle Formulierung:

$$a(u, \varphi) = (f, \varphi) \quad \forall \quad \varphi \in V \qquad \text{Ⓥ}$$

mit $a(v, w) := (\nabla v, \nabla w), \quad v, w \in V$ *und* $(v, w) := \int_\Omega v \cdot w dx, \quad v, w \in V$

Satz III.3 Ⓥ $\Leftrightarrow$ Ⓜ

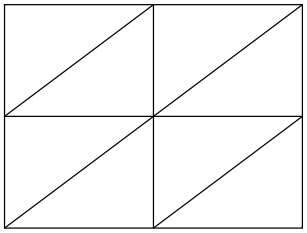

Abbildung III.2: Triangulierung eines Rechtecks

Beweis: *analog zum eindimensionalen Fall.*
Finite Elemente: *Triangulierung* $\mathbb{T}_h$, Ω *polygonal.*
Gitterparameter $h = \max_{T \in \mathbb{T}_h} diam(T)$ *mit* $diam(T)$ = *längste Seite von* T
Diskreter Teilraum $V_h \subset V$

$$V_h = \{\varphi \in V \mid \quad \varphi|_T \textit{ist linear für } T \in \mathbb{T}_h\}$$

Diskretisierung: $u_h \in V_h : \quad a(u_h, \varphi) = (f, \varphi) \quad \forall \quad \varphi \in V_h$ (V_h)

Satz III.4 (Galerkin-Eigenschaft) *Es gilt:* $a(u - u_h, \varphi) = 0 \quad \forall \varphi \in V_h$

Beweis:

$$\begin{aligned} a(u, \varphi) &= (f, \varphi) && \forall \quad \varphi \in V_h \\ a(u_h, \varphi) &= (f, \varphi) && \forall \quad \varphi \in V_h \\ \overline{a(u - u_h, \varphi)} &= \overline{0} && \forall \varphi \in V_h \end{aligned}$$

Satz III.5

$$\|\nabla(u - u_h)\| \leq \|\nabla(u - I_h u)\|$$

mit $\|w\| = \left(\int_\Omega w^2 dx\right)^{1/2}$ *und* $I_h w \in V_h$ *und* $I_h w(x_i) = w(x_i)$ *wobei* x_i *die Ecken aller Dreiecke* $T \in \mathbb{T}_h$ *durchläuft.*

ohne Beweis ■

Satz III.6 u, u_h *als Lösungen von* (V) *bzw.* (V_h)

$$\|\nabla(u - u_h)\| \leq c \cdot h$$

III.2 Natürliche und wesentliche Randbedingungen

Beispiel: *Neumann-Problem*
klassisch:

$$-\Delta u + u = f \ \textit{auf } \Omega \qquad (D)$$
$$\partial_n u = \frac{\partial u}{\partial n} = g \ \textit{auf } \Gamma = \partial\Omega$$
$$g := g(x_1, x_2) : \quad g : \mathbb{R}^2 \to \mathbb{R};\ \partial_n u = \nabla u \cdot n$$

Definition III.1

$$\frac{\partial u}{\partial n} = g \qquad \text{„Neumann-Bedingung“}$$
$$u = u_0 \qquad \text{„Dirichlet-Bedingung“}$$

Variationelle Formulierung:

$$a(u,\varphi) = (f,\varphi) + \int_\Gamma g\varphi d\Gamma \qquad \forall \quad \varphi \in V \qquad (V)$$

$$V = \{\varphi | \varphi \text{ ist stetig }, \partial_{x_i}\varphi \text{ stückweise stetig und beschränkt}\}$$
$$a(u,\varphi) := \int_\Omega \nabla u \cdot \nabla\varphi dx + \int_\Omega u\varphi dx$$
$$(f,\varphi) := \int_\Omega f\varphi dx$$

Minimum-Problem:

$$u \in V: \quad \min_{\varphi\in V}\left(\tfrac{1}{2}a(\varphi,\varphi) - (f,\varphi) - \int_\Gamma g\varphi d\Gamma\right) \qquad (M)$$

Satz III.7

$$(D) \quad \Rightarrow \quad (V)$$

<u>*Beweis*</u>*:*

1) $-\Delta u + u = f$

2) $-(\Delta u,\varphi) + (u,\varphi) = (f,\varphi)$
Greensche Formel:

3) $(f,\varphi) = \int_\Omega \nabla u \cdot \nabla\varphi dx - \int_\Gamma \frac{\partial u}{\partial n}\varphi d\Gamma + \int_\Omega u\varphi dx$
Benutze $\frac{\partial u}{\partial n} = g$*:*

4) $\int_\Omega \nabla u \cdot \nabla\varphi dx + \int_\Omega u\varphi dx = \int_\Omega f\varphi dx + \int_\Gamma g\varphi d\Gamma$

■

Satz III.8 *u als Lösung von* (V) *sei hinreichend glatt, dann*

$$(V) \quad \Rightarrow \quad (D)$$

Beweis:

$$(f,\varphi) + \int_\Gamma g\varphi d\Gamma = a(u,\varphi)$$

$$= \int_\Gamma \frac{\partial u}{\partial n}\varphi d\Gamma + \int_\Omega (-\Delta u + u)\varphi dx \qquad \textit{(Greensche Formel)}$$

$$\Leftrightarrow \quad \int_\Omega (-\Delta u + u - f)\varphi dx + \int_\Gamma \left(\frac{\partial u}{\partial n} - g\right)\varphi d\Gamma = 0 \quad \forall \quad \varphi \in V \qquad \circledast$$

Insbesondere gilt $\circledast$ *für* $\bar{\varphi} \in V$ *mit der zusätzlichen Eigenschaft* $\bar{\varphi} = 0$ *auf* Γ.
Also gilt $\int_\Omega (-\Delta u + u - f)\bar{\varphi} dx = 0$, *d. h.* $-\Delta u + u - f = 0$.
Somit reduziert sich $\circledast$ *zu* $\int_\Gamma \left(\frac{\partial u}{\partial n} - g\right)\varphi d\Gamma = 0 \quad \forall \quad \varphi \in V$.
Variationsargument $\curvearrowright$ $\frac{\partial u}{\partial n} = g$ ■

III.3 Sobolev-Räume

Bezeichnungen:

- *Gebiet* $\Omega \subset \mathbb{R}^n$ *offen, stückweise glatter Rand*
- $L_2(\Omega)$*: Menge aller Funktionen, deren Quadrat Lebesgue-integrierbar ist*
- *Skalarprodukt:* $(v,w)_0 := (v,w)_{L_2} = \int_\Omega vw dx$
- $L_2(\Omega)$ *ist ein Hilbertraum mit Norm* $\|v\|_0 = \sqrt{(v,v)_0}$

Definition III.2 (schwache Ableitung) $u \in L_2(\Omega)$ *besitzt in* $L_2(\Omega)$ *die* ***schwache Ableitung*** $v = \partial^\alpha u$, *falls* $v \in L_2(\Omega)$ *und*

$$(\varphi, v) = (-1)^{|\alpha|}(\partial^\alpha \varphi, u)_0 \qquad \forall \quad \varphi \in C_0^\infty(\Omega)$$

Multiindex $\alpha = (\alpha_1, \ldots, \alpha_n), \quad \alpha_i \in \mathbb{N}_0$

$$|a| = \sum \alpha_i$$
$$\partial^\alpha = \partial_{x_1}^{\alpha_1} \ldots \partial_{x_n}^{\alpha_n}$$

Beispiel: $\partial^\alpha u = \partial_{x_1}\partial_{x_2} u$
$\varphi \in C_0^\infty$: $\quad \varphi \in C^\infty$ *und supp* φ *ist kompakt in* Ω *enthalten.*

Definition III.3 (Sobolev-Räume) *Sei* $m \in \mathbb{N}_0$
$H^m(\Omega) = \{u \in L_2(\Omega) | u$ *besitzt schwache Ableitungen* $\partial^\alpha u \quad \forall \quad \alpha$ *mit* $|\alpha| \leq m\}$ *In* $H^m(\Omega)$ *wird durch*

$$(u, v)_m = \sum_{|\alpha| \leq m} (\partial^\alpha u, \partial^\alpha v)_0$$

ein ***Skalarprodukt*** *definiert.*
Die zugehörige ***Norm***

$$\|u\|_m = \sqrt{(u, u)_m} = \sqrt{\sum_{|\alpha| \leq m} \|\partial^\alpha u\|^2_{L_2(\Omega)}}$$

Stichwort: ***Halbnorm*** $|u|_m = \sqrt{\sum_{|\alpha|=m} \|\partial^\alpha u\|^2_0}$
Bemerkung: Mit $\|\cdot\|_m$ *ist* $H^m(\Omega)$ *vollständig.*

Satz III.9 *Sei* $m \in \mathbb{N}_0$
Dann ist $C^\infty(\Omega) \cap H^m(\Omega)$ *dicht in* $H^m(\Omega)$

Definition III.4 (Verallgemeinerung der Nullrandbedingungen) *Die Vervollständigung von* $C_0^\infty(\Omega)$ *bezüglich der Sobolev-Norm* $\|\cdot\|_m$ *wird mit* $H_0^m(\Omega)$ *bezeichnet.*

Beispiel:

$$\begin{aligned} -\Delta u &= f \qquad \text{auf } \Omega \\ u &= 0 \qquad \text{auf } \partial\Omega \end{aligned}$$

geeignete Wahl von V*:* $V = H_0^1(\Omega)$

Satz III.10 (Poincaré-Ungleichung) *Sei* Ω *in einem n-dimensionalen Würfel mit Kantenlänge s enthalten, dann*

$$\|v\|_0 \leq s|v|_1 \qquad v \in H_0^1(\Omega)$$

Beweis: Übertragung der Ideen aus dem eindimensionalen Fall. (siehe auch Abb. III.3 und Aufgabe 3.1)

III.4 Abstrakte Formulierung

Abstrakter Rahmen:

- V *ein Hilbertraum*
- $(\cdot, \cdot)_V$ *zugehöriges Skalarprodukt*
- $\|\cdot\|_V = \sqrt{(\cdot, \cdot)_V}$ *zugehörige Norm*

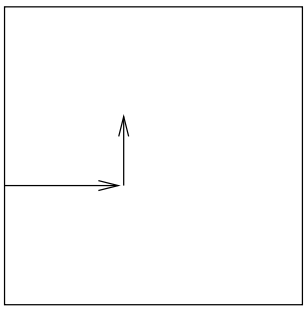

Abbildung III.3: zu Satz III.10

- *$a(\cdot,\cdot)$ eine Bilinearform auf $V \times V$*
- *$L(\cdot)$ eine Linearform auf V (bisher immer „$L(\cdot) = (f,\cdot)$“)*

Variationelle Formulierung:

$$u \in V: \quad a(u,\varphi) = L(\varphi) \quad \forall \quad \varphi \in V \qquad (V)$$

Voraussetzungen:

V i) $a(\cdot,\cdot)$ *ist symmetrisch*

V ii) $a(\cdot,\cdot)$ *ist stetig:* $|a(v,w)| < c\,\|v\|_V \cdot \|w\|_V \quad \forall \quad v,w \in V, \quad c > 0$

V iii) $a(\cdot,\cdot)$ *ist V-elliptisch:* $a(v,v) \geq \alpha\,\|v\|_V^2 \quad \forall \quad v \in V, \quad \alpha > 0$

V iv) $L(\cdot)$ *ist stetig:* $|L(v)| \leq \Lambda\,\|v\|_V \quad \forall \quad v \in V, \quad \Lambda > 0$

Satz III.11 (Existenzsatz) *Angenommen, es gelten die Bedingungen V i) bis V iv), dann existiert genau eine Lösung $u \in V$ von (V) mit der Stabilitätsabschätzung*

$$\|u\|_V \leq \frac{\Lambda}{\alpha}$$

Beweis:

i) ***Eindeutigkeit***
Annahme: $\exists \quad u_1, u_2 \in V$, u_1, u_2 *lösen* (V)

$$\begin{aligned} a(u_1,\varphi) &= L(\varphi) \\ \underline{a(u_2,\varphi)} &= \underline{L(\varphi)} \\ \Rightarrow \quad a(u_1-u_2,\varphi) &= 0 \end{aligned}$$

Wähle $\varphi = u_1 - u_2$: $\quad a(u_1-u_2, u_1-u_2) = 0$
Benutze V iii): $\alpha\,\|u_1-u_2\|_V^2 \leq a(u_1-u_2,u_1-u_2) = 0$
$\Rightarrow \quad u_1 - u_2 = 0$

ii) ***Existenz***

Idee: Reduziere Ⓥ *auf ein Fixpunktproblem*

Rieszscher Darstellungssatz:

$\exists \quad A \in \mathcal{L}(V,V)$:=*Lineare Abbildungen, stetig von* V *nach* V *und ein* $l \in V$*, so daß*

$$a(u,v) = (Au,v)_V \quad \forall u,v \in V$$
$$L(v) = (l,v)_V \quad \forall v \in V$$

Betrachte Ⓥ*:*

$$\begin{aligned} \forall \quad \varphi \in V : a(u,\varphi) - L(\varphi) &= 0 \\ \Leftrightarrow \quad (Au-l,\varphi)_V &= 0 \\ \Leftrightarrow \quad (-\rho(Au-l),\varphi)_V &= 0 \qquad \forall \quad \rho > 0 \\ \Leftrightarrow \quad (u-\rho(Au-l)-u,\varphi)_V &= 0 \\ \Leftrightarrow \quad u = u-\rho(Au-l) \quad \forall \quad \rho > 0 & \end{aligned}$$

Bretrachte $w_\rho : V \to V$ *mit* $w_\rho(v) = v - \rho(Av-l)$

Abschätzung von

$$\begin{aligned} \|w_\rho(v_1) - w_\rho(v_2)\|_V^2 &= \|v_1 - \rho(Av_1 - l) - v_2 + \rho(Av_2 - l)\|_V^2 \\ &= (v_1 - \rho Av_1 - (v_2 - \rho Av_2), v_1 - v_2 - \rho(Av_1 - Av_2))_V \\ &= (v_1 - v_2, v_1 - v_2)_V - 2\rho\,(A(v_1 - v_2), v_1 - v_2)_V + \\ &\qquad + \rho^2\,(A(v_1 - v_2), A(v_1 - v_2))_V \\ &= \|v_1 - v_2\|_V^2 - 2\rho \underbrace{a\,(v_1 - v_2, v_1 - v_2)}_{\geq \alpha \|v_1 - v_2\|_V^2,\ V\ iii)} + \rho^2 \|A(v_1 - v_2)\|_V^2 \\ &\geq \|v_1 - v_2\|_V^2 - 2\rho\alpha \|v_1 - v_2\|_V^2 + \rho^2 \|A\|_V^2 \|v_1 - v_2\|_V^2 \\ &= \left(1 - 2\rho\alpha + \rho^2 \|A\|_V^2\right) \|v_1 - v_2\|_V^2 \end{aligned}$$

Jetzt „Kurvendiskussion" für $(1 - 2\rho\alpha + \rho^2 \|A\|_V^2)$

Bedingung dafür, daß w_ρ *eine Kontraktionsabbildung ist:*

$$\begin{aligned} 1 - 2\alpha\rho + \rho^2 \|A\|_V^2 &< 1 \\ \textit{d. h.} \quad p(\rho) = -\rho(2\alpha - \|A\|_V^2\,\rho) &< 0 \\ \Rightarrow \quad \rho > 0 \quad \wedge \quad \rho &< \frac{2\alpha}{\|A\|_V^2} \\ \Rightarrow \quad w_\rho \textit{ ist für solche } \rho \textit{ eine Kontraktionsabbildung} & \end{aligned}$$

$w_\rho(v) = v - \rho(Av - l)$

$\curvearrowright$ $\exists$ *ein Fixpunkt*

$\curvearrowright$ *es gibt eine Lösung von* Ⓥ

iii) ***Stabilitätsabschätzung***

Wähle $\varphi = u$ *in* Ⓥ *und benutze „V-elliptisch" (V iii)) und die Stetigkeit von L (V*

iv))

$$\alpha \|u\|_V^2 \overset{V\ iii)}{\leq} a(u,u) \overset{\text{(V)}}{=} L(u) \overset{V\ iv)}{\leq} \Lambda \|u\|_V$$
$$\Leftrightarrow \quad \|u\|_V \leq \frac{\Lambda}{\alpha}$$

■

Bemerkung: Der Beweis gilt auch für unsymmetrische Form $a(\cdot,\cdot)$.
Abstraktes Minimierungsproblem
Finde $u \in V$, so daß

$$f(u) \leq F(\varphi) \quad \forall \quad \varphi \in V \qquad \text{(M)}$$

gilt, mit

$$F(\varphi) = \frac{1}{2} a(\varphi,\varphi) - L(\varphi)$$

Satz III.12 (M) $\Leftrightarrow$ (V)

Beweis: Übung ■

III.5 Diskretisierung

$$u \in V: \quad a(u,\varphi) = L(\varphi) \quad \forall \quad \varphi \in V$$
$$u_h \in V_h: \quad a(u_h,\varphi) = L(\varphi) \quad \forall \quad \varphi \in V_h \subset V$$
$$V_h = \langle \varphi_1, \ldots, \varphi_N \rangle$$
$$\varphi \in V_h: \quad \varphi = \sum_{i=1}^{N} \alpha_i \varphi_i, \qquad \alpha_i \in \mathbb{R}$$
$$u_h \in V_h: \quad u_h = \sum_{j=1}^{N} x_j \varphi_j, \qquad x_j \in \mathbb{R}$$
$$a(u_h,\varphi_i) = L(\varphi_i), \qquad i = 1,\ldots,N$$
$$\curvearrowright \quad \sum_{j=1}^{N} a(\varphi_j,\varphi_i) x_j = L(\varphi_i), \quad i = 1,\ldots,N$$
$$\curvearrowright \textit{ Matrixform } Ax = b, \quad A \in \mathbb{R}^{N\times N}; \quad x, b \in \mathbb{R}^N$$
$$A_{ij} = a(\varphi_j,\varphi_i), \quad b_i = L(\varphi_i) \qquad \textit{(Reihenfolge der Indizes!)}$$

Satz III.13 *Es gelte V i) - V iv), dann ist A symmetrisch und positiv definit.*

Beweis: Übung ■

Satz III.14 *Es gelte V i) - V iv), dann gilt*

$$\|u_h\|_V \leq \frac{\Lambda}{\alpha}$$

(kein Stabilitätsverlust)

Beweis: Übung ■

Satz III.15

$$\|u - u_h\|_V \leq \frac{C}{\alpha} \|u - \varphi\|_V \qquad \forall \quad \varphi \in V_h \subset V$$

III.6 Variationsungleichungen

Problem:

$$a(u, \varphi - u) \geq L(\varphi - u) \quad \forall \quad \varphi \in K \subset V$$

K ist abgeschlossen und konvex.
(Variationsungleichung 1. Art)

Lemma III.1 *$K \subset V$ sei abgeschlossen und konvex. Dann gilt*

$$\forall \quad x \in V \quad \exists! \quad y \in K, \text{ so daß } \|x - y\| = \inf_{\varphi \in K} \|x - \varphi\|$$

Der Punkt y heißt Projektion von x auf die Menge K: $y = P_K(x)$

Beweis:

i) *„Es gibt ein y“*
Sei φ_k eine „Minimalfolge“, d. h. $\lim_{k\to\infty} \|\varphi_k - x\| = d = \inf_{\varphi \in K} \|\varphi - x\|$
Durch Ausmultiplizieren verifiziert man

$$\|\varphi_k - \varphi_l\|^2 = 2\|x - \varphi_k\|^2 + 2\|x - \varphi_l\|^2 - 4\left\|x - \frac{1}{2}(\varphi_k + \varphi_l)\right\|^2$$

Da K konvex:

$$\frac{1}{2}(\varphi_k + \varphi_l) \in K \quad \curvearrowright \quad d^2 \leq \left\|x - \frac{1}{2}(\varphi_k - \varphi_l)\right\|^2$$

$$\Rightarrow \quad \|\varphi_k - \varphi_l\|^2 \leq 2\underbrace{\|x - \varphi_k\|^2}_{\to d^2} + 2\underbrace{\|x - \varphi_l\|^2}_{\to d^2} - 4d^2$$

Somit ist $\|\varphi_k - \varphi_l\| \underset{k,l\to\infty}{\longrightarrow} 0$.
Da V vollständig ist und K abgeschlossen

$$\exists \quad y \in K \quad \text{mit} \quad \lim_{k\to\infty} \varphi_k = y$$

Wegen der Stetigkeit der Norm $\|x - y\| = \lim_{k\to\infty} \|x - \varphi_k\| = d$

ii) *„Eindeutigkeit von y“*
Seien $y_1, y_2 \in K$ mit $\|x - y_1\| = \|x - y_2\| = \inf_{\varphi \in K} \|x - \varphi\|$. *Analog zu i):*

$$\|y_1 - y_2\|^2 \leq 2\|x - y_1\|^2 + 2\|x - y_2\|^2 - 4d^2 \leq 0$$

■

Satz III.16 *Sei $K \subset V$ abgeschlossen und konvex. Dann gilt $y = P_K(x)$ genau dann, wenn gilt:*

$$y \in K: \quad (y - x, \varphi - y) \geq 0 \quad \forall \quad \varphi \in K$$

Beweis:
„$\Rightarrow$“ : $x \in V$ und $y = P_K(x) \in K$
K ist konvex: $(1-t)y + t\varphi = y + t(\varphi - y) \in K$, $0 \leq t \leq 1$
Betrachte

$$\begin{aligned}\phi(t) &= \|x - y - t(\varphi - y)\|^2 \\ &= \|x - y\|^2 - 2t(x - y, \varphi - y) + t^2 \|\varphi - y\|^2\end{aligned}$$

$\phi(t)$ nimmt bei $t = 0$ das Minimum an

$$\begin{aligned}&\Rightarrow \quad \phi'(0) \geq 0 \\ &\Leftrightarrow \quad -2(x - y, \varphi - y) \geq 0 \\ &\Leftrightarrow \quad (x - y, \varphi - y) \leq 0 \\ &\Leftrightarrow \quad (y - x, \varphi - y) \geq 0\end{aligned}$$

„$\Leftarrow$“ : Wählte $\varphi \in K$ beliebig aber fest:

$$\begin{aligned}0 &\leq (y - x, \varphi - y) \\ &= (y - x, (\varphi - x) + (x - y)) \\ &= (y - x, x - y) + (y - x, \varphi - x) \\ &= -\|y - x\|^2 + (y - x, \varphi - x) \\ \Rightarrow \quad \|y - x\|^2 &\leq \|y - x\| \, \|\varphi - x\| \\ \Rightarrow \quad \|y - x\| &\leq \|\varphi - x\| \quad \forall \quad \varphi \in K\end{aligned}$$

■

Korollar III.1 *Sei $K \subset V$ abgeschlossen und konvex. Dann ist P_K nicht-expansiv:*

$$\|P_K(x) - P_K(x')\| \leq \|x - x'\| \quad \forall \quad x, x' \in V$$

Beweis: Gegeben seien $x, x' \in V$. $y = P_K(x)$, $y' = P_K(x')$
$y \in K: \quad (y, \varphi - y) \geq (x, \varphi - y) \quad \forall \quad \varphi \in K$
$y' \in K: \quad (y', \varphi - y') \geq (x', \varphi - y') \quad \forall \quad \varphi \in K$
1. Ungleichung: $\varphi = y'$; 2. Ungleichung: $\varphi = y$
„Addition“ :

$$\begin{aligned}\|y - y'\|^2 = (y - y', y - y') &\leq (x - x', y - y') \\ &\leq \|x - x'\| \, \|y - y'\| \\ \Leftrightarrow \|y - y'\| &\leq \|x - x'\|\end{aligned}$$

■

Satz III.17 *Das Problem*

$$a(u, \varphi - u) \geq L(\varphi - u) \quad \forall \quad \varphi \in K$$

hat eine eindeutige Lösung.

Beweis:

i) ***Eindeutigkeit***

$$\begin{aligned} a(u_1, \varphi - u_1) &\geq L(\varphi - u_1) \quad \forall \quad \varphi \in K \\ a(u_2, \varphi - u_2) &\geq L(\varphi - u_2) \quad \forall \quad \varphi \in K \end{aligned}$$

Testen mit $\varphi = u_2$ bzw. $\varphi = u_1$ und Addition:

$$\begin{aligned} \alpha \left\| u_1 - u_2 \right\|^2 &\overset{V\ iii)}{\leq} a(u_2 - u_1, u_2 - u_1) \leq 0 \\ &\Rightarrow \quad \left\| u_1 - u_2 \right\| \leq 0 \end{aligned}$$

ii) ***Existenz***
Mit dem Rieszschen Darstellungssatz

$$\begin{aligned} a(u, v) &= (Au, v) \quad \forall \quad u, v \in V \\ L(v) &= (l, v) \quad \forall \quad v \in V \end{aligned}$$

ergibt sich

$$\begin{aligned} & (Au, \varphi - u) \geq (l, \varphi - u) \qquad \forall \quad \varphi \in K \\ \Leftrightarrow \quad & (-(Au - l), \varphi - u) \leq 0 \\ \Leftrightarrow \quad & ((u - \rho(Au - l)) - u, \varphi - u) \leq 0 \qquad \forall \quad \rho > 0 \end{aligned}$$

Dies ist nach Satz III.16 äquivalent zu $u = P_K\left(u - \rho(Au - l)\right)$.
Betrachte $w_\rho(v) = P_K\left(v - \rho(Av - l)\right)$
Seien $v_1, v_2 \in V$

$$\begin{aligned} \left\| w_\rho(v_1) - w_\rho(v_2) \right\|^2 &\leq \left\| v_1 - v_2 \right\|^2 + \rho^2 \left\| A(v_1 - v_2) \right\|^2 - 2\rho a(v_1 - v_2, v_1 - v_2) \\ &\qquad (P_K \text{ nicht expansiv}) \\ &\leq (1 - 2\rho\alpha + \rho^2 \left\| A \right\|^2) \left\| v_1 - v_2 \right\|^2 \end{aligned}$$

$\curvearrowright$ w_ρ ist eine Kontraktion, falls $0 < \rho < \frac{2\alpha}{\|A\|^2}$ gilt.
Kontraktion $\curvearrowright$ Fixpunkt $\curvearrowright$ Fixpunkt ist Lösung. ■

III.7 Interpolation

Wiederholung: Satz III.15: $\|u - u_h\|_V \leq \frac{C}{\alpha} \|u - I_h u\|_V$
Interpolation in zwei Dimensionen für lineare Funktionen
Bezeichnungen*(Abbildung III.4)*

$$h_T = diam(T), \qquad \text{z. B. längste Seite}$$
$$\rho_T = \text{Radius des Inkreises}$$
$$h = \max_{T \in \mathbb{T}_h} \quad h_T$$

Wir betrachten $\mathbb{T}_h$ *für* $\frac{\rho_T}{h_T} \geq \beta > 0$ *(„Die Dreiecke dürfen nicht zu dünn werden.“)*

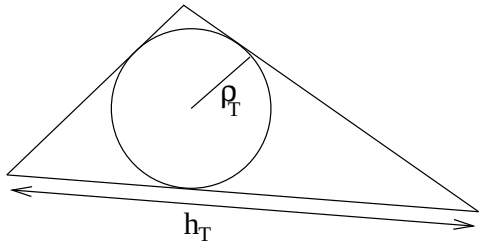

Abbildung III.4: zu den Bezeichnungen

$u \in C^0(\overline{\Omega}); \qquad I_h u(v_i) = u(v_i)$
$I_h u$ *linear auf jedem* T

Satz III.18 *Sei* $T \in \mathbb{T}_h$ *ein Dreieck mit Knoten* $a_i, \quad i = 1, 2, 3.$
Sei $v \in C^0(T)$. *Die Interpolierende* $I_h v \in P_1(T)$ *sei definiert durch*

$$I_h v(a_i) = v(a_i), \qquad i = 1, 2, 3$$

Dann gilt:

i) $\|v - I_h v\|_{L_\infty(T)} \leq 2(h_T^2) \max_{|\alpha|=2} \|D^\alpha v\|_{L_\infty(T)}$

ii) $\max_{|\alpha|=1} \|D^\alpha (v - I_h v)\|_{L_\infty(T)} \leq 6 \cdot \frac{h_T^2}{\rho_T} \max_{|\alpha|=2} \|D^\alpha v\|_{L_\infty(T)}$

<u>*Beweis*</u>*: „Taylorentwicklung“* ■

Satz III.19

i) $\|v - I_h v\|_{L_2(T)} \leq c h_T^2 |v|_{H^2(T)}$

ii) $|v - I_h v|_{H^1(T)} \leq c \frac{h_T^2}{\rho_T} |v|_{H^2(T)}$

ohne Beweis.

Satz III.20

i) $\|v - I_h v\|_{L_2} \leq ch^2 |v|_{H^2(\Omega)}$

ii) $|v - I_h v|_{H^1(\Omega)} \leq c\frac{h}{\beta} |v|_{H^2(\Omega)}$

Beweis: Übung, Summation der Integranden.

Satz III.21 (höhere Polynomansätze) $I_h v \in P_r(T)$ *mit* $r \geq 1$

i) $\|v - I_h v\|_{L^2(\Omega)} \leq ch^{r+1} |v|_{H^{r+1}(\Omega)}$

ii) $|v - I_h v|_{H^1(\Omega)} \leq ch^r |v|_{H^{r+1}(\Omega)}$

Satz III.22 („fehlende Regularität“) $1 \leq s \leq r+1$

i) $\|v - I_h v\|_{L^2(\Omega)} \leq ch^s |v|_{H^s(\Omega)}$

ii) $|v - I_h v|_{H^1(\Omega)} \leq ch^{s-1} |v|_{H^s(\Omega)}$

ohne Beweis. ***Aber bitte merken!***

IV Minimierungsalgorithmen, iterative Methoden

$Ax = b$, $A \in \mathbb{R}^{n \times n}$, *symmetrisch, positiv definit;* $x, b \in \mathbb{R}^n$
Betrachte $f(x) = \frac{1}{2} x^T A x - b^T x$*; eine Minimalstelle* $\tilde{x}$ *von* $f(x)$ *erfüllt* $A\tilde{x} - b = 0$.

IV.1 positiv definite Matrizen

Bemerkung IV.1 *Sei* $\|\cdot\|$ *eine Norm auf* $\mathbb{C}^n$ *und* $A \in M(n,n) := \mathbb{C}^{n \times n}$.
A sei regulär, dann definiert

$$\|x\|_A = \|Ax\|$$

ebenfalls eine Norm.

Definition IV.1 *Sei* $(\cdot, \cdot)$ *das euklidische Skalarprodukt auf* $\mathbb{C}^n$*, dann heißt* $A \in M(n,n)$ ***positiv definit****, wenn*

$$A = A^H \quad \text{und} \quad (Ax, x) > 0 \quad \forall \quad x \in \mathbb{C}^n, \quad x \neq 0$$

Bemerkung IV.2 $A \in M(n,n)$ *mit* $(Ax, x) > 0 \quad \forall \quad x \in \mathbb{C}^n \quad x \neq 0$
$\Leftrightarrow A = A^H$ *und alle Eigenwerte sind positiv.*
Es gibt die Darstellung $A = TDT^H$ *mit* T *unitär und* D *diagonal.*

Definition IV.2 *Sei* $A^{1/2} := TD^{1/2}T^H$, *dann heißt*

$$\|x\|_A := \left\|A^{1/2}x\right\|_2$$

die ***Energienorm***. $(\|\cdot\|_2 = \sqrt{(\cdot,\cdot)})$

Bemerkung IV.3 *Für das Skalarprodukt* $(x,y)_A$, $\quad x, y \in \mathbb{C}^n$ *gilt*

$$\|x\|_A = \sqrt{(Ax,x)}$$

Bemerkung IV.4

$$A \text{ positiv definit} \quad \Leftrightarrow \quad A^{-1} \text{ positiv definit}$$

Definition IV.3 *Die* ***Kondition*** *von* $A \in M(n,n)$, A *regulär ist definiert durch*

$$cond(A) = \|A\| \left\|A^{-1}\right\|$$

$$\left(\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}\right)$$

Bemerkung IV.5 *Vektornorm sei* $\|\cdot\|_2$
$A \in M(n,n)$, $\quad A$ *positiv definit. Dann gilt*

$$cond(A) = \frac{\lambda_{\max}}{\lambda_{\min}}$$

mit $\lambda_{\min}$ *dem kleinsten und* $\lambda_{\max}$ *dem größten Eigenwert von* A.

Lemma IV.1 *Sei* A *eine positiv definite Matrix mit Kondition* κ, *dann gilt für jeden Vektor* $x \neq 0$

$$\frac{(x^TAx)(x^TA^{-1}x)}{(x^Tx)^2} \leq \kappa \qquad (*)$$

<u>Beweis</u>: *Anordnung der Eigenwerte* $\lambda_1 \leq \ldots \leq \lambda_n$
Betrachte die Situation nach unitärer Transformation im Raum der Eigenvektoren.
Dann gilt für „linke Seite“ in (*)

$$\frac{\left(\sum_{i=1}^n \lambda_i x_i^2\right)\left(\sum_{i=1}^n \lambda_i^{-1} x_i^2\right)}{\left(\sum_{i=1}^n x_i^2\right)^2} \qquad \oplus$$

Substitution: $z_i = \frac{x_i^2}{\sum_{j=1}^n x_j^2}$

Es gilt: $\sum_{j=1}^n z_i = 1$
Einsetzen in $\oplus$:

$$\begin{aligned} \oplus &= \left(\sum_{i=1}^n \lambda_i z_i\right)\left(\sum_{i=1}^n \lambda_i^{-1} z_i\right) \\ &\leq \lambda_n \lambda_1^{-1} \end{aligned}$$

■

IV.2 Abstiegsverfahren

Aufgabe: $f : \mathbb{R}^n \to \mathbb{R}$ *stetig differenzierbar*
Gesucht: $\tilde{x} \in \mathbb{R}^n$ *mit* $f(\tilde{x}) \leq f(x) \quad \forall \quad x \in \mathbb{R}^n$

Lemma IV.2 *Unter obigen Voraussetzungen sei* $d := -\nabla f(x) \neq 0$, *dann gilt*

$$f(x + td) < f(x)$$

für hinreichend kleines $t > 0$.

Beweis: Betrachte Richtungsableitung:

$$\lim_{t \to 0} \frac{f(x + td) - f(x)}{t} = \nabla f(x)^T d < 0$$

wegen $d = -\nabla f(x)$. *Also gilt*

$$\frac{f(x + td) - f(x)}{t} < 0$$

für hinreichend kleines t.
$\Rightarrow \quad f(x + td) < f(x)$ *wegen* $t > 0$. ■

Algorithmus IV.1
Für $k = 0, 1, \ldots$

i) *Berechne* $d_k = -\nabla f(x_k)$

ii) *Liniensuche: Man sucht für* f *das Minimum auf der Linie* $\{x_k + td_k\}$ *für* $t > 0$.

IV.3 Gradientenverfahren

Spezielle Aufgabe: $f(x) = \frac{1}{2} x^T A x - b^T x$, A *positiv definit.*
Benutze Algorithmus IV.1.
Hier gilt speziell

i) $d_k = b - Ax_k$

ii) $t = \frac{d_k^T d_k}{d_k^T A d_k}$

Rechnung:

i) $\nabla f(x) = Ax - b$

ii) Minimumsuche für quadratisches f:

$$\begin{aligned} f(x_k + td_k) = & \min \\ & \Rightarrow \quad \partial_t f(x_k + td_k) = 0 \\ & \Leftrightarrow \quad \nabla f(x_k + td_k) \cdot d_k = 0 \\ & \Leftrightarrow \quad (A(x_k + td_k) - b) \cdot d_k = 0 \\ & \Leftrightarrow \quad \big(\underbrace{(Ax_k - b)}_{-d_k} + tAd_k\big) \cdot d_k = 0 \\ & \Leftrightarrow \quad d_k^T (tA - I) d_k = 0 \\ & \Leftrightarrow \quad \frac{d_k \cdot d_k}{d_k \cdot Ad_k} = t \end{aligned}$$

Lemma IV.3 *Mit $f(x) = \frac{1}{2} x^T Ax - b^T x$ gilt*

$$f(x) - f(\tilde{x}) = \frac{1}{2} \|x - \tilde{x}\|_A^2$$

Beweis:

1) „linke Seite“

$$\begin{aligned} f(x) - f(\tilde{x}) &= \frac{1}{2} x^T Ax - b^T x - \left(\frac{1}{2} \tilde{x}^T \underbrace{A\tilde{x}}_{=b} - b^T \tilde{x} \right) \\ &= \frac{1}{2} x^T Ax - b^T x + \frac{1}{2} b^T \tilde{x} \end{aligned}$$

2) „rechte Seite“

$$\begin{aligned} \frac{1}{2} \|x - \tilde{x}\|_A^2 &= \frac{1}{2} (x - \tilde{x})^T A (x - \tilde{x}) \\ &= \frac{1}{2} (x - \tilde{x})^T (Ax - b) \\ &= \frac{1}{2} x^T Ax - \frac{1}{2} \underbrace{\tilde{x}^T A}_{b^T} x - \frac{1}{2} x^T b + \frac{1}{2} \tilde{x}^T b \end{aligned}$$

Vergleiche 1) und 2) $\curvearrowright$ 1)=2). ■

Lemma IV.4 *Sei $\tilde{x}$ Lösung von $Ax = b$, dann gilt*

$$\frac{\|x_k - \tilde{x}\|_A^2}{d_k^T A^{-1} d_k} = 1$$

Beweis: Es gilt $d_k = b - Ax_k = A(\tilde{x} - x_k)$
$\Leftrightarrow \quad -A^{-1}d_k = x_k - \tilde{x}$
Betrachte: $\|x_k - \tilde{x}\|_A^2 = (x_k - \tilde{x})^T A(x_k - \tilde{x}) = (d_k^T A^{-1})A(A^{-1}d_k) = d_k^T A^{-1} d_k.$
Division leistet das Gewünschte. ■

Satz IV.1 *Sei* $\tilde{x} \in \mathbb{R}^n$, *so daß* $A\tilde{x} = b$.
Für den Iterationsfehler nach $k+1$ *Schritten des Gradientenverfahrens gilt*

$$\|x_{k+1} - \tilde{x}\|_A^2 \leq \|x_k - \tilde{x}\|_A^2 \left(1 - \frac{1}{\kappa}\right)$$

mit $\kappa = cond(A)$.

Beweis: Laut Algorithmus:

1) $d_k = b - Ax_k$

2) $t_k = \frac{d_k^T d_k}{d_k^T A d_k}$

Einsetzen:

$$\begin{aligned} f(x_{k+1}) &= f(x_k + t_k d_k) \\ &= \frac{1}{2}(x_k + t_k d_k)^T A(x_k + t_k d_k) - b^T(x_k + t_k d_k) \\ &= f(x_k) + t_k d_k^T \underbrace{(Ax_k - b)}_{-d_k} + \frac{1}{2} t_k^2 d_k^T A d_k \\ &= \ldots \text{„}t_k \text{ einsetzen“} \\ &= f(x_k) - \frac{1}{2}\frac{(d_k^T d_k)^2}{d_k^T A d_k} \end{aligned}$$

$\Leftrightarrow \quad f(x_{k+1}) - f(\tilde{x}) = f(x_k) - f(\tilde{x}) - \frac{1}{2}\frac{(d_k^T d_k)^2}{d_k^T A d_k}$
Benutze Lemma IV.3

$$\frac{1}{2}\|x_{k+1} - \tilde{x}\|_A^2 = \frac{1}{2}\|x_k - \tilde{x}\|_A^2 - \frac{1}{2}\frac{(d_k^T d_k)^2}{d_k^T A d_k} \cdot 1$$

Benutze Lemma IV.4:

$$\begin{aligned} \|x_{k+1} - \tilde{x}\|_A^2 &= \|x_k - \tilde{x}\|_A^2 - \frac{(d_k^T d_k)^2 \|x_k - \tilde{x}\|_A^2}{d_k^T A d_k d_k^T A^{-1} d_k} \\ &= \|x_k - \tilde{x}\|_A^2 \left(1 - \frac{(d_k^T d_k)^2}{(d_k^T A d_k)(d_k^T A^{-1} d_k)}\right) \end{aligned}$$

Benutze Lemma IV.1:
$\curvearrowright \quad \|x_{k+1} - \tilde{x}\|_A^2 \leq \|x_k - \tilde{x}\|_A^2 \left(1 - \frac{1}{\kappa}\right)$ ■

Lemma IV.1

$$\frac{(x^T Ax)(x^T A^{-1} x)}{(xTx)^2} \leq \left(\frac{1}{2}\sqrt{\kappa} + \frac{1}{2}\sqrt{\kappa^{-1}}\right)^2$$

Ohne Beweis.
Einstieg in Beiweis von Satz IV.1:

$$\begin{aligned}\|x_{k+1}-\tilde{x}\|_A^2 &\le \|x_k-\tilde{x}\|_A^2\left(1-\frac{(d_k^T d_k)}{(d_k^T A d_k)(d_k^T A^{-1} d_k)^2}\right)\\ &\le \|x_k-\tilde{x}\|_A^2\left(1-\frac{4}{\left(\sqrt{\kappa}+\sqrt{\kappa^{-1}}\right)^2}\right)\\ &= \|x_k-\tilde{x}\|_A^2\,\frac{(\kappa-1)^2}{(\kappa+1)^2}\end{aligned}$$

$$\Rightarrow \quad \boxed{\|x_k-\tilde{x}\|_A \le \left(\frac{\kappa-1}{\kappa+1}\right)^k \|x_0-\tilde{x}\|_A}$$

$\frac{\kappa-1}{\kappa+1}=\frac{\kappa+1}{\kappa+1}-\frac{2}{\kappa+1}\approx 1-\frac{2}{\kappa}$ *(vgl. $(1-1/\kappa)$ aus Satz IV.1)*

Beispiel zu 2D-Poisson

$A\sim$ „$-\Delta$“ (Lineare FE-Diskretisierung in 2D)

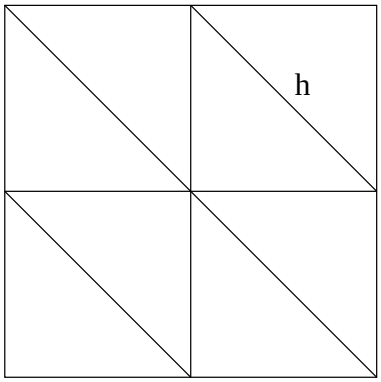

Abbildung IV.1: Gitter zur linearen FE-Diskretisierung

$$\begin{aligned}\lambda_{min} &= 8h^{-2}\sin^2\left(\frac{\pi}{2}h\right)\\ \lambda_{max} &= 8h^{-2}\cos^2\left(\frac{\pi}{2}h\right)\end{aligned}$$

Einsetzen in

$$\begin{aligned}1-\frac{2}{\kappa} &= 1-2\frac{\sin^2\frac{\pi}{2}h}{\cos^2\frac{\pi}{2}h} \le 1-2\sin^2\frac{\pi}{2}h\\ &= 1-(1-\cos\pi h)=\cos\pi h\\ &\approx 1-\frac{1}{2}\pi^2h^2\end{aligned}$$

IV.4 Projiziertes Gradientenverfahren

$K = \{x \in \mathbb{R}^n \,|\, x(i) \geq 0; i = 1, \ldots, n\}$
Aufgabe: Finde $u \in K$ *mit* $\min = \frac{1}{2}u^T Au - f^T u =: J(u)$
$A \in M(n,n)$ *symmetrisch, positiv definit,* $u, f \in \mathbb{R}^n$

Algorithmus IV.2 *P_K bezeichne die Projektion auf K.*

i) *Initialisierung: wähle* $u_0 \in K$

ii) *Iteration:*

$$\begin{aligned} &\text{for } k = 0, 1, \ldots \\ &u_{k+1} = P_K(u_k - \alpha_k J'(u_k)); \quad \alpha_k > 0 \end{aligned}$$

mit $J'(x) = (Ax - f)$
Schritt ii) zerfällt wie folgt:

$$\begin{aligned} &u_{k+\frac{1}{2}} = u_k + \alpha_k(f - Au_k) \;\; \text{„Gradientenschritt “} \\ &u_{k+1} = P_K(u_{k+\frac{1}{2}}), \text{ d. h. } u_{k+1}(i) = \max\{0, u_{k+\frac{1}{2}}\} \end{aligned}$$

Satz IV.2 $\exists \quad \alpha, \beta > 0$, *so daß mit* $\alpha \leq \alpha_k \leq \beta$ *der Algorithmus IV.2 gegen die Lösung u konvergiert.*

Beweis:

i) *Wir zeigen:* $u = P_K(u - \alpha_k J'(u))$
Aus Abschnitt II.5: $(J'(u), \varphi - u) \geq 0 \quad \forall \quad \varphi \in K$
Mit $\alpha_k > 0$ *gilt:*

$$\begin{aligned} & (\alpha_k J'(u), \varphi - u) \geq 0 \quad \forall \quad \varphi \in K \\ \Leftrightarrow \quad & (u - (u - \alpha_k J'(u)), \varphi - u) \geq 0 \quad \forall \quad \varphi \in K \\ \Leftrightarrow \quad & u = P_K(u - \alpha_k J'(u)) \end{aligned}$$

ii)

$$\begin{aligned} \|u_{k+1} - u\| &= \|P_K(u_k - \alpha_k J'(u_k)) - P_K(u - \alpha_k J'(u))\| \\ &\leq \|u_k - u - \alpha_k(J'(u_k) - J'(u))\| \qquad \text{„nicht expansiv“} \end{aligned}$$

Quadrieren: $\|u_{k+1} - u\|^2 \leq \|u_k - u\|^2 - 2\alpha_k (u_k - u, J'(u_k) - J'(u)) + \alpha_k^2 \|J'(u_k) - J'(u)\|^2$
Nun gilt $J'(u_k) - J'(u) = (Au_k - f) - (Au - f) = A(u_k - u)$
Weiterhin A symmetrisch, positiv definit:
$(u_k - u)^T A(u_k - u) \geq \lambda_{min} \|u_k - u\|^2$
Einsetzen: $\|u_{k+1} - u\|^2 \leq \|u_k - u\|^2 - 2\alpha_k \lambda_{min} \|u_k - u\|^2 + \alpha_k^2 \|A\|^2 \|u_k - u\|^2$
Mit üblicher Faktordiskussion:
$\Rightarrow \quad \alpha_k > 0 \qquad \alpha_k < \frac{2\lambda_{min}}{\|A\|^2}$

■

IV.5 Konjugiertes Gradientenverfahren (cg)

Idee bisher: $x_{i+1} = x_i + \alpha_i d_i$

$$f(x_{i+1}) = \min_{z \in span[d_i]} f(x_i + z) \qquad \text{„eindimensionale Minimierung“}$$

Verbesserung: $f(x_{i+1}) = \min_{z \in \langle d_{i-1}, g_i \rangle} f(x_i + z)$

$d_{i-1} = x_i - x_{i-1}$ „Richtung der letzten Korrektur“

$g_i = Ax_i - b$ Gradientenrichtung

Definition IV.4 *Sei $A \in M(n,n)$ symmetrisch, positiv definit.*
*Zwei Vektoren $x, y \in \mathbb{R}^n$ heißen **konjugiert** oder **A-orthogonal**, falls $x^T Ay = 0$ ist.*

Bemerkung: $x_1, \ldots, x_k \in \mathbb{R}^n$ *paarweise konjugiert* $\Rightarrow$ $x_1, \ldots, x_k$ *linear unabhängig*
Rechnung:

$$\begin{aligned} & \sum_{i=1}^{k} \alpha_i x_i = 0 \\ \Leftrightarrow \quad & \sum_{i=1}^{k} \alpha_i x_j^T A x_i = 0 \\ \Leftrightarrow \quad & \alpha_j \underbrace{x_j^T A x_j}_{>0} = 0 \quad \Rightarrow \quad \alpha_j = 0 \end{aligned}$$

Lemma IV.5 („konjugierte Richtungen sind gut“) *Seien $d_0, \ldots, d_{n-1}$ konjugierte Richtungen.*
Weiterhin: $\tilde{x} = A^{-1}b$ (Lösung)
Dann gilt

$$\tilde{x} = \sum_{i=0}^{n-1} \alpha_i d_i \qquad \text{mit } \alpha_i = \frac{d_i^T b}{d_i^T A d_i}$$

(„Lösung ist direkt hinschreibbar“)

Beweis: Ansatz:

$$\begin{aligned} & \tilde{x} = \sum_{k=0}^{n-1} \alpha_k d_k \\ \Leftrightarrow \quad & d_i^T A \tilde{x} = \sum_{k=0}^{n-1} \alpha_k \underbrace{d_i^T A d_k}_{=0, \quad i \neq k} = \alpha_i d_i^T A d_i, \quad i = 0, \ldots, n-1 \\ \Leftrightarrow \quad & \alpha_i = \frac{d_i^T A \tilde{x}}{d_i^T A d_i} = \frac{d_i^T b}{d_i^T A d_i} \quad i = 0, \ldots, n-1 \end{aligned}$$

■

Lemma IV.6 (Hilfssatz über konjugierte Richtungen) *$d_0, \ldots, d_{n-1}$ konjugierte Richtungen.*
Für jedes $x_0 \in \mathbb{R}^n$ liefert die durch

$$x_{i+1} = x_i + \alpha_i d_i, \quad \alpha_i = \frac{-g_i^T d_i}{d_i^T A d_i}, \quad g_i = Ax_i - b$$

für $i > 0$ erzeugte Folge nach höchstens n Schritten die Lösung $\tilde{x} = A^{-1}b$.

Beweis: Betrachte $A(\tilde{x} - x_0) = (b - Ax_0)$.
Mit Lemma IV.5: $(\tilde{x} - x_0) = \sum_{i=0}^{n-1} \alpha_i d_i$, $\alpha_i = \frac{d_i^T(b - Ax_0)}{d_i^T A d_i}$
Bleibt zu zeigen:

$$\frac{-g_i^T d_i}{d_i^T A d_i} \stackrel{!}{=} \frac{d_i^T(b - Ax_0)}{d_i^T A d_i}$$

Rechnung:

$$\begin{aligned}\alpha_i &= \frac{-d_i^T(Ax_0 - b)}{d_i^T A d_i} = \frac{-d_i^T(Ax_0 - Ax_i + Ax_i - b)}{d_i^T A d_i} \\ &= \frac{-d_i^T(Ax_i - b)}{d_i^T A d_i} - \underbrace{\frac{d_i^T(Ax_0 - Ax_i)}{d_i^T A d_i}}_{=0?}\end{aligned}$$

Laut Algorithmus:

$$\begin{aligned} x_i &= x_0 + \sum_{j=0}^{j<i} \alpha_j d_j \\ \Leftrightarrow \quad (x_i - x_0) &= \sum_{j=0}^{j<i} \alpha_j d_j \\ \Leftrightarrow \quad d_i^T A(x_i - x_0) &= \sum_{j=0}^{j<i} \alpha_j d_i^T A d_j = 0 \end{aligned}$$

Insgesamt: $\alpha_i = \frac{-d_i^T g_i}{d_i^T A d_i}$ ■

Korollar IV.1 *Voraussetzungen wie in Lemma IV.6.*
Dann minimiert die k-te Iterierte x_k die Funktion f in $x_0 + V_k$ mit $V_k = span[d_0, \ldots, d_{k-1}]$.
Insbesondere gilt $d_i^T g_k = 0$ für $i < k$.

Beweis:

1) Es genügt $d_i^T g_k = 0, \quad i < k$ zu zeigen.

$$\begin{aligned} f(x_k) = \min_{\alpha_i} \quad & f\left(x_0 + \sum_{i=0}^{i<k} \alpha_i d_i\right) \\ \Leftrightarrow \quad & \frac{\partial}{\partial \alpha_i} f(x_k) = 0 \\ \Leftrightarrow \quad & \nabla f(x_k)^T d_i = 0 \\ \Leftrightarrow \quad & (Ax_k - b)^T d_i = 0 \\ \Leftrightarrow \quad & g_k^T d_i = 0 \end{aligned}$$

2)

$$\begin{aligned} 0 \overset{!}{=} & d_k^T g_{k+1} && \text{zu zeigen} \\ = & d_k^T (Ax_{k+1} - b) \\ = & d_k^T \left(A\left(x_k - \frac{g_k^T d_k}{d_k^T A d_k} d_k\right) - b\right) && \text{Iterationsschrit aus Lemma IV.6} \\ = & d_k^T (Ax_k - b) - \frac{d_k^T A d_k}{d_k^T A d_k} g_k^T d_k \\ = & d_k^T g_k - g_k^T d_k = 0 \end{aligned}$$

3) Induktion (zu zeigen $d_i^T g_k = 0, \quad i < k$)
Anfang: $k = 1$ $d_0^T g_1 = 0$ (wegen 2)
Annahme: $d_i^T g_{k-1} = 0, \quad i < k-1$
Schritt: (zeige für k)
Benutze Algorithmus aus Lemma IV.6

$$\begin{aligned} & x_k - x_{k-1} = \alpha_{k-1} d_{k-1} \\ \Rightarrow \quad & A(x_k - x_{k-1}) = \alpha_{k-1} A d_{k-1} \\ \Rightarrow \quad & Ax_k - b - (Ax_{k-1} - b) = \alpha_{k-1} A d_{k-1} \\ \Rightarrow \quad & g_k - g_{k-1} = \alpha_{k-1} A d_{k-1} \end{aligned}$$

Also gilt für $i < k-1$
$d_i^T (g_k - g_{k-1}) = 0$ „konjugiert sein“
Wegen Induktionsannahme: $d_i^T g_{k-1} = 0$
$\Rightarrow \quad d_i^T g_k = 0$ für $i < k-1$
für $i = k-1$ benutze 2)
Insgesamt: $d_i^T g_k = 0$ für $i < k$ ■

Algorithmus IV.3 (cg-Verfahren)

1) Initialisierungsschritt: $x_0 \in \mathbb{R}^n$ als Startwert
Setze $d_0 = -g_0 = b - Ax_0$

2) Iteration über $k = 0, 1, \ldots$

$$\alpha_k = \frac{-g_k^T d_k}{d_k^T A d_k}$$
$$x_{k+1} = x_k + \alpha_k d_k$$
$$g_{k+1} = g_k + \alpha_k A d_k$$
$$\beta_k = \frac{g_{k+1}^T A d_k}{d_k^T A d_k}$$
$$d_{k+1} = -g_{k+1} + \beta_k d_k$$

Satz IV.3 (Eigenschaften des cg-Verfahrens) *Solange $g_{k-1} \neq 0$ gilt:*

1) $d_{k-1} \neq 0$

2) $V_k := span[g_0, Ag_0, A^2 g_0, \ldots, A^{k-1} g_0]$ („Krylov-Raum“)
Es ist $V_k = span[g_0, \ldots, g_{k-1}] = span[d_0, \ldots, d_{k-1}]$

3) $d_0, \ldots, d_{k-1}$ sind paarweise konjugiert

4) Es ist $f(x_k) = \min_{z \in V_k} (f(x_0 + z))$

Beweis: Induktion
Anfang: k=1 - klar (nach Rechnungen zu Gradientenverfahren)
Annahme: Satz IV.3 gelte bereits für k.
Schritt: (Verifiziere für $k+1$)
Zunächst: $g_k = g_{k-1} + \alpha_{k-1} A d_{k-1}$ (Algorithmus)
Wegen $span[g_0, Ag_0, A^2 g_0, \ldots, A^{k-1} g_0] = span[d_0, \ldots, d_{k-1}]$

$$d_{k-1} = \sum_{j=0}^{k-1} \gamma_j A^j g_0$$

Einsetzen:

$$g_k = \underbrace{g_{k-1}}_{\in V_k} + \alpha_{k-1} \sum_{j=1}^{k} \gamma_{j-1} A^j g_0$$

Somit: $span[g_0, \ldots, g_k] \subset V_{k+1}$.
***Nach Annahme**: $d_0, \ldots, d_{k-1}$ konjugiert*
wegen Optimalität von x_k: $d_i^T g_k = 0$ (Korollar IV.1)
Falls $g_k \neq 0 \quad \Rightarrow \quad g_k$ linear unabhängig von $(d_0, \ldots, d_{k-1}) \quad \Rightarrow \quad g_k \neq V_k$
Also ist $span[g_0, \ldots, g_k]$ ein $k+1$-dimensionaler Unterraum und kein echter Unterraum von V_{k+1}.
Zusammen mit $span[g_0, \ldots, g_k] \subset V_{k+1}$ folgt $span[g_0, Ag_0, \ldots, A^k g_0] = span[g_0, \ldots, g_k]$
Zeige nun, daß $V_{k+1} = span[d_0, \ldots, d_k]$
Algorithmus: $g_k + d_k = g_k - g_k + \beta_{k-1} d_{k-1} \Leftrightarrow \quad g_k + d_k = \beta_{k-1} d_{k-1}$

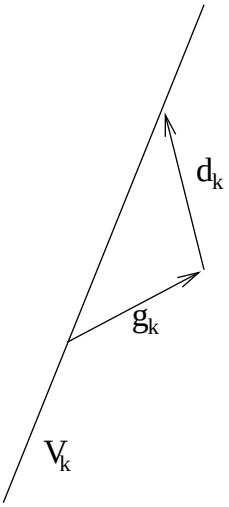

Abbildung IV.2: $g_k + d_k \in V_k$

Also: $g_k + d_k \in V_k$
Also: $span[g_0, \ldots, g_{k-1}, g_k] = span[g_0, \ldots, g_{k-1}, d_k] \overset{IA}{=} span[d_0, \ldots, d_{k-1}, d_k]$
Also: $V_{k+1} = span[d_0, \ldots, d_k]$
Zeige nun: $d_0, \ldots, d_k$ *sind paarweise konjugiert.*

$$\begin{aligned} \textit{Algorithmus:} \quad & d_k = -g_k + \beta_{k-1} d_{k-1} \\ \Rightarrow \quad & d_i^T A d_k = -d_i^T A g_k + \beta_{k-1} d_i^T A d_{k-1} \end{aligned}$$

1. Fall: $i < k-1$
Nach Induktionsannahme: $\beta_{k-1} d_i^T A d_{k-1} = 0$

$$\begin{aligned} \textit{Weiterhin:} \quad d_i \in V_{k-1} \quad & \Rightarrow \quad A d_i \in V_k \\ & \Rightarrow \quad A d_i = \sum_{j=0}^{k-1} \delta_j d_j \\ \textit{und damit} \quad & d_i^T A d_k = -(A d_i)^T g_k \\ & = -\sum_{j=0}^{k-1} \delta_j \underbrace{d_j^T g_k}_{=0,\ wg.\ Orthogonalität} \\ \curvearrowright \quad & d_i^T A g_k = 0 \end{aligned}$$

2. Fall: $i = k-1$
Wegen Algorithmus: $\beta_{k-1} = \frac{g_k^T A d_{k-1}}{d_{k-1}^T A d_{k-1}}$

$$d_{k-1}^T A d_k = -d_{k-1}^T A g_k + \frac{g_k^T A d_{k-1}}{d_{k-1}^T A d_{k-1}} d_{k-1}^T A d_{k-1} = 0$$

zu 4) Minimaleigenschaft:
Anwendung von Korollar IV.1 ■

Prinzip der Verfahren bisher:
Im Iterationsschritt k ist x_k in $x_0 + V_k$ enthalten.
Satz *Unter all diesen Verfahren liefert das cg-Verfahren den kleinsten Fehler $\|x_k - \tilde{x}\|_A$.*

Vorbereitung: Sei $p \in P_k$ (Polynome mit Maximalgrad k)

$$z \in \mathbb{R}: \quad p(z) = \sum_{i=0}^{k} \alpha_i z^i$$

$$\curvearrowright \text{ Übertragung auf } A \in M(n,n): \quad p(A) = \sum_{i=0}^{k} \alpha_i A^i$$

Satz IV.4 *Es gebe ein Polynom $p \in P_k$ mit $p(0) = 1$ und $|p(z)| \leq r \quad \forall \quad z \in \sigma(A)$ ($\sigma(A)$ =Menge aller Eigenwerte von A). Dann gilt für das cg-Verfahren:*

$$\|x_k - \tilde{x}\|_A \leq r \, \|x_0 - \tilde{x}\|_A$$

Beweis:

1) *Darstellung von $(y - \tilde{x})$; $\quad y \in x_0 + V_k$*

$$\text{Setze} \quad q(z) = \frac{p(z) - 1}{z} = \sum_{i=1}^{k} \gamma_i z^{i-1}$$

$$\curvearrowright \quad q(A) = \sum_{i=1}^{k} \gamma_i A^{i-1}$$

$$y = x_0 + \left(\sum_{i=1}^{k} \gamma_i A^{i-1} \right) g_0 \in x_0 + V_k$$

$$y = x_0 + q(A) g_0$$

Betrachte:

$$\begin{aligned} y - \tilde{x} &= x_0 - \tilde{x} + y - x_0 \\ &= (x_0 - \tilde{x}) + q(A) g_0 \\ &= (x_0 - \tilde{x}) + (p(A) - 1) A^{-1} A (x_0 - \tilde{x}) \end{aligned}$$

$$\Leftrightarrow \quad \boxed{y - \tilde{x} = p(A)(x_0 - \tilde{x})}$$

2) *Darstellung von $\|y - \tilde{x}\|_A^2$ und $\|x_0 - \tilde{x}\|_A^2$*
Sei $\{z_j\}_{j=1}^n$ ein ON-System aus Eigenvektoren zu A, d. h. $Az_j = \lambda_j z_j$.

Entwicklung: $x_0 - \tilde{x} = \sum_{j=1}^{n} c_j z_j$

Dann gilt

$$\begin{aligned} y - \tilde{x} = p(A)(x_0 - \tilde{x}) &= \sum_{j=1}^{n} c_j p(A) z_j \\ &= \sum_{j=1}^{n} c_j p(\lambda_j) z_j \end{aligned}$$

Berechne

$$\begin{aligned} \|x_0 - \tilde{x}\|_A^2 &= (x_0 - \tilde{x})^T A (x_0 - \tilde{x}) \\ &= \left(\sum_{j=1}^{n} c_j z_j \right)^T \left(\sum_{j=1}^{n} c_j \underbrace{A z_j}_{=\lambda_j z_j} \right) \\ &= \sum_{j=1}^{n} \lambda_j c_j^2 \qquad \text{wegen } z_j^T z_i = \begin{cases} 0; & i \neq j \\ 1; & i = j \end{cases} \end{aligned}$$

und analog

$$\|y - \tilde{x}\|_A^2 = \sum_{j=1}^{n} \lambda_j |c_j p(\lambda_j)|^2$$

Abschätzung:

$$\begin{aligned} \|y - \tilde{x}\|_A^2 &= \sum_{j=1}^{n} \lambda_j |c_j p(\lambda_j)|^2 \\ &\leq r^2 \sum_{j=1}^{n} \lambda_j |c_j|^2 \\ &= r^2 \|x_0 - \tilde{x}\|_A^2 \end{aligned}$$

3) Abschluss

$$\begin{aligned} & \|y - \tilde{x}\|_A \leq r \|x_0 - \tilde{x}\|_A , \quad y \in x_0 + V_k \\ \curvearrowright \quad & \|x_k - \tilde{x}\|_A \leq \|y - \tilde{x}\|_A \qquad \textit{(Minimaleigenschaft des cg-Verfahrens)} \\ & \leq r \|x_0 - \tilde{x}\|_A \end{aligned}$$

■

Bemerkung IV.6 (Tschebyscheff-Polynome)

übliche Definition: $T_k(x) := \cos(k \cdot \arccos x)$

äquivalent dazu: $T_k(x) := \frac{1}{2} \left\{ (x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k \right\}$

Eigenschaften:

- $T_k(1) = 1$
- $|T_k(x)| \leq 1$ *für* $-1 \leq x \leq 1$

Für cg-Analyse:

$$p(z) := \frac{T_k\left(\frac{(b+a)-2z}{b-a}\right)}{T_k\left(\frac{b+a}{b-a}\right)} \quad mit \quad 0 < a < b$$

Bemerke:

i) *Es gilt* $p(0) = 1$

ii) *für* $x \in [a,b]$ *gilt* $-1 \leq \frac{(b+a)-2z}{b-a} \leq 1$

Ab jetzt: $a = \lambda_{\min}$, $b = \lambda_{\max}$, $\kappa = \frac{b}{a}$

Satz IV.5 (Konvergenz des cg-Verfahrens)

$$\|x_k - \tilde{x}\|_A \leq 2\left(\frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}\right)^k \|x_0 - \tilde{x}\|_A$$

Beweis: Benutze $p(z)$*!*

$$\textit{Satz IV.4} \quad \curvearrowright \quad \|x_k - \tilde{x}\|_A \leq \max_{z\in[a,b]} p(z)\, \|x_0 - \tilde{x}\|_A$$

$$\curvearrowright \quad \|x_k - \tilde{x}\|_A \leq \max_{x\in[a,b]} \frac{\overbrace{T_k\left(\frac{b+a-2z}{b-a}\right)}^{\leq 1}}{T_k\left(\frac{b+a}{b-a}\right)} \|x_0 - \tilde{x}\|_A$$

Betrachte $T_k\left(\frac{(b+a)\frac{1}{a}}{(b-a)\frac{1}{a}}\right) = T_k\left(\frac{\kappa+1}{\kappa-1}\right)$

Für $z \geq 1$ *gilt:*

$$T_k(z) \geq \frac{1}{2}\left(z + \sqrt{z^2-1}\right)$$

$$\curvearrowright \quad \|x_k - \tilde{x}\|_A \leq \frac{1}{\frac{1}{2}\left(\left(\frac{\kappa+1}{\kappa-1}\right) + \sqrt{\left(\frac{\kappa+1}{\kappa-1}\right)^2 - 1}\right)^k} \|x_0 - \tilde{x}\|_A$$

wegen $\kappa - 1 = (\sqrt{\kappa}-1)(\sqrt{\kappa}+1)$

$$\frac{\kappa+1}{\kappa-1} + \sqrt{\frac{(\kappa+1)^2 - (\kappa-1)^2}{(\kappa-1)^2}} = \frac{\kappa+1+\sqrt{4\kappa}}{\kappa-1} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}$$

zusammen:

$$\|x_k - \tilde{x}\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \|x_0 - \tilde{x}\|_A$$

■

IV.6 Vorkonditionierung

Grundidee:

i) *Transformiere* $Ax = b \quad \curvearrowright \quad \tilde{A}\tilde{x} = \tilde{b}$ *mit* $cond(\tilde{A}) < cond(A)$

ii) *löse* $\tilde{A}\tilde{x} = \tilde{b}$

iii) *Rücktransformation* $\tilde{x} \to x$

IV.6.1 Transformation

Sei C symetrisch, positiv definit
$C = HH^T$ *(z. B. wegen Cholesky)*
Betrachte

$$\begin{aligned} & Ax = b \\ \Leftrightarrow \quad & \underbrace{H^{-1}AH^{-T}}_{\tilde{A}} \underbrace{H^T x}_{\tilde{x}} = \underbrace{H^{-1}b}_{\tilde{b}} \end{aligned}$$

Mit dieser Setzung: $\tilde{A}\tilde{x} = \tilde{b}$

IV.6.2 Zur Wahl von C

Betrachte Ähnlichkeitstransformation:

$$H^{-T}\tilde{A}H^T = H^{-T}(H^{-1}AH^{-T})H^T = C^{-1}A$$

Somit ist $\tilde{A}$ ähnlich zu $C^{-1}A$.
Falls $C = A$, dann ist $\tilde{A}$ ähnlich zu (I), das hieße $cond(\tilde{A}) = 1$.
Einfaches Beispiel für praktikables C: $C = \mathrm{diag}(A)$

Algorithmus IV.4 (Vorkonditioniertes cg-Verfahren)

1)

$$\begin{aligned} x_0 \in \mathbb{R}^n \quad & \curvearrowright \quad g_0 = Ax_0 - b \\ & \curvearrowright \quad d_0 = -h_0 = -C^{-1}g_0 \end{aligned}$$

2) *Iteration* $k = 0, 1, \ldots$

$$\begin{aligned} & x_{k+1} = x_k + \alpha_k d_k, && \alpha_k = \frac{g_k^T h_k}{d_k^T A d_k} \\ & g_{k+1} = g_k + \alpha_k A d_k \\ & h_{k+1} = C^{-1} g_{k+1} && \text{„Vorkonditionierung“} \\ & d_{k+1} = -h_{k+1} + \beta_k d_k, && \beta_k = \frac{g_{k+1}^T h_{k+1}}{d_k^T h_k} \end{aligned}$$

M Mehrgitteralgorithmus

Schlüsselpunkte: Löser für lineare Gleichungssysteme, wie sie bei der Diskretisierung partieller Differentialgleichungen auftreten.

- *Konvergenzrate ist* ***unabhängig*** *von h*
- *Zahl der Operationen $\mathcal{O}(n^1)$*

M.1 Idee

M.1.1 1D-Beispiel: $-u'' = f$

Diskretisierung: $A_h x_h = f_h$

$$\frac{1}{h}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \sim A_h \in \mathbb{R}^{n\times n}, \quad x_h, f_h \in \mathbb{R}^n$$

Iterative Löser, Jacobi-Verfahren:

$$x_h^{i+1} = x_h^i + D_h^{-1}(f_h - A_h x_h^i) \qquad \textit{(Defektkorrektur)}$$

i: Iterationsindex
$D_h = \mathrm{diag}(A_h)$
Beobachtung:
Betrachte $e_h^i = x_h - x_h^i$
„Der Fehler wird „glatter“ mit Anwendung des Jacobi-Verfahrens, aber nicht schnell kleiner.“

M.1.2 Gittertransfer

Prolongation*:* $p : \mathbb{R}^{n/2} \to \mathbb{R}^{n+1}$
n bezeichne die Zahl der inneren Punkte.
Definition via Abbildung M.1
Restriktion*:* $r : \mathbb{R}^{n+1} \to \mathbb{R}^{n/2}$
Definition via Abbildung M.2

M.1.3 Grobgitterkorrektur

Wir möchten lösen: $A_h x_h = f_h$
Nach i Schritten des Jacobi-Verfahrens x_h^i

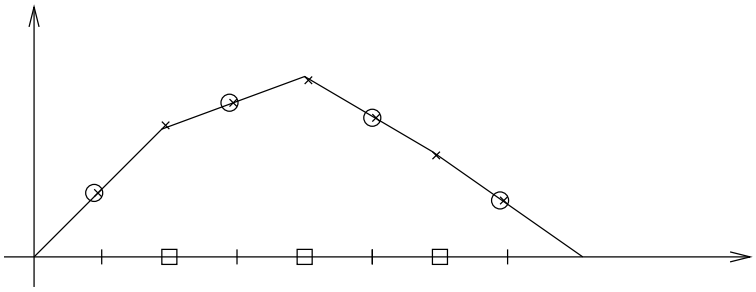

Abbildung M.1: 3 $\square$ werden abgebildet auf 7 |

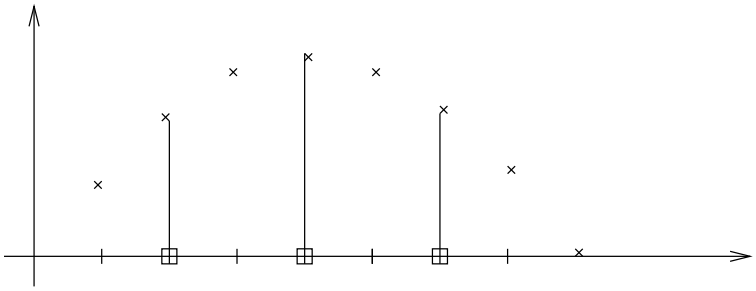

Abbildung M.2: 7 | werden abgebildet auf 3 $\square$

Wir rechnen:

$$A_h x_h - A_h x_h^i = \underbrace{f_h - A_h x_h^i}_{d_h^i} \qquad \textit{(Defekt)}$$

$$\Leftrightarrow \quad A_h \underbrace{(x_h - x_h^i)}_{e_h^i} = d_h^i$$

$$\Leftrightarrow \quad \boxed{A_h e_h^i = d_h^i}$$

Einschub: $x_h = x_h^i + (x_h - x_h^i) = x_h^i + e_h^i$
Restriktion liefert: $A_{2h} e_{2h} = r(d_h^i)$
Lösung dieses Problems ist „billiger“ !
Nun „Verbesserung der Iterierten“ gemäß

$$x_h^{i+1} = x_h^i + p(e_{2h})$$

M.1.4 Algorithmus

o) Anfangswerte: ${}^0x_h^0$
i) for $k = 0, 1, 2, \ldots$
for $i = 0, \ldots, \nu$ (ν *Glättungsschritte, z. B. 3-4)*
${}^k x_h^{i+1} = {}^k x_h^i + D_h^{-1}(f_h - A({}^k x_h^i))$
end

$A_{2h}e_{2h} = r(f_h - A_h{}^k x_h^\nu)$
${}^{k+1}x_h^0 = {}^k x_h^\nu + p(e_{2h})$

end

M.2 Glättung

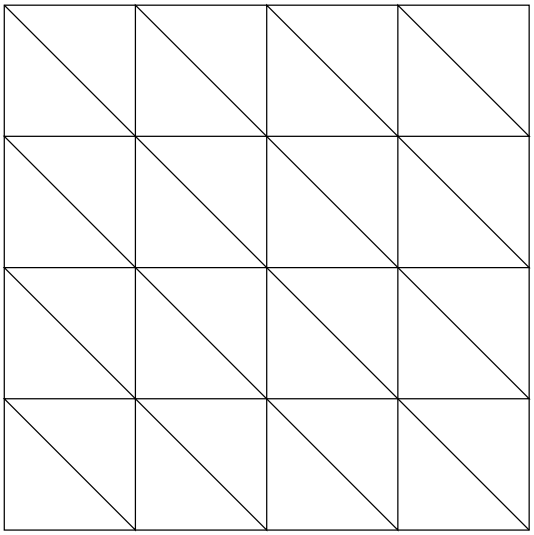

Abbildung M.3: 2D-Modellproblem

Jacobi-Verfahren

$$\begin{aligned} x^{i+1} &= x^i + D^{-1}(f - Ax^i) \\ &= (Id - D^{-1}A)x^i + D^{-1}f \end{aligned}$$

Iterationsmatrix $C = Id - D^{-1}A$
$h = \frac{1}{N+1}$ N^2 *Zahl der inneren Punkte*
Eigenwerte von C

$$\mu^{(k,l)} = \frac{1}{2}\left(\cos\frac{k\pi}{N+1} + \cos\frac{l\pi}{N+1}\right)$$

Eigenvektoren: z^{kl}

$$\begin{aligned} x^i - x &= \sum \alpha^{kl} z^{kl} \\ C(x^i - x) &= \sum \alpha^{kl}\mu^{kl} z^{kl} \end{aligned}$$

Schlechtes Konvergenzverhalten ist verursacht durch $\mu^{(1,1)}$
Hohe und niedrige Frequenzen
Eigenvektoren z^{kl} *haben die Einträge:*

$$z_{ij}^{kl} = \sin\frac{k\pi i}{N+1} \cdot \sin\frac{l\pi j}{N+1}$$

Der „schlechte" Eigenvektor ergibt sich für $k = l = 1$.
wie im 1D-Beispiel: „Niedrige Frequenzen machen Probleme."
***Frage**: Wie verhält sich das Jacobi-Verfahren auf Unterräumen mit „hohen" Freqzenzen? Definiere*

$$X_{OSC} = span\{z^{kl} : 1 \leq k, l \leq N, \quad \max(l,k) > N/2\}$$

„Mindestens eine hohe Frequenz"

***Lemma M.1** Angenommen $x^0 - x \in X_{OSC}$. Dann ist $e^m = x^m - x \in X_{OSC}$ und es gilt*

$$\|x^m - x\|_2 \leq \frac{1}{\sqrt{2}} \left\|e^{m-1}\right\|_2$$

Die Konvergenzrate ist h-unabhängig bei Einschränkung auf den hochschwingenden Raum X_{OSC}.

Beweis: „$\sum$" und „max" : $1 \leq k \leq N; \quad N/2 < l \leq N$
Darstellung des Iterationsfehlers e^m:

$$e^m = \sum \alpha_{kl} z^{kl} \quad \curvearrowright \quad \|e^m\|_2^2 = \sum \alpha_{kl}^2$$

Nun: $Cz^{kl} = \mu^{kl} z^{kl}$
Somit: i) $e^m \in X_{OSC}$
ii) Abschätzung des Fehlers $\|e^{m+1}\|_2$

$$\begin{aligned} \left\|e^{m+1}\right\|_2^2 &= \sum (\mu^{kl})^2 (\alpha_{kl})^2 \\ &\leq \max(\mu^{kl})^2 \sum (\alpha_{kl})^2 = \max(\mu^{kl})^2 \left\|e^m\right\|_2^2 \end{aligned}$$

Nun gilt wegen $l > N/2$

$$\begin{aligned} \max \mu^{kl} &\leq \frac{1}{2} \cos \frac{k\pi}{N+1} \leq \frac{1}{2} \\ \Rightarrow \quad \left\|e^{m+1}\right\|_2^2 &\leq \frac{1}{2} \left\|e^m\right\|_2^2 \end{aligned}$$

■

***Korollar M.1** Schreibe den Startfehler als $e^0 = e^0_{OSC} + e^0_{glatt}$ mit $e^0_{OSC} \in X_{OSC}$ und $e^0_{glatt} \in X_{glatt} = X^{\perp}_{OSC}$.*
Dann gilt nach m Schritten des Jacobi-Verfahrens:

$$e^m = e^m_{OSC} + e^m_{glatt} \quad mit \quad e^m_{OSC} = C^m e^0_{OSC} \in X_{OSC}; \quad e^m_{glatt} = C^m e^0_{glatt} \in X_{glatt}$$

und

$$\|e^m_{OSC}\|_2 \leq \left(\frac{1}{\sqrt{2}}\right)^m \left\|e^0_{OSC}\right\|$$

$\left\|e^m_{glatt}\right\|_2$ *„schrumpft" nur langsam. Daher interpretieren wir e^m als glatter.*

M.3 Hierarchie der Gleichungssysteme

Einbettung von $Ax = b$ in eine Familie von Gleichungssystemen.
Jeweils abhängig von $h = \frac{1}{N+1}$
Grobgitter habe die Schrittweite h_0.
Durch l-fache reguläre Verfeinerung: $h_l = \frac{h_0}{2^l}$
Der Index l heißt Stufenzahl.
Es gilt: $h_0 > h_1 > \ldots > h_{l-1} > h_l > \ldots$ mit $\lim_{l\to\infty} h_l = 0$
Jeder Schrittweite h_l, d. h. jeder Stufe l entspricht ein System: $A_l x_l = b_l$.

M.4 Gittertransfer

i) $T_{ij} = (ih_l, (i+1)h_l) \times (jh_l, (j+1)h_l)$ *(Abbildung M.4)*

ii) $V_l = \{\varphi \in C^0 | \quad \varphi|_{T_{ij}}$ *ist bilinear, d. h.* $\varphi|_{T_{ij}} = a + bx + cy + dxy\}$

iii) $f_l : \mathbb{R}^{n_l} \to V_l; \quad x_l \to u_l$ *mit* $u_l(ih_l, jh_l) = x_{ij}$

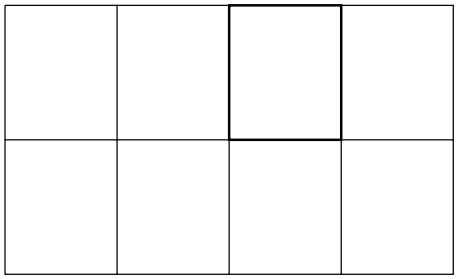

Abbildung M.4: $T_{ij} = (ih_l, (i+1)h_l) \times (jh_l, (j+1)h_l)$

Die Prolongation sei eine lineare injektive Abbildung vom groben in das feine Gitter:

$$
\begin{array}{lccc}
p: & \mathbb{R}^{n_{l-1}} & \to & \mathbb{R}^{n_l} \\
 & x_{l-1} & \to & x_l \\
f_{l-1} & \downarrow & & \uparrow \quad f_l^{-1} \\
 & V_{l-1} & \overset{I}{\to} & V_l
\end{array}
$$

$x_l = f_l^{-1} I f_{l-1} x_{l-1}$ *(I: Interpolation)*

<u>*Beispiel*</u>: $p : \mathbb{R}^1 \to \mathbb{R}^9$ *(Abbildung M.5)*

$$
(2) \to \begin{pmatrix} 0,5 \\ 1 \\ 0,5 \\ 1 \\ 2 \\ 1 \\ 0,5 \\ 1 \\ 0,5 \end{pmatrix}
$$

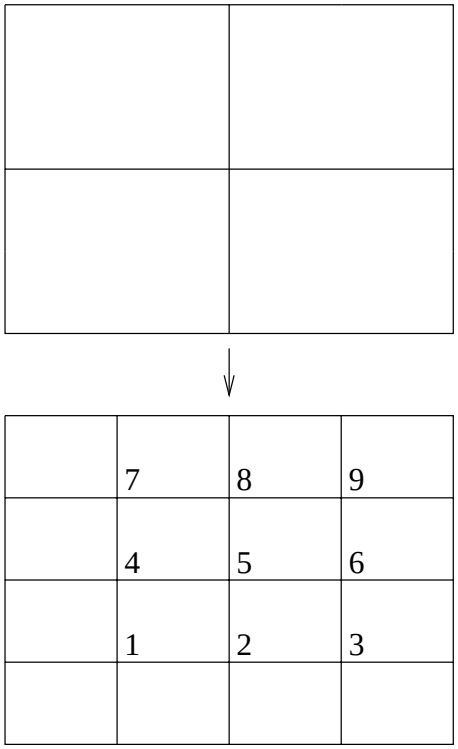

Abbildung M.5: $p : \mathbb{R}^1 \to \mathbb{R}^9$

Die Restriktion r ist eine lineare, surjektive Abbildung

$$
\begin{array}{lccc}
r: & \mathbb{R}^{n_l} & \to & \mathbb{R}^{n_{l-1}} \\
 & \mathbb{R}^{n_l} & \to & \mathbb{R}^{n_{l-1}} \\
f_l & \downarrow & & \uparrow \quad f_{l-1}^{-1} \\
 & V_l & \overset{r_{triv.}}{\to} & V_{l-1}
\end{array}
$$

$$x_{l-1} = f_{l-1}^{-1} r_{triv.} f_l x_l$$

M.5 Grobgitterkorrektur

$\phi_l^{GGK}: \quad x_l^{n+1} = x_l^n + pA_{l-1}^{-1} r(b_l - A_l x_l^n)$
Iterationsmatrix: $M_l^{GGK} = (Id - pA_{l-1}^{-1} r A_l)$

Bemerkung M.1 *Die Grobgitterkorrektur ϕ^{GGK} allein ist* ***nicht*** *konvergent.*

Beweis: Wegen $\mathbb{R}^{n_l} > \mathbb{R}^{n_{l-1}}$ ist der Kern von r nicht trivial.
Sei $0 \neq \xi \in Kern(r)$
Setze $\eta = A_l^{-1}\xi$

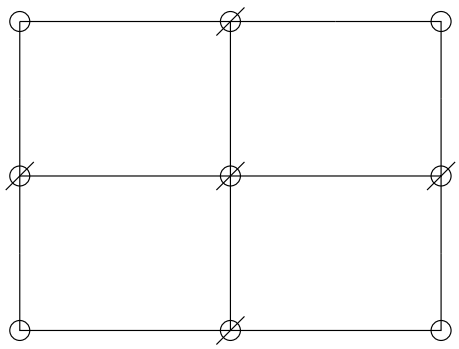

Abbildung M.6: zur Restriktion

Nun gilt

$$\begin{aligned} M_l^{GGK}\eta &= (Id - pA_{l-1}^{-1} r A_l)(A_l^{-1}\xi) \\ &= (Id)\eta - 0 = \eta \\ \Rightarrow \quad M_l^{GGK}\eta = \eta \quad &\Rightarrow \quad \rho(M_l^{GGK}) \geq 1 \qquad\qquad (\rho = \textit{Kontraktionszahl}) \end{aligned}$$

■

M.6 Zweigitterverfahren

Glättungsiteration (Jacobi)

$$\begin{aligned} S_l : x_l^{m+1} &= x_l^m + D_l^{-1}(b_l - A_l x_l^m) \\ \phi_l^{2GM} &:= \phi_l^{GGK} \circ S_l^\nu \end{aligned}$$

Algorithmus M.1

$$\begin{aligned} for \quad & i = 0, 1, 2, \ldots \\ & for \quad k = 0, 1, \ldots, \nu - 1 \\ & \qquad {}^i x_l^{k+1} = S_l({}^i x_l^k) \\ & d_{l-1} = r(b_l - A_l {}^i x_l^{\nu-1}) \\ & \boxed{e_{l-1} = A_{l-1}^{-1} d_{l-1}} \\ & {}^{i+1}x_l^0 = {}^i x_l^{\nu-1} + p(e_{l-1}) \end{aligned}$$

Algorithmus M.2 (Mehrgitteriteration)

$$
\begin{aligned}
&\textit{procedure} \quad \phi_l^{MGM(\nu)}(x_l, b_l)\\
&if(l=0) \quad then \quad x_0 := A_0^{-1} b_0; else\\
&\qquad begin\\
&\qquad\quad for \quad i := 1 \quad to \quad \nu \quad do \quad x_l := S_l(x_l, b_l);\\
&\qquad\quad d_{l-1} := r(b_l - A_l x_l);\\
&\qquad\quad e_{l-1}^0 := 0;\\
&\qquad\quad for \quad i := 1 \quad to \quad \gamma \quad do\\
&\qquad\qquad e_{l-1}^i := \phi_{l-1}^{MGM(\nu)}(e_{l-1}^{i-1}, d_{l-1});\\
&\qquad\quad x_l := x_l + p e_{l-1}^{\gamma};\\
&\qquad end;
\end{aligned}
$$

M.7 Rechenaufwand

„Bestandsaufnahme“:

$C_s n_l$ *Operationen zum Glätten (nur für dünn besetzte A)*

$\left.\begin{array}{l} C_D n_l \quad \textit{Operationen für die Restriktion} \\ C_C n_l \quad \textit{Operationen für die Prolongation} \end{array}\right\}$ *weil lokale Operationen*

Wir brauchen:

$$n_{l-1} \leq \frac{n_l}{C_h}$$

„Aus einer Zelle werden vier neue.“

Satz M.1 *Es gelten obige Voraussetzungen und es sei $\gamma < C_h$.*
Dann ist der Aufwand des Mehrgitterverfahrens $\mathcal{O}(n_l)$.

Beweis: Sei $C_l n_l$ der Aufwand für einen Schritt ϕ_l^{MGM}.
*(C_l ist a-priori **keine** Konstante)*
Es gilt $C_l n_l \leq (\nu C_s + C_D + C_C) \cdot n_l + \gamma C_{l-1} n_{l-1}$.
Wegen $n_{l-1} \leq \frac{n_l}{C_h}$:

$$
\begin{aligned}
& C_l n_l \leq (\nu C_S + C_D + C_C) n_l + \frac{\gamma}{C_h} C_{l-1} n_l\\
\curvearrowright \quad & C_l \leq (\nu C_S + C_D + C_C) + \frac{\gamma}{C_h} C_{l-1}
\end{aligned}
$$

Rekursion liefert

$$C_l \leq (\nu C_S + C_C + C_D) \left(1 + \frac{\gamma}{C_h} + \left(\frac{\gamma}{C_h}\right)^2 + \ldots + \left(\frac{\gamma}{C_h}\right)^{l-1}\right) + \gamma^l \frac{C_0}{n_l}$$

Bemerkung:

$$\frac{\gamma^l}{n_l} \leq \frac{\gamma^l}{C_h^l n_0}$$

Argumentation via „geometrische Reihe“. ■

Konvergenz? ***Ja!***

Stichworte:

- Glättungseigenschaft

- Approximationseigenschaft

V Adaptivität

V.1 Laplace-Problem

$$\begin{aligned} -\Delta u &= f \qquad auf \quad \Omega \\ u &= 0 \qquad auf \quad \partial\Omega \end{aligned} \tag{*}$$

$V = H_0^1(\Omega)$, Triangulierung $\mathbb{T}_h$ aus Dreiecken, $V_h \subset V$
linearer Ansatz:

$$\begin{aligned} u \in V &: (\nabla u, \nabla\varphi) = (f, \varphi) \quad \forall \quad \varphi \in V \\ u_h \in V_h &: (\nabla u_h, \nabla\varphi) = (f, \varphi) \quad \forall \quad \varphi \in V_h \end{aligned}$$

Satz V.1 (Energiefehlerschätzer) *Für den Diskretisierungsfehler $e = u - u_h$ in (*) gilt die a-posteriori-Abschätzung*

$$\|\nabla e\|^2 \leq C \sum_{T \in \mathbb{T}_h} \left(h_T^2 \rho_{1,T}^2 + h_T \rho_{2,T}^2 \right)$$

mit

$$\begin{aligned} \rho_{1,T} &= \|f + \Delta u_h\|_T \\ \rho_{2,T} &= \sum_{j=1}^{3} \int_{\partial T_j} \frac{1}{2} [\partial_n u_h]_{T_j}^2 \, d\Gamma \end{aligned}$$

und für $x \in \partial T_j$:

$$[\partial_n u_h](x) = (\partial_n u_h(x - \varepsilon n) - \partial_n u_h(x + \varepsilon n)) \qquad \text{für } \varepsilon \to 0$$

(siehe Abb. V.1)

<u>Einschub</u>: 2 Aspekte bei „Adaptivität“:

i. Fehlerschätzung

ii. „Gitterverfeinerung“ (und auch Vergröberung)

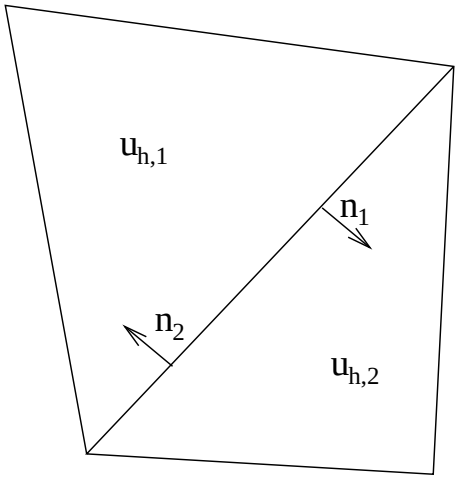

Abbildung V.1: $\partial_{n_1} u_{h,1} = \nabla u_{h,1} \cdot n_1; \quad \partial_{n_2} u_{h,2} = \nabla u_{h,2} \cdot n_2 = -\nabla u_{h,2} \cdot n_1$

Beweis: (Satz V.1)

i) Rechnung:

$$\begin{aligned}
\|\nabla e\|^2 &= (\nabla u - \nabla u_h, \nabla e) \\
&= (\nabla u - \nabla u_h, \nabla e - \nabla(I_h e)) \qquad \textit{(Galerkin-Orthogonalität)} \\
&= (f, e - I_h e) - \sum_{T \in \mathbb{T}_h} (\nabla u_h, \nabla e - \nabla(I_h e))_T \\
&= (f, e - I_h e) - \sum_{T \in \mathbb{T}_h} \left((-\Delta u_h, e - I_h e)_T + \int_{\partial T} (\partial_n u_h)(e - I_h e) d\Gamma \right) \\
&= \sum_{T \in \mathbb{T}_h} (f + \Delta u_h, e - I_h e)_T - \sum_{T \in \mathbb{T}_h} \sum_{j=1}^{3} \int_{\partial T_j} \frac{1}{2} [\partial_n u_h](e - I_h e) d\Gamma
\end{aligned}$$

ii) Wiederholung: Interpolation: Satz III.22

$$\|v - I_h v\|_{L_2(T)} \leq c h_T |v|_{H^1(T)}$$

***Spursatz**: (siehe Übung)*

$$\int_{\Gamma_i} v^2 d\Gamma \leq \frac{2}{r} \|v\|_0^2 + r |v|_1^2$$

mit dem Durchmesser des Gebietes r (hier: $r \approx h_T$)

iii) Abschätzung:

$$\|\nabla e\|^2 \leq \sum_{T\in\mathbb{T}_h} \|f+\Delta u_h\|_T \, ch_T \, \|\nabla e\|_T + \sum_{T\in\mathbb{T}_h} \left\|\frac{1}{2}[\partial_n u]\right\|_{\partial T} \|e - I_h e\|_{\partial T}$$

Anwendung des Spursatzes:

$$\begin{aligned}\|e-I_he\|_{\partial T}^2 &\leq \frac{2}{h_T}\|e-I_he\|_T^2 + h_T\|\nabla(e-I_he)\|_T^2 \\ &\leq \frac{2c}{h_T}\cdot h_T^2\|\nabla e\|_T^2 + h_T\|\nabla e\|_T^2 \quad \textit{(Stabilität der Interpolation)} \\ &\leq c\cdot h_T\|\nabla e\|_T^2\end{aligned}$$

Einsetzen:

$$\|\nabla e\|^2 \leq \sum_{T\in\mathbb{T}_h} \|f+\Delta u_h\|_T \, ch_T \, \|\nabla e\|_T + \sum_{T\in\mathbb{T}_h} \left\|\frac{1}{2}[\partial_n u_h]\right\|_{\partial T} \cdot c\sqrt{h_T}\,\|\nabla e\|_T$$

$$\curvearrowright \quad \|\nabla e\|^2 \leq \sqrt{\sum_{T\in\mathbb{T}_h} ch_T^2\|f+\Delta u_h\|_T^2}\cdot\sqrt{\sum_{T\in\mathbb{T}_h}\|\nabla e\|_T^2} + \sqrt{\sum_{T\in\mathbb{T}_h}\left\|\frac{1}{2}[\partial_n u_h]\right\|_{\partial T}^2 ch_T}\sqrt{\sum_{T\in\mathbb{T}_h}\|\nabla e\|_T^2}$$

(Teilen durch $\|\nabla e\|$, dann quadrieren)

$$\curvearrowright \quad \|\nabla e\|^2 \leq C\left(\sum_{T\in\mathbb{T}_h} h_T^2\|f+\Delta u_h\|_T^2 + \sum_{T\in\mathbb{T}_h}\left\|\frac{1}{2}[\partial_n u_h]\right\|_{\partial T}^2 h_T\right)$$

■

V.5 Dualitätsargument

Wieder Laplace:
Fehlerabschätzungen

$$\begin{aligned}\|u-u_h\|_{H^1(\Omega)} &\leq c\cdot h|u|_{H^2(\Omega)} \\ \|u-u_h\|_{L^2} &\leq c\cdot h|u|_{H^2(\Omega)} \qquad (\textit{„suboptimal“})\end{aligned}$$

Andererseits: $\|u-I_hu\|_{L^2(\Omega)} \leq ch^2|u|_{H^2(\Omega)}$

V.5.1 A-priori-Abschätzung

Satz V.4 *Sei Ω ein konvexes, polygonales Gebiet.*

$$\begin{aligned} u: &\quad (\nabla u, \nabla\varphi) = (f,\varphi) \quad \forall \quad \varphi\in V \\ u_h: &\quad (\nabla u_h, \nabla\varphi) = (f,\varphi) \quad \forall \quad \varphi\in V_h\end{aligned}$$

Dann gibt es ein $c>0$, unabhängig von u und h, so daß gilt:

$$\|u-u_h\|_{L^2(\Omega)} \leq c\cdot h^2|u|_{H^2(\Omega)}$$

Beweis: Betrachte das Hilfsproblem („duales Problem“)

$$-\Delta z = e := u - u_h \quad auf \quad \Omega$$
$$z = 0 \qquad\qquad auf \quad \partial\Omega$$

***Bemerkung**: (Stabilität des dualen Problems) Man kann zeigen, falls Ω konvex, daß gilt: $\|z\|_{H^2(\Omega)} \leq C_s \|e\|_{L^2(\Omega)}$ und C_s ist unabhängig von e.*
Schreibe duales Problem in variationeller Formulierung:

$$-(\varphi, \Delta z) = (\varphi, e) \quad \forall \quad \varphi \in V$$
$$\Leftrightarrow \quad (\nabla\varphi, \nabla z) = (\varphi, e) \quad \forall \varphi \in V$$

Wähle nun speziell $\varphi = e$:

$$\begin{aligned}
(e,e) &= (\nabla e, \nabla z) \\
&= (\nabla e, \nabla z - \nabla I_h z) && \textit{(Galerkin-Orthogonalität)} \\
&\leq \|\nabla e\| \, \|\nabla z - \nabla I_h z\| && \textit{(Cauchy-Schwarz)} \\
&\leq ch|u|_{H^2(\Omega)} c \cdot h|z|_{H^2(\Omega)} && \textit{(Energiefehler; Interpolation)} \\
&\leq ch|u|_{H^2(\Omega)} c \cdot hC_s \|e\|_{L^2(\Omega)} && \textit{(Stabilität des dualen Problems)}
\end{aligned}$$

$$\curvearrowright \quad \|e\|_{L^2(\Omega)} \leq ch^2 |u|_{H^2(\Omega)}$$

■

V.5.2 A-posteriori-Abschätzung

Ausgangspunkt: duales Problem

$$z \in V: \quad J(\varphi) = (\nabla\varphi, \nabla z) \qquad \forall \quad \varphi \in V$$

J lineares Funktional
Beispiel:

$$J(\varphi) = \int_\Omega \varphi dx; \qquad J(\varphi) = \int_\Gamma \varphi dx$$

Wähle $\varphi = e$: $J(e) = (\nabla e, \nabla z)$
Approximation des dualen Problems:

$$\begin{aligned}
\tilde{z} \in \tilde{V} \subset V: \quad J(\varphi) &= (\nabla\varphi, \nabla\tilde{z}) \qquad \forall \quad \varphi \in \tilde{V} \\
J(e) &= (\nabla e, \nabla z - \nabla\tilde{z}) + (\nabla e, \nabla\tilde{z} - \nabla I_h\tilde{z}) \\
&\leq \underbrace{\|\nabla e\| \cdot \|\nabla z - \nabla\tilde{z}\|}_{\textit{rigoros kontrollierbar mit Energiefehlerschätzer}} + (\nabla e, \nabla\tilde{z} - \nabla I_h\tilde{z})
\end{aligned}$$

Behandlung von

$$\begin{aligned}
&(\nabla e, \nabla\tilde{z} - \nabla I_h\tilde{z}) \\
&\quad = (f, \tilde{z} - I_h\tilde{z}) - (\nabla u_h, \nabla\tilde{z} - \nabla I_h\tilde{z}) \qquad \textit{(für Schätzung)} \\
&\quad \leq \sum_T \|f + \Delta u_h\|_T \, \|\tilde{z} - I_h\tilde{z}\|_T + \sum_T \left\| \frac{1}{2}[\partial_n u_h] \right\|_{\partial T} \|\tilde{z} - I_h\tilde{z}\|_{\partial T} \qquad \textit{(für die Gittersteuerung)}
\end{aligned}$$

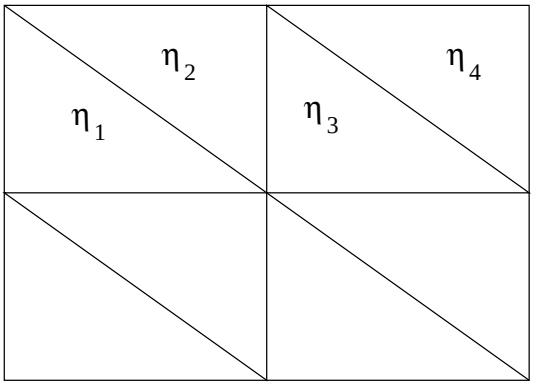

Abbildung V.2: Fehler auf den Gitterdreiecken

Verschiedene Verfeinerungsstrategien (Abb. V.2)

i) *Bestimme* $\max_i \eta_i =: \eta_{max}$
Verfeinere alle Zellen mit $\eta_i > r\eta_{max}$, $0 < r < 1$

ii) *Sortiere Zellen nach Fehlerbeitrag* η_i
Verfeinere dann zum Beispiel die ersten 30% *der Zellen*

VI Parabolische Probleme

Modellbeispiel:

$$\partial_t u - \Delta u = f \qquad auf \quad \Omega \times I$$

I ist das Zeitinterval $[0;T]$
$u = u(t,x), \quad f = f(t,x)$
Randbedingungen und Anfangsbedingungen:

$$u(t,x) = 0, \qquad t \in I, \quad x \in \partial\Omega$$
$$u(0,x) = u_0(x), \qquad x \in \Omega$$

***Numerische Behandlung**:*
***1) Zeitdiskretisierung**:*
Zerlegung von I in Stützstellen
$(t_0 = 0, t_1, t_2, \ldots, t_N = T)$
Schreibweise: $u^n = u(t_n, x)$
$I_n = (t_{n-1}, t_n) \qquad k_n = t_n - t_{n-1}$
$a(v,w) = (\nabla v, \nabla w)$
Schwache Formulierung:

$$\int_{I_n} (\partial_t u, \varphi) dt + \int_{I_n} a(u,\varphi) dt = \int_{I_n} (f,\varphi) dt \qquad \forall \quad \varphi \in V$$

Approximation:

$$\int_{I_n} a(u,\varphi)dt \approx k_n \left((1-\alpha)a(u^{n-1},\varphi) + \alpha a(u^n,\varphi)\right)$$

Analog für $\int_{I_n}(f,\varphi)dt$
Insgesamt:

$$(u^n - u^{n-1},\varphi) + k_n\alpha a(u^n,\varphi) = \int_{I_n}(f,\varphi)dt - k_n(1-\alpha)a(u^{n-1},\varphi)$$

$$\curvearrowright \quad (u^n,\varphi) + k_n\alpha a(u^n,\varphi) = \alpha k_n(f^n,\varphi) + (1-\alpha)(f^{n-1},\varphi) + (u^{n-1},\varphi) - k_n(1-\alpha)a(u^{n-1},\varphi)$$

***2) Ortsdiskretisierung**: Benutze FE zur Approximation von u^n*
Numerische Schemata:

i) Implizites Rückwärts-Euler-Verfahren

$$\left(\frac{u_h^n - u_h^{n-1}}{k_n},\varphi\right) + a\left(u_h^n,\varphi\right) = \left(f^n,\varphi\right) \qquad \forall \quad \varphi \in V_h \subset V$$

Genauigkeit in der Zeit: $\mathcal{O}(\Delta t)$ *mit* $\Delta t = \max_n k_n$

ii) Crank-Nicholson:

$$\left(\frac{u_h^n - u_h^{n-1}}{k_n},\varphi\right) + a\left(\frac{u_h^n + u_h^{n-1}}{2},\varphi\right) = \left(\frac{f^n + f^{n-1}}{2},\varphi\right) \qquad \forall \quad \varphi \in V_h$$

Genauigkeit: $\mathcal{O}(\Delta t^2)$

iii) Implizites Zweischrittverfahren:

$$\left(\frac{\frac{3}{2}u_h^n + \frac{3}{2}u_h^{n-1}}{k_n} - \frac{\frac{1}{2}u_h^{n-1} - \frac{1}{2}u_h^{n-2}}{k_{n-1}},\varphi\right) + a(u_h^n,\varphi) = (f^n,\varphi) \qquad \forall \quad \varphi \in V_h$$

$u_h^0 = u_{0h}$*;* u_h^1 *gegeben durch einen „Euler-Schritt“*
Genauigkeit: $\mathcal{O}(\Delta t^2)$

Übertragung auf Variationsungleichungen:
Aufgabe:

$$u(t) \in K \subset V, \qquad \textit{so daß } (\partial_t u, \varphi - u) + a(u, \varphi - u) \geq (f, \varphi - u) \quad \forall \quad \varphi \in K$$

+ Anfangs- und Randbedingungen
Schema:

$$\left(\frac{u_h^n - u_h^{n-1}}{k_n},\varphi - u_h^n\right) + a(u_h^n, \varphi - u_h^n) \geq (f^n, \varphi - u_h^n)$$

(„Rückwärts Euler“)

Verwende NICHT Crank-Nicolson

$$a\left(\frac{u_h^n + u_h^{n-1}}{2},\varphi\right) \quad \curvearrowright \quad a\left(\frac{u_h^n + u_h^{n-1}}{2},\varphi - \frac{u_h^n + u_h^{n-1}}{2}\right)$$

gefährlich, falls $K = K(t)$

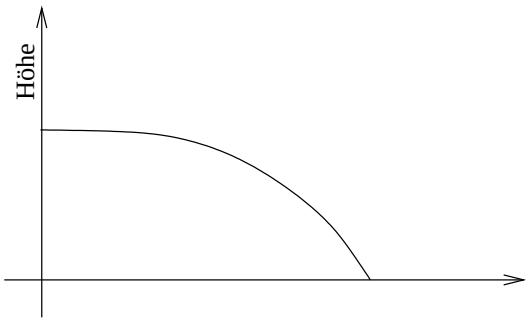

Abbildung VI.1: Beispiel: Gletscherentwicklung

VII Sattelpunktprobleme

Motivation:

$$\begin{aligned} \Delta u &= -f \qquad \text{auf } \Omega \\ u &= 0 \qquad \text{auf } \partial\Omega \end{aligned}$$

Substitution: $\sigma = \nabla u$
σ*: Spannung*
$\sigma = (\sigma_1, \sigma_2) = (\partial_{x_1} u, \partial_{x_2} u)$
Bemerkung: $\Delta u = div\ \nabla u$
Somit als System:

$$\boxed{\begin{array}{ll} \sigma & = \nabla u \\ div\ \sigma & = -f \end{array}} \quad \begin{array}{ll} \leftarrow & \textit{Materialgesetz} \\ \leftarrow & \textit{Kräftegleichgewicht} \end{array}$$

Umformulierung: $\frac{1}{2}\int \sigma^2 = \min$ *unter der Nebenbedingung* $div\ \sigma = -f$.
Aufgabe: Variationsprobleme mit Nebenbedingungen
Bezeichungen: X, M *Hilberträume*
$\left.\begin{array}{ll} a: & X \times X \to \mathbb{R} \\ b: & X \times M \to \mathbb{R} \end{array}\right\}$ *stetige Bilinearformen*
$X', \quad M'$ *Dualräume zu* X *und* M
entsprechende Paarungen werden mit $\langle \cdot, \cdot \rangle$ *bezeichnet*
Weiterhin gegeben $f \in X', \quad g \in M'$
Problem: Gesucht wird in X *das Minimum von*

$$J(u) = \tfrac{1}{2} a(u,u) - \langle f, u \rangle \qquad \textit{(PM)}$$

unter der Nebenbedingung $b(u,q) = \langle g, q \rangle \quad \forall \quad q \in M$
(Bezug zu Anwendung: $u \sim \sigma$*)*
Umformulierung: Betrachte die Lagrange-Funktion

$$\mathcal{L}(u, \lambda) := J(u) + [b(u, \lambda) - \langle g, \lambda \rangle]$$

Dies führt auf das Sattelpunktproblem: Problem (PS) (vgl. Übung)
Gesucht wird $(u, \lambda) \in X \times M$ *mit*

$$\begin{aligned} a(u,\varphi) + b(\varphi,\lambda) &= \langle f, \varphi\rangle \qquad \forall \quad \varphi \in X \\ b(u,q) &= \langle g, q\rangle \qquad \forall \quad q \in M \end{aligned}$$

Für jede Lösung (u, λ) *von (PS) kann man die „Sattelpunkteigenschaft“*

$$\mathcal{L}(u,q) \le \mathcal{L}(u,\lambda) \le \mathcal{L}(\varphi,\lambda)$$

nachrechnen. (vgl. Übung)
Herleitung einer „schwachen Formulierung“ für

$$\begin{aligned} \sigma &= \nabla u \\ div\ \sigma &= -f \end{aligned}$$

Wir „testen“ :

$$\begin{aligned} (\sigma,\tau) &= (\nabla u,\tau) \qquad \forall \quad \tau \in L_2(\Omega)^2 \\ (div\ \sigma,\varphi) &= (-f,\varphi) \qquad \forall \quad \varphi \in H_0^1(\Omega) \end{aligned}$$

Partielle Integration:

$$\begin{aligned} (\sigma,\tau) - (\nabla u,\tau) &= 0 \qquad \forall \quad \tau \in L^2(\Omega)^2 \\ -(\sigma,\nabla\varphi) &= -(f,\varphi) \qquad \forall \quad \varphi \in H_0^1(\Omega) \end{aligned}$$

In diesem Beispiel: $u \sim \sigma$; $\lambda \sim u$
$X = (L_2(\Omega))^2, \quad M = H_0^1(\Omega)$
$a(\sigma,\tau) = (\sigma,\tau) = \int\limits_\Omega (\sigma_1\tau_1 + \sigma_2\tau_2)dx$
$b(\tau,\varphi) = -(\tau,\nabla\varphi)$
(Im Skript: Bsp VII.2)

Beispiel VII.1

$$\begin{aligned} -\Delta u &= f \qquad auf\ \Omega \\ u &= g \qquad auf\ \partial\Omega \end{aligned}$$

Mit $X = H^1(\Omega)$ *und* $M = L_2(\partial\Omega)$
Definiere $a(u,\varphi) := \int\limits_\Omega \nabla u \nabla\varphi dx$

$$\langle f,\varphi\rangle = \int\limits_\Omega f\cdot\varphi dx = (f,\varphi)$$

$$b(\varphi,q) = \int\limits_{\partial\Omega} \varphi\cdot q d\Gamma$$

$$\langle g,q\rangle = \int\limits_{\partial\Omega} gq d\Gamma$$

Nebenbedingung: $\int\limits_{\partial\Omega} uq d\Gamma = \int\limits_{\partial\Omega} g\cdot q d\Gamma$

VII.1 Hilfsmittel aus der Funktionalanalysis

VII.1.1 Adjungierte Operatoren

***Bezeichnungen**:*

- *X, Y Banachräume*
- *X', Y' Dualräume*
- *$\langle\cdot,\cdot\rangle$ Paarungen zwischen $\begin{matrix} X \\ Y \end{matrix}$ und $\begin{matrix} X' \\ Y' \end{matrix}$*
- *$L: \quad X \to Y$ beschränkter, linearer Operator*

„Bastelaufgabe“ : $L': \quad Y' \to X'$ a la

$$\begin{array}{ccc} X & \overset{L}{\to} & Y \\ \big|_{\langle\cdot,\cdot\rangle} & & \big|_{\langle\cdot,\cdot\rangle} \\ X' & \leftarrow & Y' \end{array}$$

Konstruktion:

1) *Vorgabe $y^* \in Y'$*
2) *$x \in X$ wird abgebildet durch Einsetzen in $\langle y^*, L\cdot\rangle$*
   $X \ni x \to \langle y^*, Lx\rangle \in \mathbb{R}$

Wir haben also nach Vorgabe eines $y^* \in Y'$ durch $\langle y^*, L\cdot\rangle$ ein Element aus X' konstruiert. Dieses Element wird als $L'y^* \in X'$ bezeichnet.

$$\langle L'y^*, x\rangle := \langle y^*, Lx\rangle$$

***Definition VII.1** $L': Y' \to X'$ heißt der zu L **adjungierte Operator**.*

***Definition VII.2** Sei V ein abgeschlossener Unterraum von X. Dann heißt*

$$V^0 := \{l \in X' | \langle l, v\rangle = 0 \quad \forall \quad v \in V\}$$

*die **Polare** von V.*

***Definition VII.3** Sei X ein Hilbertraum. Dann heißt*

$$V^\perp := \{x \in X | (x, v) = 0 \quad \forall \quad v \in V\}$$

***orthogonales Komplement**.*

***Satz VII.1 (closed range-theorem)** Mit obigen Voraussetzungen gilt:*

i) $L(X)$ abgeschlossen in Y
   $\Leftrightarrow$

ii) $L(X) = (Kern \quad L')^0$

(ohne Beweis)

VII.1.2 Abstrakter Existenzsatz

***Bezeichnungen**:*

- *U, V Hilberträume*
- *U', V' Dualräume*
- *$a : U \times V \to \mathbb{R}$ eine Bilinearform*

Definiere einen Operator: $L : U \to V'$

$$U \ni u \to \langle Lu, \underset{\text{Einsetzen von } v \in V}{\cdot} \rangle := a(u, \cdot)$$

Variationsprobleme hatten die Struktur:

$$\begin{aligned} & a(u,v) = \langle f, v \rangle \quad \forall \quad v \in V \qquad (*) \\ \to \quad & \langle Lu, v \rangle = \langle f, v \rangle \quad \forall \quad v \in V \end{aligned}$$

formal $u = L^{-1} f$

Definition VII.4 *Seien U, V normierte Räume. Eine bijektive, lineare Abbildung $L : U \to V'$ heißt **Isomorphismus**, wenn L und L^{-1} stetig sind.*

Satz VII.2 *Seien U, V Hilberträume. Eine lineare Abbildung $L : U \to V'$ ist ein Isomorphismus, wenn die zugehörige Bilinearform $a : U \times V \to \mathbb{R}$ folgende Bedingung erfüllt:*

i) *(Stetigkeit) $\exists c > 0$ mit $|a(u,v)| \leq c \|u\|_U \|v\|_V$*

ii) *(Inf-Sup-Bedingung)*

$$\exists \quad \alpha > 0 \quad mit \quad \sup_{v \in V; v \neq 0} \frac{a(u,v)}{\|v\|_V} \geq \alpha \|u\|_U \qquad \forall \quad u \in U$$

bzw.

$$\boxed{\inf_{u \in U} \sup_{v \in V} \frac{a(u,v)}{\|u\|_U \|v\|_V} \geq \alpha > 0}$$

iii) $\forall \quad v \in V \qquad \exists \quad u \in U$ *mit* $a(u,v) \neq 0$

<u>*Beweis*</u>*:*

(1) *aus ii) folgt „L injektiv“ :*
$Lu_1 = Lu_2 \quad \curvearrowright \quad a(u_1, v) = a(u_2, v) \quad \forall \quad v \in V$
Also ist $\sup_{v \in V; v \neq 0} \frac{a(u_1 - u_2, v)}{\|v\|_V} = 0 \overset{ii)}{\geq} \alpha \|u_1 - u_2\|_U$
$\Rightarrow \quad u_1 = u_2$

(2) *„L^{-1} ist stetig auf dem Bild von L“*
zu $f \in L(U)$ gibt es wegen (1) ein eindeutiges $u = L^{-1}f$.

$$\text{Rechnung: } \alpha \left\| L^{-1}f \right\|_U = \alpha \|u\|_U \leq \sup_{v \in V} \frac{a(u,v)}{\|v\|_V}$$
$$= \sup_{v \in V} \frac{\langle Lu, v \rangle}{\|v\|_V} = \sup_{v \in V} \frac{\langle f, v \rangle}{\|v\|_V} = \|f\|_{V'}$$
$$\Rightarrow \quad \frac{\|L^{-1}f\|_U}{\|f\|_{V'}} \leq \frac{1}{\alpha}$$

$f \in L(U)$ war beliebig: d. h. $\|L^{-1}\| = \sup_{f \in V'} \frac{\left\|L^{-1}f\right\|_U}{\|f\|_{V'}} \leq \frac{1}{\alpha}$

(3) *„$L(U)$ ist abgeschlossen“*
L^{-1} ist stetig nach (2). L ist stetig (siehe Übung 11.3 und i))
$\Rightarrow$ $L(U)$ abgeschlossen

(4) *„L ist surjektiv“*
$L' : V \to U'$
$V \ni v \to \langle L\cdot, v \rangle = a(\cdot, v)$
Wegen ii) liegt nur $v = 0$ in Kern(L').
Satz VII.1 sagt: $L(U) = (\text{Kern } L')^0 = \{l \in V' | \langle l, v \rangle = 0 \quad \forall v \in \text{Kern } L'\} = V'$

■

VII.1.3 Abstrakter Konvergenzsatz

Seien $U_h \subset U$ und $V_h \subset V$ endlichdimensionale Räume.
Diskrete Aufgabe: $a(u_h, v) = \langle f, v \rangle \qquad \forall \quad v \in V_h \qquad (**)$

Satz VII.3 *Bezeichnungen und Bedingungen wie in Satz VII.2. Weiterhin seien $U_h \subset U$, $V_h \subset V$ so gewählt, daß gilt*

ii_h) $\inf_{u_h \in U_h} \sup_{v_h \in V_h} \frac{a(u_h, v_h)}{\|u_h\|_U \|v_h\|_V} \geq \alpha$ *mit h-unabhängigem α*

iii_h) $\forall \quad v_h \in V_h \quad \exists \quad u_h \in U_h$ *mit* $a(u_h, v_h) \neq 0$ *(Es gibt ein diskretes u_h.)*

Dann gilt die Fehlerabschätzung

$$\|u - u_h\|_U \leq \left(1 + \frac{c}{\alpha}\right) \inf_{w_h \in U_h} \|u - w_h\|_U$$

Beweis: (*)-(**) liefert die Galerkin-Eigenschaft:

$$a(u - u_h, v) = 0 \qquad \forall \quad v \in V_h$$

$$\curvearrowright \quad a(u-w_h,v)=a(u_h-w_h,v) \quad \forall \quad v\in V \qquad \oplus$$

$w_h \in U_h$ beliebig gewählt.
Wir schätzen ab:

$$\begin{aligned}
\alpha \left\|u_h-w_h\right\| &\overset{ii_h)}{\leq} \sup_{v_h\in V_h} \frac{a(u_h-w_h,v_h)}{\left\|v_h\right\|_V} \\
&\overset{\oplus}{=} \sup_{v\in V_h} \frac{a(u-w_h,v_h)}{\left\|v_h\right\|_V} \\
&\overset{V_h\subset V}{\leq} \sup_{v\in V} \frac{a(u-w_h,v)}{\left\|v_h\right\|_V} \\
&\overset{\text{Satz VII.2 i)}}{\leq} \frac{c\left\|u-w_h\right\|_U \left\|v\right\|_V}{\left\|v\right\|_V} \\
&= c\left\|u-w_h\right\|_U \\
\curvearrowright \quad \left\|u_h-w_h\right\| &\leq \frac{c}{\alpha}\left\|u-w_h\right\|_U
\end{aligned}$$

Schließlich „Dreiecksungleichung“:

$$\begin{aligned}
\left\|u-w_h+w_h-u_h\right\|_U &\leq \left\|u-w_h\right\|_U + \left\|w_h-u_h\right\|_U \\
&\leq \left(1+\frac{c}{\alpha}\right)\left\|u-w_h\right\|_U
\end{aligned}$$

w_h war beliebig: Übergang zu

$$\left\|u-u_h\right\|_U \leq \inf_{w_h\in U_h} \left(1+\frac{c}{\alpha}\right)\left\|u-w_h\right\|_U$$

■

VII.2 Die „Inf-Sup-Bedingung“

Aufgabe:

$$\begin{aligned}
(u,\lambda)\in X\times M: \quad a(u,\varphi)+b(\varphi,\lambda) &= \langle f,\varphi\rangle \quad \forall \quad \varphi\in X \\
b(u,q) &= \langle g,q\rangle \quad \forall \quad q\in M
\end{aligned} \tag{PS}$$

Abstrakte Situation:

$$\begin{array}{ccccc}
L: & \underbrace{X\times M}_{\| \atop U} & \to & \underbrace{X'\times M'}_{\| \atop V'} \\
& \ni (u,\lambda) & \to & (f,g)\in
\end{array} \tag{PA}$$

Satz VII.4 *Durch (PS) wird (PA) genau dann ein Isomorphismus $L: X\times M \to X'\times M'$ erklärt, wenn die beiden folgenden Bedingungen erfüllt sind:*

i) Die Bilinearform a ist elliptisch auf dem Raum $V=\{v\in X | b(v,q)=0 \quad \forall q\in M\}$, also $a(v,v)\geq \alpha\left\|v\right\|^2$ für $v\in V\subset X$ und $\alpha>0$

ii) $\boxed{\exists \quad \beta > 0 \quad mit \inf_{q \in M} \sup_{v \in X} \frac{b(v,q)}{\|v\| \, \|q\|} \geq \beta}$

Beweisbemerkung: Die abstrakte inf-sup-Bedingung in Satz VII.2 kann ausgedrückt werden durch $a(\cdot,\cdot)$ und $b(\cdot,\cdot)$.

Definition VII.5 (Gemischte FE-Methoden) *Wie immer: $X_h \subset X$, $M_h \subset M$. Man löst:*

$$a(u_h, \varphi) + b(\varphi, \lambda_h) = \langle f, \varphi \rangle \quad \forall \varphi \in X_h$$
$$b(u_h, q) = \langle g, q \rangle \quad \forall q \in M_h$$

Definition VII.5 *Eine Familie von FE-Räumen $X_h \times M_h$ erfüllt die **Babuška-Brezzi-Bedingung**, wenn es von h unabhängige Zahlen $\alpha > 0$ und $\beta > 0$ mit den folgenden eigenschaften gibt:*

i_h) *$a(\cdot,\cdot)$ ist V_h-elliptisch, d. h. mit $\alpha > 0$ und $V_h = \{v_h \in X_h | b(v_h, q_h) = 0 \quad \forall q_h \in M_h\}$ gilt: $a(v_h, v_h) \geq \alpha \|v_h\|^2 \quad \forall v_h \in V_h$*

ii_h) *Mit $\beta > 0$ gilt: $\inf_{q_h \in M_h} \sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\| \|q_h\|} \geq \beta$*

Satz VII.5 *Voraussetzungen wie in Satz VII.4 und X_h, M_h erfüllen die BB-Bedingung. Dann gilt:*

$$\|u - u_h\|_X + \|\lambda - \lambda_h\|_M \leq c \left\{ \inf_{v_h \in X_h} \|u - v_h\|_X + \inf_{q_h \in M_h} \|\lambda - q_h\|_M \right\}$$

Beweisbemerkung: Anwendung des abstrakten Konvergenzsatzes VII.3.

VII.3 Diskrete Sattelpunktprobleme

Die Diskretisierung von (PS) führt auf das System

$$Au + B^T \lambda = f$$
$$Bu = g$$

$A \in M(n,n), \quad f, u \in \mathbb{R}^n$
$B \in M(m,n), \quad g, \lambda \in \mathbb{R}^m$
Annahme: A positiv definit, also A^{-1} existiert.
Betrachte die Umformung

$$Au + B^T \lambda = f$$
$$\curvearrowright \quad u = A^{-1} f - A^{-1} B^T \lambda$$
$$\curvearrowright \quad B(A^{-1} f - A^{-1} B^T \lambda) = g$$
$$\boxed{\underbrace{(BA^{-1}B^T)}_{Schurkomplement} \lambda = BA^{-1} f - g}$$

Man kann die Schurkomplementgleichung zum Beispiel mit dem Gradientenverfahren oder dem cg-Verfahren lösen. (siehe Übung)

VII.4 Laplace-Gleichung als gemischtes Problem

$$\begin{aligned} \Delta u &= -f \quad auf\ \Omega \\ u &= 0 \quad auf\ \partial\Omega \end{aligned}$$

„formales System“ :

$$\sigma = \nabla u \qquad u : \Omega \to \mathbb{R} div\ \sigma \quad = \partial_{x_1}\sigma_1 + \partial_{x_2}\sigma_2 = -f \qquad \sigma = \begin{pmatrix}\sigma_1\\ \sigma_2\end{pmatrix}, \quad \sigma_i : \Omega \to \mathbb{R}$$

Variationeller Ansatz:

$$\begin{aligned} (\sigma,\tau) - (\nabla u,\tau) &= 0 \qquad \forall\tau \\ (div\ \sigma,\varphi) &= (-f,\varphi) \qquad \forall\varphi \end{aligned}$$

VII.4.1 Primal-gemischte Formulierung

Gesucht ist $(\sigma,u) \in L_2(\Omega)^2 \times H_0^1(\Omega)$, so daß gilt:

$$\begin{aligned} (\sigma,\tau) - (\nabla u,\tau) &= 0 \quad \forall\tau \in L_2(\Omega)^2 \\ -(\sigma,\nabla\varphi) &= (-f,\varphi) \quad \forall\varphi \in H_0^1(\Omega) \end{aligned}$$

Im abstrakten Rahmen: $X = L_2(\Omega)^2, \quad M = H_0^1(\Omega)$

$$\begin{aligned} a(\sigma,\tau) &= (\sigma,\tau) := (\sigma_1,\tau_1) + (\sigma_2,\tau_2) \\ b(\tau,\varphi) &= -(\tau,\nabla\varphi) \end{aligned}$$

Nachweis:

i) (Bedingung i) aus Satz VII.4)
Wegen $\|\sigma\|^2 = a(\sigma,\sigma)$, ist $a(\cdot,\cdot)$ elliptisch auf ganz X.

ii) (Bedingung ii) aus Satz VII.4)
Sei $v \in H_0^1(\Omega)$ gegeben.
Wähle $\tau = -\nabla v \in L_2(\Omega)^2$
Es gilt

$$\begin{aligned} \frac{b(\tau,v)}{\|\tau\|} &= \frac{-(\tau,\nabla v)}{\|\tau\|} \\ &= \frac{(\nabla v,\nabla v)}{\|\tau\|} = \frac{(\nabla v,\nabla v)}{\|\nabla v\|} \\ &= \|\nabla v\| \geq \underbrace{\frac{1}{c}}_{Poincaré\text{-}Ungleichung} \|v\|_1 \end{aligned}$$

Also $\sup\limits_{\tau\in L_2(\Omega)^2} \frac{b(\tau,v)}{\|\tau\|} \geq \frac{1}{c}\|v\|_1$
Passende Finite Elemente?
Triangulierung $\mathbb{T}_h$ mit Dreiecken
Wähle $k \geq 1$ und $X_h = \{\sigma_h \in L_2(\Omega)^2 |\quad \sigma_h|_T \in P_{k-1}\}$ (σ_k global unstetig)
$M_h = \{\varphi_h \in H_0^1(\Omega) |\quad \varphi_h|_T \in P_k\}$ (φ_h global stetig)
Nachweis der „BB-Bedingung“ analog zum kontinuierlichen Fall.

VII.4.2 Dual-gemischte Formulierung

Vorbereitung: $H_{div} := \{\tau \in L_2(\Omega)^2 | div\ \tau \in L_2(\Omega)\}$
Zugehörige Norm: $\|\tau\|_{div}^2 := \|\tau\|^2 + \|div\ \tau\|^2$
Für $v \in H_0^1(\Omega)$ *und* $\sigma \in H_{div}$

$$(\sigma, \nabla v) = -(div\ \sigma, v)$$

„schwache Formulierung“ : Gesucht ist $(\sigma, u) \in H_{div} \times L_2(\Omega)$ *mit*

$$(\sigma, \tau) + (u, div\ \tau) = 0 \quad \forall \tau \in H_{div}$$
$$(div\ \sigma, \varphi) = (-f, \varphi) \quad \forall \varphi \in L_2(\Omega)$$

Abstrakt: $X := H_{div}\,, \quad M := L_2(\Omega)$

$$a(\sigma, \tau) = (\sigma, \tau), \qquad b(\tau, v) = (div\ \tau, v)$$

Nachweis zu i) aus Satz VII.4:

$$a(\tau, \tau) = (\tau, \tau) = (\tau, \tau) + \underbrace{(div\ \tau, div\ \tau)}_{=0 \text{ für } \tau \in V = \{\tau \in H_{div} | (div\ \tau, v) = 0, v \in L_2(\Omega)\}}$$
$$= \|\tau\|^2 + \|div\ \tau\|^2 = \|\tau\|_{div}^2 \qquad \forall \tau \in V$$

Nachweis zu ii) aus Satz VII.4:
zu zeigen: $\sup\limits_{\tau \in H_{div}} \frac{(div\ \tau, v)}{\|\tau\|_{div}} \geq \beta \|v\|$

I) Sei $v \in L_2(\Omega)$ *beliebig: Wähle dazu* $w \in C_0^\infty(\Omega)$ *mit*

$$\|v - w\| \underset{Dichtheitsargument}{\leq} \tfrac{1}{2}\|v\| \qquad (A1)$$

II) Man setze $\xi := inf\{x_1 | x \in \Omega\}$.

Ansatz: $\tau_1(x) = \int\limits_\xi^{x_1} w(t, x_2) dt, \quad \tau_2 = 0$

Damit gilt: $div\ \tau = \partial_{x_1} \tau_1 = w$
Ähnlich wie beim Beweis der Poincaré-Abschätzung:

$$|\tau(x)|^2 = |\tau_1(x)|^2 \leq \int\limits_\xi^{x_1} |w(t, x_2)|^2 dt$$
$$\leq \int\limits_\xi^{s} |w(t, x_2)|^2 dt \qquad mit\ s = \max\{x_1 | x \in \Omega\}$$

Integration über x_1*:*

$$\int\limits_\xi^s |\tau_1(x)|^2 dx_1 \leq c \int\limits_\xi^s |w(t, x_2)|^2 dt$$
$$= c \int\limits_\xi^s |w(x_1, x_2)|^2 dx_1$$

Integration über x_2:

$$\|\tau\|^2 \leq c\,\|w\|^2 \qquad \textit{(A2)}$$

III) Ausgangspunkt ist (A1)

$$\begin{aligned}\|v-w\|^2 = (v-w, v-w) &\leq \frac{1}{4}\|v\|^2 \\ \Leftrightarrow \quad 2(v,w) &\geq \frac{3}{4}\|v\|^2 + \|w\|^2 \\ \Leftrightarrow \quad (v,w) &\geq \frac{3}{8}\|v\|^2 + \frac{1}{2}\|w\|^2 \geq c\,\|v\|^2\end{aligned}$$

IV) Auswertung von $\frac{b(\tau,v)}{\|\tau\|_{div}}$:

$$\begin{aligned}\frac{b(\tau,v)}{\|\tau\|_{div}} &= \frac{(div\ \tau, v)}{\left(\|\tau\|^2 + |div\ \tau|^2\right)^{1/2}} \\ &\geq \frac{(div\ \tau, v)}{c\,\|w\|} \quad \textit{wegen (A2) und } div\ \tau = w \\ &= \frac{(w,v)}{c\,\|w\|} \\ &\geq \frac{c\,\|v\|^2}{d\,\|v\|} \quad \textit{wegen } (v,w) \geq c\,\|v\|^2 \\ &= c\,\|v\|\end{aligned}$$

Passende Diskretisierung: (Raviart-Thomas-Elemente)

$$X_h = \left\{ \tau \in L_2(\Omega)^2 \,\middle|\, \tau|_T = \begin{pmatrix} a_T \\ b_T \end{pmatrix} + c_T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, a_T, b_T, c_T \in \mathbb{R}, \tau \cdot n \textit{ stetig an Kanten} \right\}$$

$$M_h = \{ v \in L_2(\Omega) | \quad v|_T = d_T, \quad d_T \in \mathbb{R} \}$$

Bemerke: $div\ \tau = \partial_{x_1}\tau_1 + \partial_{x_2}\tau_2 = c_T + c_T = const. \in M_h$