Lösung von PDGL mit der Finite Elemente Methode

THOMAS WICK

Mitschrift WS 06/07

gelesen von
Prof. Dr. F.-T. Suttmeier

Fachbereich 6 - Mathematik
UNIVERSITÄT SIEGEN

Dienstag, den 10. April 2007

Zusammenfassung

Dieser Kurs behandelt das Lösen von partiellen Differentialgleichungen mit der *Methode der finiten Elemente*.

Inhaltsverzeichnis

1 Numerische Simulation

Zunächst eine Übersicht

Realität

Praxis Theorie

Negative Faktoren:
Praxis:

- zu teuer (crash-tests),
- zu gefährlich,
- zu weit weg (Mond),
- zu klein (Nano)

Theorie:

- Gleichungen schwer lösbar

Die Beschreibung der Realität führt zur Modellierung von PDGL
→ Näherungslösungen des Modells, Diskretisierung $u \approx u_h$
→ Auswertung $u \approx u_h$
→ Berwertung, Einordnung von u_h bzw. u
⟶ numerische Experimente mit akzeptierten Modell

Abschließend wird die Theorie der PDGL und der numerischen Analyse zusammengefasst

Theorie PDGL
→ PDGL
→ Diskretisierung FDM, FEM
→ diskretisiertes System von Gleichungen

Numerische Analyse
→ diskretisiertes System von Gleichungen

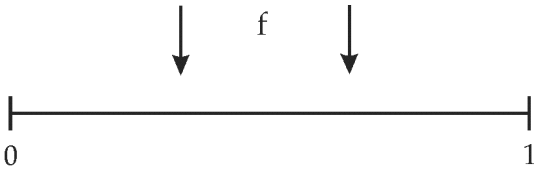
Das **diskretisierte System** mündet dann in
→ Lösung der diskreten Systeme
→ Theorie der Iterationsverfahren

2 Einleitung zur FEM (1D)

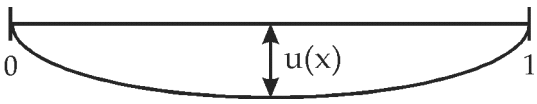
2.1 Das Modell-Problem

Fast die gesamte Vorlesung basiert auf der Untersuchung der inhomogenen Laplace-Gleichung, der sog. *Poisson-Gleichung*. Anhand dieser Gleichung können wesentliche Aspekte der Finite Element Methode erarbeitet werden.

In der Praxis charakterisiert die Poisson-Gleichung im 1D-Fall beispielsweise einen eingespannten elastischen Draht, auf den die Kraft f wirkt



Der Draht wird nach unten ausgelenkt



Es soll nun eine mathematische Formulierung hergeleitet werden. Dazu betrachte man die Längenänderung

$$\Delta l = \int_0^1 \sqrt{1 + (\partial_x u)^2}\, dx - \int_0^1 1\, dx$$

Das erste Integral beschreibt die Bogenlänge und kann mit der Taylor-Entwicklung noch vereinfacht werden. Für kleine Auslenkungen erhält man

$$\sqrt{1+y} = \sqrt{1+0} + \frac{1}{2 \cdot \sqrt{1+0}}(y-0) + \mathcal{O}((y-0)^2)$$

mit $y = (\partial_x u)^2$ und $\mathcal{O}((y-0)^2)$ als Fehlerordnung. Die Näherung ergibt

$$\sqrt{1+y} \approx 1 + \frac{1}{2}y$$

Für die Längenänderung Δl folgt damit

$$\Delta l \approx \frac{1}{2} \int_0^1 (\partial_x u)^2 dx$$

Die elastische Energie des Drahts ist proportional zu Δl:

$$U_E \sim \Delta l$$

Durch Einwirken der Kraft f besitzt der Draht außerdem die potentielle Energie U_f mit

$$U_f = -\int_0^1 f \cdot u \, dx \quad (E = F \cdot s \quad \text{bzw.: Energie = Kraft} \times \text{Weg})$$

Aufgrund der Energieerhaltung werden elastische- und potentielle Energie addiert und erhält so

$$U(u) = U_E + U_f = \frac{1}{2}\int_0^1 (\partial_x u)^2 \, dx - \int_0^1 f \cdot u \, dx$$

In der stabilen Gleichgewichtslage wird die Gesamtenergie minimiert, so dass man folgende Aufgabe formulieren kann.

Aufgabe.
Gesucht ist $u \in V$, so dass

$$U(u) \leq U(v) \quad \forall v \in V, \quad V \text{ ist der Raum der Vergleichsfkt.}$$

EIGENSCHAFTEN VON V

1. alle stetigen Funktionen mit „Nullrandwerten "(Auslenkung an den Rändern ist Null)
2. stückweise stetige, beschränkte 1. Ableitung („Integrale müssen Sinn machen ")

Lösung der Aufgabe.
Wir nehmen an, dass Lösung u gefunden ist. Dazu wird $\varphi \in V$ als „Störung"gewählt und

$$v = u + \varepsilon \cdot \varphi, \quad \varepsilon \in \mathbb{R}$$

betrachtet. Wegen Minimaleigenschaft gilt dann

$$U(u) \leq U(v) = U(u + \varepsilon\varphi)$$

Die notwendige Bedingung lautet

$$\begin{aligned}
& 0 = \frac{d}{d\varepsilon} U(u + \varepsilon\varphi)\Big|_{\varepsilon=0} \\
\Leftrightarrow \quad & 0 = \frac{d}{d\varepsilon}\left[\frac{1}{2}\int_0^1 \left(\partial_x(u + \varepsilon\varphi)\right)^2 dx - \int_0^1 f \cdot (u + \varepsilon\varphi)\, dx\right]_{\varepsilon=0} \\
\overset{\text{Kettenregel}}{\Leftrightarrow} \quad & 0 = \left[\frac{1}{2}\int_0^1 2 \cdot (\partial_x(u + \varepsilon\varphi))\partial_x\varphi \, dx - \int_0^1 f \cdot \varphi \, dx\right]_{\varepsilon=0} \\
\Leftrightarrow \quad & 0 = \int_0^1 \partial_x u \partial_x \varphi \, dx - \int_0^1 f \cdot \varphi \, dx \qquad (2.1)
\end{aligned}$$

Man spricht bei (2.1) von der variationellen Formulierung. Hieraus kann die klassische Formulierung hergeleitet werden:

Satz 2.1.1 *Sei $u(x)$ Lösung von* (2.1)*. Zusätzlich existiere $\partial_x^2 u(x)$ und sei stetig. Dann gilt*

$$\begin{aligned}
-\partial_x^2 u(x) &= f(x) \quad x \in (0,1) \\
u(0) &= u(1) = 0 \quad \textit{Randbedingungen}
\end{aligned}$$

Beweis.
Partielle Integration angewendet auf (2.1).

$$[\partial_x u \cdot \varphi]_0^1 - \int_0^1 \partial_x^2 u(x) \cdot \varphi(x)\, dx - \int_0^1 f(x) \cdot \varphi(x)\, dx = 0$$

Beachte

$$\varphi(0) = \varphi(1) = 0 \Rightarrow [\partial_x u \cdot \varphi] = 0$$

Also

$$\begin{aligned} & 0 = -\int_0^1 (\partial_x^2 u(x) + f(x)) \cdot \varphi(x)\, dx \quad \forall \varphi(x) \in V \\ \Rightarrow \quad & 0 = -\partial_x^2 u(x) - f(x) \quad \text{auf } (0,1) \end{aligned}$$

Zusätzlich erhält man $u(0) = u(1) = 1$

□

Als Kurzschreibweise wird die nächste Formulierung oft verwendet werden

$$\begin{aligned} -u'' &= f \quad \text{auf } (0;1) \\ u(0) &= u(1) = 0 \end{aligned}$$

2.2 Klassische und variationelle Formulierung

Zum klassischen Poisson-Problem betrachte man die Dirichlet-Formulierung $\mathcal{D}$

$$\begin{aligned} -u'' &= f \\ u(0) &= u(1) = 0 \end{aligned} \tag{2.2}$$

Die variationelle Formulierung $\mathcal{V}$ dazu lautet

$$u \in V: \quad (u', \varphi') = (f, \varphi) \quad \forall \varphi \in V \tag{2.3}$$

mit $(v, w) := \int_I v \cdot w\, dx$. Der Raum V ist der Raum der Vergleichsfunktionen (auch Testfunktionen genannt) mit den Eigenschaften

i) alle stetigen Funktionen mit Nullrandwerten und

ii) stückweise stetige beschränkte 1. Ableitungen sind enthalten

Als dritte Formulierung sei die Minimalcharakterisierung $\mathcal{M}$ genannt, die in der vorherigen Sektion hergeleitet wurde

$$U(\varphi) = \frac{1}{2} \int_I (\varphi')^2\, dx - \int_I f \cdot \varphi\, dx \quad \forall \varphi \in V \tag{2.4}$$

Gesucht ist das Minimum

$$u \in V: \quad U(u) \leq U(\varphi) \quad \forall \varphi \in V$$

Es gilt

Satz 2.2.1 *Damit sind äquivalent*

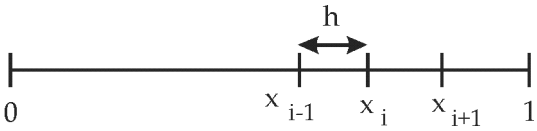
i) Die variationelle Formulierung $\mathcal{V}$ (2.3),

ii) die Dirichlet-Formulierung $\mathcal{D}$ (2.2),

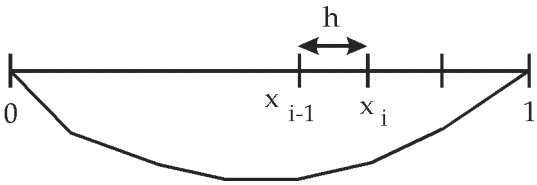
iii) die Minimumcharakterisierung $\mathcal{M}$ (2.4)

2.3 Diskretisierung

Vorbereitungen:
Einteilung von $(0;1)$ in Teilintervalle der Größe h mit n inneren Punkten.



Mit der Näherung $u \approx u_h$. Dabei sollte u_h stetig sein und stückweise linear.



Notation: $u_h^i = u_h(x_i)$, $f^i = f(x_i)$.
Für die Ableitungen gilt:

$$u_h'(x) = \begin{cases} V^{i+} = \frac{u_h^{i+1} - u_h^i}{h} & \text{auf } (x_i, x_{i+1}) \\ V^{i-} = \frac{u_h^i - u_h^{i-1}}{h} & \text{auf } (x_{i-1}, x_i) \end{cases}$$

Idee:

$$-u''(x_i) = f(x_i) \approx -\left(\frac{V^{i+} - V^{i-}}{h}\right) = f^i$$

Einsetzen liefert:

$$-u''(x_i) = f(x_i) \approx -\left(\frac{u_h^{i+1} - 2u_h^i + u_h^{i-1}}{h^2}\right) = f^i \tag{2.5}$$

Für jeden Gitterpunkt erhält man eine Gleichung. Somit entsteht ein lineares Gleichungssystem.
Also: $Ax = b$ mit $x, b \in \mathbb{R}^n, A \in \mathbb{R}^{n \times n}$.

$$x = \begin{pmatrix} u_h^1 \\ \vdots \\ \vdots \\ \vdots \\ u_h^n \end{pmatrix}, \; b = \begin{pmatrix} f^1 \\ \vdots \\ \vdots \\ \vdots \\ f^n \end{pmatrix}, \; A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{pmatrix} \cdot \frac{1}{h^2} \tag{2.6}$$

In die Matrix A werden die Koeffizienten aus (2.5) eingetragen.

Bemerkung zu den Randbedingungen
Setze in den beiden Gleichungen von (2.5), die die Randwerte 0 und 1 enthalten direkt Null ein. Denn die Randbedingung fordert $u(0) = u(1) = 0$. Folglich ist für $i = 1$ und $n - 1$: $u_h^0 = 0$ und $u_h^n = 0$.

Die nächsten Schritte lauten:

- Löse $Ax = b$ und erhalte x als Lösung
- (Graphische) Darstellung der Lösung

Verallgemeinerung.
Betrachte

$$Lu := -u'' + b(x)u' + c(x)u = f(x) \tag{2.7}$$

mit $x \in (0;1), u(0) = u(1), u = u(x)$. Außerdem $c(x) \geq 0$, welches die Existenz und Eindeutigkeit der Lösung sichert.

Schon dieses Problem ist i. Allg. nicht „exakt “lösbar.

2.4 Differenzenverfahren

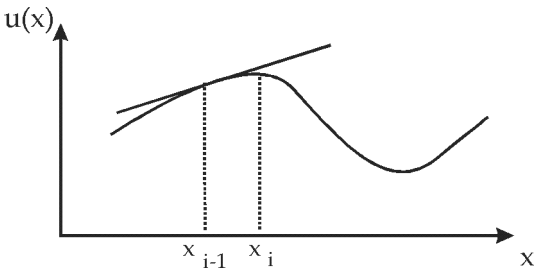


Abbildung 2.1: Diskretisierung der Funktion u

Ziel soll sein, die Tangente in x_i näherungsweise zu bestimmen. Dies geschieht mit Hilfe einer Sekante durch die Punkte x_{i-1} und x_i:

$$u'(x_i) = \frac{u(x_i) - u(x_{i-1})}{x_i - x_{i-1}}$$

Dieses Verfahren wird analog für höhere Ableitungen verwendet. Zur Diskretisierung betrachte man eine Menge von „Gitterpunkten “:

$$x_i = i \cdot h,\ i = 0, \ldots, N,\ h = \frac{1}{N},\ N \text{ ist Zahl der Teilintervalle}$$

Definiere:

$$\omega_h = \{x_i = ih | i = 1, \ldots, N-1\} \quad \text{Menge der inneren Gitterpunkte}$$
$$\gamma_h = \{x_0, x_N\} \quad \text{Randpunkte}$$
$$\overline{\omega_h} = \omega_h \cup \gamma_h \quad \text{alle Gitterpunkte}$$

Eine Gitterfunktion $\overline{u_h}$ ist auf $\overline{\omega_h}$ definiert, also im Prinzip ein Vektor:

$$\overline{u_h} = (u_h(x_0), u_h(x_1), \ldots, u_h(x_N))^T = (u_0, \ldots, u_N)^T$$

Approximation der 1. Ableitungen:

$$\begin{aligned}
&\text{Vorwärtsdifferenz:} \quad (D^+u)(x) := \frac{u(x+h) - u(x)}{h} \\
&\text{Rückwärtsdifferenz:} \quad (D^-u)(x) := \frac{u(x) - u(x-h)}{h} \\
&\text{symm. Differenz:} \quad (D^0u)(x) := \frac{u(x+h) - u(x-h)}{2h}
\end{aligned}$$

Zweite Ableitung:

$$(D^+D^-u)(x) := \frac{u(x+h) - 2u(x) + u(x+h)}{h^2}$$

Ansatz zur Diskretisierung:
mit (2.7):

$$-D^+D^-u_i + b_iD^0u_i + c_iu_i = f_i, \quad i = 1, \ldots, N-1 \tag{2.8}$$

mit den schon oben benutzten Kurzschreibweisen: $b_i = b(x_i), c_i = c(x_i), f_i = f(x_i)$.
Aus Aufstellung (2.8) folgt, dass für jeden Gitterpunkt eine Gleichung steht. Also ein System von $N-1$ Gleichungen mit $N-1$ Unbekannten.

Beispiel.
Seien $b(x) = c(x) = 0$, es treten also keine 1. Ableitungen auf.

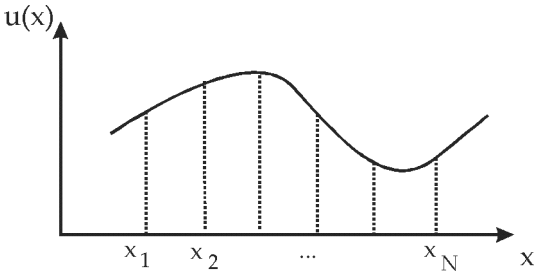
$$\frac{u_{i+1} + 2u_i - u_{i-1}}{h^2} = f_i, \quad i = 1, \ldots, N-1$$

Schreibe dieses System in eine Matrix:

$$\begin{pmatrix}
2 & -1 & & & \\
-1 & 2 & -1 & & \\
 & -1 & \ddots & \ddots & \\
 & & \ddots & \ddots & -1 \\
 & & & -1 & 2
\end{pmatrix}
\begin{pmatrix} u_1 \\ \vdots \\ \vdots \\ \vdots \\ u_{N-1} \end{pmatrix}
=
\begin{pmatrix} f_1 \\ \vdots \\ \vdots \\ \vdots \\ f_{N-1} \end{pmatrix}$$

Dieses System wird auch Tridiagonales Gleichungssystem genannt. Aufgabe ist nun, dieses Gleichungssystem zu mit dem Lösungsvektor $(u_1, \ldots, u_{N-1})^T$ lösen.

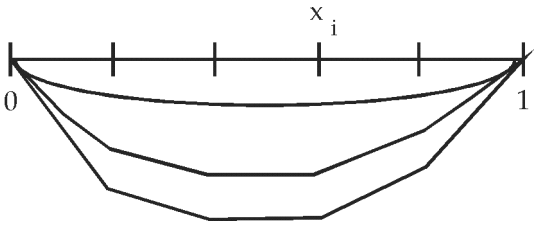
Die Näherungslösung mit Hilfe der Diskretisierung ist in folgender Skizze dargestellt.



Die nächste Frage wird sein, wie solche Näherungslösungen weiter verbessert werden können. Immerhin ist obige Lösung lediglich eine Approximation. Kann evtl. durch die Verwendung einer größeren Anzahl von Gitterpunkten die Lösung verbessert werden?

Verbesserung der Lösung

Betrachtung des Drahtproblems



Untersuchung des Fehlers mit der Schreibweise $u_i := u(x_i)$

$$\max_i |u_i - u_h^i|$$

Definition 2.4.1 *Das Differenzenverfahren heißt* konvergent *von der Ordnung k, wenn gilt*

$$\max_i |u_i - u_h^i| \leq C \cdot h^k, \quad C = const$$

Beispiel.
Für die vorgestellte Diskretisierung in Abschnitt (2.3) für $-u'' = f$ ist $k = 2$.
Die prinzipielle Untersuchung auf Konvergenz geschieht durch die Begriffe *Konsistenz* und *Stabilität*. Betrachte dazu die folgende „Fehlergleichung".

$$L_h(R_h u - u_h) = L_h R_h u - L_h u_h = L_h R_h u - f_h = L_h R_h u - R_h L u$$

Hier ist $R_h u - u_h$ die i-te Zeile des Vektors x. Nämlich $u(x_i) - u_h^i$.

Definition 2.4.2 *Das Differenzenverfahren heißt* konsistent *von der Ordnung k, falls gilt*

$$\begin{aligned} ||L_h R_h u - R_h L u||_{\infty,\omega_h} &\leq C \cdot h^k \\ = \max_i |L_h R_h u - R_h L u| &\leq C \cdot h^k \end{aligned}$$

Definition 2.4.3 *Folgt aus $L_h \omega_h = f_h$ die Ungleichung*

$$\max_i |\omega_h| \leq C \cdot \max_i |f_h|$$

so heißt das Verfahren stabil.

BEWERTUNG IN BEZUG AUF u_r

$$\begin{aligned} ||u_r - u_h|| &= ||u_r - u + u - u_h|| \\ &= \underbrace{||u_r - u||}_{\text{Modellfehler}} + \underbrace{||u - u_h||}_{\text{Diskretisierungsfehler}} \end{aligned}$$

Verbesserung des Modells mit

$$-u'' = f \rightarrow -\mu\, u'' = f$$

wobei μ die Elastizitätskonstante ist. Damit wird

$$-0,75\, u'' = f$$

als akzeptiertes Modell angenommen.

2.5 Näherungslösungen, Ritz-Galerkin-Verfahren

Es sollen die Vorzüge der variationellen Formulierung ausgenutzt werden. Idee ist, den Raum V durch einen endlich-dimensionalen Teilraum V^N mit dim $V^N = N$ zu approximieren. Man betrachte dazu die schwache Formulierung $\mathcal{V}$

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

Gesucht ist die Lösung $u^N \in V^N$. Im weiteren Verlauf wird die folgende Schreibweise verwendet

$$a(v, w) := (\partial_x v, \partial_x w), \quad v, w \in V$$

Also für unser Problem

$$a(u^N, \varphi^N) = (f, \varphi^N) \quad \forall \varphi^N \in V^N \tag{2.9}$$

Für den endlichen Vektorraum $V^N =< \varphi_1, \ldots, \varphi_N >$ wird eine Basis genommen und für einen beliebigen Vektor $\varphi \in V^N$ gilt dann

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

Es genügt

$$a(u^N, \varphi_i) = (f, \varphi_i) \quad \forall i = 1, \ldots, N \tag{2.10}$$

zu erfüllen.

Rechnung.
Umstellung von (2.9) zeigt

$$\begin{aligned} 0 &= a\left(u^N, \sum_{i=1}^{N} v_i \varphi_i\right) - \left(f, \sum_{i=1}^{N} v_i \varphi_i\right) \\ &\overset{(2.10)}{=} \sum_{i=1}^{N} v_i \underbrace{\Big(a(u^N, \varphi_i) - (f, \varphi_i)\Big)}_{=0} \end{aligned}$$

Für die Berechnung der Näherungslösung u^N wird

$$u^N = \sum_{j=1}^{N} u_j \varphi_j, \quad u_j \in \mathbb{R}$$

in (2.10) eingesetzt. Dann folgt

$$\begin{aligned} & a(u^N, \varphi_i) = (f, \varphi_i) \\ \Leftrightarrow\ & a\left(\sum_{j=1}^{N} u_j \varphi_j, \varphi_i\right) = (f, \varphi_i) \\ \Leftrightarrow\ & \sum_{j=1}^{N} u_j a(\varphi_j, \varphi_i) = (f, \varphi_i), \quad i = 1, \ldots, N \end{aligned}$$

Die Bestimmung von $u^N \in V^N$ wird somit auf das Lösen eines linearen Gleichungssystems $Ax = b$ zurückgeführt mit

$$\begin{aligned} x^T &= (u_1, \ldots, u_N) \\ b^T &= (b_1, \ldots, b_N), \quad b_i = (f, \varphi_i) \\ A &= a(\varphi_j, \varphi_i), \quad A \in \mathbb{R}^{N \times N} \end{aligned}$$

ERSTE FEHLERABSCHÄTZUNG, GALERKIN-EIGENSCHAFT

$$\begin{aligned} a(u, \varphi) &= (f, \varphi) \quad \forall \varphi \in V \\ a(u^N, \varphi^N) &= (f, \varphi^N) \quad \forall \varphi^N \in V^N \end{aligned}$$

Wir arbeiten mit dem konformen Ansatz. D.h. es gilt $V^N \subset V$. Diese Wahl ist naheliegend aber nicht zwingend. Für ein Element $\varphi^N \in V^N$ und Subtraktion der beiden obigen Gleichungen folgt

$$a(u - u^N, \varphi^N) = 0 \quad \forall \varphi^N \in V^N$$

Diese Beziehung wird *Galerkin-Orthogonalität* genannt. Der Fehler der Näherungslösung u^N kann nun abgeschätzt werden.

$$\begin{aligned} ||u - u^N||_V^2 &= a(u - u^N, u - \varphi^N + \varphi^N - u^N) \\ &= a(u - u^N, u - \varphi^N) + \underbrace{a(u - u^N, \varphi^N - u^N)}_{=0 \text{ (wg. Galerkin-Eigenschaft)}} \\ &\leq c \cdot ||u - u^N||_V \cdot ||u - \varphi^N||_V \end{aligned}$$

Division zeigt

$$||u - u^N||_V \leq c\, ||u - \varphi^N||_V \quad \forall \varphi^N \in V^N$$

Übergang zum Infimum liefert

$$||u - u^N||_V \leq \frac{c}{\alpha} \inf_{\varphi^N \in V^N} ||u - \varphi^N||_V$$

wobei α die Elliptiziätskonstante ist.

2.6 Einfache Finite Elemente

Nachdem wir erste Ansätze der Diskretisierung in den ersten Abschnitten kennen gelernt haben, werden nun Finite Element Methoden eingeführt. Wir beginnen mit den linearen Ansätzen und geben am Ende der Sektion eine Idee zur Konstruktion von quadratischen Finiten Elementen.

Die ausgeführten Schritte werden ausschließlich für das Modell-Problem gemacht. Vorweg die allgemeinen Überlegungen

Lösung u sollte gut approximierbar sein

Leichte Berechnung der Einträge von A, also

$$a(\varphi_j, \varphi_i) = \int \partial_x \varphi_j \cdot \partial_x \varphi_i \, dx$$

A sollte für die Numerik günstige Eigenschaften haben, wie zum Beispiel moderate Kondition, dünnbesetzt, usw.

2.6.1 Lineare Finite Elemente

Idee: V^N besteht aus stückweise linearen Funktionen, die global stetig sind.

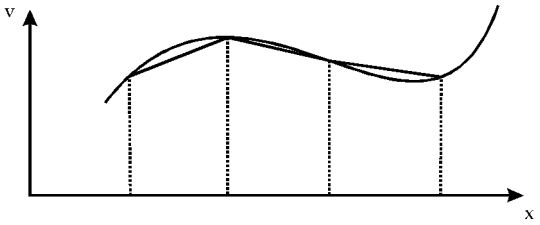


Abbildung 2.2: Beispiel einer Funktion $v \in V^N$

Dazu werden die passenden Basisfunktionen $\varphi_i \in V^N$ gewählt, definiert durch

$$\varphi_i(x_j) = \begin{cases} 1, & \text{falls } j = i \\ 0, & \text{falls } j \neq i \end{cases} \tag{2.11}$$

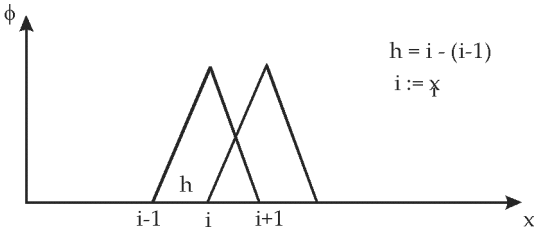


Abbildung 2.3: Die Basisfunktionen φ_i und φ_{i+1}, sog. Hutfunktionen

Eine Funktion $v \in V^N$ kann somit geschrieben werden als

$$v(x) = \sum_{j=1}^{N} v_j \, \varphi_j(x), \quad x \in [0,1], \, v_j := v(x_j)$$

Durch die geschickte Wahl der Basisfunktionen überschneiden sich nur die direkten Nachbarn. Alle weiteren schneiden sich nicht. Daher sind die gebildeten Integrale gleich Null

$$\int \partial_x \varphi_{i-1} \cdot \partial_x \varphi_{i+1} \, dx = 0 \tag{2.12}$$

Die so erzeugte Matrix ist also dünnbesetzt. Es stehen lediglich in der Hauptdiagonalen und den beiden ersten Nebendiagonalen Werte, die im folgenden berechnet werden.
Es gilt nach (2.11):

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

Weiter ist

$$\partial_x \varphi_i = -\frac{1}{h} \quad \text{auf } (x_i, x_{i+1})$$

Also folgt für die Hauptdiagonale $i = j$

$$\begin{aligned} a(\varphi_i, \varphi_i) &= \int_I \partial_x \varphi_i \cdot \partial_x \varphi_i \, dx \\ &\overset{(2.12)}{=} \int_{i-1}^{i+1} \partial_x \varphi_i \cdot \partial_x \varphi_i \, dx \\ &= 2 \cdot \int_i^{i+1} \partial_x \varphi_i \cdot \partial_x \varphi_i \, dx \\ &= 2 \cdot h \cdot \frac{1}{h^2} = \frac{2}{h} = A_{ii} \end{aligned}$$

Die A_{ii} stellen die Einträge der Diagonalen. Bestimmen der Nebendiagonalwerte liefert

$$\begin{aligned} \int_I \partial_x \varphi_i \partial_x \varphi_{i+1} \, dx &= \int_i^{i+1} \partial_x \varphi_i \partial_x \varphi_{i+1} \, dx \\ &= -\frac{1}{h^2} \cdot h = -\frac{1}{h} = A_{i,i+1} = A_{i-1,i} \end{aligned}$$

Also

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

Berechnung der rechten Seite

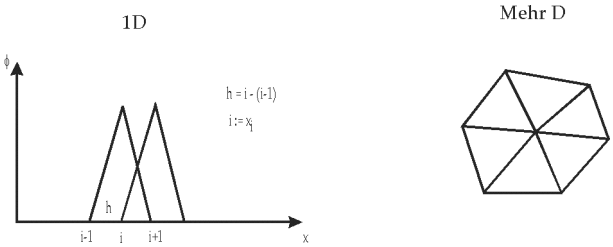
$$b_i = \int_I f \cdot \varphi_i \, dx$$

Falls f konstant ist, folgt

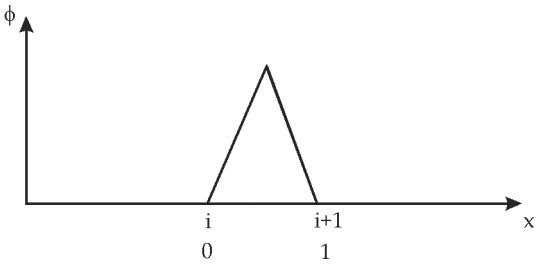
$$b_i = f \cdot \int_{i-1}^{i+1} \varphi_i \, dx = h f_i$$

ALTERNATIVES VORGEHEN

Bisher wurde zum Aufstellen der Matrix A eine Schleife der Basisfunktionen gebildet. Alternativ dazu kann *zellweise* vorgegangen werden. D.h. es wird erst eine Zelle komplett abgearbeitet. Also der Hauptdiagonalwert und die beiden Nebendiagonalwerte bestimmt werden, ehe anschließend zur nächsten Zelle weitergegangen wird.



Betrachten wir nun den 1D Fall, und der Einfachheit halber nur innere Punkte. Es sei die Zelle (x_i, x_{i+1}) vorgegeben.



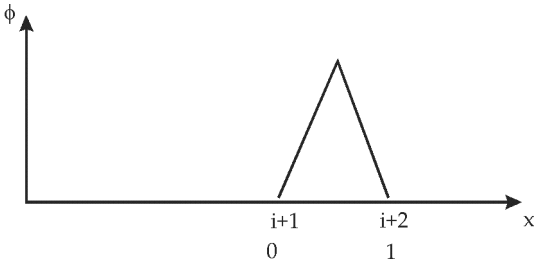
Man bilde die zugehörigen Integrale

$$\int_i^{i+1} \partial_x \varphi_i \cdot \partial_x \varphi_{i+1}\, dx = -\frac{1}{h}$$
$$\int_i^{i+1} \partial_x \varphi_i \cdot \partial_x \varphi_i\, dx = \int_i^{i+1} \partial_x \varphi_{i+1} \cdot \partial_x \varphi_{i+1}\, dx = \frac{1}{h}$$

und stelle die Matrix auf

$$A = \begin{pmatrix} & \\ \frac{1}{h} & -\frac{1}{h} \\ -\frac{1}{h} & \frac{1}{h} \\ & \end{pmatrix}$$
$$\begin{matrix} \uparrow & \uparrow \\ i & i+1 \end{matrix}$$

Für die weiteren Daten wird nun die nächste Zelle betrachtet.



Folgende Einträge werden in Matrix A gemacht.

$$A = \begin{pmatrix} & & \\ \frac{1}{h} & -\frac{1}{h} & \\ -\frac{1}{h} & \frac{1}{h}+\frac{1}{h} & -\frac{1}{h} \\ & -\frac{1}{h} & \frac{1}{h} \\ & & \end{pmatrix}$$

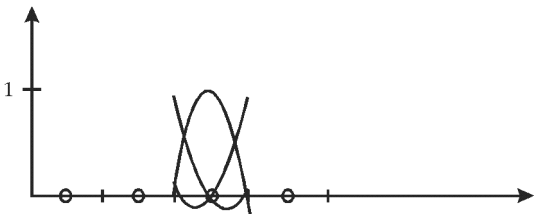
$$\begin{matrix} \uparrow & \uparrow & \uparrow \\ i & i+1 & i+2 \end{matrix}$$

Für nicht äquidistante Abstände erfolgt die Berechnung nach dem gleichem Schema.

2.6.2 Quadratische Finite Elemente

Aufgabe.
Konstruiere 3 Basisfunktionen, die jeweils durch drei Punkte gehen. Der Ansatz liegt auf der Hand, die Lösungen sind quadratische Funktionen der Bauart

$$ax^2 + bx + c$$



Es sei folgende Aufgabe gestellt:
Man konstruiere einen endlich-dim. Teilraum $V_h \subset V$, der aus Funktionen besteht, die quadratisch sind auf jedem Teilintervall I_j einer Zerlegung von $I = (0, 1)$.

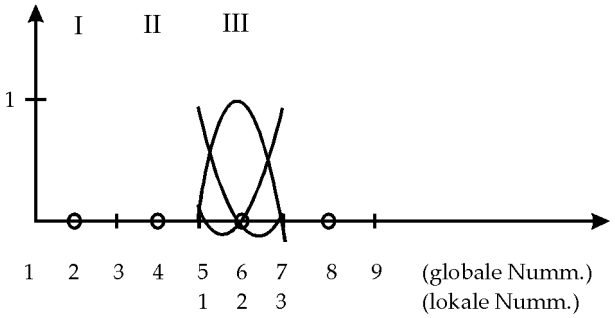
Die Beschreibung der Parameter der entsprechenden Funktionen geschieht durch die Wahl der Basisfunktionen. Außerdem formuliere man die entsprechende FE-Methode zu

$$u \in V: \quad \int_I \partial_x u \partial_x \varphi \, dx = \int_I f \cdot \varphi \, dx \quad \forall \varphi \in V$$

und bestimme das zugehörige Gleichungssystem im Falle einer gleichförmigen Zerlegung.

Lösung.
Es gilt

$$A_{ij} = a(\varphi_j, \varphi_i) = \int_I \partial_x \varphi_j \partial_x \varphi_i \, dx$$



Im Gegensatz zu den linearen Basisfunktionen, gehen bei einer quadratischen Annahme durch einen Stützpunkt drei Basisfunktionen (bei linearen Hutfunktionen sind es zwei).

Lokale Nummerierung:
φ_1 bestimmt durch

$$\begin{aligned} \varphi_1(x_1) &= 1 \\ \varphi_1(x_2) &= 0 \\ \varphi_1(x_3) &= 0 \end{aligned}$$

φ_2 bestimmt durch

$$\begin{aligned} \varphi_2(x_1) &= 0 \\ \varphi_2(x_2) &= 1 \\ \varphi_2(x_3) &= 0 \end{aligned}$$

φ_1 bestimmt durch

$$\begin{aligned} \varphi_3(x_1) &= 0 \\ \varphi_3(x_2) &= 0 \\ \varphi_3(x_3) &= 1 \end{aligned}$$

Ansatz einer quadratischen Funktion

$$\varphi_i(x) = a_i + b_i x + c_i x^2$$

Hier können aber auch die Lagrange-Polynome angesetzt werden. Diese Variante ist in diesem Fall kürzer.

$$\varphi_1 = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)}, \quad \varphi_2 = \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)}$$

Analog für φ_3. Bei einer äquidistanten Unterteilung ergeben die Integrale für jedes lokale Intervall denselben Wert. Daher können die Überlegungen auf $[0,h]$ gemacht werden. Also

$$\varphi_1(x) = \frac{2}{h^2}\left(x^2 - \frac{3}{2}h\cdot x + \frac{h^2}{2}\right), \quad \partial_x\varphi_1(x) = \frac{1}{h^2}(4x-3h)$$
$$\varphi_2(x) = -\frac{4}{h^2}(x^2 - hx), \quad \partial_x\varphi_2(x) = -\frac{4}{h^2}(2x-h)$$
$$\varphi_3(x) = \frac{2}{h^2}\left(x^2 - \frac{hx}{2}\right), \quad \partial_x\varphi_3(x) = \frac{1}{h^2}(4x-h)$$

Berechnung der Integrale

$$\int_0^h \partial_x\varphi_1\partial_x\varphi_1\,dx = \frac{1}{h}\cdot\frac{14}{6} = 14\cdot\frac{1}{6h}$$
$$\int_0^h \partial_x\varphi_1\partial_x\varphi_2\,dx = -\frac{1}{h}\cdot\frac{16}{6} = -16\cdot\frac{1}{6h}$$
$$\int_0^h \partial_x\varphi_1\partial_x\varphi_3\,dx = \frac{1}{h}\cdot\frac{2}{6} = 2\cdot\frac{1}{6h}$$
$$\int_0^h (\partial_x\varphi_2)^2\,dx = \frac{1}{h}\cdot\frac{32}{6} = 32\cdot\frac{1}{6h}$$

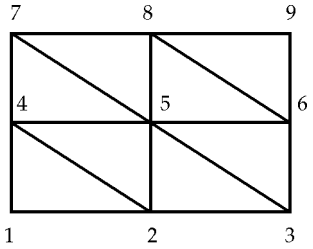
Zur globalen Betrachtung benutze man die in der Zeichnung dargestellte globale Nummerierung mit

$$\int_I \partial_x\varphi_j\partial_x\varphi_i\,dx = \sum_T \int_T \partial_x\varphi_j\partial_x\varphi_i\,dx$$

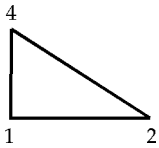
Nun kann die Matrax A aufgestellt werden

$$A = \frac{1}{6}\cdot\begin{pmatrix} & & & & \\ & & & & \\ \frac{14}{h_{III}} & -\frac{16}{h_{III}} & \frac{2}{h_{III}} & & \\ -\frac{16}{h_{III}} & \frac{32}{h_{III}} & -\frac{16}{h_{III}} & & \\ \frac{2}{h_{III}} & -\frac{16}{h_{III}} & \frac{14}{h_{III}}+\frac{14}{h_{IV}} & -\frac{16}{h_{IV}} & \frac{2}{h_{IV}} \\ & & -\frac{16}{h_{IV}} & -\frac{32}{h_{IV}} & -\frac{16}{h_{IV}} \\ & & \frac{2}{h_{IV}} & -\frac{16}{h_{IV}} & \frac{14}{h_{IV}} \\ & & & & \\ & & & & \end{pmatrix}$$
$$\begin{matrix} & \uparrow & \uparrow & \uparrow & & \uparrow & \uparrow \\ & 5 & 6 & 7 & & 8 & 9 \end{matrix}$$

Im Mehrdimensionalen können folgende Nummerierungen verwendet werden:



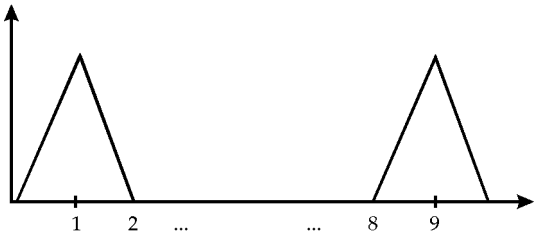
Beim Durchlaufen der Zellen können Lücken in der Nummerierung auftreten



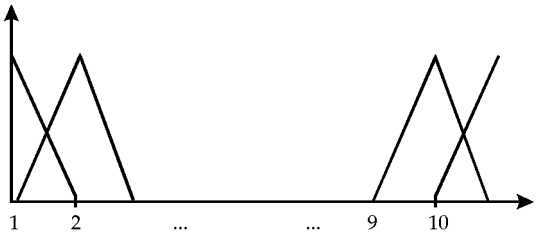
Ziel in der Praxis: Einmal berechnete Basisfunktionen (wie oben gezeigt) immer wieder verwenden können, und nicht für jede Zelle neu berechnen zu müssen.

2.6.3 Einbau von Randwerten

Kurzer Anriss: Aus



folgt für Randwerte



Wir erhalten die Matrix A

$$\frac{1}{h} \cdot \begin{pmatrix} 1 & -1 & & & \\ -1 & 1+1 & -1 & & \\ & -1 & & & \\ & & & & \ddots \end{pmatrix}$$

also

$$\frac{1}{h} \cdot \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}$$

Insgesamt erhält man so das vollständige Gleichungssystem

$$\begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ \vdots \\ \vdots \\ u_{11} \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ \vdots \\ \vdots \\ f_{11} \end{pmatrix}$$

Weiter folgt dann

$$\begin{pmatrix} 1 & 0 & \dots & & 0 \\ & 2 & & & \\ & & \ddots & & \\ & & & 2 & -1 \\ 0 & \dots & & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ \vdots \\ \vdots \\ u_{11} \end{pmatrix} = \begin{pmatrix} 0 \\ f_2 \\ \vdots \\ f_{10} \\ 0 \end{pmatrix}$$

Die erste und die zweite Zeile werden also gelöscht. Nur in der ersten Zeile erste Stelle und letzte Zeile letzte Stelle wird jeweils eine 1 gesetzt. Ebenso werden f_1 und f_{11} gleich Null gesetzt. Damit sind die Randbedingungen erfüllt.

Falls andere Randwerte, bsp. $u_1 = a$ und $u_{11} = b$ gefordert sind, so wird lediglich $f_1 = a$ und $f_{11} = b$ gesetzt.

2.7 Fehlerbetrachtungen für lineare FE

In diesem Abschnitt werden Fehlerapproximationen für lineare FE hergeleitet. Motivation sei die Abschätzung

$$\begin{aligned} ||\partial_x(u - u_h)|| &\leq \inf_{\varphi \in V_h} ||\partial_x(u - \varphi)|| \\ &\leq ||\partial_x(u - I_h u)|| \end{aligned}$$

Dabei bezeichne $I_h u$ die lineare Interpolierende von u. Dementsprechend ist $I_h u \in V_h$.

2.7.1 Interpolationsfehler

Zunächst wird der Interpolationsfehler hergeleitet, ehe dieser auf die Energiefehlerabschätzung übertragen werden kann. Es gilt der folgende

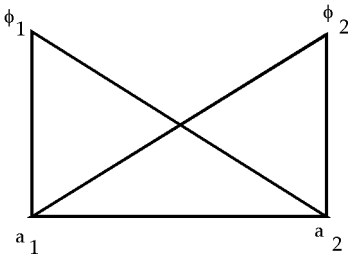
Satz 2.7.1 *Auf einem Teilintervall T, $T = (a_1, a_2)$ mit lokaler Gitterweite $h_T = a_2 - a_1$, der Zerlegung von $I \subset \mathbb{R}$ gilt*

i) $||v - I_h v||_{L^\infty(T)} \leq c \cdot h_T^2 \cdot ||\partial_x^2 v||_{L^\infty(T)}$,

ii) $||\partial_x(v - I_h v)||_{L^\infty(T)} \leq c \cdot h_T \cdot ||\partial_x^2 v||_{L^\infty(T)}$

mit $v \in V$

Skizze



Beweis.
Es bezeichne $P_n(T)$ den Vektorraum der Polynome vom Grad $\leq n$. Zunächst seien φ_1, φ_2 Basisfunktionen für $P_1(T)$ vorgegeben. Allgemein sei eine Funktion $w \in P_1(T)$ bestimmt durch

$$w(x) = \sum_{i=1}^{2} w(a_i)\, \varphi_i(x), \quad x \in T$$

Also

$$I_h v(x) = \sum_{i=1}^{2} v(a_i)\, \varphi_i(x), \quad x \in T \tag{2.13}$$

Man betrachte als nächstes

$$v(a_i) = v(x) + \partial_x v(x)(a_i - x) + \frac{1}{2}\partial_x^2 v(\xi_i)(a_i - x)^2 \tag{2.14}$$

Einsetzen von (2.14) in (2.13) liefert

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

Umsortieren

$$I_h v(x) = v(x) \cdot \sum_{i=1}^{2} \varphi_i(x) + \sum_{i=1}^{2} \partial_x v(x)(a_i - x) \cdot \varphi_i(x) + \sum_{i=1}^{2} \frac{1}{2}\partial_x^2 v(\xi_i)(a_i - x)^2 \cdot \varphi_i(x) \tag{2.15}$$

Wir zeigen später die Beziehungen

$$\sum_{i=1}^{2} \varphi_i(x) = 1$$

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

Was bleibt zunächst?

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

Wegen $\varphi_i(x) \leq 1$ und $|(a_i - x)| \leq h_T$ folgt

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

Es bleibt zu zeigen

$$\sum_{i=1}^{2} \varphi_i(x) = 1$$

Rechnung dazu. Betrachte $v(x) \equiv 1$, dann ist

$$\partial_x v(x) = \partial_x^2 v(x) = 0$$

Die erste und die zweite Ableitung fallen also weg, speziell ist

$$I_h v(x) \equiv 1$$

Einsetzen in (2.15):

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

Es bleibt noch zu zeigen, dass

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

Rechnung. Sei $v \in V$ gegeben. Für festes $\overline{x} \in T$ wird

$$d = \partial_x v(\overline{x})$$

gesetzt. Ansatz,

$$w(x) = d \cdot x \quad \text{(Gerade mit Steigung } d\text{)}$$

Die folgenden Relationen gelten dann wegen linearer Interpolation einer linearen Funktion $w(x)$.

$$\begin{aligned} I_h w &= w \\ \partial_x w &= d \\ \partial_x^2 w &= 0 \end{aligned}$$

Einsetzen in (2.15) zeigt

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

Dann folgt letztendlich

$$0 = \sum_{i=1}^{2} d \cdot (a_i - x) \cdot \varphi_i(x) \qquad \square \text{ (Teil 1)}$$

Beweis (Teil 2).
Zu zeigen ist

$$||\partial_x (v - I_h v)||_{L^\infty(T)} \leq c \cdot h_T \, ||\partial_x^2 v||_{L^\infty(T)}$$

Wir betrachten

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

Einsetzen der Entwicklung für $v(a_i)$ liefert die Gleichung

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

Nun gilt

$$\begin{aligned} \partial_x \varphi_1 &= -\frac{1}{h_T} \\ \partial_x \varphi_2 &= \frac{1}{h_T} \end{aligned}$$

Damit folgt

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

Weiterhin

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

Damit gilt

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

Somit ist der zweite Teil gezeigt und insbesondere der vollständige Beweis geführt.

□

Der nächste Satz behandelt den Interpolationsfehler auf dem Gesamtintervall I.

Satz 2.7.2 *(Interpolationsfehler auf I)*
Auf $I \subset \mathbb{R}$ gilt bei gegebener Zerlegung in Teilintervalle $T \subset I$ mit maximaler Größe h; also $diam(T) \leq h$:

$$||\partial_x^i (v - I_h v)||_{L^2(I)} \leq c \cdot h^{2-i} \, ||\partial_x^2 v||_{L^\infty(I)}, \quad i = 0, 1$$

Man bemerke, dass die Zerlegung nicht notwendigerweise äquidistant sein muß. Weiter sei angemerkt, dass die L^2-Norm auf der linken Seite mit der Supremumsnorm auf der rechten Seite abgeschätzt wird.

Beweis.

$$\begin{aligned}
||\partial_x^i (v - I_h v)||_{L^2(I)}^2 &= \sum_T \int_T \left(\partial_x^i (v - I_h v) \right)^2 \, dx \\
&\leq \sum_T \int_T \left(c^2 \cdot h^{2(2-i)} \, ||\partial_x^2 v||_{L^\infty(T)}^2 \right) \, dx \\
&\leq \sum_T c^2 \, h_T^{2(2-i)} \, ||\partial_x^2 v||_{L^\infty(T)}^2 \int_T 1 \, dx \\
&\leq c^2 \, h^{2(2-i)} \, ||\partial_x^2 v||_{L^\infty(I)}^2 \cdot \underbrace{\mu(I)}_{\text{Maß}}
\end{aligned}$$

In der letzten Ungleichung wurde mit der größten Schrittweite h über das gesamte Intervall abgeschätzt.

□

2.7.2 Energiefehler

Hier wird der Hauptsatz des Abschnitts gezeigt und bewiesen.

Satz 2.7.3 *(Energiefehler)*
Auf $I \subset \mathbb{R}$ gilt bei gegebener Zerlegung $T \subset I$ mit maximaler Größe h

$$||\partial_x(u - u_h)||_{L^2(I)} \leq c \cdot h \, ||\partial_x^2 u||_{L^\infty(I)}$$

Beweis.

$$\begin{aligned} ||\partial_x(u - u_h)||_{L^2(I)} &\leq \inf_{\varphi \in V_h} ||\partial_x(u - \varphi)||_{L^2(I)} \\ &\leq ||\partial_x(u - I_h u)||_{L^2(I)} \\ &\leq c \cdot h \, ||\partial_x^2 u||_{L^\infty(I)} \end{aligned}$$

Es handelt sich hier um eine a priori-Abschätzung (wegen u). Dagegen spricht man von einer a posteriori-Abschätzung für u_h.

2.8 Variationsungleichungen

2.8.1 Minimumsuche in 1D

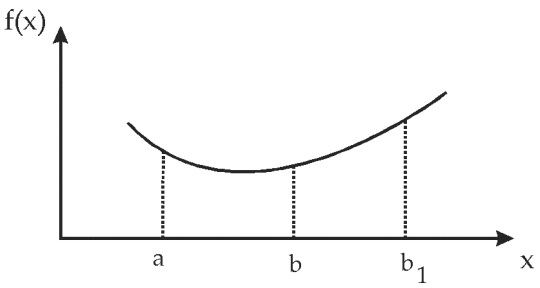


Abbildung 2.4: Absolutes Minimum in $[a, b]$ und relatives Minimum in $[b, b_1]$

Voraussetzung an die Funktion f ist die Differenzierbarkeit. Es folgt eine Fallunterscheidung, auch für Minima am Rand.

$$\begin{array}{rcll} f(a) & \leq & f(x) \quad \forall x \in [a,b] & \rightarrow f'(a) \geq 0 \\ f(b) & \leq & f(x) \quad \forall x \in [a,b] & \rightarrow f'(b) \leq 0 \\ f(x_i) & \leq & f(x) \quad \forall x \in [a,b] & \rightarrow f'(x_i) = 0 \end{array}$$

Kompakte Version der Fallunterscheidung

$$f'(x_0) \cdot (x - x_0) \geq 0 \quad \forall x \in [a, b]$$

Dies kann als erste Variationsungleichung bezeichnet werden.

2.8.2 Minimierung auf konvexer Menge $K \subset \mathbb{R}^n$

Wir betrachten die Funktion $f : K \to \mathbb{R}$. Gesucht ist

$$x_0 \in K : \quad f(x_0) \leq f(x) \quad \forall x \in K$$

Ein Weg in K wird durch die folgende Gleichung beschrieben

$$F(\xi) = f(x_0 + \xi(x - x_0)) \quad \text{mit } \xi \in [0;1]$$

Nun wird die notwendige Bedingung diskutiert. Aus dem 1D-Fall folgt

$$\underbrace{\frac{d}{d\xi} F(\xi)\Big|_{\xi=0}}_{\leftrightarrow f'(x_0)(1D)} \cdot \underbrace{\xi}_{\leftrightarrow (x-x_0)(1D)} \geq 0 \quad \forall \xi \in [0,1]$$

Mit Kettenregel rechnet man aus

$$\begin{aligned} & \nabla f(x_0)(x - x_0) \cdot \xi \geq 0 \quad \forall \xi \in [0,1] \\ \Leftrightarrow \quad & \nabla f(x_0)(x - x_0) \geq 0 \quad \forall x \in K \end{aligned}$$

2.8.3 Elliptische VU 1. Art

Die vorherigen Tatsachen werden nun auf das Drahtproblem mit Tisch als Hinderniss angewendet. Daher auch die Bezeichnung Hindernissproblem.

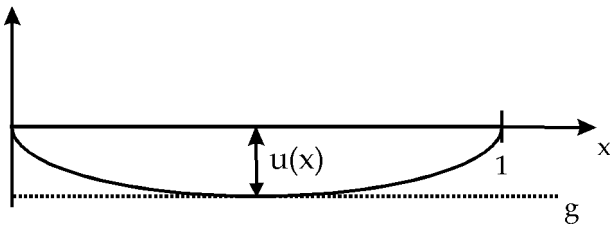


Abbildung 2.5: Hindernissproblem, mit g als Tisch

Ausgangsgleichung soll die Minimalcharakterisierung sein

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

Definition der konvexen Menge K

$$K = \{v \in V := H_0^1(I) | v \geq g\}$$

Für $v_1, v_2 \in K$, $\alpha \in (0,1)$ rechnet man nach

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

Formuliere zu diesem Problem wieder den Weg und wende darauf die notwendige Bedingung an

$$F(\xi) = U(u + \xi(v - u))$$

aus 1D ergibt sich

$$\frac{d}{d\xi}F(\xi)\Big|_{\xi=0} \cdot \xi \geq 0 \quad \forall \xi \in [0,1]$$

Einsetzen liefert

$$\begin{aligned} & 0 \leq \xi \cdot \left(\int_I \partial_x(u + \xi(v-u)) \cdot \partial_x(v-u)dx - \int_I f \cdot (v-u)dx|_{\xi=0} \right) \quad \forall \xi \in [0,1] \\ \Rightarrow \quad & 0 \leq \xi \cdot \left(\int_I \partial_x u \cdot \partial_x(v-u)dx - \int_I f \cdot (v-u)dx \right) \quad \forall \xi \in [0,1] \\ \Rightarrow \quad & 0 \leq \int_I \partial_x u \cdot (\partial_x(v-u))dx - \int_I f \cdot (v-u)dx \quad \forall v \in K \end{aligned}$$

Zusammengefasst erhält man als Ergebnis den Prototypen einer *elliptischen Variationsungleichung 1. Art*

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

Die Diskretisierung von (2.8.3) mit linearen Finiten Elementen ergibt

$$\begin{aligned} & I) \quad a(u, \varphi - u) \geq (f, \varphi - u) \quad \forall \varphi \in K \\ & II) \quad a(u_h, \varphi - u_h) \geq (f, \varphi - u_h) \quad \forall \varphi \in K_h = V_h \cap K,\ K_h \subset K \end{aligned}$$

Der nächste Satz leitet eine Fehlerdarstellung für Variationsungleichungen her. Die eigentliche Aussage ähnelt Satz (2.7.3).

Satz 2.8.1 *(Energiefehler)*
Auf $I \subset \mathbb{R}$ gilt bei gegebener Zerlegung $T \subset I$ mit maximaler Größe h die Abschätzung in der Ableitungsnorm

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

Beweis.
Mit der Bezeichnung $u_i = I_h u$ geht man von der folgenden Gleichung aus

$$\begin{aligned} a(u-u_h, u-u_h) &= a(u-u_h, u-u_i+u_i-u_h) \\ &= a(u-u_h, u-u_i) + a(u-u_h, u_i-u_h) \end{aligned} \tag{2.16}$$

Bei Variationsgleichungen war der zweite Term gleich Null. Hier ist das nicht der Fall. Es gilt

$$\begin{aligned} a(u-u_h, u_i-u_h) &= (f, u_i-u_h) - a(u_h, u_i-u_h) && (2.17) \\ &\quad + a(u, u_i-u) - (f, u_i-u) \\ &\quad + a(u, u-u_h) - (f, u-u_h) && (2.18) \end{aligned}$$

Es folgt

$$\begin{aligned} (2.17) &\leq 0, \quad \text{wegen Test mit } \varphi = u_i \text{ in II)} \\ (2.18) &\leq 0, \quad \text{wegen Test mit } \varphi = u_h \text{ in I)} \end{aligned}$$

Weiter ergibt sich

$$\begin{aligned} & a(u, u_h - u) \geq (f, u_h - u) \\ \Leftrightarrow \quad & a(u, u_h - u) - (f, u_h - u) \geq 0 \\ \Leftrightarrow \quad & a(u, u - u_h) - (f, u - u_h) \leq 0 \end{aligned}$$

Nun weiter in (2.16):

$$\begin{aligned}
a(u-u_h, u-u_h) &= a(u-u_h, u-u_i+u_i-u_h) \\
&= a(u-u_h, u-u_i) + a(u-u_h, u_i-u_h) \\
&\leq ||\partial_x(u-u_h)|| \cdot ||\partial_x(u-u_i)|| + a(u, u_i-u) - (f, u_i-u) \\
&\leq \frac{1}{2}\,||\partial_x(u-u_h)||^2 + \frac{1}{2}\,||\partial_x(u-u_i)||^2 \\
&\quad - \int_I (\partial_x^2 u)(u_i-u)\,dx - (f, u_i-u) \\
&\leq \frac{1}{2}\,||\partial_x(u-u_h)||^2 + \frac{1}{2}\,||\partial_x(u-u_i)||^2 \\
&\quad + ||\partial_x^2 u||\,||u-u_i|| + ||f|| \cdot ||u-u_i||
\end{aligned}$$

Somit erhält man als Ergebnis

$$\begin{aligned}
\frac{1}{2}\,||\partial_x(u-u_h)||^2 &\leq \mathcal{O}(h^2) + \left(||\partial_x^2 u|| + ||f||\right) \cdot ||u-u_i|| \\
&\leq \mathcal{O}(h^2) + \mathcal{O}(h^2)
\end{aligned}$$

□

2.9 A posteriori Fehlerschätzer

Bisher haben wir für lineare FE die Abschätzung

$$||\partial_x(u-u_h)||_{L^2(I)} \leq c \cdot h\,||\partial_x^2 u||_{L^\infty(I)}$$

mit unbekannter Lösung u benutzt; siehe Satz (2.7.3). Ziel soll eine Abschätzung sein, deren rechte Seite komplett berechenbar ist:

$$||\partial_x(u-u_h)||_{L^2(I)} \leq \eta(u_h, f)$$

Als Vorbereitung zum Hauptsatz dient die folgende Aussage.

Satz 2.9.1 *Auf einem Teilintervall $T = (x_i, x_{i+1})$ gilt für $v \in V$ und $v(x_i) = 0$ die Ungleichung*

$$||v||_{L^2(T)} \leq h \cdot ||\partial_x v||_{L^2(T)} \tag{2.19}$$

mit $h = x_{i+1} - x_i$

Beweis.
Für $y \in (x_i, x_{i+1})$ gilt nach dem Hauptsatz der Differential- und Integralrechnung

$$\begin{aligned}
v(y) &= \int_{x_i}^{y} \partial_x v(x)\,dx \\
&\overset{\text{Hölder}}{\leq} \left(\int_{x_i}^{y} 1^2\right)^{\frac{1}{2}} \left(\int_{x_i}^{y} (\partial_x v)^2\,dx\right)^{\frac{1}{2}} \\
&\leq \sqrt{h} \cdot ||\partial_x v||_{L^2(T)}
\end{aligned}$$

Quadrieren bringt

$$v^2(y) \leq \underbrace{h \cdot ||\partial_x v||^2_{L^2(T)}}_{=\text{ const bzgl. } y}$$

Integration

$$\begin{aligned} & \int_T v^2 \, dy \leq \int_T h \cdot ||\partial_x v||^2_{L^2(T)} \, dy \\ \Leftrightarrow \quad & ||v||^2_{L^2(T)} \leq h^2 \, ||\partial_x v||^2_{L^2(T)} \end{aligned}$$

□

Satz 2.9.2 *(Stabilität der Interpolation)*
Auf $I \subset \mathbb{R}$ gilt bei gegebener Zerlegung in Zellen T

$$||\partial_x I_h v||_{L^2(T)} \leq ||\partial_x v||_{L^2(T)} \tag{2.20}$$

Dabei ist wie bisher I_h der Interpolationsoperator im Raum der linearen finiten Elemente.

Beweis.
Es sei $y \in (x_i, x_{i+1})$. Weiter folgt

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

Quadrieren und Integrieren

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

□

Satz 2.9.3 *(Energie-Fehlerschätzer)*
Auf $I \subset \mathbb{R}$ mit Zerlegung in T gilt

$$||\partial_x (u - u_h)||_{L^2(I)} \leq \left(\sum_T h_T^2 \, \varrho_T^2 \right)^{\frac{1}{2}}$$

mit dem Residuum $\varrho_T = 2 \cdot ||f + \partial_x^2 u_h||_{L^2(T)}$.

Beweis.
Folgende Schreibweisen werden verwendet:

$$e = u - u_h$$
$$e_i = I_h e$$

Dabei steht e für error. Es gilt wegen der Galerkin-Orthogonalität $(\partial_x u - \partial_x u_h, \partial_x e_i) = 0$, daher die erste Gleichheit in der folgenden Gleichungskette

$$\begin{aligned}
||\partial_x(u-u_h)||^2_{L^2(I)} &= (\partial_x u - \partial_x u_h, \partial_x e - \partial_x e_i), \quad e_i \in V_h \\
&= (f, e - e_i) - (\partial_x u_h, \partial_x e - \partial_x e_i) \\
&= (f, e - e_i) - \sum_T (\partial_x u_h, \partial_x e - \partial_x e_i)_T \\
&\overset{\text{part. Int.}}{=} (f, e - e_i) - \sum_T \left((-\partial_x^2 u_h, e - e_i)_T + \underbrace{[\partial_x u_h (e - e_i)]_{x_i}^{x_{i+1}}}_{=0} \right)
\end{aligned}$$

Da wir das Problem im 1-dim betrachten, ist der letzte Summand gleich Null. Denn e stimmt mit der Interpolation e_i auf den Punkten x_{i+1} und x_i überein. Insgesamt folgt für die lokale Betrachtung

$$||\partial_x(u-u_h)||^2_{L^2(I)} = \sum_T \left(f + \partial_x^2 u_h, e - e_i\right)_T$$

Weiter gilt mit der Hölderischen Ungleichung

$$\begin{aligned}
(f + \partial_x^2 u_h, e - e_i)_T &\leq ||f + \partial_x^2 u_h||_{L^2(T)} \cdot ||e - e_i||_{L^2(T)} \\
&\overset{(2.19)}{\leq} ||f + \partial_x^2 u_h||_{L^2(T)} \cdot h_T \, ||\partial_x(e - e_i)||_{L^2(T)} \\
&\overset{(2.20)}{\leq} ||f + \partial_x^2 u_h||_{L^2(T)} \cdot h_T \left(||\partial_x e||_{L^2(T)} + ||\partial_x e_i||_{L^2(T)} \right) \\
&\leq \underbrace{||f + \partial_x^2 u_h||_{L^2(T)} \cdot 2}_{=\varrho_T} \cdot h_T \, ||\partial_x e||_{L^2(T)}
\end{aligned}$$

Einsammeln der Resultate auf jeder Zelle

$$\begin{aligned}
||\partial_x(u-u_h)||^2_{L^2(I)} &\leq \sum_T h_T \, \varrho_T \, ||\partial_x e||_{L^2(T)} \\
&\overset{\text{C.-S.}}{\leq} \left(\sum_T h_T^2 \, \varrho_T^2 \right)^{\frac{1}{2}} \cdot \underbrace{\left(\sum_T ||\partial_x e||^2_{L^2(T)} \right)^{\frac{1}{2}}}_{=||\partial_x e||_{L^2(I)}}
\end{aligned}$$

Also

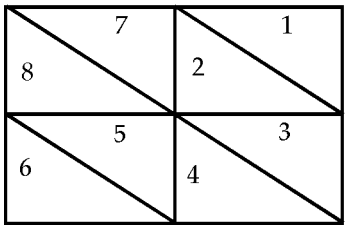
$$\begin{aligned}
&||\partial_x(u-u_h)||^2_{L^2(I)} \leq \left(\sum_T h_T^2 \, \varrho_T^2 \right)^{\frac{1}{2}} \cdot ||\partial_x e||_{L^2(I)} \\
\Leftrightarrow \quad &||\partial_x(u-u_h)||_{L^2(I)} \leq \left(\sum_T h_T^2 \, \varrho_T^2 \right)^{\frac{1}{2}}
\end{aligned}$$

□

Algorithmus 2.9.4 *(Adaptiv)*

1) Gegeben sei ein Gitter T_h
2) Berechne u_h auf T_h
3) Berechne alle ϱ_T
4) Berechne $\eta = \left(\sum \varrho_T^2 \, h_T^2\right)^{\frac{1}{2}}$
5) Falls $\eta < TOL$ (Fehlertoleranz)
$\rightarrow$ fertig, sonst
6) Verfeinere T_h und gehe zu 1)

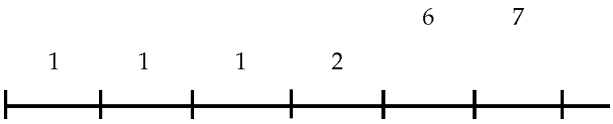
Bemerkung zu 6). Welche Strategien gibt es zum Verfeinern? -Man strebe eine Gleichverteilung der lokalen Schätzerbeiträge an. Dazu die folgende Abbildung.



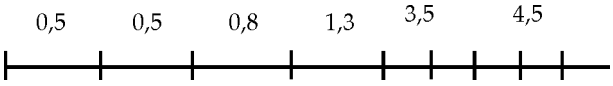
großer Fehler in Feld 1. Kleiner Fehler in den übrigen Feldern.

Schritt: Man teilt Zelle 1, um den Fehler zu verkleinern und den anderen Zellen anzupassen (Gleichverteilung).

Dieser Schritt wird in den nachfolgenden Abbildungen exemplarisch für den 1-dim. Fall gezeigt.



Die beiden hinteren Zellen mit den großen Fehlerwerten werden verkleinert und dann folgt



Wegen der globalen Kopplung verändern sich bei Zellteilung der beiden hinteren Zellen AUCH die Fehlerwerte der übrigen Zellen.

Zusammenfassung
Der A posteriori Fehlerschätzer lässt sich darstellen als

$$||\partial_x(u - u_h)|| \leq \eta(u_h, f),$$

$$\text{mit } \eta(u_h, f) = c \cdot \left(\sum_T \varrho_T^2 \, h_T^2\right)^{\frac{1}{2}}, \quad \varrho_T = ||f + u_h''||_T$$

Bei linearem Ansatz ist $u_h'' = 0$. Klar!

2.10 Referenzelement, Gebietstransformation

Ziel: Fast alle Rechnungen auf einem Referenzelement durchzuführen (z. B. dem Einheitsintervall)

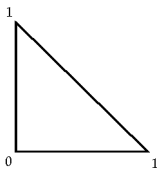


Abbildung 2.6: Skizze eines Referenzelements

Was spricht für die Betrachtung des Problems auf einem Referenzelement?

- Basisfunktionen nur einmal ausrechnen,
- numerische Integrationsformeln werden nur auf dem Referenzelement benötigt.

Wir betrachten die Funktion $T_h : x$ mit $x \in [x_0, x_1]$

und transformieren diese auf das Einheitsintervall mit $T_1 : \xi$, $\xi \in [0, 1]$

Vorbereitungen für die Substitutionsregel

$$\begin{aligned} F_h : T_1 &\to T_h \\ \xi &\mapsto x = x_0 + \xi \cdot (x_1 - x_0) \end{aligned}$$

Weiter

$$\begin{aligned} \frac{d}{dx} : \quad & 1 = (x_1 - x_0)\,\frac{d\xi}{dx} \\ \Rightarrow \quad & dx = (x_1 - x_0) \cdot d\xi \end{aligned}$$

Dabei wird $J := x_1 - x_0$ als Flächentransformation definiert. Im 1-dim gilt $J = h$. Im Allg. gilt das nicht mehr. Man konstruiere nun die Umkehrfunktion

$$\begin{aligned} F_h^{-1} : T_h &\to T_1 \\ x &\mapsto \xi = \frac{x - x_0}{x_1 - x_0} \end{aligned}$$

mit der Ableitung

$$\xi_x = \frac{d\xi}{dx} = \frac{1}{x_1 - x_0}$$

Eine Basisfunktion φ_i^h auf T_h wird angesetzt mit

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

und für die Ableitung folgt mit der Kettenregel

$$\partial_x \varphi_i^h(x) = \partial_\xi \varphi_i^1(\xi) \cdot \partial_x F_h^{-1}(x) = \partial_\xi \varphi_i^1(\xi) \cdot \xi_x$$

mit $F_h^{-1}(x) = \xi$.

Beispiele.
1)

$$\int_{T_h} f(x)\, \varphi_i^h(x)\, dx \overset{\text{Sub.}}{=} \int_{T_1} f(F_h(\xi)) \cdot \varphi_i^1(\xi) \cdot J \cdot d\xi \tag{2.21}$$

2)

$$\int_{T_h} \partial_x \varphi_i^h(x) \cdot \partial_x \varphi_j^h(x)\, dx = \int_{T_1} \left(\partial_\xi \varphi_i^1(\xi)\right) \cdot \xi_x \cdot \left(\partial_\xi \varphi_j^1(\xi)\right) \cdot \xi_x \cdot J\, d\xi \tag{2.22}$$

Die Integrale werden mit Hilfe der numerische Integration berechnet, dazu die allgemeine Formel

$$\int_{T_1} g(\xi) \approx \sum_{k=1}^{q} \omega_k g(\xi_k)$$

mit Integrationsgewichten ω_k und Stützstellen ξ_k.

Beispiel.
Für $q = 2$ erhält man die Trapezregel mit

$$\xi_1 = 0,\ \xi_2 = 1, \quad \omega_1 = \frac{1}{2},\ \omega_2 = \frac{1}{2}$$

Die numerische Integration wird nun auf die beiden obigen Integrale angewendet,

$$\int_{T_h} f(x)\, \varphi_i^h(x)\, dx \overset{(2.21)}{\approx} \sum_{k=1}^{q} \omega_k f(F_h(\xi_k))\, \varphi_i^1(\xi_k) \cdot J$$

und für das zweite Beispiel

$$\int_{T_h} \partial_x \varphi_j^h(x) \partial_x \varphi_i^h(x)\, dx \approx \sum_{k=1}^{q} \omega_k \left(\partial_\xi \varphi_j^1(\xi_k) \cdot \xi_x\right) \cdot \left(\partial_\xi \varphi_i^1(\xi_k) \cdot \xi_x\right) \cdot J$$

2.11 Rechentechnische Betrachtungen

Hier wird kurz die Implementierung in einer beliebigen Programmiersprache gezeigt. Dazu wird die Konstruktion der Matrix A besprochen

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

NICHT:

$$\begin{aligned} &\text{for } i = 1 \ldots n \\ &\quad \text{for } j = 1 \ldots n \\ &\qquad A_{ij} = \int_I \partial_x \varphi_j \partial_x \varphi_i \, dx \end{aligned}$$

PRAXIS:

$$\begin{aligned} &\text{forall } T \\ &\text{for } i = 1 \ldots n \\ &\quad \text{for } j = 1 \ldots n \\ &\qquad A_{ij} += \int_T \partial_x \varphi_j \partial_x \varphi_i \, dx \end{aligned}$$

Noch besser ist die Betrachtung auf einer Zelle.

$$\begin{aligned} &\text{for } k = 1 \ldots q \quad // \text{ über alle Integrationspunkte} \\ &\text{Berechne Basisfunktionen auf } T^1 \text{ in } \xi_k \\ &\text{for } i = 1 \ldots \text{local}_n \\ &\quad \text{for } j = 1 \ldots \text{local}_n \\ &\qquad A += \omega_k \cdot J \cdot \partial_\xi \varphi_j^1 \cdot \xi_x \cdot \partial_\xi \varphi_j^1 \xi_x \end{aligned}$$

mit $A = A_{\text{global}(i),\, \text{global}(i)}$ wegen der Transformation. Man beachte weiterhin, dass der Kern des Problems im Lösen der Differentialgleichung

$$\partial_\xi \varphi_j^1 \cdot \xi_x \cdot \partial_\xi \varphi_j^1 \xi_x$$

besteht.

3 FEM für elliptische Probleme (2D)

In diesem Kapitel werden die bis hierher gemachten Überlegungen in 2D übertragen. Dazu wird die Modellgleichung in den drei bekannten Formulierungen angegeben. Im Mehrdimensionalen spricht man von der *Poisson-Gleichung*. Anschließend werden die Sätze aus dem 1D-Fall im abstrakteren Rahmen mit Hilfe der Funktionalanalysis diskutiert.

3.1 Typeinteilung PDGL 2. Ordnung

Typische Vertreter von partiellen Differentialgleichungen 2. Ordnung sind die Laplace- und Wellengleichung. Allerdings unterscheiden sich schon die Lösungen, von $\partial_{xx} + \partial_{yy} = 0$ (Laplace-Gleichung) und $\partial_{xx} - \partial_{yy} = 0$ (Wellengleichung), grundsätzlich.

Wir betrachten die allgemeinste Darstellung einer PDGL 2. Ordnung. Der Faktor 2 bei der gemischten Ableitung wird aus rein praktischen Gründen hinzugefügt.

$$a_{11}\partial_x^2 u + 2a_{12}\partial_x\partial_y u + a_{22}\partial_y^2 u + a_1\partial_x u + a_2\partial_y u + a_0 u = 0 \tag{3.1}$$

mit $a_{ij} \in \mathbb{R}, u = u(x, y)$.

Satz 3.1.1 *Gleichung* (3.1) *kann durch lineare Transformation der unabhängigen Variablen x, y auf eine der drei folgenden Formen gebracht werden.*

1. Elliptisch, *falls* $a_{12}^2 < a_{11}a_{22}$.
   $\rightarrow \partial_x^2 u + \partial_y^2 u + \ldots = 0 \quad \overset{allg.}{\rightarrow} \Delta u = 0$
2. Hyperbolisch, *falls* $a_{12}^2 > a_{11}a_{22}$.
   $\rightarrow \partial_x^2 u - \partial_y^2 u + \ldots = 0 \quad \overset{allg.}{\rightarrow} \partial_t^2 u - \Delta u = 0$
3. Parabolisch, *falls* $a_{12}^2 = a_{11}a_{22}$.
   $\rightarrow \partial_x^2 u + \ldots = 0 \quad \overset{allg.}{\rightarrow} \partial_t u - \Delta u = 0$
   Beispiel hierzu: $\partial_t u - \partial_x^2 = 0$ *Wärmeleitung.*

Die . . .(Pünktchen) im obigen Satz stellen Terme der Ordnung 1 und 0 dar.

Bevor der Beweis geführt wird, werden die Analogien zur analytischen Geometrie dargestellt.

1. $x^2 + y^2 = 1$ Kreis/ Ellipse

2. $x^2 - y^2 = 1$ Hyperbel

3. $y = x^2$ Parabel

Beweis. Für den elliptischen Fall.
Die anderen Fälle werden analog geführt. Es sei o.B.d.A: $a_{11} = 1, a_1 = a_2 = a_0 = 0$.
Mit quadratischer Ergänzung erhält man aus (3.1) folgenden Ausdruck:

$$\begin{aligned} 0 &= (\partial_x + a_{12}\partial_y)^2 u + (a_{22} - a_{12}^2)\partial_y^2 u \\ &= (\partial_x^2 + 2a_{12}\partial_x\partial_y + a_{12}^2\partial_y^2)u + (a_{22} - a_{12}^2)\partial_y^2 u \end{aligned} \tag{3.2}$$

mit Voraussetzung: $a_{12}^2 < a_{11} \cdot a_{22}$. Setze $b := \sqrt{a_{22} - a_{12}^2} > 0$.
Führe nun eine Koordinatentransformation durch:

$$x = \xi, \quad y = a_{12} \cdot \xi + b \cdot \eta \quad \text{(linear in } \xi \text{ und } \eta\text{)}$$

Frage: Wie transformieren sich die Ableitungen im neuen Koordinatensystem?
Man definiere nun

$$v(\xi, \eta) = u(x, y)$$

und interpretiere x und y als Funktionen:

$$x(\xi, \eta), y(\xi, \eta) \to v(\xi, \eta) = u(x(\xi, \eta), y(\xi, \eta))$$

Man berechne $(\partial_\xi v, \partial_\eta v)$:

$$\begin{aligned} (\partial_\xi v, \partial_\eta v) &= (\partial_x u, \partial_y u) \begin{pmatrix} \partial_\xi x & \partial_\eta x \\ \partial_\xi y & \partial_\eta y \end{pmatrix} \\ &= (\partial_x u, \partial_y u) \begin{pmatrix} 1 & 0 \\ a_{12} & b \end{pmatrix} \\ &= (\partial_x u + a_{12}\partial_y u, 0 \cdot \partial_x u + b \cdot \partial_y u) \end{aligned}$$

Damit folgt: $\partial_\xi = \partial_x + a_{12}\partial_y$ und $\partial_\eta = b \cdot \partial_y$.
Setze jetzt in (3.2) ein. Die resultierende Gleichung

$$\partial_\xi^2 + \partial_\eta^2 = 0$$

ist die Laplace-Gleichung.

□

Bemerkung.
Die oben durchgeführten Untersuchungen können gleichwohl mit Hilfe der *linearen Algebra* behandelt werden. Hier soll kurz die grobe Struktur angerissen werden.

Für die einzelnen Gleichungen mit den Variablen und Koeffizienten läßt sich ein lineares System aufstellen. Die resultierende Koeffizientenmatrix ist symmetrisch und kann daher gut mit der *Eigenwerttheorie* angegangen werden. Symmetrische Matrizen lassen sich in Diagonalmatrizen konvertieren. Hier stehen dann in der Diagonalen die Eigenwerte. Anhand derer lassen sich die Eigenschaften - elliptisch, hyperbolisch, parabolisch - bestimmen.

Nun zwei Beispiele zur Typeinteilung PDgl 2.Ordnung.
Beispiel 1.
Klassifiziere die folgenden Gleichungen.

1. $\partial_{xx}u - 5\partial_{xy}u = 0$
2. $4\,\partial_{xx}u - 12\partial_{xy}u + 9\partial_{yy}u + \partial_y u = 0$
3. $4\,\partial_{xx}u + 6\partial_{xy}u + 9\partial_{yy}u = 0$

Wie oben besprochen wird die „Diskriminante“$D = a_{12}^2 - a_{11}a_{22}$ zur Entscheidung herangezogen.

1. $D = (-\frac{5}{2})^2 - 1 \cdot 0 = \frac{25}{4} > 0 \quad \rightarrow$ hyperbolisch
2. $D = (-6)^2 - 4 \cdot 9 = 36 - 36 = 0 \quad \rightarrow$ parabolisch
3. $D = 3^2 - 4 \cdot 9 = 9 - 36 = -25 < 0 \quad \rightarrow$ elliptisch.

Beispiel 2.
Bestimme Teilbereiche in der xy-Ebene, in denen die Gleichung

$$y\partial_{xx}u - 2\partial_{xy}u + x\partial_{yy}u = 0$$

elliptisch, hyperbolisch oder parabolisch ist.

Zunächst gilt: $D = (-1)^2 - y \cdot x = 1 - yx$.

Parabolisch auf Hyperbel $xy = 1$. Denn $1 - 1 = 0$. Die elliptischen Bereiche sind konvex: $xy > 1$. Denn $D = 1 - yx < 0$. Im zusammenhängenden Bereich $xy < 1$ ist die Gleichung hyperbolisch.

3.2 Poisson-Problem

Die Poisson-Gleichung entspricht der inhomogenen Laplace-Gleichung.

KLASSISCHES POISSON-PROBLEM

Auch als Dirichlet-Problem $\mathcal{D}$ bekannt.

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

MINIMALCHARAKTERISIERUNG $\mathcal{M}$

Das zugrundeliegende Rechengebiet $\Omega \subset \mathbb{R}$ sei beschränkt. Es ist

$$\min \frac{1}{2} \int_\Omega (\nabla u)^2 \, dx - \int_\Omega f \cdot u \, dx$$
$$\Leftrightarrow \quad \min \frac{1}{2} a(u,u) - (f,u)$$

mit $u = u(x), f = f(x), x = (x_1, x_2) \in \mathbb{R}^2 \supset \Omega$. Dabei sind f, u Funktionen mit

$$f, u : \Omega \to \mathbb{R}, \quad \nabla u = (\partial_{x_1} u, \partial_{x_2} u)$$

Gesucht ist die Lösung von $\mathcal{M}$ in

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

Im höherdimensionalen spielt die Green´sche Formel eine wesentliche Rolle. Diese ist mit der partiellen Integration in 1D zu vergleichen.

Satz 3.2.1 *(Green'sche Formel)*
Für hinreichend glatte Funktionen v, w gilt

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

mit $\Delta = \partial_{x_1}^2 + \partial_{x_2}^2$, $\partial_n w = (\nabla w) \cdot n$

Beweis in der Übung.

Satz 3.2.2 *Für hinreichend glattes u gilt: Aus dem Dirichlet-Problem $\mathcal{D}$ folgt die Minimum-Gleichung $\mathcal{M}$. Die Umkehrung gilt erstmal nicht.*

Beweis. Analog zum 1D Fall unter Anwendung der Green'schen Formel.

VARIATIONELLE FORMULIERUNG

Wir betrachten nun die variationelle Formulierung $\mathcal{V}$

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

mit $a(v,w) := (\nabla v, \nabla w) = \int_\Omega \nabla v \cdot \nabla w dx$ (Integral ist Lebesgue-Integrierbar). Allgemein ist

$$(v, w) := \int_\Omega v \cdot w \, dx$$

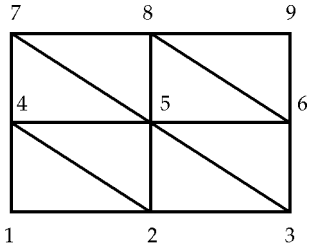
mit $v, w \in V$.

Satz 3.2.3 *Die variationelle Gleichung $\mathcal{V}$ gilt genau dann, wenn das Minimum-Problem $\mathcal{M}$ gilt.*

Beweis. Analog zum 1D Fall!

Finite Elemente

Es sei die Triangulierung $\mathbb{T}_h$ gegeben und weiter sei Ω polygonal.



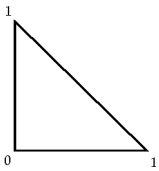
Es sei h der Gitterparameter mit

$$h = \max_{T \in \mathbb{T}} \operatorname{diam}(T)$$

mit $\operatorname{diam}(T)$ = „Längste Seite von T“. Desweiteren sei ein diskreter Teilraum $V_h \subset V$ gegeben mit

$$V_h = \{\varphi | \varphi \in V, \varphi|_T \text{ ist linear für } T \in \mathbb{T}_h \}$$

Ansatz: $a + b\xi + c\eta$ auf T_1.



Diskretisierung

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

Satz 3.2.4 *(Galerkin-Eigenschaft)*
Es gilt

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

Diese Eigenschaft gilt ausdrücklich nur für Funktionen aus dem diskretisierten Teilraum V_h.

Der nächste Satz gibt eine erste Fehlerabschätzung an.

Satz 3.2.5 *Es gilt die Abschätzung*

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

wobei $I_h u$ definiert ist durch

$$u \in V : \quad I_h u \in V_h \quad \textit{und} \quad I_h u(x_i) = u(x_i)$$

mit den Ecken x_i aller $T \in \mathbb{T}_h$.

Beweis.
Wir beginnen mit

$$\begin{aligned} ||\nabla(u-u_h)||^2 &= a(u-u_h, u-\varphi_h+\varphi_h-u_h) \\ &= a(u-u_h, u-\varphi_h) + \underbrace{a(u-u_h, \varphi_h-u_h)}_{=0 \text{ wg. Gal. Eig.}} \end{aligned}$$

Es folgt

$$\begin{aligned} & ||\nabla(u-u_h)||^2 \leq ||\nabla(u-u_h)|| \cdot ||\nabla(u-\varphi)|| \\ \Rightarrow \quad & ||\nabla(u-u_h)|| \leq ||\nabla(u-\varphi_h)|| \end{aligned}$$

mit $\varphi_h = I_h u$. Dieses Ergebnis führt zu dem folgenden Satz:

Satz 3.2.6 *Die linke Seite des vorherigen Satzes kann mit h abgeschätzt werden. Also*

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

Beweis. Nicht geführt.

3.3 Natürliche und wesentliche Randbedingungen

KLASSISCH
Das Dirichlet-Problem $\mathcal{D}$

$$-\Delta u + u = f \quad \text{auf } \Omega \tag{3.3}$$

mit Normalableitung

$$\partial_n u = g \quad \text{auf } \Gamma = \partial\Omega$$

speziell ist $g : \Gamma \to \mathbb{R}, g = g(x_1, x_2)$

Definition 3.3.1 *Die verschiedenen Randbedingungen dienen zur Eindeutigkeit der Lösung und erhalten spezielle Bezeichnungen*

$$\begin{aligned} \partial_n u &= g \quad \textit{Neumann-Bedingung, auch natürliche Randbedingung genannt} \\ u &= u_0 \quad \textit{Dirichlet-Bedingung, wesentliche Randbedingung} \end{aligned}$$

VARIATIONELLE FORMULIERUNG $\mathcal{V}$

$$a(u,\varphi) = (f,\varphi) + \int_\Gamma g \cdot \varphi d\Gamma \quad \forall \varphi \in V \tag{3.4}$$

$V = \{\varphi | \varphi$ ist stetig; $\partial_{x_i}\varphi$ ist stückweise stetig und beschränkt $\}$. Bemerke, dass in diesem Raum V im Gegensatz zu oben, keine Abhängigkeit von der Randbedingung gefordert wird.

$$\begin{aligned} a(u,\varphi) &= \int_\Omega \nabla u \cdot \nabla\varphi dx + \int_\Omega u \cdot \varphi dx \\ (f,\varphi) &= \int_\Omega f \cdot \varphi dx \end{aligned}$$

MINIMUM-PROBLEM $\mathcal{M}$

$$u \in V: \quad \min_{\varphi \in V} \frac{1}{2} a(\varphi,\varphi) - (f,\varphi) - \int_\Gamma g \cdot \varphi \, d\Gamma \tag{3.5}$$

Satz 3.3.2 *Aus der Dirichlet-Formulierung $\mathcal{D}$ folgt die variationelle Gleichung $\mathcal{V}$. Die Umkehrung gilt hier noch nicht.*

Beweis.
1) $-\Delta u + u = f$ (gilt punktweise)

2) $(-\Delta u, \varphi) + (u, \varphi) = (f, \varphi) \quad \forall \varphi \in V$ (gilt Integralweise). Es gilt i. Allg. Aus punktweise $\Rightarrow$ Integralweise.

Green'sche Formel
3) $(f, \varphi) = \int_\Omega \nabla u \cdot \nabla \varphi dx - \int_\Gamma \partial_n u \cdot \varphi d\Gamma + \int_\Omega u \cdot \varphi dx$. Beachte im mittleren Integral die Beziehung $\partial_n u = g$

4) $\int_\Omega \nabla u \cdot \nabla \varphi dx + \int_\Omega u \cdot \varphi dx = (f, \varphi) + \int_I g \cdot \varphi dx \quad \forall \varphi \in V$

□

Satz 3.3.3 *Lösung u von $\mathcal{V}$ sei hinreichend glatt. Dann gilt $\mathcal{V} \Rightarrow \mathcal{D}$.*

Beweis.
Wir verwenden die Green'sche Formel

$$\begin{aligned}(f, \varphi) + \int_\Gamma g \cdot \varphi \, d\Gamma &= a(u, \varphi) \\ &= \int_\Gamma \partial_n u \cdot \varphi \, d\Gamma + \int_\Omega (-\Delta u + u) \cdot \varphi \, dx\end{aligned}$$

$$\Leftrightarrow \quad \int_\Omega (-\Delta u + u - f) \cdot \varphi \, dx + \int_\Gamma (\partial_n u - g) \cdot \varphi \, d\Gamma = 0 \quad \forall \varphi \in V \tag{3.6}$$

Insbesondere gilt (3.6) für $\overline{\varphi} \in V$ mit der zusätzlichen Bedingung $\overline{\varphi} = 0$ auf Γ. Also gilt

$$\begin{aligned}&\int_\Omega (-\Delta u + u - f) \cdot \overline{\varphi} \, dx = 0 \\ \Rightarrow \quad &-\Delta u + u - f = 0 \quad \text{auf } \Omega \quad \text{punktweise}\end{aligned}$$

Benutze dieses Resultat, um festzustellen

$$\int_\Gamma (\partial_n u - g) \cdot \varphi \, d\Gamma = 0 \quad \forall \varphi \in V$$

Wiederum Variationsargument

$$\Rightarrow \partial_n u - g = 0$$

□

Merke: Die klassische Bedingung taucht in der variationellen Gleichung nirgends explizit auf.

ZUSAMMENFASSUNG
Kurz dargestellt:

$$\text{Dirichlet } \mathcal{D} \Longleftrightarrow \text{Variationell } \mathcal{V} \Longleftrightarrow \text{Minimum } \mathcal{M}$$

3.4 Sobolev-Räume

Gegeben sei das Gebiet $\Omega \subset \mathbb{R}^n$, offen, stückweise glatter Rand. Wir definieren den Raum der Quadrat-Integrablen Funktionen mit

$$L_2(\Omega) = \{v | v \text{ ist definiert auf } \Omega \text{ und } \int_\Omega v^2 \, dx < \infty\}$$

Definiere Skalarprodukt

$$(v, w)_0 := (v, w)_{L_2(\Omega)} = \int_\Omega v \cdot w \, dx$$

Damit ist $L_2(\Omega)$ ein *Hilbertraum* mit der Norm

$$||v||_0 = \sqrt{(v, v)_0}$$

Insbesondere ist der Raum $L_2(\Omega)$ mit der L_2-Norm vollständig.

Definition 3.4.1 *(schwache Ableitung)*
Die Funktion $u \in L_2(\Omega)$ besitzt in $L_2(\Omega)$ die schwache Ableitung

$$v = \partial^\alpha u, \quad \textit{falls } v \in L_2(\Omega) \quad \textit{und}$$

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

Mit Multiindex $\alpha = (\alpha_1, \ldots, \alpha_n), \alpha_i \in \mathbb{N}_0$ und
$|\alpha| = \sum \alpha_i$, außerdem ist $\partial^\alpha = \partial_{x_1}^{\alpha_1} \cdot \partial_{x_2}^{\alpha_2} \cdots \partial_{x_n}^{\alpha_n}$. Beachte: $C_0^\infty(\Omega)$ ist der Raum der unendlich oft differenzierbaren Funktionen - mit kompakten Träger Ω - und Nullrandwerten. Damit gilt $C_0^\infty(\Omega) \subset C^\infty(\Omega)$.

Beispiel.
Betrachte $|\alpha| = 1$, dann folgt

$$(\varphi, \partial_{x_i} u) = -(\partial_{x_i} \varphi, u) \quad \forall \varphi \in C_0^\infty(\Omega)$$

Definition 3.4.2 *(Sobolev-Räume)*
Sei $m \in \mathbb{N}_0$ vorgegeben, dann definiert

$$H^m(\Omega) = \{u \in L_2(\Omega) | u \textit{ besitzt schwache Ableitungen } \partial^\alpha u \textit{ für alle } |\alpha| \leq m\}$$

einen Funktionenraum, in dem durch

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

ein Skalarprodukt bestimmt wird. Mit Hilfe des Skalarprodukts werden zwei (Halb-)Normen definiert durch

$$\textit{Norm} \quad ||u||_m = \sqrt{(u, u)_m}$$

$$\textit{Halbnorm} \quad |u|_m = \sqrt{\sum_{|\alpha| = m} ||\partial^\alpha u||_0^2}$$

Bemerkung.
Mit $||\cdot||_m$ ist $H^m(\Omega)$ vollständiger Hilbertraum

Satz 3.4.3 *Sei $m \in \mathbb{N}_0$. Dann liegt*

$$C^\infty(\Omega) \cap H^m(\Omega) \text{ dicht in } H^m(\Omega)$$

bzw.

$$\forall \varphi \in H^m \ \exists \varphi_\varepsilon \in C^\infty, \quad \text{so dass } ||\varphi - \varphi_\varepsilon||_m \leq \varepsilon$$

Das Ziel der nächsten Definition ist es, den Raum $C_0^\infty(\Omega)$ zu vervollständigen.

Definition 3.4.4 *(Verallgemeinerung von Nullrandwerten)*
Die Vervollständigung von $C_0^\infty(\Omega)$ bzgl. der Sobolev-Norm $||\cdot||_m$ wird mit $H^m(\Omega)$ bezeichnet.

Beispiel 1.
Geeignete Wahl von V bei

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

lautet $V = H_0^1(\Omega)$ und man hat

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in V := H_0^1(\Omega)$$

Beispiel 2.

$$\begin{aligned} -\Delta u + u &= f \\ \partial_n u &= g \end{aligned}$$

$$\rightarrow \quad (\nabla u, \nabla \varphi) + (u, \varphi) = (f, \varphi) + \int_\Gamma g \cdot \varphi \, d\Gamma \quad \forall \varphi \in V = H^1(\Omega)$$

Aus den beiden Beispielen folgt somit

$$H_0^1(\Omega) \subset H^1(\Omega)$$

Satz 3.4.5 *(Poincaré-Ungleichung)*
Es gilt die Abschätzung

$$||v||_0 \leq C \cdot |v|_1 \quad \text{mit } v \in H_0^1(\Omega)$$

Beweis. Wurde in der Übung nachgerechnet.
Zunächst gilt für die Norm

$$||v||_0 = \sqrt{(v, v)_0}$$

Außerdem ist nach Voraussetzung $v \in [0, 1], v(0) = 0$. Man schreibe die Norm als Integral

$$||v||_0 = \int_0^1 v \cdot v \, dx = \int_0^1 v^2(x) \, dx$$

Für eine Funktion u folgert man

$$u(x) = \int_0^x u'(t)\,dt$$

Quadrieren zeigt

$$\begin{aligned} u^2(x) &= \left(\int_0^x u'(t)\,dt\right)^2 \\ &\leq \int_0^1 (u'(t))^2\,dt \\ &= ||u'||^2 \end{aligned}$$

Also damit

$$\int_0^1 u^2(x)\,dx \leq ||u'||^2$$

□

3.5 Abstrakte Formulierung

V sei ein Hilbertraum. Insbesondere ist V vollständig und es ist ein Skalarprodukt definiert.

- Skalarprodukt $(\cdot,\cdot)_V$
- Norm $||\cdot||_V = \sqrt{(\cdot,\cdot)_V}$
- Bilinearform $a(\cdot,\cdot) : V \times V \to \mathbb{R}$
- Linearform $L(\cdot) : V \to \mathbb{R}$

VARIATIONELLE FORMULIERUNG

$$u \in V : \quad a(u,\varphi) = L(\varphi) \quad \forall \varphi \in V \tag{3.7}$$

Beispiel 1.

$$\begin{aligned} a(u,\varphi) &:= (\nabla u, \nabla\varphi) \quad \forall \varphi \in V := H_0^1(\Omega) \\ L(\varphi) &:= (f,\varphi) \quad \forall \varphi \in V := H_0^1(\Omega) \end{aligned}$$

Beispiel 2.

$$\begin{aligned} a(u,\varphi) &:= (\nabla u, \nabla\varphi) + (u,\varphi) \\ L(\varphi) &:= (f,\varphi) + \int_\Gamma g\cdot\varphi d\Gamma \end{aligned}$$

mit $V := H^1(\Omega)$.

ABSTRAKTES MINIMIERUNGSPROBLEM

$$F(u) \leq F(\varphi) \quad \forall \varphi \in V \quad \text{mit } F(\varphi) = \frac{1}{2}a(\varphi,\varphi) - L(\varphi)$$

Damit ergeben sich i. Allg. die folgenden Voraussetzungen

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

V ii) $a(\cdot,\cdot)$ ist stetig. Es gilt nämlich

$$|a(v,w)| \leq C\,||v||_V\,||w||_V \quad \forall v,w \in V,\, C>0$$

V iii) $a(\cdot,\cdot)$ ist V-Elliptisch

$$\alpha\,||v||_V^2 \leq a(v,v) \quad \forall v \in V,\, \alpha>0$$

V iv) $L(\cdot)$ ist stetig

$$|L(v)| \leq \Lambda\,||v||_V \quad \forall v \in V,\, \Lambda>0$$

Satz 3.5.1 *(Existenzsatz)*
Es gelten V i) bis V iv). Dann existiert genau eine Lösung $u \in V$ mit der Stabilitätsbedingung

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

Beweis.
i) Eindeutigkeit

Annahme: $u_1, u_2 \in V$ und u_1, u_2 lösen

$$\begin{aligned} a(u_1,\varphi) &= L(\varphi) \quad \forall \varphi \in V \\ a(u_2,\varphi) &= L(\varphi) \quad \forall \varphi \in V \end{aligned}$$

Subtrahiere:

$$a(u_1-u_2,\varphi)=0 \quad \forall \varphi \in V$$

Wähle speziell $\varphi = u_1 - u_2$

$$a(u_1-u_2,u_1-u_2)=0$$

Benutze V iii)

$$0 \leq \alpha\,||u_1-u_2||_V^2 \leq a(u_1-u_2,u_1-u_2)=0$$

Dann folgt

$$||u_1-u_2||_V=0 \quad \Rightarrow u_1=u_2$$

ii) Existenz

Idee: Reduziere $a(u,\varphi)=L(\varphi)$ auf ein Fixpunktproblem.

Riesz'scher Darstellungssatz (für Hilberträume):

$$\exists l \in V: \quad L(v)=(l,v)_V$$

Außerdem sei $A \in L(V,V)=L(V)$ (Raum der stetigen linearen Abbildungen). Dann gilt

$$a(u,v)=(Au,v)_V$$

Rechne

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

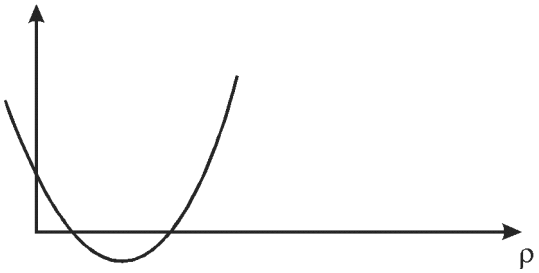
Betrachte nun $W_\varrho : V \to V$ mit $W_\varrho(v) = v - \varrho(Av - l)$. Zeige, dass W_ϱ eine Kontraktion ist, um somit auf einen Fixpunkt zu schließen.

$$\begin{aligned} & ||W_\varrho(v_1) - W_\varrho(v_2)||_V^2 \\ = \quad & ||v_1 - \varrho(Av_1 - l) - v_2 + \varrho(Av_2 - l)||_V^2 \\ = \quad & (v_1 - \varrho A v_1 - (v_2 - \varrho A v_2), v_1 - v_2 - \varrho(Av_1 - Av_2))_V \\ = \quad & (v_1 - v_2, v_1 - v_2)_V - 2\varrho(A(v_1 - v_2), v_1 - v_2)_V + \varrho^2(A(v_1 - v_2), A(v_1 - v_2))_V \\ = \quad & ||v_1 - v_2||_V^2 - 2\varrho a(v_1 - v_2, v_1 - v_2) + \varrho^2 \, ||A(v_1 - v_2||_V^2 \\ \overset{Viii)}{\leq} \quad & ||v_1 - v_2||_V^2 - 2\varrho\alpha \, ||v_1 - v_2||_V^2 + \varrho^2 ||A||_V^2 \, ||v_1 - v_2||_V^2 \\ = \quad & (1 - 2\varrho\alpha + \varrho^2 ||A||^2) \cdot ||v_1 - v_2||_V^2 \end{aligned}$$

Scharfes Hinsehen liefert Kontraktion für $(1 - 2\varrho\alpha + \varrho^2||A||^2) < 1$. Also

$$-2\varrho\alpha + \varrho^2 \, ||A||^2 < 0$$

Bestimme zu dieser Gleichung die Nullstellen von ϱ. Dazu die Skizze



Es folgt

$$0 < \varrho < \frac{2\alpha}{||A||^2},$$

dann ist

$$0 < 1 - 2\varrho\alpha + \varrho^2 < 1$$

Damit ist ϱ so wählbar, dass W_ϱ eine Kontraktionsabbildung ist. Es existiert also ein Fixpunkt

$$\begin{aligned} u &= W_\varrho(u) \\ u &= u - \varrho(Au - l) \end{aligned}$$

Dieser Fixpunkt ist Lösung von

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

iii) Stabilität

$$\begin{aligned} & \alpha||u||_V^2 \leq a(u,u) = L(u) \leq \Lambda||u||_V \\ \Rightarrow \quad & ||u||_V \leq \tfrac{\Lambda}{\alpha} \end{aligned}$$

□

3.6 Diskretisierung

In diesem Abschnitt soll auf die Diskretisierung eingegangen werden. Wir haben wie gewohnt

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

Der endlich-dimensionale Raum V_h wird durch die Basisvektoren
$< \varphi_1, \ldots, \varphi_N >$ aufgespannt. Dementsprechend lassen sich die folgenden Linearkombinationen bilden

$$\begin{aligned} \varphi \in V_h: \quad & \varphi = \sum_{i=1}^{N} \alpha_i \varphi_i, \quad \alpha_i \in \mathbb{R} \\ u_h \in V_h: \quad & u_h = \sum_{j=1}^{N} x_j \varphi_j, \quad x_j \in \mathbb{R} \end{aligned}$$

In der abstrakten Schreibweise lässt sich unser Problem so formulieren

$$a(u_h, \varphi_i) = L(\varphi_i), \quad i = 1, \ldots, N$$

Einsetzen der Linearkombination für u_h zeigt

$$\sum_{j=1}^{N} a(\varphi_j, \varphi_i)\, x_j = L(\varphi_i), \quad i = 1, \ldots, N$$

Dann lassen sich die einzelnen Komponenten des Gleichungssystems $Ax = b$ wie folgt darstellen

$$\begin{aligned} A_{ij} &= a(\varphi_j, \varphi_i) \\ b_i &= L(\varphi_i) \\ x &= (u_1, \ldots, u_N)^T \end{aligned}$$

Die diskretisierte Lösung u_h genügt den nächsten Sätzen

Satz 3.6.1 *Unter den Voraussetzungen V i) bis V iv) des vorherigen Abschnitts ist die Matrix A symmetrisch und positiv definit.*

Beweis. Nicht geführt.

Satz 3.6.2 *Es gelten V i) bis V iv), dann gilt die Stabilitätsbedingung*

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

Beweis. Übung.

Zuletzt kann eine Approximation des Fehlers angegeben werden, der weiter unten noch näher bestimmt wird.

Satz 3.6.3 *Für den Diskretisierungsfehler gilt*

$$||u - u_h||_V \leq \frac{c}{\alpha} ||u - \varphi||_V \quad \forall \varphi \in V_h$$

Beweis. Übung.

3.7 Variationsungleichungen

Kurze Wiederholung:
elliptische Variationsungleichung 1. Art in der variationellen Formulierung

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

Dabei ist der Raum K abgeschlossen und konvex, und V ein Hilbertraum.

Lemma 3.7.1 *Es sei $K \subset V$ abgeschlossen und konvex. Dann existiert genau ein $y \in K$ für alle $x \in V$, so dass*

$$||x - y|| = \inf_{\varphi \in K} ||x - \varphi||$$

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

Beweis.
i) Es gibt ein y. Sei φ_k eine Minimalfolge, d.h.

$$\lim_{k \to \infty} ||\varphi_k - x|| = d = \inf_{\varphi \in K} ||\varphi - x||$$

Ausmultiplizieren liefert

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

Bemerke

$$a^2 \leq ||x - \frac{1}{2}(\varphi_k + \varphi_l||^2$$

Zusammengefasst

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

für $k, l \to \infty$ folgt:

$$\lim_{k,l\to\infty} ||\varphi_k - \varphi_l|| = 0$$

Da V vollständig und abgeschlossen $\exists y \in K$ mit

$$\lim_{k\to\infty} \varphi_k = y$$

Wegen der Stetigkeit der Norm ist letztendlich

$$||x - y|| = \lim_{k\to\infty} ||x - \varphi_k|| = d$$

ii) Eindeutigkeit von y

Seien $y_1, y_2 \in K$ mit

$$||x - y_1|| = ||x - y_2|| = \inf_{\varphi\in K} ||x - \varphi||$$

Durch ähnliche Rechnung wie in i) ergibt sich

$$\begin{aligned} ||y_1 - y_2||^2 &\leq 2\,||x - y_1||^2 + 2\,||x - y_2||^2 - 4\,||x - \frac{1}{2}(y_1 + y_2)||^2 \\ &\leq 2\,d^2 + 2\,d^2 - 4\,d^2 \\ &= 0 \end{aligned}$$

Daraus folgt

$$||y_1 - y_2||^2 \leq 0$$

□

Satz 3.7.2 *Sei K abgeschlossen und konvex. Dann ist $y = P_K(x)$ genau dann eine Projektion, wenn*

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

Man beachte die Linearität: $(y, \varphi - y) \geq (x, \varphi - y)$.

Beweis. $\Longrightarrow$ (Hinrichtung)
Seien $x \in V$ und $y = P_K(x) \in K$. Da K konvex, so kann eine Konvexkombination gebildet werden

$$(1 - t)\,y + t \cdot \varphi = y + t\,(\varphi - y) \quad \varphi \in K$$

Betrachte

$$\varphi(t) = ||x - y - t\,(\varphi - y)||^2, \quad t \in [0, 1]$$

Die Funktion $\varphi(t)$ nimmt bei $t = 0$ das Minimum an. Das heißt

$$\varphi'(0) \geq 0 \Leftrightarrow -2(x - y, \varphi - y) \geq 0$$

Denn

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

Ableiten nach t:

$$-2(x-y,\varphi-y)+2t(\varphi-y,\varphi-y)$$

Bei $t=0$ folgt die Behauptung. Letztlich ist dann

$$-2(x-y,\varphi-y)\geq 0 \Leftrightarrow (x-y,\varphi-y)\leq 0 \Leftrightarrow (y-x,\varphi-y)\geq 0$$

$\Longleftarrow$ (Rückrichtung)

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

Korollar 3.7.3 *Sei $K\subset V$ abgeschlossen und konvex. Dann ist P_K nicht-expansiv, d.h.*

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

Beweis.
Seien $x,x'\in V$ gegeben und $y=P_K(x), y'=P_K(x')$. Dann ist

$$\begin{aligned} y\in K: \quad & (y,\varphi-y)\geq (x,\varphi-y) \\ y'\in K: \quad & (y',\varphi-y')\geq (x',\varphi-y') \end{aligned}$$

Testen mit $\varphi=y'$ in 1. Ungleichung. Dementsprechend setze $\varphi=y$ in 2. Ungleichung. Anschließend addieren

$$||y-y'||^2=(y-y',y-y')\leq (x-x',y-y') \overset{\text{C.S.}}{\leq} ||x-x'||\,||y-y'||$$

$$\Leftrightarrow ||y-y'||\leq ||x-x'|| \Leftrightarrow ||P_K(x)-P_K(x')||\leq ||x-x'||$$

□

Satz 3.7.4 *(Existenzsatz)*
Das Problem

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

hat eine eindeutige Lösung.

Beweis.
i) Eindeutigkeit

Annahme u_1,u_2 seien Lösungen

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

Teste mit $\varphi = u_2$ bzw. $\varphi = u_1$. Addition liefert

$$a(u_2 - u_1, u_2 - u_1) \leq 0$$

und

$$\alpha \, ||u_1 - u_2||^2 \overset{\text{V iii}}{\leq} a(u_2 - u_1, u_2 - u_1)$$

Es folgt die Gleichheit: $u_1 = u_2$.

ii) Existenz

Benutze Riez'schen Darstellungssatz

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

Es folgt

$$\begin{aligned} & (Au, \varphi - u) \geq (l, \varphi - u) \\ \Leftrightarrow \quad & (-(Au - l), \varphi - u) \leq 0 \quad | \cdot \varrho \quad | + u - u \\ \Leftrightarrow \quad & ((u - \varrho(Au - l)) - u, \varphi - u) \leq 0 \quad \forall \varphi \in K \end{aligned}$$

Dies ist äquivalent zu

$$u = P_K(u - \varrho(Au - l)), \quad \varrho > 0$$

Betrachte die Abbildung $W_\varrho : V \to V$ mit der Vorschrift

$$W_\varrho(v) = P_K(v - \varrho(Av - l))$$

Zu zeigen ist, dass W_ϱ eine Kontraktionsabbildung ist.

Rechnung.
Seien $v_1, v_2 \in V$. Dann

$$\begin{aligned} ||W_\varrho(v_1) - W_\varrho(v_2)||^2 &= ||P_K(v_1 - \varrho(Av_1 - l)) - P_K(v_2 - \varrho(Av_2 - l))||^2 \\ &\overset{\text{nicht-expansiv}}{\leq} ||(v_1 - \varrho(Av_1 - l)) - (v_2 - \varrho(Av_2 - l))||^2 \end{aligned}$$

Durch Copy-Paste ergibt sich

$$||W_\varrho(v_1) - W_\varrho(v_2)||^2 \leq (1 - 2\varrho\alpha + \varrho^2 \, ||A||^2) \cdot ||v_1 - v_2||^2$$

Scharfes Hinsehen auf der rechten Seite zeigt, dass W_ϱ eine Kontraktion ist, falls gilt

$$0 < \varrho < \frac{2\alpha}{||A||^2}.$$

Somit garantiert die Kontraktion die Existenz eines Fixpunktes. Dieser ist die gesuchte Lösung.

□

3.8 Lineare Funktionale

Es seien X, Y normierte $\mathbb{R}$-Vektorräume. Wir untersuchen im Folgenden lineare Operatoren der Form

$$T : X \to Y, \quad T \text{ ist linear und stetig}$$

Weiter soll die schon vorher häufig verwendete Supremumsnorm hier mathematisch sauber aufgeführt werden

Definition 3.8.1 *(Supremumsnorm einer Funktion)*
Es sei X eine Menge und $L^\infty(X)$ der Vektorraum aller beschränkten Funktionen von X nach $\mathbb{R}$. Für eine Funktion $v \in L^\infty(X)$ setze man

$$||v||_{L^\infty(X)} = \sup_{x \in X} |v(x)|$$

Lemma 3.8.2 *Ist $T : X \to Y$ linear, so sind äquivalent:*

1. *T ist stetig*
2. *T ist stetig in x_0, $x_0 \in X$*
3. *$\sup_{||x|| \leq 1} ||Tx|| < \infty$*
4. *$\exists C > 0 \quad mit \ ||Tx|| \leq C\,||x|| \quad \forall x \in X$*

Der Raum der stetigen Abbildungen von X nach Y wird mit

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

bezeichnet.

Definition 3.8.3 *(Operatornorm von T)*

$$||T|| := \sup_{||x|| \leq 1} ||Tx||$$

$||T||$ ist die kleinste Zahl mit der Eigenschaft

$$||Tx|| \leq ||T||\,||x||$$

Definition 3.8.4 *(Dualraum, Nullraum)*

i) $X' := L(X, \mathbb{R})$ ist der Dualraum von X. Die Elemente x' heißen lineare Funktionale.

ii) Für $T \in L(X, Y)$ ist $N(T) := \{x \in X | Tx = 0\}$ der Nullraum von T.

Bemerkung 3.8.5 *$N(T)$ ist ein abgeschlossener Unterraum von X.*

Beweis.
i) Abgeschlossenheit

Betrachte $x_k \to x$ für $k \to \infty$. Sei $x_k \in N(T), x \in X$. Es gilt

$$\lim_{k\to\infty} Tx_k = 0 \overset{T \text{ stetig}}{=} T(x) \Rightarrow x \in N(T)$$

ii) Zeige Unterraum. Seien $x_1, x_2 \in N(T)$. Dann ist

$$T(x_1 + x_2) = T(x_1) + T(x_2) = 0 \Rightarrow x_1 + x_2 \in N(T)$$

□

Satz 3.8.6 ***(Riez'scher Darstellungssatz)***
Es sei X ein Hilbertraum. Dann definiert die lineare Abbildung $J : X \to X'$ mit der Vorschrift

$$(Jx)(y) = (y, x)$$

einen linearen isometrischen Isomorphismus.

Beweis.
i) J ist linear.

$$J(x_1) = (y, x_1),\ J(x_2) = (y, x_2)$$
$$J(x_1) + J(x_2) = (y, x_1) + (y, x_2) = (y, x_1 + x_2) = J(x_1 + x_2)$$

ii) $J(x) \in X'$. Es ist

$$|(Jx)(y)| = |(y, x)| \overset{\text{C.S.}}{\leq} ||x||\,||y|| \to \sup_{||y||\leq 1} |(Jx)(y)| \leq ||x||$$

iii) J ist injektiv. Siehe

$$\left|(Jx)\left(\frac{x}{||x||}\right)\right| = \left(\frac{x}{||x||}, x\right) = \frac{(x, x)}{||x||} = \frac{||x||^2}{||x||} = ||x||$$

Also

$$\sup_{||y||\leq 1} |(Jx)(y)| \geq ||x|| \quad \to ||Jx|| \geq ||x||$$

Das heißt $x \neq 0$, somit ist Jx nicht das Nullfunktional. Weiterhin wurde hiermit gezeigt, dass J isometrisch ist: $||Jx|| = ||x||$.

iv) J ist surjektiv

Vorgehensweise: Konstruiere zu gegebenen $x_0' \in X', x_0' \neq 0$ ein $w \in X$ mit $x_0'(x) = (x, w)\,\forall x \in X$. P bezeichne die Projektion auf den abgeschlossenen Unterraum $N(x_0')$. Wähle $e \in X$ mit $x_0'(e) = 1$ und setze $x_0 = e - Pe$. Es gilt dann

$$x_0'(x_0) = x_0'(e) - x_0'(Pe) = 1 - 0 = 1$$

Insbesondere ist $x_0 \neq 0$. Erinnere an die Definition der Projektion:

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

Hier

$$\begin{aligned}&(\tilde{y} - Pe, e - Pe) \leq 0 \quad \forall \tilde{y} \in N(x_0')\\ =&(\tilde{y} - Pe, x_0)\end{aligned}$$

Sei $y \in N(x_0'), Pe \in N(x_0')$. Konstruiere

$$\begin{aligned}\tilde{y} &= y + Pe\\ \tilde{y} &= -y + Pe\end{aligned}$$

Einsetzen liefert

$$\begin{aligned}&(y, x_0) \leq 0 \wedge (-y, x_0) \leq 0\\ \Rightarrow &(y, x_0) = 0 \quad \forall y \in N(x_0')\end{aligned}$$

Damit gilt für alle $x \in X$

$$x = x - x_0'(x)\, x_0 + x_0'(x)\, x_0$$

Weiterhin

$$x_0' \underbrace{(x - x_0'(x) \cdot x_0)}_{\in N(x_0')} = x_0'(x) - x_0'(x) \cdot \underbrace{x_0'(x_0)}_{=1} = 0$$

Und

$$(x, x_0) = (x - x_0'(x) \cdot x_0 + x_0'(x) \cdot x_0, x_0) = (x_0'(x) \cdot x_0, x_0) = x_0'(x) \cdot ||x_0||^2$$

D.h.

$$x_0'(x) = \left(x, \frac{x_0}{||x_0||^2}\right) = J\left(\frac{x_0}{||x_0||^2}\right)(x)$$

□

3.9 Fehlerapproximationen

Als Motivation betrachten wir die Fehlerapproximation

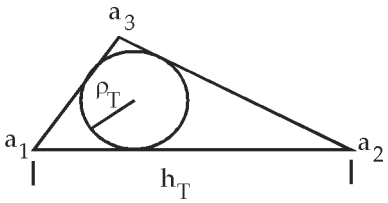
$$\begin{aligned}||u - u_h||_V &\leq \frac{c}{\alpha}||u - \varphi||_V \quad \forall \varphi \in V_h\\ &\leq \frac{c}{\alpha}||u - I_h u||_V\end{aligned}$$

Diese wurde in Kapitel 2 für den 1D-Fall vollständig diskutiert. Nun wird eine Verallgemeinerung auf den 2D-Fall angestrebt. Zunächst werden die Interpolationsabschätzungen angegeben. Anschließend werden im Kapitel *Adaptivität* die eigentlichen Fehlerabschätzungen gezeigt.

3.9.1 Interpolationsfehler in 2D für lineare Funktionen

Die folgenden Bezeichungen werden verwendet werden:

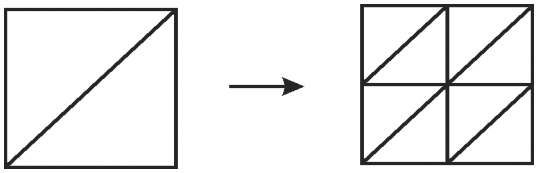
$$\begin{aligned} h_T &= \operatorname{diam}(T) \\ \varrho_T &= \text{ Inkreisradius} \\ h &= \max_{T\in\mathbb{T}} h_T \end{aligned}$$



Wie betrachten Familien von $\mathbb{T}_h$ für die unabhängig von h gilt

$$\frac{\varrho_T}{h_T} \geq \beta > 0 \quad \forall T \in \mathbb{T}_h$$

Die positive Zahl β ist demnach ein Maß für den kleinsten Winkel in T. Die Dreiecke dürfen also nicht zu dünn werden. Bei der nachfolgend skizzierten Triangulierung gibt es mit der Bedingung kein Problem:



Seien nun a_i, $i = 1, \ldots, N$ die Knoten von $\mathbb{T}_h$. Für $u \in C^0(\overline{\Omega})$ definieren wir

$$I_h u(a_i) = u(a_i), \quad i = 1, \ldots, N$$

und $I_h u$ sei zellweise linear.

Satz 3.9.1 *Sei $T \in \mathbb{T}_h$ mit den Knoten $a_i, i = 1,2,3$. Sei $v \in C^0(T)$. Die Interpolierende $I_h v \in P_1(T)$ sei definiert durch*

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

Dann folgt

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

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

Man erwähne die Analogie der beiden Abschätzungen mit dem 1D-Fall.

Bemerkung 3.9.2 *Die Bedingung $\frac{h_T^2}{\varrho_T}$ des obigen Satzes kann noch umgeschrieben werden zu*

$$\frac{h_T^2}{\varrho_T} = h_T \cdot \frac{h_T}{\varrho_T} \sim \frac{1}{\beta} \cdot h_T$$

Außerdem ist zu bemerken, dass

$$\begin{aligned} &zu\ i) \quad ||v - I_h v|| = \mathcal{O}(h_T^2) \\ &zu\ ii) \quad \max_{|\alpha|=1} ||D^\alpha (v - I_h v)||_{L^\infty(T)} = \mathcal{O}(h_T) \end{aligned}$$

Der nächtste Satz bildet den Übergang von der Supremumsnorm zum quadratischen Mittel.

Satz 3.9.3 *Es gilt*

$$\begin{aligned} ||v - I_h v||_{L^2(T)} &\leq c \cdot h_T^2 |v|_{H^2(T)} \\ |v - I_h v|_{H^1(T)} &\leq c \cdot \frac{h_T}{\varrho_T} \cdot h_T \, |v|_{H^2(T)} \end{aligned}$$

Beweis. Nicht geführt.

Der nächste Satz ersetzt T durch Ω. Wir leiten somit den globalen Fehlerterm her. Dazu muß über alle $T \in \mathbb{T}_h$ summiert werden. Der Ansatz dazu lautet

$$||\cdot||^2_{L^2(\Omega)} = \sum_{T \in \mathbb{T}_h} ||\cdot||^2_{L^2(T)}$$

Nun folgt

Satz 3.9.4 *Unter obigen Voraussetzungen zeigt man*

$$\begin{aligned} ||v - I_h v||_{L^2(\Omega)} &\leq c \cdot h^2 |v|_{H^2(\Omega)} \\ |v - I_h v|_{H^1(\Omega)} &\leq c \cdot \frac{h}{\beta} \, |v|_{H^2(\Omega)} \end{aligned}$$

Beweis. Übung.

Für höhere Polynomansätze ist nachfolgendes Resultat von Bedeutung.

Satz 3.9.5 *(höhere Polynomansätze)*
Es sei P_r der Raum der Polynome vom Grad $\leq r$. Weiter sei $I_h v \in P_r(T)$ mit $r \geq 1$. Dann gelten die Approximationen

$$\begin{aligned} &i) \quad ||v - I_h v||_{L^2(\Omega)} \leq c \cdot h^{r+1} |v|_{H^{r+1}(\Omega)} \\ &ii) \quad |v - I_h v|_{H^1(\Omega)} \leq c \cdot h^r |v|_{H^{r+1}(\Omega)} \end{aligned}$$

Insbesondere folgt

Satz 3.9.6 *(fehlende Regularität)*
Für $1 \leq s \leq r+1$ folgert man

$$\begin{aligned} &i) \quad ||v - I_h v||_{L^2(\Omega)} \leq c \cdot h^s |v|_{H^s(\Omega)} \\ &ii) \quad |v - I_h v|_{H^1(\Omega)} \leq c \cdot h^{s-1} |v|_{H^s(\Omega)} \end{aligned}$$

Praxis: Notwendige Bedingung ist das Wissen der Regularität (Glattheit) der Lösung v, da die höheren Ableitungen die Beschränkheit bestimmen.

3.9.2 Fehlerabschätzung für elliptische FE

Wir erhalten den Diskretisierungsfehler für elliptische Finite Elemente. Ausgehend von

$$\begin{aligned} ||u - u_h||_V &\leq \frac{c}{\alpha}||u - \varphi||_V \quad \forall \varphi \in V_h \\ &\leq \frac{c}{\alpha}||u - I_h u||_V \end{aligned}$$

gilt mit den vorherigen Resultaten für den Raum $V_h := H_0^1(\Omega)$

$$||u - u_h||_{H^1(\Omega)} \leq c \cdot h \, |u|_{H^2(\Omega)}$$

wobei V_h aus linearen Funktionen besteht. Es sei aber angemerkt, dass die Abschätzung mit Hilfe von Satz (3.9.5) auf beliebige Räume V_h ausgeweitet werden kann.

4 Adaptivität

Der Begriff der Adaptivität steht für die zusätzliche Verfeinerung bei gewissen Stellen im Gebiet Ω, an denen Singularitäten auftreten. Schwerpunktmäßig werden in diesem Kapitel die Fehlerabschätzungen vervollständigt.

4.1 Laplace-Problem

Es sei wiederum die Poisson-Gleichung gegeben

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

Zu diesem Problem wurde eine Lösung mit linearen Ansätzen $u \in V = H_0^1(\Omega)$ gesucht, also

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

Die beiden Variationsgleichungen genügen dem folgenden Energiefehlerschätzer.

Satz 4.1.1 *(Energiefehlerschätzer)*
Für den Diskretisierungsfehler $e = u - u_h$ für die zwei obigen Gleichungen gilt die a-posteriori Abschätzung

$$||\nabla e||^2 \leq c \cdot \sum_{T \in \mathbb{T}} \left(h_T^2 \varrho_{1,T}^2 + h_T \varrho_{2,T}^2 \right)$$

mit

$$\begin{aligned} \varrho_{1,T} &= ||f + \Delta u_h||_T \\ \varrho_{2,T} &= \int_{\partial T} \frac{1}{2} \left[\partial_n u_h \right]_{\partial T} \, d\Gamma \end{aligned}$$

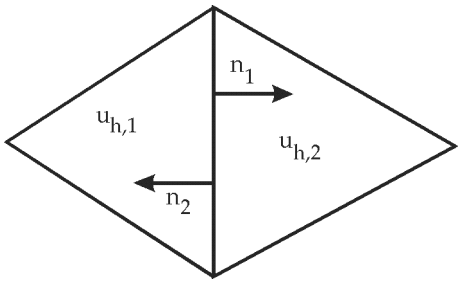
Beweis.
Ansatz:

$$\begin{aligned} ||\nabla e||^2 &= (\nabla u - \nabla u_h, \nabla e) \\ &= (\nabla u - \nabla u_h, \nabla e - \nabla(I_h e)) \quad \text{Galerkin-Eig.} \\ &= (f, e - I_h e) - \sum_T (\nabla u_h, \nabla(e - I_h e))_T \\ &= (f, e - I_h e) - \sum_T \left((-\Delta u_h, e - I_h e)_T + \int_{\partial T} (\partial_n u_h) \cdot (e - I_h e) \, d\Gamma \right) \\ &= \sum_T (f + \Delta u_h, e - I_h e)_T - \sum_T \sum_{j=1}^{3} \int_{\partial T_j} \frac{1}{2} \left[\partial_n u_h \right] (e - I_h e) \, d\Gamma \end{aligned}$$

Im letzten Gleichungsschritt wurden die Beziehungen

$$\partial_{n_1} u_{h,1} = \nabla u_{h,1} \cdot n_1$$
$$\partial_{n_2} u_{h,2} = \nabla u_{h,2} n_2 = -\nabla u_{h,2} n_1$$

ausgenutzt.



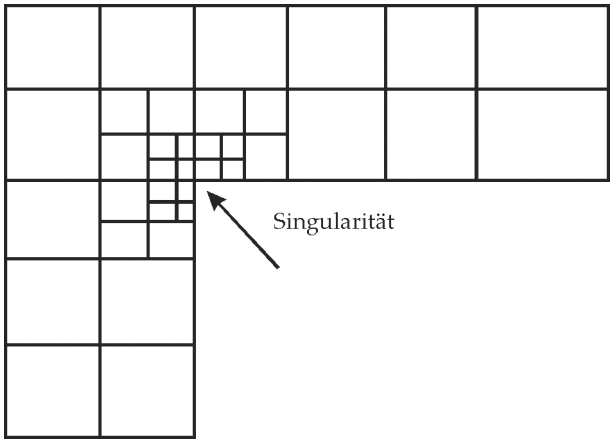
Kontinuierliches Modell

$$a(u, \varphi) + s(u, \varphi) = (f, \varphi) \quad \forall \varphi \in V$$

mit dem Störungsterm

$$s(u, \varphi) = \sum_{T \in \mathbb{T}_h} s_T(u, \varphi)$$
$$\text{bzw.} \quad s_\alpha(u, \varphi) = \sum_{T \in \mathbb{T}_h} \alpha_T s_T(u, \varphi), \quad \alpha_T \in \{0, 1\}$$

Der Parameter α_T steuert, welche Zellen zusätzlich verfeinert werden.



Zur weiteren Untersuchung werden zwei Räume eingeführt

$$\mathcal{H} = \{T \in \mathbb{T}_h | \alpha_T = 1\}$$
$$\mathcal{N} = \{T \in \mathbb{T}_h | \alpha_T = 0\}$$

Damit lassen sich zwei zusätzliche Störungsterme konstruieren

$$s_h(u, \varphi) = \sum_{T \in \mathcal{H}} s_T(u, \varphi)$$
$$\text{und} \quad s_\mathcal{N}(u, \varphi) = \sum_{T \in \mathcal{N}} s_T(u_h, \varphi)$$

Diskretes Modell

Folgt analog zu den obigen Überlegungen. Gesucht ist die diskrete Lösung $u_h \in V_h$:

$$\begin{aligned} & a(u_h, \varphi) + s_h(u_h, \varphi) = (f, \varphi) \quad \forall \varphi \in V_h \\ \Leftrightarrow \quad & a(u_h, \varphi) + s(u_h, \varphi) = (f, \varphi) + s_{\mathcal{N}}(u, \varphi) \end{aligned}$$

Kombination beider Modelle

Wir subtrahieren das diskrete von dem kontinuierlichen Modell. Dann ergibt sich

$$a(u - u_h, \varphi) + s(u - u_h, \varphi) + \underbrace{s_{\mathcal{N}}(u_h, \varphi)}_{\text{Störungsterm}} = 0$$

Mit $e = u - u_h$ folgt

$$\begin{aligned} & a(e,e) + s(e,e) \\ &= a(u - u_h, e - e_i) + s(u - u_h, e - e_i) - s_{\mathcal{N}}(u_h, e_i) \\ &= (f, e - e_i) - a(u_h, e - e_i) - s_{\mathcal{H}}(u_h, e - e_i) - s_{\mathcal{N}}(u_h, e - e_i) - s_{\mathcal{N}}(u_h, e_i) \\ &= \underbrace{(f, e - e_i)}_{\text{kontinuierlich}} - \underbrace{a(u_h, e - e_i) - s_{\mathcal{H}}(u_h, e - e_i)}_{\text{diskret}} - s_N(u_h, e) \end{aligned}$$

Die ersten drei Summanden entsprechen formal dem diskreten Modell.

Beispiel.

$$s_{\mathcal{N}}(u_h, e) = \sum_{T \in \mathcal{N}} \left((\mu(x) - \overline{\mu}) \cdot \nabla u_h, \nabla e \right)_T$$

Mit

$$\underbrace{\overline{\mu}(\nabla u, \nabla \varphi)}_{a(u,\varphi)} + \underbrace{((\mu(x) - \overline{\mu}) \cdot \nabla u, \nabla e)}_{s(u,\varphi)} = (f, \varphi)$$

Als betragliche Abschätzung ergibt sich

$$\begin{aligned} |s_{\mathcal{N}}(u, \varphi)| &\leq \sum_{T \in \mathcal{N}} ||(\mu(x) - \overline{\mu}) \cdot \nabla u_h||_T \cdot ||\nabla e||_T \\ &\leq \left(\sum_{T \in \mathcal{N}} ||(\mu(x) - \overline{\mu}) \cdot \nabla u_h||_T \right)^{\frac{1}{2}} \cdot ||\nabla e||_\Omega \end{aligned}$$

4.2 A posteriori Energiefehlerschätzer für VU

Es seien die beiden Variationsungleichungen nochmal genannt.

$$\begin{aligned} (\nabla u, \nabla(\varphi - u)) &\geq (f, \varphi - u) \quad \forall \varphi \in K \subset V, \, K \text{ konvex} \\ (\nabla u_h, \nabla(\varphi - u_h)) &\geq (f, \varphi - u_h), \quad K_h = K \cap V_h \end{aligned}$$

Hierzu soll der a posteriori Energiefehlerschätzer hergeleitet werden. Für die eigentliche Abschätzung ist das nächste Lemma von Interesse

Lemma 4.2.1 *Es gilt*

$$(\nabla e, \nabla e_i) = (f, e_i) - (\nabla u_h, \nabla e_i) + (\nabla u, \nabla(e_i - e)) - (f, e_i - e) + (\nabla u, \nabla e) - (f, e)$$

mit

$$(f, e_i) - (\nabla u_h, \nabla e_i) \leq 0 \tag{4.1}$$
$$(\nabla u, \nabla e) - (f, e) \leq 0 \tag{4.2}$$

Beweis.
zu (4.1): mit $e = u - u_h$ und $e_i = u_i - u_h$ folgt

$$(f, u_i - u_h) - (\nabla u_h, \nabla(u_i - u_h)) \leq 0 \tag{4.3}$$
$$\Leftrightarrow \quad (f, u_i - u_h) - (\nabla u_h, \nabla u_i) + (\nabla u_h, \nabla u_h) \leq 0$$

Weiter gilt nach dem Beispiel (Voraussetzung)

$$(\nabla u, \nabla(u_h - u)) \geq (f, u_h - u)$$

Einsetzen in (4.3) zeigt

$$(\nabla u_h, \nabla(u_i - u_h)) \geq (f, u_i - u_h)$$

zu (4.2): Z.z.

$$(\nabla u, \nabla e) - (f, e) \leq 0$$
$$\Leftrightarrow \quad (\nabla u, \nabla(u - u_h)) - (f, u - u_h) \leq 0 \tag{4.4}$$

Mit eingangs angeführtem Beispiel folgt

$$(\nabla u, \nabla(\varphi - u)) \geq (f, \varphi - u)$$

Testen mit u_h liefert

$$(\nabla u, \nabla(u_h - u)) \geq (f, u_h - u)$$

Gehe mit diesem Ergebnis in die Ungleichung (4.4):

$$(\nabla u, \nabla(u - u_h)) - (\nabla u, \nabla(u_h - u)) \leq 0$$

□

Nun folgt die eigentliche Abschätzung

Satz 4.2.2 *Für Variationsungleichungen gilt die a-posteriori Fehlerabschätzung*

$$\begin{aligned} ||\nabla e||^2 &= (\nabla e, \nabla e) = (\nabla e, \nabla(e - e_i)) + (\nabla e, \nabla e_i) \\ &\overset{(4.2.1)}{\leq} (\nabla u, \nabla(e - e_i)) - (\nabla u_h, \nabla(e - e_i)) + (\nabla u, \nabla(e_i - e)) - (f, e_i - e) \\ &= (f, e - e_i) - (\nabla u_h, \nabla(e - e_i)) \end{aligned}$$

Der weitere Verlauf ist analog zu den Variationsgleichungen des vorherigen Abschnitts.

ABER: suboptimal in den Kontaktbereichen. Denn aus

$$||\Delta u_h + f|| \geq 0 \quad \text{im Kontaktbereich}$$

folgt

$$\Delta u + f \geq 0$$

4.3 Dualitätsargument

Es sei wie im vorherigen Abschnitt die Ausgangssituation

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

gegeben. Dazu wurde in Abschnitt (3.9.2) der Energiefehlerschätzer hergeleitet

$$||u - u_h||_{H^1(\Omega)} \leq c \cdot h \, |u|_{H^2(\Omega)} \tag{4.5}$$

In dieser Sektion soll eine vergleichbare Abschätzung in der L_2-Norm hergeleitet werden

A priori Abschätzung

Satz 4.3.1 *Sei Ω ein konvexes, polygonales Gebiet. Für die Lösung u gelten die bekannten Gleichungen*

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

Dann existiert ein positives c, das unabhängig von u und h ist, so dass

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

In dem Beweis wird Stabilität des dualen Problems ausgenutzt. Dazu das Lemma

Lemma 4.3.2 *(Stabilität des dualen Problems)*
Für konvexes Ω mit dem dualen Problem $-\Delta z = e$ auf Ω und $z = 0$ auf $\partial\Omega$ gilt die Ungleichung

$$||z||_{H^2(\Omega)} \leq c_s \cdot ||e||_{L^2(\Omega)}$$

Der Faktor c_s ist unabhängig von e.

Beweis von (4.3.1).
Wir betrachten das folgende Hilfsproblem, auch duales Problem genannt.

$$\begin{aligned} -\Delta z = e := u - u_h \quad &\text{auf } \Omega \\ z = 0 \quad &\text{auf } \partial\Omega \end{aligned}$$

Hier wird also zunächst die duale Lösung z gesucht. Dazu wird mit variationellen Formulierung gearbeitet. Hier

$$\begin{aligned} & -(\varphi, \Delta z) = (\varphi, e) \quad \forall \varphi \in V = H_0^1(\Omega) \\ \Leftrightarrow \quad & (\nabla \varphi, \nabla z) = (\varphi, e) \quad \forall \varphi \in V \end{aligned}$$

Für die spezielle Wahl $\varphi = e$ rechnet man nach

$$\begin{aligned}(e,e) &= (\nabla e, \nabla z - \nabla I_h z) \quad \text{wg. Galerkin-Eig.}\\ &\overset{\text{C.S}}{\leq} ||\nabla e||_{L^2(\Omega)} \cdot ||\nabla z - \nabla I_h z||_{L^2(\Omega)}\\ &\overset{\text{IP}}{\leq} ||\nabla e||_{L^2(\Omega)}\, c \cdot h\, ||z||_{H^2(\Omega)}\\ &\overset{(4.3.2)}{\leq} ||\nabla e||_{L^2(\Omega)}\, c \cdot h \cdot c_s\, ||e||_{L^2(\Omega)}\\ &= \mathcal{O}(h^2) \cdot c_s \cdot ||e||_{L^2(\Omega)}\end{aligned}$$

Weiter folgt

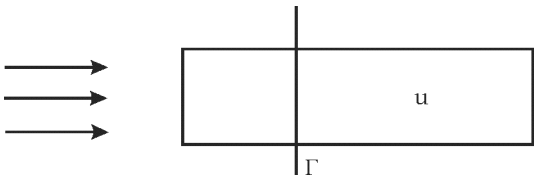
$$\begin{aligned}||e||_{L^2(\Omega)} &\leq c \cdot h \cdot c_s \cdot ||\nabla e||_{L^2(\Omega)}\\ &\overset{(4.5)}{\leq} c \cdot c_s \cdot h \cdot h \cdot |u|_{H^2(\Omega)}\\ &= \mathcal{O}(h^2)\end{aligned}$$

□

A posteriori Abschätung

Die a posteriori Fehlerabschätzung für den Diskretisierungsfehler wird mit linearen Funktionalen angegangen.

Beispiel.
Mittlere Konzentration beim Durchfluss



Dazu die mathematische Formulierung

$$\int_\Gamma u\, d\Gamma, \quad \delta_{x_0}(u) \quad \text{Dirac-Distr.}$$

mit dem Fehler

$$J(\varphi) = \int_\Gamma \varphi\, d\Gamma$$

Wir arbeiten nun wie oben mit dem dualen Problem. Die Betrachtungen werden hier allerdings auf unterschiedlichen Gittern gemacht. Dazu wird eine weitere Gitterweite $\tilde{h}$ eingeführt, wobei der zugehörige Raum $V_{\tilde{h}} \subset V$ wie gewohnt eine endlichdimensionale Basis aus linearen Funktionen besitzt. Also

$$\begin{aligned}z \in V: &\quad J(\varphi) = (\nabla\varphi, \nabla z) \quad \forall \varphi \in V\\ z_{\tilde{h}} \in V_{\tilde{h}}: &\quad J(\varphi) = (\nabla\varphi, \nabla z_{\tilde{h}}) \quad \forall \varphi \in V_{\tilde{h}}\end{aligned}$$

Abschätzung

$$\begin{aligned}J(e) &= (\nabla e, \nabla z - \nabla z_{\tilde{h}}) + (\nabla e, \nabla z_{\tilde{h}}) \qquad (4.6)\\ &\leq \underbrace{||\nabla u - \nabla u_h||}_{=\mathcal{O}(h)} \cdot \underbrace{||\nabla z - \nabla z_{\tilde{h}}||}_{=\mathcal{O}(h)} + (\nabla e, \nabla z_{\tilde{h}})\end{aligned}$$

Die beiden Faktoren der ersten Summe sind aus der Energiefehlerschätzung bekannt. Der zweite Summand $(\nabla e, \nabla z_{\tilde{h}})$ wird bei gleichem Gitter von h und $\tilde{h}$ gleich Null. Umschreiben des zweiten Summanden zeigt

$$\begin{aligned}
(\nabla e, \nabla z_{\tilde{h}}) &= (\nabla e, \nabla z_{\tilde{h}} - \nabla I_h z_{\tilde{h}}) \\
&= (\nabla u - \nabla u_h, \nabla z_{\tilde{h}} - \nabla I_h z_{\tilde{h}}) \\
&= (f, \tilde{z} - I_h z_{\tilde{h}}) - \sum_{T \in \mathbb{T}_h} (\nabla u_h, \nabla (z_{\tilde{h}} - I_h z_{\tilde{h}}))_T \\
&= (f, z_{\tilde{h}} - I_h z_{\tilde{h}}) - \sum_{T \in \mathbb{T}_h} [(-\Delta u_h, z_{\tilde{h}} - I_h z_{\tilde{h}})_T \\
&\quad + \int_{\partial T} \frac{1}{2} (\partial_n u_h) \cdot (z_{\tilde{h}} - I_h z_{\tilde{h}}) \Big] \\
&= \sum_{T \in \mathbb{T}_h} \left[(f + \Delta u_h, z_{\tilde{h}} - I_h z_{\tilde{h}})_T + \int_{\partial T} \frac{1}{2} (\partial_n u_h) \cdot (z_{\tilde{h}} - I_h z_{\tilde{h}}) d\Gamma \right]
\end{aligned}$$

In dem letzten Term sind alle Einträge bekannt. Die Methode des dualen Problems wird auch als DWR-Zugang (engl. Dual-Weighted-Residual) bezeichnet.

Postprocessing

Gleichung (4.6) kann weiter umgeschrieben werden zu

$$\begin{aligned}
J(u) - J(u_h) - (\nabla e, \nabla z_{\tilde{h}}) &= (\nabla e, \nabla z - \nabla z_{\tilde{h}}) \\
\Leftrightarrow \quad J(u) - (J(u_h) + (f, z_{\tilde{h}}) - (\nabla u_h, \nabla z_{\tilde{h}})) &= (\nabla e, \nabla z - \nabla z_{\tilde{h}})
\end{aligned}$$

mit $J(u) - J(u_h) = J(e)$.

Primale Probleme:

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

Duale Probleme:

$$\begin{aligned}
a(\varphi, z) &= J(\varphi) \quad \forall \varphi \in V \\
a(\varphi, z_{\tilde{h}}) &= J(\varphi) \quad \forall \varphi \in V_{\tilde{h}}
\end{aligned}$$

Pure:

$$\begin{aligned}
J(u - u_h) = J(u) - J(u_h) &= a(u - u_h, z - z_{\tilde{h}}) + a(u - u_h, z_{\tilde{h}}) \\
&= a(u - u_h, z - z_{\tilde{h}}) + (f, z_{\tilde{h}}) - a(u_h, z_{\tilde{h}})
\end{aligned}$$

Postprocessing

$$\tilde{J}(u_h) = J(u_h) + (f, z_{\tilde{h}}) - a(u_h, z_{\tilde{h}})$$

Dann folgt für die betragliche Abschätzung

$$\begin{aligned}
|J(u) - \tilde{J}(u_h)| &= |a(u - u_h, z - z_{\tilde{h}})| \\
&\leq C \cdot ||u - u_h||_V \cdot ||z - z_{\tilde{h}}||_V \\
&\leq \eta_p \cdot \eta_d
\end{aligned}$$

5 Iterative Methoden, Minimierungsalgorithmen

Bei Anwendung der Finite Element Methode auf partielle Differentialgleichungen erhält man sehr große Gleichungssysteme der Form $Ax = b$. Um diese zu lösen gibt es verschiedene Ansätze. Zunächst einmal die direkten Methoden von *Gauß-Seidel* und *Cholesky*. Diese benötigen aber bisweilen sehr lange Rechenzeiten.

Vorteilhafter bei symmetrischen und dünnbesetzten Matrizen sind iterative Verfahren. Einige davon werden in diesem Kapitel vorgestellt. Motivation sei

$$\begin{aligned} Ax = b, \quad & A \in \mathbb{R}^{n\times n}, \text{symmetrisch, positiv definit} \\ & x, b \in \mathbb{R}^n \end{aligned}$$

Wir betrachten das Minimierungsproblem

$$f(x) = \frac{1}{2} x^T A x - b^T x$$

Die Minimalstelle $\tilde{x}$ von $f(\tilde{x})$ erfüllt

$$A\tilde{x} - b = 0$$

5.1 Positiv definite Matrizen

Es folgen einige Bemerkungen, Tatsachen und Vorzüge positiv definiter Matrizen

Bemerkung 5.1.1 *Sei $||\cdot||$ eine Norm auf $\mathbb{C}^n$. Weiter sei $A \in M(n,n) := \mathbb{C}^{n\times n}$. Außerdem sei A regulär, dann definiert*

$$||x||_A = ||Ax||, \quad x \in \mathbb{C}^n$$

ebenfalls eine Norm.

Definition 5.1.2 *Sei $(\cdot,\cdot)$ das euklidische Skalarprodukt auf $\mathbb{C}^n$. Dann heißt $A \in M(n,n)$ positiv definit , wenn*

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

Bemerkung 5.1.3 *Es gilt $A \in M(n,n)$ mit $(Ax,x) > 0, \forall x \in \mathbb{C}^n, x \neq 0$ genau dann, wenn $A = A^H$ und alle Eigenwerte positiv sind. Es gibt dann die Darstellung*

$$A = T\,D\,T^H, \quad T \textit{ unitär und } D \textit{ diagonal}$$

Definition 5.1.4 *Sei $A^{\frac{1}{2}} := T\, D^{\frac{1}{2}}\, T^H$, dann heisst*

$$||x||_A := ||A^{\frac{1}{2}}x||_2, \quad mit\ ||\cdot||_2 = \sqrt{(\cdot,\cdot)}$$

die Energienorm.

Bemerkung 5.1.5 *Für das Skalarprodukt*

$$(x,y)_A := (Ax,y), \quad x,y \in \mathbb{C}^n$$

gilt

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

Bemerkung 5.1.6 *A ist genau dann positiv definit, wenn A^{-1} positiv definit ist.*

Definition 5.1.7 *Die Kondition der regulären Matrix $A \in M(n,n)$ ist gegeben durch*

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

Bemerkung 5.1.8 *Die zugrunde liegende Vektornorm sei $||\cdot||_2$. Für eine positv definite Matrix $A \in M(n,n)$ werde $cond(A)$ bestimmt in der zugeordneten Matrixnorm. Dann ist*

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

dabei ist λ_{max} der größte Eigenwert von A und dementsprechend λ_{min} der kleinste Eigenwert.

Lemma 5.1.9 *Sei A eine positiv definite Matrix mit Spektralkondition κ. Dann gilt für jeden Vektor $x \neq 0$*

$$\frac{(x^TAx)\,(x^TA^{-1}x)}{(x^Tx)^2} \leq \kappa \tag{5.1}$$

Beweis.
Anordnung der Eigenwerte $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Wie betrachten die Situation nach unitärer Transformation im Raum der Eigenvektoren. Dann schreibt sich die linke Seite von (5.1) als

$$\frac{\left(\sum_{i=1}^n \lambda_i x_i^2\right)\,\left(\sum_{i=1}^n \lambda_i^{-1} x_i^2\right)}{\left(\sum_{j=1}^n x_j^2\right)^2} \tag{5.2}$$

Substitution

$$z_i = \frac{x_i^2}{\sum_{j=1}^n x_j^2} \quad \text{mit} \sum_{i=1}^n z_i = 1$$

Einsetzen in (5.2) zeigt

$$\frac{\left(\sum_{i=1}^n \lambda_i x_i^2\right)\,\left(\sum_{i=1}^n \lambda_i^{-1} x_i^2\right)}{\left(\sum_{j=1}^n x_j^2\right)^2} = \underbrace{\left(\sum_{i=1}^n \lambda_i z_i\right)}_{\leq \lambda_n} \cdot \underbrace{\left(\sum_{i=1}^n \lambda_i^{-1} z_i\right)}_{\leq \lambda_1^{-1}} \leq \frac{\lambda_n}{\lambda_1} = \kappa$$

wegen

$$\sum_{i=1}^{n} \lambda_i z_i \leq \sum_{i=1}^{n} \lambda_n z_i = \lambda_n$$
$$\text{und} \quad \sum_{i=1}^{n} \lambda_i^{-1} z_i \leq \sum_{i=1}^{n} \frac{1}{\lambda_1} z_i = \frac{1}{\lambda_1}$$

□

Im vorherigen Lemma werden die Summen im Beweis recht grob abgeschätzt. Daher kann die Abschätzung noch etwas verbessert werden.

Lemma 5.1.10 *Mit den gleichen Voraussetzungen wie im vorherigen Lemma gilt*

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

Beweis. Nicht geführt.

5.2 Abstiegsverfahren

Beim Abstiegsverfahren wird eine Folge

$$x_0 \quad \to \quad x_1 \quad \to \quad \dots$$

durch eindimensionale Minimierung der Funktion f in Richtung des negativen Gradienten erzeugt.

Aufgabe.
Sei $f : \mathbb{R}^n \to \mathbb{R}$ stetig differenzierbar. Gesucht ist eine Lösung $\tilde{x} \in \mathbb{R}^n$, so dass $f(\tilde{x}) \leq f(x)$ für alle $x \in \mathbb{R}^n$ ist.

Lemma 5.2.1 *Unter obigen Voraussetzungen sei $d := -\nabla f(x) \neq 0$. Dann gilt*

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

für hinreichend kleines $t > 0$.

Beweis.
Man betrachte die Richtungsableitung

$$\lim_{t \to 0} \frac{f(x+td) - f(x)}{t} = \nabla f(x)^T \cdot d < 0$$
$$\Rightarrow \quad \frac{f(x+td) - f(x)}{t} < 0 \quad \text{für hinreichend kleines } t$$
$$\overset{t>0}{\Rightarrow} \quad f(x+td) < f(x)$$

□

Algorithmus 5.2.2 *Für $k = 0, 1, 2, \dots$ und gegebenes x_0 folgt*

1. *Berechne $d_k = -\nabla f(x_k)$*
2. *Liniensuche: Man suche für f das Minimum auf der Linie $\{x_k + t \cdot d_k\}$ für $t > 0$*
3. *$x_{k+1} = x_k + t \cdot d_k$*

5.3 Gradientenverfahren

Spezielle Aufgabe

$$f(x) = \frac{1}{2}x^T Ax - b^T x, \quad A \text{ positiv definit}$$

Ein Minimum liegt vor, falls $Ax - b = 0$ ist. Wir benutzen Algorithmus (5.2.2). Hier gilt speziell

i $d_k = b - Ax_k$

ii $t_k = \frac{d_k^T \cdot d_k}{d_k^T (A\, d_k)}$

Rechnung.
zu i): $\nabla f(x) = Ax - b$

zu ii): Minimumsuche für quadratisches f.

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

□

Lemma 5.3.1 *Mit $f(x) = \frac{1}{2}x^T Ax - b^T x$ gilt*

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

wobei $A\tilde{x} = b$ gilt.

Beweis.
1) linke Seite:

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

2) rechte Seite:

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

Vergleich beider Seiten liefert die Behauptung.

□

Lemma 5.3.2 *Sei $A\tilde{x} = b$. Dann ist*

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

mit der k-ten iterierten Lösung x_k und der Suchrichtung d_k.

Beweis.
Es gilt $d_k = b - Ax_k = A(\tilde{x} - x_k)$ genau dann, wenn $-A^{-1}d_k = x_k - \tilde{x}$. Man betrachte nun

$$\begin{aligned}||x_k-\tilde{x}||_A^2 &= (x_k-\tilde{x})^T\,A\,(x_k-\tilde{x}) \\ &= (d_k^T A^{-1})\,A\,(A^{-1}d_k) \\ &= d_k^T\,A^{-1}\,d_k\end{aligned}$$

Division liefert die Behauptung.

□

Satz 5.3.3 *Sei $A\tilde{x} = b$. Für den Iterationsfehler des Gradientenverfahrens zeigt man*

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

mit $\kappa = cond(A)$.

Beweis.
Laut Algorithmus gelten

$$\begin{aligned}1)\quad & d_k = b - Ax_k \\ 2)\quad & t_k = \frac{d_k^T d_k}{d_k^T A d_k}\end{aligned}$$

Einsetzen in

$$\begin{aligned}
f(x_{k+1}) &\overset{\text{Alg.}}{=} f(x_k + t_k d_k) \\
&\overset{(5.3.1)}{=} \frac{1}{2}(x_k + t_k d_k)^T A\,(x_k + t_k d_k) - b^T\,(x_k + t_k d_k) \\
&\overset{\text{umsortieren}}{=} f(x_k) + t_k d_k \underbrace{(Ax_k - b)}_{=-d_k} + \frac{1}{2} t_k^2 d_k^T\, A d_k \\
&\overset{\text{mit } t_k}{=} f(x_k) + \frac{d_k^T d_k}{d_k^T A d_k} \cdot d_k^T\,(-d_k) + \frac{1}{2}\frac{(d_k^T d_k)^2}{(d_k^T A d_k)^2} \cdot d_k^T A d_k \\
&= f(x_k) - \frac{1}{2}\frac{(d_k^T d_k)^2}{d_k^T A d_k}
\end{aligned}$$

Damit folgen die Äquivalenzen

$$\begin{aligned}
& f(x_{k+1}) = f(x_k) - \frac{1}{2}\frac{(d_k^T d_k)^2}{d_k^T A d_k} \\
\Leftrightarrow \quad & f(x_{k+1}) - f(\tilde{x}) = f(x_k) - f(\tilde{x}) - \frac{1}{2}\frac{(d_k^T d_k)^2}{d_k^T A d_k} \\
\overset{(5.3.1)}{\Leftrightarrow} \quad & \frac{1}{2}\,||x_{k+1} - \tilde{x}||_A^2 = \frac{1}{2}\,||x_k - \tilde{x}||_A^2 - \frac{1}{2}\frac{(d_k^T d_k)^2}{d_k^T A d_k} \cdot 1 \\
\overset{(5.3.2)}{\Leftrightarrow} \quad & ||x_{k+1} - \tilde{x}||_A^2 = ||x_k - \tilde{x}||_A^2 - \frac{(d_k^T d_k)^2}{d_k^T A d_k} \cdot \frac{||x_k - \tilde{x}||_A^2}{d_k^T A^{-1} d_k} \\
& \qquad\qquad = ||x_k - \tilde{x}||_A^2 \cdot \left(1 - \frac{(d_k^T d_k)^2}{(d_k^T A d_k)\,(d_k^T A^{-1} d_k)}\right)
\end{aligned}$$

Letztendlich ist

$$\begin{aligned}
& f(x_{k+1}) = f(x_k) - \frac{1}{2}\frac{(d_k^T d_k)^2}{d_k^T A d_k} \\
\overset{(5.1.10)}{\Leftrightarrow} \quad & ||x_{k+1} - \tilde{x}||_A^2 \leq ||x_k - \tilde{x}||_A^2 \cdot \left(1 - \frac{4}{(\sqrt{\kappa} + \sqrt{\kappa^{-1}})^2}\right) \\
& \qquad\qquad = ||x_k - \tilde{x}||_A^2 \cdot \frac{(\kappa - 1)^2}{(\kappa + 1)^2}
\end{aligned}$$

□

Außerdem wurde mit dem Beweis gezeigt, dass

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

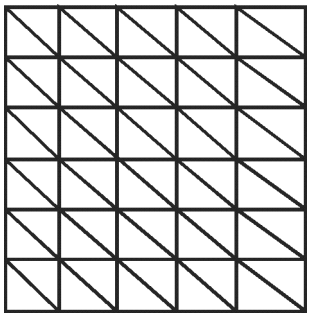
mit

$$\begin{aligned}
& \frac{(\kappa - 1)}{(\kappa + 1)} < 1 \\
\text{und} \quad & \frac{(\kappa - 1)}{(\kappa + 1)} = \frac{(\kappa + 1)}{(\kappa + 1)} - \frac{2}{\kappa + 1} \approx \left(1 - \frac{2}{\kappa}\right)
\end{aligned}$$

Wie bereits oben bemerkt, dauert die Iteration umso länger, je schlechter (größer) die Kondition der Matrix ist.

5.4 Beispiel

Gegeben sei das Poisson-Problem in 2D. Es wird mit linearen Finiten Elementen diskretisiert.



mit zugehöriger Matrix

$$\begin{pmatrix}
4 & -1 & & & & -1 & & & & \\
-1 & \ddots & \ddots & & & & \ddots & & & \\
 & \ddots & \ddots & \ddots & & & & \ddots & & \\
 & & \ddots & \ddots & -1 & & & & \ddots & \\
 & & & -1 & 4 & & & & & -1 \\
-1 & & & & & & & & & \\
 & \ddots & & & & & & & & \\
 & & \ddots & & & & & & & \\
 & & & \ddots & & & & & & \\
 & & & & -1 & & & & &
\end{pmatrix}$$

Der größte und der kleinste Eigenwert sind gegeben durch

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

Einsetzen in $1 - \frac{2}{\kappa}$:

$$\begin{aligned}
1 - \frac{2}{\kappa} &= 1 - 2\,\frac{\sin^2\left(\frac{pi}{2}h\right)}{\cos^2\left(\frac{pi}{2}h\right)} \\
&\leq 1 - 2\,\sin^2\left(\frac{pi}{2}h\right) \\
&= 1 - (1 - \cos \pi h) \\
&= \cos \pi h
\end{aligned}$$

Frage: Wie viele Gradientenschritte i sind nötig, um eine Reduktion um den Faktor r zu erreichen?
Dazu arbeitet man mit den Taylorentwicklungen

$$i) \quad \cos \pi h = 1 - \frac{\pi^2}{2} \cdot h^2 + \mathcal{O}(h^4)$$

$$ii) \quad \ln 1 - z = -z - \frac{z^2}{2} + \mathcal{O}(z^3)$$

Zwischen den beiden Funktionen gibt es die Beziehung

$$(\cos \pi h)^i = r \quad \Leftrightarrow \quad i = \frac{\ln r}{\ln \cos \pi h}$$

Den Nenner der rechten Seite schätzt man ab mit

$$\begin{aligned} \ln \cos \pi h &\overset{i)}{\approx} \ln \left(1 - \frac{\pi^2}{2} \cdot h^2\right) \\ &\approx -\frac{\pi^2}{2} \cdot h^2 + \mathcal{O}(h^2) \end{aligned}$$

Daraus folgt

$$\begin{aligned} i &= \ln r \cdot \frac{2}{-\pi^2} \cdot \frac{1}{h^2} \\ &= c \cdot \frac{1}{h^2}, \quad \text{mit } c > 0 \end{aligned}$$

Also ist

$$i \sim \frac{1}{h^2}$$

Dabei steht i für die Anzahl der Iterationen und h wie gewohnt für die Gitterweite.
Verfeinerung
Bei Halbierung der Gitterweite

$$h \to \frac{h}{2}$$

folgt für die Anzahl der Iterationen

$$i = c \cdot \frac{1}{\left(\frac{h}{2}\right)^2} = c \cdot 4 \cdot \frac{1}{h^2}$$

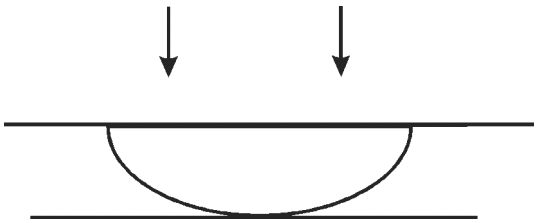
Erklärung.
Bei Halbierung der Gitterweite vervierfacht sich die Anzahl der Iterationen!
Praxis.
Darüberhinaus geht weitere Rechenzeit verloren, da bei halbierter Gitterweite die zugehörige Matrix dementsprechend größer ist.
Konsequenz.
Einfache Halbierung ist für die Praxis zunächst einmal schlecht.

Bemerkung 5.4.1 *Die Kondition einer Matrix ist ausschließlich von der Ordnung des Differentialoperators abhängig. Nicht von der Dimension des betrachteten Problems.*

Beispiel sei das Poisson-Problem $u'' = f$. Egal ob man dieses Problem im 2-dim oder 3-dim berechnet. Für die Kondition gilt stets $\kappa \sim \frac{1}{h^2}$. Also bei 1-mal verfeinern werden schon 4-mal mehr Schritte gebraucht.

5.5 Projiziertes Gradientenverfahren

Motivation sei wieder das Drahtproblem mit Hinderniss.



Rechnerisch erhalten wir aus der Abbildung die schon bekannte Variationsungleichung

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

Die gestellte Aufgabe lautet: Finde eine Lösung $u \in K \subset \mathbb{R}^n$, so dass

$$\min \frac{1}{2} u^T A u - f^T u =: J(u)$$

mit positiv definiter Matrix $A \in M(n,n)$ und $f \in \mathbb{R}^n$. Der Raum K ist definiert durch

$$\mathbb{K} = \{x \in \mathbb{R}^n | x(i) \geq 0, i = 1, \ldots, n\}$$

Algorithmus 5.5.1 *Es bezeichne $P_{\mathbb{K}}$ die Projektion auf $\mathbb{K}$.*

i) Initialisierung: Man wähle $u_0 \in \mathbb{K}$

ii) Iteration: für $k = 0, 1, 2, \ldots$

$$u_{k+1} = P_{\mathbb{K}}(u_k - \alpha_k J'(u_k)) \quad \textit{mit } J'(u) = Au - f \textit{ und } \alpha_k > 0$$

Schritt ii) zerfällt in zwei Teilschritte:

$$\begin{aligned} u_{k+\frac{1}{2}} &= u_k + \alpha_k (f - A u_k) \quad \text{(Gradientenschritt)} \\ u_{k+1} &= P_{\mathbb{K}}(u_{k+\frac{1}{2}}) \end{aligned}$$

wobei im letzten Schritt $u_{k+1}(i) = \max(0, u_{k+\frac{1}{2}}(i))$ gilt.

Satz 5.5.2 *Zu dem obigen Algorithmus existieren $\alpha, \beta > 0$, so dass mit $\alpha < \alpha_k < \beta$ dieser gegen die Lösung u konvergiert. Die Konvergenzgeschwindigkeit wird hier nicht ermittelt.*

Beweis.
i) Wir zeigen die Fixpunkteigenschaft von u, also $u = P_{\mathbb{K}}(u - \alpha_k J'(u))$. Aus Abschnitt (2.8) folgt

$$(J'(u), \varphi - u) \geq 0 \quad \forall \varphi \in \mathbb{K}$$

Mit $\alpha_k > 0$ gilt weiterhin

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

Damit ist der erste Teil gezeigt.

ii) Wir rechnen

$$\begin{aligned} ||u_{k+1} - u|| &= ||P_{\mathbb{K}}(u_k - \alpha_k J'(u_k)) - P_{\mathbb{K}}(u - \alpha_k J'(u))|| \\ &\overset{\text{nicht-expansiv}}{\leq} ||u_k - u - \alpha_k (J'(u_k) - J'(u))|| \end{aligned}$$

Quadrieren bringt

$$||u_{k+1} - u||^2 \leq ||u_k - u||^2 - 2\,\alpha_k (u_k - u, J'(u_k) - J(u)) + \alpha_k^2 ||J'(u_k) - J'(u)||^2$$

Nun gilt

$$J'(u_k) - J'(u) = (Au_k - f) - (Au - f) = A(u_k - u)$$

Außerdem ist A positiv definit nach Voraussetzung

$$(u_k - u)^T A(u_k - u) \overset{\text{Rayleigh-Quotient}}{\geq} \lambda_{\min} ||u_k - u||^2$$

Einsetzen liefert

$$\begin{aligned} ||u_{k+1} - u||^2 &\leq ||u_k - u||^2 - 2\alpha_k \lambda_{\min} ||u_k - u||^2 + \alpha_k^2 ||A||^2 ||u_k - u||^2 \\ &= ||u_k - u||^2 \cdot \underbrace{(1 - 2\alpha_k \lambda_{\min} + \alpha_k^2 ||A||^2)}_{=:D} \end{aligned}$$

Ziel: Wähle α_k so, dass $D \in (0,1)$ ist. Diese Überlegung führt zu einer Faktordiskussion. Soll hier nicht weiter ausgeführt werden, da die gleichen Überlegungen in den Existenzsätzen (3.5.1) und (3.7.4) bereits ausgearbeitet wurden.

Es folgt als Ergebnis für α_k:

$$\alpha_k > 0 \quad \wedge \quad \alpha_k < \frac{2 \cdot \lambda_{\min}}{||A||^2}$$

□

5.6 Konjugiertes Gradientenverfahren (cg-Verfahren)

5.6.1 Hintergrund

Das konjugierte Gradientenverfahren gehört zu den sog. *Krylovraum-Methoden*. Es sei $A \in M(n,n)$ eine symmetrische positiv definite Matrix A, die der Gleichung $Ax = b$ genügt.

Krylovraum-Methoden erzeugen iterativ, ausgehend von einer Näherungslösung $x_0 \in \mathbb{R}^n$ mit dem zugehörigen Residuum $d_0 := b - Ax_0$ eine Folge weiterer Näherungslösungen x_k , also

$$x_0 \quad \rightarrow \quad x_1 \quad \rightarrow \quad \ldots \quad \rightarrow \quad x_m$$

Bei exakter Rechnung bricht das Verfahren spätestens nach n Schritten mit der gesuchten Lösung ab. Also

$$x_m = \tilde{x} := A^{-1}b, \quad m \leq n$$

Dabei wird für alle $k > 0$

$$x_k \in x_0 + V_k(d_0, A)$$

verlangt, mit dem Krylovraum $V_k(d_0, A)$ zur Matrix A und dem Startresiduum d_0.

Infolge von Rundungsfehlern sind diese Verfahren aber nicht endlich und so ist das Konvergenzverhalten von entscheidender Relevanz. Der Arbeitsaufwand wird durch die Multiplikation einer Matrix mit einem Vektor bestimmt.

Das cg-Verfahren eignet sich gut für sehr große Matrizen, die dünn und unregelmäßig besetzt sind.

5.6.2 Das cg-Verfahren

Idee. Bisher wurde das Funktional $f(x) = \frac{1}{2}x^T Ax - b^T x$ betrachtet, wobei die Komponenten des Vektors x positiv sind. Hierzu wurde die Lösung im 1D-Fall mit Hilfe der Liniensuche approximiert. Also

$$x_{i+1} = x_i + \alpha_i d_i$$
$$f(x_{i+1}) = \min_{z \in < d_i >} f(x_i + z)$$

Für die lineare Hülle wird die Bezeichung

$$< d_i >= \text{span}(d_i) \quad \dim < d_i >= 1$$

verwendet.

In dieser Sektion wird nun eine Verbesserung des Verfahrens angestrebt. Dazu wird eine 2D-Minimierung, also eine Ebenensuche, verwendet.

$$f(x_{i-1}) = \min f(x_i + z), \quad z \in < d_{i-1}, g_i >$$

$$\begin{aligned} \text{mit} \quad & d_{i-1} = x_i - x_{i-1} \quad \text{(Richtung der letzten Korrektur)} \\ \text{und} \quad & g_i = Ax_i - b \quad \text{(Gradient des Funktionals)} \end{aligned}$$

Zunächst werden einige Hilfssätze bereitgestellt.

Definition 5.6.1 *Sei $A \in M(n,n)$ positiv definit. Die Vektoren $x, y \in \mathbb{R}^n$ heißen konjugiert oder A-orthogonal, falls*

$$x^T A\, y = 0$$

Beweis. Nicht geführt.

Bemerkung 5.6.2 *Wenn die Vektoren $x_1, \ldots, x_k \in \mathbb{R}^n$ paarweise konjugiert sind, dann sind diese sogar linear unabhängig.*

Beweis/ Rechnung.
Multiplikation der Gleichung

$$\sum_{i=1}^{k} \alpha_i x_i = 0$$

mit $(Ax_j)^T$ $(j = 1, \ldots, k)$ führt zu

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

□

Das nachfolgende Lemma zeigt die Vorteile der konjugierten Richtungen.

Lemma 5.6.3 *Gegeben seien die konjugierten Richtungen $d_0, \ldots, d_{n-1}$. Weiterhin sei $\tilde{x} = A^{-1} \cdot b$. Dann gilt*

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

Die Lösung ist also direkt hinschreibbar!

Beweis.
Wir arbeiten mit dem Ansatz

$$\tilde{x} = \sum_{k=0}^{n-1} \alpha_k d_k$$

Dieser wird mit $(Ad_i)^T, (i = 0, \ldots, n-1)$ multipliziert. Dann gilt

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

□

Lemma 5.6.4 *Es seien die konjugierten Richtungen $d_0, \ldots, d_{n-1}$ vorgegeben. Für jedes $x_0 \in \mathbb{R}^n$ liefert die durch*

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

mit $i \geq 0$ erzeugte Folge nach (höchstens) n-Schritten die Lösung $\tilde{x}$. Weiter wird α_i berechnet durch

$$\alpha_i = \frac{-g_i^T d_i}{d_i^T A d_i}$$
$$\text{und} \quad g_i = Ax_i - b$$

Beweis.
Wir beginnen mit $A(\tilde{x} - x_0) = (b - Ax_0)$. Unter Verwendung von Lemma (5.6.3) ergibt sich

$$(\tilde{x} - x_0) = \sum_{i=0}^{n-1} \alpha_i d_i \quad \text{mit } \alpha_i = \frac{d_i^T (b - Ax_0)}{d_i^T A d_i}$$

Es bleibt zu zeigen:

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

Rechnung.

$$\alpha_i = \frac{-d_i^T (Ax_0 - b)}{d_i^T A d_i} = \frac{-d_i^T (Ax_0 - Ax_i + Ax_i - b)}{d_i^T A d_i}$$

$$\Leftrightarrow \quad \alpha_i = \frac{-d_i^T \overbrace{(Ax_i - b)}^{=g_i}}{d_i^T A d_i} - \underbrace{\frac{d_i^T (Ax_0 - Ax_i)}{d_i^T A d_i}}_{=0\,?}$$

Von Interesse ist die Frage, wann der zweite Summand verschwindet. Diese Frage wird mit dem Algorithmus näher untersucht: Multiplikation mit $(Ad_i)^T$ führt auf das Ergebnis

$$d_i^T A(x_i - x_0) = \sum_{j=0}^{j<i} \alpha_j d_i^T A d_j$$

Dabei ist $d_i^T A d_j = 0$ wegen der Konjugiertheit.

□

Korollar 5.6.5 *Mit den gleichen Voraussetzungen wie im vorangegangen Lemma folgt, dass x_k das Funktional f in dem Raum $x_0 + V_k$, mit $V_k =< d_0, \ldots, d_{k-1} >$, minimiert. Insbesondere gilt*

$$d_i^T g_k = 0 \quad \text{für } i < k$$

Beweis.
1) Es genügt $d_i^T g_k = 0$ für $i < k$ zu zeigen. Wir beginnen mit

$$\begin{aligned} f(x_k) &= \min_{\alpha_i} f\left(x_0 + \sum_{i=0}^{k-1} \alpha_i d_i\right) \\ \Leftrightarrow \quad \frac{\partial}{\partial \alpha_i} f(x_k) &= 0 \\ \Leftrightarrow \quad \nabla f(x_k)^T \cdot d_i &= 0 \\ \Leftrightarrow \quad (Ax_k - b)^T \cdot d_i &= 0 \\ \Leftrightarrow \quad g_k^T \cdot d_i &= 0 \end{aligned}$$

Die Äquivalenzkette gilt, weil f quadratisch ist.
2) Es ist $0 = d_k^T g_{k+1}$ zu zeigen. Also

$$\begin{aligned} 0 &= d_k^T g_{k+1} \\ &= d_k^T (Ax_{k+1} - b) \\ &= d_k^T \left(A\left(x_k - \frac{g_k^T d_k}{d_k^T A d_k} \cdot d_k\right) - b\right) \\ &= d_k^T (Ax_k - b) - \frac{d_k^T A d_k}{d_k^T A d_k} \cdot g_k^T d_k \\ &= d_k^T g_k - g_k^T d_k = 0 \end{aligned}$$

3) Induktion.
Zu zeigen $d_i^T g_k = 0$ für $i < k$.
Ind.Anfang: $k = 1 : d_0^T g_1 = 0$ (erfüllt wegen 2))
Ind.Annahme: $d_i^T g_{k-1} = 0$ für $i < k - 1$.
Ind.Schritt: $k - 1 \to k$.
Der Algorithmus liefert

$$x_k - x_{k-1} = \alpha_{k-1} d_{k-1}$$

Multiplikation mit A und anschließendes Einfügen von b bringt

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

Für g_k gilt

$$g_k - g_{k-1} = \alpha_{k-1} A d_{k-1}$$

Diese Gleichung wird mit d_i^T multipliziert

$$d_i^T (g_k - g_{k-1}) = 0 \quad \text{für } i < k - 1$$

Unter Benutzung der Induktionsannahme folgt

$$d_i g_k = 0 \quad \text{für } i < k - 1$$

Für den Fall $i = k - 1$ wird 2) verwendet. Damit folgt insgesamt

$$d_i g_k = 0 \quad \text{für } i < k$$

□

Algorithmus 5.6.6 *(cg-Verfahren)*

1. *Initialisierung:* $x_0 \in \mathbb{R}^n$.
   Setze $d_0 = -g_0 = b - Ax_0$.
2. *Iteration über* $k = 0, 1, 2, \ldots$

$$\alpha_k = \frac{-g_k^T d_k}{d_k^T A d_k} \quad \textit{(Lemma (5.6.4))}$$
$$x_{k+1} = x_k + \alpha_k d_k$$

Sukzessiver Aufbau der konjugierten Richtungen:

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

Dazu die kurze Erläuterung. Der Update-Schritt ist

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k d_k \\ &= x_k + \alpha_k(-g_k + \beta_{k-1} d_{k-1}) \end{aligned}$$

wobei g_k *der Gradient und* d_{k-1} *die Suchrichtung ist, und alles im 2-dim. betrachtet wird.*

Bemerkung.

- Bei der Durchführung des cg-Verfahrens müssen lediglich die vier Vektoren x_k, g_k, d_k, Ad_k gespeichert werden. Dabei muß nur eine einzige Matrix-Vektor-Multiplikation realisiert werden.
- Das cg-Verfahren wird nur bei symmetrischen Matrizen verwendet. Bei nicht symmetrischen Matrizen A rechnet man

  $$A^T \cdot Ax = A^T \cdot b$$

  Hier die Lösung zu berechnen ist genauso aufwendig wie das Gradientenverfahren und somit für die Praxis nicht interessant.

Der nächste Satz listet weitere Eigenschaften des cg-Verfahrens auf.

Satz 5.6.7 *(Eigenschaften des cg-Verfahrens)*
Solange $g_{k-1} \neq 0$ *ist (für* $g_{k-1} = 0$ *hat man die Lösung nämlich gefunden) gilt*

1. $d_{k-1} \neq 0$
2. $V_k := < g_0, Ag_0, A^2 g_0, \ldots, A^{k-1} g_0 >$
   Den so konstruierten Raum nennt man Krylov-Raum.
   Es ist $V_k = < g_0, g_1, \ldots, g_{k-1} > = < d_0, \ldots, d_{k-1} >$. *Damit folgt natürlich nicht* $g_1 = Ag_0$ *, usw. Es sei an den Austauschsatz von Steinitz erinnert.*

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.
Via Induktion.
Ind.Anfang: $k = 1$ (alle 4 Punkte erfüllt.)
Ind.Schritt: $k \to k+1$
Zunächst

$$g_k \overset{\text{Alg.}}{=} g_{k-1} + \alpha_{k-1} A d_{k-1}$$

Wegen

$$< g_0, Ag_0, \ldots, A^{k-1} g_0 > = < d_0, \ldots, d_{k-1} >$$

gibt es die Darstellung

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

Also bringt Einsetzen

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

und damit hat man $< g_0, \ldots, g_k > \subset V_{k-1}$ gezeigt. Nach Annahme seien $d_0, \ldots, d_{k-1}$ konjugiert und wegen der Optimalität von x_k folgt

$$d_i^T g_k = 0, \quad i < k$$

Falls $g_k \neq 0$ (für $g_k = 0$ hat man die Lösung gefunden) verifiziert man die Folgerungen

$$\begin{aligned} & g_k \text{ linear unabhängig von } (d_0, \ldots, d_{k-1}) \\ \Rightarrow \quad & g_k \notin V_k \end{aligned}$$

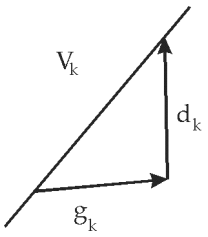
Demnach ist $< g_0, \ldots, g_k >$ ein $(k-1)$-dimensionaler Raum und kein echter UR von V_{k+1}. Also gilt

$$< g_0, Ag_0, \ldots, A^{k-1} g_0 > = < g_0, \ldots, g_k >$$

Damit wurde aus dem Satz Punkt 2 Teil 1) bestätigt. Zeige nun $V_{k+1} = < d_0, \ldots, d_k >$. Mit diesem Beweisteil werden gleichzeitig Punkt 1 und Punkt 2 Teil 2) bewiesen. Betrachten wir dazu

$$\begin{aligned} & g_k + d_k = g_k - \underbrace{g_k + \beta_{k-1} d_{k-1}}_{= d_k} \\ \Leftrightarrow \quad & g_k + d_k = \beta_{k-1} d_{k-1}, \quad d_{k-1} \in V_k \\ \Rightarrow \quad & g_k + d_k \in V_k \end{aligned}$$

Dazu die erklärende Abbildung



Es wurde gezeigt

$$< g_0, \ldots, g_{k-1}, g_k > = < g_0, \ldots, g_{k-1}, d_k > = < d_0, \ldots, d_{k-1}, d_k >$$

Also $V_{k+1} = < d_0, \ldots, d_k >$. Wir rechnen nun Punkt 3 des obigen Satzes nach. Zu zeigen ist, dass $d_0, \ldots, d_k$ paarweise konjugiert sind. Nach dem Algorithmus ist

$$d_k = -g_k + \beta_{k-1} d_{k-1}$$

Multiplikation mit $(Ad_i)^T$ liefert

$$d_i^T A d_k = \underbrace{-d_i^T A g_k}_{\text{Wann} = 0\,?} + \beta_{k-1} d_i^T A d_{k-1} \tag{5.3}$$

i) Fall $i < k-1$
Nach Annahme ist $\beta_{k-1} d_i^T A d_{k-1} = 0$ und weiterhin ist $d_i \in V_{k-1}$ woraus $Ad_i \in V_k$ folgt. Damit bleibt die Frage, wann in der Gleichung (5.3) der Term $-d_i^T A g_k$ verschwindet. Dazu

$$Ad_i = \sum_{j=0}^{k-1} \delta_j d_j$$

und damit

$$\begin{aligned} d_i^T A g_k &= (Ad_i)^T g_k \\ &= \sum_{j=0}^{k-1} \delta_j d_j^T g_k \\ &= 0, \quad \text{wg. Optimalitätsbedingung} \\ \Rightarrow \quad d_i^T A g_k &= 0 \end{aligned}$$

ii) Fall $i = k-1$
Wir betrachten wieder (5.3):

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

Für den Bruch gilt laut Algorithmus $\frac{g_k^T A d_{k-1}}{d_{k-1}^T A d_{k-1}} = 0$.

Punkt 4 des Satzes:
Minimaleigenschaft. Anwendung von Korollar (5.6.5). Dort waren die Punkte 1 - 3 Voraussetzungen!

□

Satz 5.6.8 *(Konvergenz des cg-Verfahrens)*
Es gilt

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

Beweis. Nicht gezeigt.

5.7 Vorkonditionierung

Dient der Verbesserung der Konvergenz der Iterationsverfahren.

Grundidee:

i) Transformiere $Ax = b$ in $\tilde{A}\tilde{x} = \tilde{b}$ mit $\text{cond}(\tilde{A}) < \text{cond}(A)$

ii) Löse $\tilde{A}\tilde{x} = \tilde{b}$

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

1) Transformation
Sei C symmetrisch und positiv definit. Mit Hilfe von Cholesky gilt die Zerlegung $C = H \cdot H^T$. Betrachten wir nun

$$Ax = b \quad \Leftrightarrow \quad \underbrace{H^{-1}AH^{-T}}_{=:\tilde{A}} \cdot \underbrace{H^T x}_{=\tilde{x}} = \underbrace{H^{-1}b}_{=\tilde{b}}$$

und schreiben nochmal zur Verdeutlichung

$$\begin{aligned} \tilde{A} &= H^{-1}AH^{-T} \\ \tilde{x} &= H^T x \\ \tilde{b} &= H^{-1}b \end{aligned}$$

Also folgt

$$\tilde{A}\tilde{x} = \tilde{b}$$

2) Löse $\tilde{A}\tilde{x} = \tilde{b}$
Ähnlichkeitstransformation.

$$\begin{aligned} H^{-T}\tilde{A}H^T &= H^{-T} \left(H^{-1}AH^{-T} \right) H^T \\ &= C^{-1}A \end{aligned}$$

Somit ist $\tilde{A}$ ähnlich zu $C^{-1}A$. Falls $C = A$ liegt Ähnlichkeit zu Id vor. Dies hieße, dass $\text{cond}(\tilde{A}) = 1$, was numerisch aber zu aufwendig herbeizuführen ist. Daher wird der Kompromiss angestrebt, dass A in irgendeiner Beziehung zu C steht. Beispielsweise $C = \text{diag}(A)$.

Es wird nun der nächste Algorithmus präsentiert, der die Vorkonditionierung (engl. preconditional) berücksichtigt.

Algorithmus 5.7.1 *(pcg-Verfahren)*
Startwert sei $x_0 \in \mathbb{R}^n$. Damit folgt

$$g_0 = Ax_0 - b, \qquad d_0 = -h_0 = -C^{-1}g_0$$

Es folgt somit die Iteration für $k = 0, 1, 2, \ldots$

$$\alpha_k = \frac{g_k^T h_k}{d_k^T A d_k}$$
$$x_{k+1} = x_k + \alpha_k d_k$$

Berechnung der weiteren Vektoren

$$g_{k+1} = g_k + \alpha_k A d_k$$
$$h_{k+1} = C^{-1} g_{k+1} \quad \textit{(Vorkonditionierungsschritt)}$$
$$d_{k+1} = -h_{k+1} + \beta_k d_k, \quad \textit{mit } \beta_k = \frac{g_{k+1}^T h_{k+1}}{d_k^T h_k}$$

Der Update-Schritt $h_{k+1} = C^{-1} g_{k+1}$ wird wieder durch Lösen eines linearen Gleichungssystems bestimmt. Also

$$C\, h_{k+1} = g_{k+1}$$

Bemerkung.
Erweiterung der Matrix C zu einer oberen Dreiecksmatrix macht das pcg-Verfahren kaputt, da dann die Bedingung, C symmetrisch, nicht mehr erfüllt ist. Lösungsansatz: Vorwärts-Rückwärtseinsetzen. Wähle C einmal als obere - und andererseits als untere Dreiecksmatrix.

Stichwort: Unvollständige Cholesky-Zerlegung

$$A = H^T H + R \quad \rightarrow \quad C = H^T H$$

Abschließend gilt der Satz

Satz 5.7.2 *Es gilt*

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

mit $\kappa = cond(C^{-1}A)$.

5.8 Defektkorrektur

Hier wird kurz der Begriff der Defektkorrektur erläutert. Ausgehend von

$$x_{k+1} = x_k + \alpha_k(b - Ax_k)$$

oder allgemeiner

$$x_{k+1} = x_k + \alpha_k \underbrace{C^{-1}(b - Ax_k)}_{=d_k}$$

nach Umformung mit $Cd_k = b - Ax_k$ folgt eine Fallunterscheidung

1. Annahme $C = A$:

$$x_k + \alpha_k A^{-1} b - \alpha_k x_k = (1 - \alpha_k)\, x_k + \alpha_k \tilde{x}$$

2. Annahme $C = \mathrm{diag}(A)$:

$$x_{k+1} = x_k + \alpha_k \cdot (\mathrm{diag}(A))^{-1} \cdot (b - Ax_k)$$

Komponentenweise Betrachtung führt auf das *Jacobi-Verfahren*:

$$x_{k+1}(i) = x_k(i) + \alpha_k \frac{1}{A_{ii}} \left(b(i) - \sum_{j=1}^{n} A_{ij} x_k(j) \right)$$

Eine Verbesserung des Verfahrens wurde nach *Gauß-Seidel* benannt:

$$x_{k+1}(i) = x_k(i) + \alpha_k \frac{1}{A_{ii}} \left(b(i) - \sum_{j=1}^{i-1} A_{ij} x_{k+1}(j) - \sum_{j=i}^{n} A_{ij} x_k(j) \right)$$

Dieses Verfahren konvergiert für $0 < \alpha_k < 2$.

Projiziertes Gauß-Seidel Verfahren

$$x_{k+1}(i) = \max \left(0, x_k(i) + \alpha_k \frac{1}{A_{ii}} \left(b(i) - \sum_{j=1}^{i-1} A_{ij} x_{k+1}(j) - \sum_{j=i}^{n} A_{ij} x_k(j) \right) \right)$$

Konvergenz liegt wie beim vorherigen Verfahren für $0 < \alpha_k < 2$ vor.

5.9 Vergleich der Verfahren

Dieses Kapitel wurde in einer Übungsstunde erarbeitet und vergleicht das Gradientenverfahren mit dem konjugierten Gradientenverfahren (cg-Verfahren).

Gradientenverfahren
Wie bereits gezeigt, bedeutet eine Halbierung der Gitterweite h, dass 4-mal so viele Schritte bearbeitet werden müssen. Dazu die Gleichung

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

cg-Verfahren
In dem Fehlerterm des adjungierten Gradientenverfahrens erhält man eine Wurzel in der Kardinalität und somit benötigt man hier bei einer Halbierung von h nur 2-mal so viele Schritte:

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

Das bedeutet eine deutliche Verbesserung!

In der Praxis kann es trotzdem vorkommen, dass bei Anwendung des cg-Verfahrens der Schrittfaktor vervierfacht wird, bei Halbierung der Schrittweite. Dazu die Begründung:

Problem: Löse $Ax = b$. Wobei aber A nicht symmetrisch ist. In diesem Fall ist

$$L(\nabla u, \nabla \varphi) + \beta_n(\partial_x u, \varphi)$$

Also wird $Ax = b$ mit A^T multipliziert, um A symmetrisch zu machen:

$$A^T Ax = A^T b$$

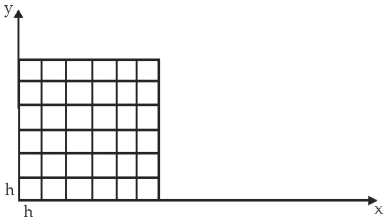
Somit ist die linke Seite symmetrisch und die rechte Seite positiv definit. Allerdings werden bei $A^T \cdot A$ die Konditionszahlen ebenfalls multipliziert und die Wurzel in dem Fehlerterm fällt weg. Also hat man dann $\kappa \sim \frac{1}{h^2}$, was dem Faktor 4 bei halbierter Gitterweite entspricht.

5.10 Aufwand $\mathcal{O}(\cdot)$ der Verfahren

In dieser Sektion soll ein kurzer Einblick in den Rechenaufwand der einzelnen Verfahren gegeben werden. Dazu sei folgende Situation gegeben

Problem
Löse $Ax = b$. Wir betrachten die Laplace-Gleichung im 2D. Die Ordnung der Ableitung (hier 2) bestimmt die Matrix A. Weiter sei $\dim(A) = n$. Die Einträge pro Zeile entsprechen $\mathcal{O}(1)$.

Frage: Wie hängt n mit der Gitterweite h zusammen? - Dazu als Erklärung die Abbildung



Demnach ist wegen $x \cdot y = h \cdot h$ die Dimension $n \sim \frac{1}{h^2}$. Da wir das Laplace-Problem in 2D betrachten gilt für die Kondtion $\kappa \sim \frac{1}{h^2}$.

Wir versuchen nun die Zahl der Iterationen bei vorgegebener Reduktion am zu ermitteln:

Beispiel. Gradientenverfahren

$$\left(1 - \frac{2}{\kappa}\right)^m = \text{red} < 1$$

$$\Rightarrow \quad m \cdot \ln\left(1 - \frac{2}{\kappa}\right) = ln(\text{red})$$

Demnach folgt

$$m = \frac{\ln(\text{red})}{-\frac{2}{\kappa}} \quad \Rightarrow \quad m \sim \kappa$$

Aufwand pro Schritt kann aus dem Algorithmus (5.7.1) bestimmt werden:

$$\text{Matrix} \cdot \text{Vektor} \triangleq \mathcal{O}(n^2)$$
$$\text{Vektor} \cdot \text{Vektor} \triangleq \mathcal{O}(n)$$

In unserem Beispiel folgt wegen den Einträgen pro Zeile $= \mathcal{O}(1)$ dann

$$\text{Matrix} \cdot \text{Vektor} \triangleq \mathcal{O}(n)$$

Also ist der Aufwand pro Schritt: $\mathcal{O}(n)$. Demnach

$$m \sim \kappa \sim \frac{1}{h^2} \sim n$$
$$m \cdot n \sim n^2 \quad \text{(Gradientenverfahren in 2D)}$$

Die Zahl m zählt die Schritte und n gibt die Kosten an. Bemerke, dass dieses Verfahren gegenüber Gauß-Seidel eine Verbesserung um eine Potenz darstellt. Denn Gauß-Seidel $\triangleq \mathcal{O}(n^3)$.

Beispiel. cg-Verfahren

$$m \sim \sqrt{\kappa} \sim \frac{1}{h} \sim \sqrt{n}$$
$$m \cdot n \sim n^{1.5} \quad \text{(cg-Verfahren in 2D)}$$

Ein optimales Verfahren würde demnach für $\mathcal{O}(n^1)$ gelten. Diese Art von Lösern nennt man *Mehrgitterverfahren* und werden im nächsten Kapitel vorgestellt.

6 Mehrgitterverfahren

6.1 Einleitung

- Löser für lineare Gleichungssysteme bei PDGL
- Konvergenzrate unabhängig von der Gitterweite h, also Zahl der Iterationsschritte konstant
- Aufwand lediglich $\mathcal{O}(n)$, dim(Gl.systems) = n.

Die Mehrgitterverfahren stellen eine Klasse von Algorithmen, die mehr-dimensionale Gleichungssysteme „optimal"lösen können. So können PDGL, wie beispielsweise die Poisson-Gleichung, mit Rechenaufwand $\mathcal{O}(n)$ gelöst werden, wobei n für die Anzahl der Unbekannten steht.

6.2 Grundidee

Die Auflösung des Gitters wird zunächst vergröbert, um so die Problemgröße zu reduzieren. Auf dem groben Gitter werden jeweils nur Korrekturen der Fehler auf den feineren Gitter mittels Glättern (Gauß-Seidel) approximiert.

Hochfrequente Anteile werden so gedämpft, was eine Glättung des Fehlers nach sich zieht. Siehe dazu die Abbildung (6.1). Niederfrequente Anteile auf dem feineren Gitter werden daher zu hochfrequenten Anteilen auf dem gröberen Gitter und können so beseitigt werden.

6.2.1 Beispiel

Zu dem bekannten Problem $-u'' = f$ gibt es die Diskretisierung $A_h x_h = f_h$ mit

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

Iterativer Löser, Jacobi-Verfahren

$$x_h^{i+1} = x_h^i + D_h^{-1}(f_h - A_h x_h^i), \quad D_h = \operatorname{diag}(A_h)$$

Dazu folgende Abbildung. Diese zeigt, dass der Fehler zwar nicht unbedingt kleiner werden muß in der Summe, dieser aber auf jeden Fall mit zunehmender Anzahl von Iterationen glatter wird. Sprich, der Iterationsfehler $e_h^i = x_h - x_h^i$ wird glatt bei Anwendung der Jacobi-Verfahrens.

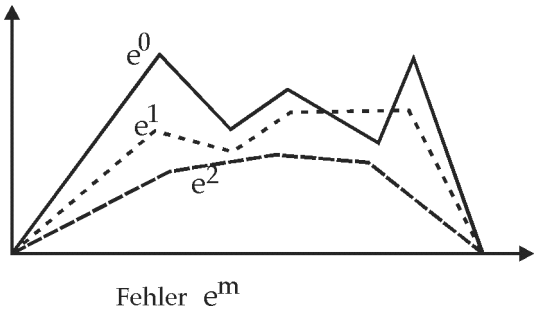


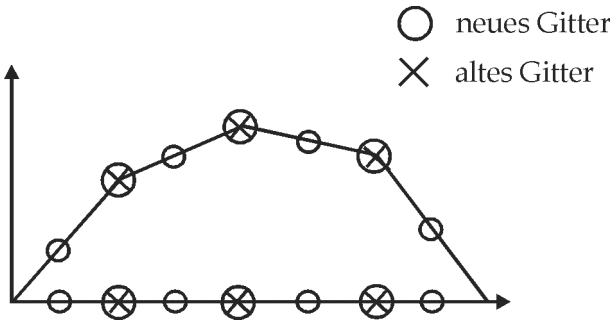
Abbildung 6.1: Glättung des Fehlers

6.2.2 Gittertransfer

Prolongation

$$p : \mathbb{R}^{\frac{n}{2}} \to \mathbb{R}^{n+1}$$

Definition 6.2.1 *Zu einer gegebenen Diskretisierung wird ein neues verfeinertes Gitter gewählt. Dazu die Abbildung*



Restriktion

$$v : \mathbb{R}^{n+1} \to \mathbb{R}^{\frac{n}{2}}$$

Dabei werden in vorangegangen Abbildung die Rollen des Kreises und der Kreuzes getauscht.

6.2.3 Grobgitter-Korrektur

Motivation ist die Fragestellung, wie man zwischen zwei verschiedenen Gittern während der Berechnung „wechselt “. Zu lösen ist $A_h x_h = f_h$. Nach i Schritten des Jacobi-Verfahrens erhält man näherungsweise die Lösung x_h^i. Wir rechnen

$$\begin{aligned} & A_h x_h - A_h x_h^i = f_h - A_h x_h^i \\ \Leftrightarrow \quad & A_h \underbrace{(x_h - x_h^i)}_{=:e_h^i} = \underbrace{f_h - A_h x_h^i}_{=:d_h^i} \\ \Leftrightarrow \quad & A_h e_h^i = d_h^i \end{aligned}$$

Man beachte

$$\begin{aligned} x_h &= x_h^i + (x_h - x_h^i) \\ &= x_h^i + e_h^i \end{aligned}$$

Also

$$A_h e_h^i = d_h^i \quad \rightarrow \quad A_{2h} e_{2h} = r \cdot (d_h^i)$$

In Worten. Weil der Fehler glatt ist, kann zu einem größeren Gitter übergegangen werden. Es gehen kaum Informationen verloren und die Rechenzeit verkürzt sich erheblich. Daher ist die rechte Seite der Äquivalenz „billiger “, weil die Dimension halbiert ist.

Algorithmus 6.2.2 *(Mehrgitter-Verfahren)*

1. *Startwert* $x_h^{0,0}$
2. *for* $k = 0, 1, 2, \ldots$

   for $i = 1, \ldots, \nu$

   $x_h^{k,i+1} = x_h^{k,i} + D_h^{-1} \left(f_h - A_h^k x_h^i \right)$

   end

   $A_{2h} e_{2h} = r \cdot \left(f - A_h^k x_h^\nu \right)$ *(Grobgitter)*

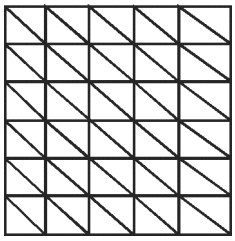
   $x_h^{k+1,0} = x_h^{k,\nu} + p(e_{2h})$

   end (of k)

Im obigen Algorithmus steht ν für die Anzahl der Glättungsschritte. In der Regel 3,4,5.

6.3 Glättung

2D-Modell



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 wird definiert als

$$C = Id - D^{-1}A, \quad C \in M(n,n)$$

Die Gitterweite wird beschrieben durch

$$h = \frac{1}{N+1}, \quad n = N^2 \quad \text{(Zahl der inneren Punkte)}$$

Eigenwerte der Matrix C

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

Zugehörige Eigenvektoren seien $z^{(k,l)}$. Schlechtes Konvergenzverhalten wird durch

$$\mu^{(1,1)} = \frac{1}{2}\left(\cos\frac{\pi}{N+1} + \cos\frac{\pi}{N+1}\right) \tag{6.1}$$

Wir betrachten abschließend den Fehler. Dazu

$$\begin{aligned} x^i - x &= \sum_{k,l=1}^{N} \alpha^{kl} z^{kl} \\ C\,(x^i - x) &= \sum_{k,l=1}^{N} \alpha^{kl} \mu^{kl} z^{kl} \end{aligned}$$

Hohe und niedrige Frequenzen
Einträge der Eigenvektoren

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

Ein „schlechter "war bereits für $k = l = 1$ gefunden, siehe (6.1). Stichwort: langwelliges EV. Für $k,l > 1$ würde der Sinus schneller oszillieren.

Im folgenden wird die Fragestellung behandelt, wie sich das Jacobi-Verfahren auf Teilräumen mit Eigenvektoren höherer Frequenz verhält. Definiere

$$\mathbb{R}^{N^2} = \mathbb{R}^n \supset X_{\text{OSC}} = \text{span}\{z^{kl} | 1 \le k,l \le N, \max(l,k) > \frac{N}{2}\}$$

Lemma 6.3.1 *Für das Modellproblem sei $x^0 - x \in X_{OSC}$. Dann ist $e^m = x^m - x \in X_{OSC}$ und es gilt*

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

D.h. die Konvergenzrate ist h-unabhängig bei Einschränkung auf X_{OSC}.

Beweis.
Zunächst eine Notation zur vereinfachten Schreibweise

$$\sum := \sum_{\substack{1 \leq k \leq N \\ \frac{N}{2} < l \leq N}}, \quad \max := \max_{\substack{1 \leq k \leq N \\ \frac{N}{2} < l \leq N}}$$

Darstellung von e^m:

$$e^m = \sum \alpha^{kl} z^{kl}$$

mit der Norm

$$||e^m||_2^2 = \sum \left(\alpha^{kl}\right)^2 \tag{6.2}$$

Nun wird die Eigenvektor-Eigenschaft $Cz^{kl} = \mu^{kl} z^{kl}$ ausgenutzt. Damit folgt

i) $e^m \in X_{\text{OSC}}$

ii) Abschätzung von $||e^{m+1}||_2^2$:

$$\begin{aligned} ||e^{m+1}||_2^2 &\leq \sum \left(\mu^{kl}\right)^2 \cdot \left(\alpha^{kl}\right)^2 \\ &\overset{(6.2)}{\leq} \max \left(\mu^{kl}\right)^2 \underbrace{\sum \left(\alpha^{kl}\right)^2}_{=||e^m||_2^2} \end{aligned}$$

wegen $l > \frac{N}{2}$ folgt $\cos \frac{l\pi}{N+1} \in (-1, 0)$ und

$$\begin{aligned} ||e^{m+1}||_2^2 &\leq \frac{1}{2} \max \cos \frac{k\pi}{N+1} \cdot ||e^m||_2^2 \\ &\leq \frac{1}{2} \cdot ||e^m||_2^2 \end{aligned}$$

□

Literaturverzeichnis

[1] Claes Johnson, Numerical solution of partial differential equations by the finite element method
Studentlitteratur, 1987

[2] Monika Dücker, Theorie und Numerik von Variationsungleichungen
Skript, WS 2003/04

[3] Walter A. Strauss, Partielle Differentialgleichungen. Eine Einführung
Vieweg, Lehrbuch Mathematik, 1995

[4] Wolfgang Hackbusch, Theorie und Numerik elliptischer Differentialgleichungen
Teubner Studienbücher, Mathematik, 1986

[5] Wikipedia
Internet, März 2007