Numerische Lösung von Differentialgleichungen I

Prof. Dr. Christian Lubich

Unautorisierter Vorlesungsmitschrieb
vom Wintersemester 2005/2006

Universität Tübingen

Gesetzt in LaTeX

Valentin Schwamberger*

Dank für Mitschriebe und Fehlerkorrekturen an:
Alexander Edele – Ludwig Gauckler – Gabriele Kallert
Christian Krüger – Moritz Nadler – Katharina Reiß
Jaron Treutlein – Mirko Windhoff

Letzte Änderung am 27. Mai 2007 um 23:09:17
(Revision 210)

*v.schwamberger@gmx.de

Inhaltsverzeichnis

Kapitel I

Randwertprobleme gewöhnlicher Differentialgleichungen

§ 1 Einführung

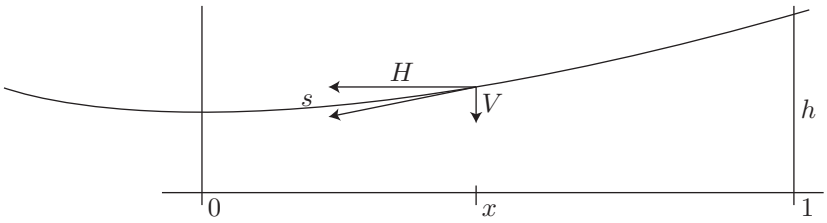
§ 1.1 Kettenlinie (Leibniz, Johannes Bernoulli 1691)

Beispiel. Sei H die horizontale Komponente der Seilkraft

$$H = a \quad (= \text{const})$$

und V die vertikale

$$V = bs \quad (V \propto \text{Seilmasse zwischen 0 und } x \propto s).$$



Im Gleichgewicht greift die Seilkraft tangential am Seil an:

$$\Rightarrow \quad \frac{dy}{dx} = \frac{V}{H} = \frac{b}{a} \cdot s =: c \cdot s \quad \text{mit } \frac{ds}{dx} = \sqrt{1 + \left(\frac{dy}{dx}\right)^2}\,, \quad ds = \sqrt{dx^2 + dy^2}\,.$$

Erhalte

$$\frac{d^2y}{dx^2} = c\sqrt{1 + \left(\frac{dy}{dx}\right)^2} \quad \text{bzw.} \quad y'' = c\sqrt{1 + (y')^2}\,, \qquad \text{Differentialgleichung,}$$

$$y'(0) = 0\,, \quad y(1) = h \qquad \text{Randbedingungen}$$

und daraus die Lösung

$$y(x) = K + \frac{1}{c}\cosh(cs)$$

mit K so, dass $y(1) = h$.

Bemerkung. Eine Differentialgleichung 2. Ordnung kann in ein System von Differentialgleichungen 1. Ordnung umgewandelt werden:

$$\begin{pmatrix} y \\ s \end{pmatrix}' = \begin{pmatrix} cs \\ \sqrt{1 + (cs)^2} \end{pmatrix}\,.$$

§ 1.2 Allgemeine Form eines Randwertproblems

$$y' = f(t,y)\ , \quad t \in [a,b]\ , \quad f : U \subset \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d\ ,$$
$$r\big(y(a), y(b)\big) = 0\ , \quad r : V \subset \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d\ .$$

Gesucht:

$$y : [a,b] \to \mathbb{R}^d\ , \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix}\ .$$

Anwendungsprobleme in Physik, Chemie, Biologie, WiWi: siehe in [AschMat88].

Bemerkung. Gegeben sei das Anfangswertproblem $r\big(y(a), y(b)\big) = y(a) - y_0$. Dann existiert genau eine Lösung des Anfangswertproblems (z. B. falls f Lipschitz-stetig).

Komplizierter bei Randwertproblemen:

Beispiel. $y'' + y = 0$,

$$z = y' : \quad \begin{pmatrix} y \\ z \end{pmatrix}' = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix}\ , \quad d = 2\ .$$

Allgemeine Lösung:

$$y(t) = C_1 \sin t + C_2 \cos t\ .$$

Randbedingungen:

(a) $y(0) = 0$, $y\big(\frac{\pi}{2}\big) = 0 \quad \Rightarrow \quad y(t) = 0$, eindeutige Lösung.

(b) $y(0) = 0$, $y(\pi) = 0 \quad \Rightarrow \quad y(t) = C \sin t$, $C \in \mathbb{R}$ beliebig, unendlich viele Lösungen.

(c) $y(0) = 0$, $y(\pi) = 1 \quad \Rightarrow \quad \nexists$ Lösung.

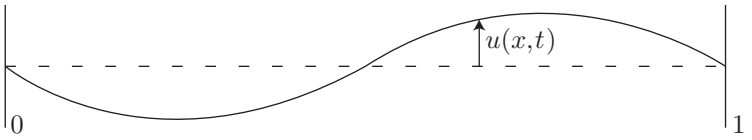
Betrachte wichtige Beispielklassen, die sich auf obige Standardform zurückführen lassen:

§ 1.3 Eigenwertprobleme

Beispiel (eingespannte Saite)**.** Auslenkung $u = u(x,t)$, $x \in [0,1]$ (Ort), $t > 0$ (Zeit). Wellengleichung

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial x^2}$$

mit Randbedingungen $u(0,t) = 0$, $u(1,t) = 0\ \forall t$.



Suche Lösungen der Form

$$u(x,t) = y(x)\sin\omega t\ , \quad \omega \text{ unbekannt, „stehende Welle“.}$$

Einsetzen in die Wellengleichung:

$$-y(x)\omega^2\sin\omega t = y''(x)\sin\omega t \quad \forall t\ .$$

Mit $\lambda = \omega^2$ erhalte:

$$y'' + \lambda y = 0\ , \quad y(0) = 0\ , \quad y(1) = 0\ .$$

Möchte $y \neq 0$, weitere Randbedingung: $y'(0) = 1$. $\lambda \in \mathbb{R}$ unbekannter Eigenwert, y Eigenfunktion. Führe $\lambda' = 0$ als weitere Differentialgleichung ein, damit: 3 Differentialgleichungen, 3 Randbedingungen in Standardform:

$$\begin{cases} \lambda' = 0\ , & y' = z\ , & z' = -\lambda y\ , \\ y(0) = 0\ , & y(1) = 1\ , & z(0) = 1\ . \end{cases}$$

($\lambda = (k\pi)^2$, $k \in \mathbb{Z}$ lokal eindeutige Lösung)
Allgemeiner:

$$\begin{aligned} y' &= f(y,\lambda)\ , && \lambda \in \mathbb{R}^l \text{ unbekannte Parameter,} \\ r\big(y(a), y(b), \lambda\big) &= 0 && \Rightarrow \quad \lambda' = 0 \text{ hinzufügen.} \end{aligned}$$

§ 1.4 Probleme mit freiem Rand

Beispiel. $y'' + \omega^2 y = 0$, ω gegeben. $y(0) = 0$, $y(b) = 0$, $y'(0) = 1$. (Bestimme die Länge b der Saite so, dass ω Eigenfrequenz.) Setze $\xi = x/b$, $0 \le \xi \le 1$, $z(\xi) = y(\xi b)$. Mit $' = d/d\xi$ erhalte:

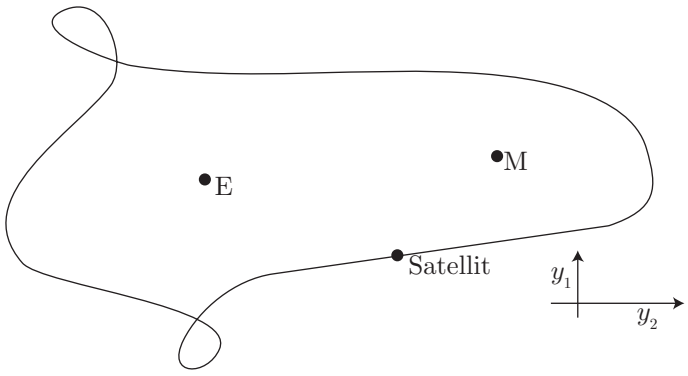
$$\begin{cases} \dfrac{1}{b^2} z'' + \omega^2 z = 0\ , & b' = 0\ , \\ z(0) = z(1) = 0\ , & z'(0) = b\ . \end{cases}$$

(Lösung: $b = k\pi/\omega$, $k \in \mathbb{Z}$.)

§ 1.5 Periodische Lösungen

$$\begin{aligned} y' &= f(y)\ , \\ t &\in [0,T]\ , \\ y(0) &= y(T)\ , \\ T' &= 0\ . \end{aligned}$$

T ist die unbekannte Periode.



Zusätzliche Randbedingung: Beispielsweise $y_2(0) = 0$ (aus physikalischen Überlegungen).

§ 2 Theoretische Grundlagen

Betrachte zunächst lineare Randwertprobleme

$$\begin{cases} y' = C(t)y + q(t) \ , \quad t \in [a,b] \ , \\ Ay(a) + By(b) = r \ . \end{cases} \tag{I.1}$$

Hier gegeben

$$A, B, C(t) \in \mathbb{R}^{d\times d} \ , \quad r, q(t) \in \mathbb{R}^d$$

mit q, C stetig auf $[a,b]$.
Sei $R(t,s) \in \mathbb{R}^{d\times d}$ *Resolvente*: Für das homogene Anfangswertproblem

$$y' = C(t)y \ , \quad y(a) = y_0$$

hängt die Lösung *linear* von y_0 ab.
Existiert die Matrix $R(t,s) \in \mathbb{R}^{d\times d}$, dann ist $y(t) = R(t,s)\eta$ Lösung des Anfangswertproblems

$$y' = C(t)y \ , \quad y(s) = \eta \ .$$

$R(t,s)$ ist Abbildungsmatrix der linearen Abbildung $y(s) \mapsto y(t)$.

Eigenschaften der Resolvente:

(a) $R(s,s) = I$,

(b) $R(t,\tau)R(\tau,s) = R(t,s)$ wegen

$$y(t) = R(t,s)y(s) = R(t,\tau)y(\tau) = R(t,\tau)R(\tau,s)y(s) \ \forall \ y(s) = y \in \mathbb{R}^d \ .$$

Satz 1. *Das lineare Randwertproblem hat genau dann eine eindeutige Lösung, wenn* $E := A + BR(b,a)$ *invertierbar.*

Beweis. Führe das Problem zunächst auf die homogene Differentialgleichung ($q(t) = 0$) zurück. Sei z Lösung des Anfangswertproblems:

$$\begin{cases} z' = C(t)z + q(t) \ , \\ z(a) = 0 \ . \end{cases}$$

Wegen der Linearität ist $y = z+w$ eine Lösung von (I.1), wobei w Lösung des homogenen Randwertproblems ist:

$$\begin{cases} w' = C(t)w \ , \\ Aw(a) + Bw(b) = r - Bz(b) \ . \end{cases}$$

Lösung von $w' = C(t)w$ mit Anfangswert $w(a) =: v$ ist $w(t) = R(t,a)v$. Wähle v:

$$\mathbb{R}^d \ni v \mapsto Av + BR(b,a)v =: Ev \in \mathbb{R}^d \ .$$

Diese Abbildung ist genau dann bijektiv, wenn E invertierbar ist:

$$Ev = r - Bz(b) \ .$$

Die Lösung von (I.1) lautet dann $y(t) = z(t) + R(t,a)v$. □

Bemerkung.

$$\begin{cases} y' = C(t)y + q(t) + \varphi\big(t, y(t)\big) \ , \\ Ay(a) + By(b) + \rho\big(y(a), y(b)\big) = r \ , \end{cases}$$

wobei φ, ρ „klein“ sind.
Die Existenz lässt sich mit dem Banachschen Fixpunktsatz zeigen.

Im nichtlinearen Fall ist es oft schwierig, die Existenz zeigen. Hier nur *lokale Eindeutigkeit.*

§ 2.1 Lokale Eindeutigkeit

$$\begin{cases} y' = f(t, y) \ , \\ r\big(y(a), y(b)\big) = 0 \ , \end{cases}$$

wobei f, r stetig differenzierbar. Sei $y^*(t)$ Lösung des Randwertproblems.
Bezeichnung:

$$A^* = \frac{\partial r}{\partial y_a}\big(y^*(a), y^*(b)\big) \ ,$$
$$B^* = \frac{\partial r}{\partial y_b}\big(y^*(a), y^*(b)\big) \ ,$$

außerdem linearisierte Differentialgleichung (Variationsgleichung)

$$v' = \frac{\partial f}{\partial y}\big(t, y^*(t)\big)v$$

mit Resolvente $R^*(t, s)$.

Satz 2. *Falls*

$$E^* := A^* + B^* R^*(b, a)$$

invertierbar ist, dann ist y^* lokal eindeutig. Das heißt, es existiert $V^* \subset U\big(y^*(a)\big)$, sodass $\forall \eta \in V^*$ mit $\eta \neq y^*(a)$ die Lösung des Anfangswertproblems

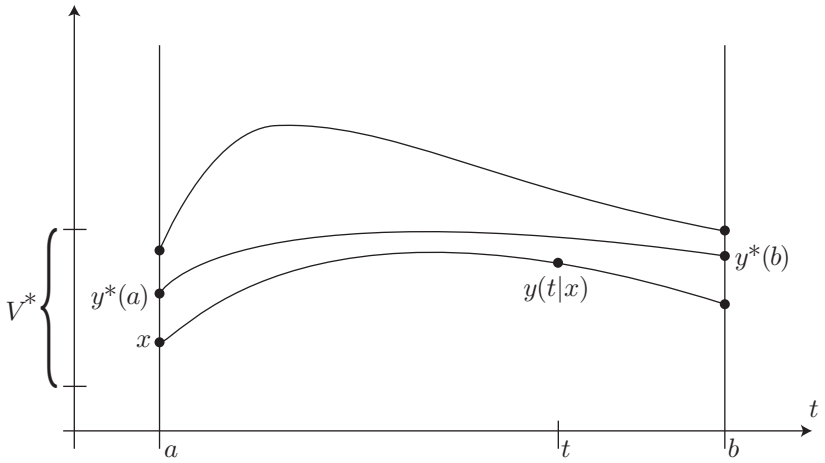
$$\begin{cases} y' = f(t, y) \ , \\ y(a) = \eta \end{cases}$$

keine Lösung des Randwertproblems ist.

Beweis. Betrachte das Anfangswertproblem

$$\begin{cases} y' = f(t, y) \ , \\ y(a) = x \ . \end{cases}$$

Es existiert eine Umgebung $V^* \subset U\big(y^*(a)\big)$, sodass das Anfangswertproblem für jedes $x \in V^*$ eine Lösung in $[a, b]$ besitzt. Wir nennen diese Lösung $y(t|x)$.



$F : x \in V^* \to \mathbb{R}^d$ ist eine Abbildung

$$x \mapsto F(x) := r\big(x, y(b|x)\big) \ .$$

F erfüllt die Gleichung

$$F\Big(y^*(a)\Big) = r\Big(y^*(a), y^*(b)\Big) = 0 \ ,$$

weil y^* Lösung des Randwertproblems ist.
Ableiten nach dem Anfangswert x liefert

$$\left\{ \begin{array}{l} \dfrac{\partial y'}{\partial x} = \dfrac{\partial f(t,y)}{\partial x} \overset{\text{Kettenregel}}{=} \dfrac{\partial f(t,y)}{\partial y}\dfrac{\partial y}{\partial x} \ , \\ \dfrac{\partial y(a)}{\partial x} = \dfrac{\partial x}{\partial x} = I \ , \end{array} \right.$$

$$\overset{\text{Übung}}{\Downarrow} \quad \Rightarrow \quad \frac{\partial y(t|x)}{\partial x} = R^*(t,a) \ ,$$

$$F'(x) = \frac{\partial r}{\partial y_a}\big(x, y(b|x)\big) + \frac{\partial r}{\partial y_b}\big(x, y(b|x)\big)\frac{\partial y(b|x)}{\partial x} \ ,$$

$$\Rightarrow \quad F'\big(y^*(a)\big) = A^* + B^*R^*(b,a) = E^* \ .$$

Laut Voraussetzung ist E^* invertierbar. Mit dem Satz über die Umkehrfunktion folgt, dass F in einer Umgebung von $y^*(a)$ bijektiv ist. Aus der Bijektivität und aus

$$F\big(y^*(a)\big) = 0$$

folgt dann

$$F(x) \neq 0 \quad \Leftrightarrow \quad x \neq y^*(a) \ .$$

Somit ist $y^*(t)$ eindeutig. □

§ 2.2 Störung des Randwerts

Lineare Randwertprobleme

$$\text{Ungestörtes Randwertproblem:} \quad \begin{cases} y' = C(t)y + q(t) \ , \\ Ay(a) + By(b) = r \ , \end{cases}$$

$$\text{Gestörtes Randwertproblem:} \quad \begin{cases} \tilde{y}' = C(t)\tilde{y} + q(t) \ , \\ A\tilde{y}(a) + B\tilde{y}(b) = \tilde{r} \ . \end{cases}$$

Setze

$$\Delta y(t) := \tilde{y}(t) - y(t) \ , \qquad \Delta r := \tilde{r} - r \ .$$

$$\Rightarrow \quad \begin{cases} (\Delta y)' = C(t)\Delta y \ , \\ A\Delta y(a) + B\Delta y(b) = \Delta r \ . \end{cases}$$

Habe $\Delta y(b) = R(b,a)\Delta y(a)$ und damit

$$\Delta r = A\Delta y(a) + BR(b,a)\Delta y(a) = E\Delta y(a) \ .$$

Nun mit $\Delta y(a) = E^{-1}\Delta r$:

$$\Delta y(t) = R(t,a)\Delta y(a) = R(t,a)E^{-1}\Delta r = R(t,a)E^{-1}\Big(A\Delta y(a) + B\Delta y(b)\Big) \ ,$$

setze

$$E(t) := ER(t,a)^{-1} = [A + BR(b,a)]R(a,t) = AR(a,t) + BR(b,t)$$

und damit schließlich

$$\Delta y(t) = E^{-1}(t)\Delta r = E^{-1}(t)\Big(A\Delta y(a) + B\Delta y(b)\Big) \ .$$

Konditionszahl für Randwertprobleme:

$$\rho := \max_{t\in[a,b]} \Big(\|E^{-1}(t)A\| + \|E^{-1}(t)B\|\Big) \ .$$

Konditionszahl für Anfangswertprobleme: $\Big(\Delta y(t) = R(t,a)\Delta y(a)\Big)$

$$\alpha := \max_{t\in[a,b]} \|R(t,a)\| \ .$$

Beide Konditionszahlen können von völlig unterschiedlicher Größenordnung sein.

Bemerkung. Nichtlineares Problem

$$\begin{cases} \tilde{y}' = f(t,\tilde{y}) \ , \\ r\Big(\tilde{y}(a), \tilde{y}(b)\Big) = r \ . \end{cases}$$

Linearisierung (wie in Satz 2) $\Delta y = \tilde{y} - y^*$ erfüllt

$$\Delta y(t) = E^{-1^*}(t)\Delta r + \mathcal{O}\big(\|\Delta r\|^2\big) \ .$$

§ 3 Einfaches Schießverfahren

(englisch: single shooting)
Randwertproblem

$$\begin{cases} y'(t) = f(t,y), \quad t \in [a,b] \\ r\big(y(a), y(b)\big) = 0 \end{cases}$$

unter Voraussetzung von § 2 (lokale Eindeutigkeit). Sei

$$y^* : [a,b] \to \mathbb{R}^d$$

lokal eindeutige Lösung des Randwertproblems.

Idee. Rückführung auf eine Folge von Anfangswertproblemen.

Man wählt einen Anfangswert $x \in \mathbb{R}^d$, löst das Anfangswertproblem

$$\begin{cases} y' = f(t,y) \\ y(a) = x \end{cases}$$

und erhält $y(b|x)$. Dann bestimmt man iterativ (Newton-Verfahren) den Anfangswert x so, dass

$$F(x) := r\big(x, y(b|x)\big) \stackrel{!}{=} 0 \qquad \text{(Nichtlineares Gleichungssystem im } \mathbb{R}^d\text{)}.$$

Ableitungsmatrix im Lösungspunkt $x^* = y^*(a)$ (siehe Beweis von Satz 2, § 2):

$$F'(x^*) = \underbrace{A^* + B^* R^*(b,a)}_{\text{invertierbar}} = E^*$$

invertierbar. Das Newton-Verfahren ist lokal konvergent (Numerik I).

$$x^{k+1} = x^k + \Delta x^k \quad \text{mit} \quad F'(x^k)\Delta x^k = -F(x^k) \ .$$

Hier:

$$F'(x^k) = A^k + B^k R^k(b,a) =: E^k \ .$$

Linearisierung um $y(\cdot|x^k)$:

$$A^k = \frac{\partial r}{\partial y_a}(x^k, y(b|x^k))$$
$$B^k = \frac{\partial r}{\partial y_b}(x^k, y(b|x^k)) \ .$$

$R^k(b,a)$ ist die Resolvente zu $v' = C^k(t)v$ mit $C^k(t) = \frac{\partial f}{\partial y}\big(t, y(t|x^k)\big)$. Somit:

$$E^k \Delta x^k = -r\big(x^k, y(b|x^k)\big) \ .$$

Berechnung von $y(b|x^k)$: durch numerische Lösung des Anfangswertproblems, z. B. mit dem Runge-Kutta-Verfahren.

Berechnung von E^k: 2 Möglichkeiten.

(a) „Äußere“ numerische Differentiation:

$$\begin{aligned}(E^k)_{ij} &= \frac{\partial}{\partial x_j} r_i\big(x, y(b|x)\big)\Big|_{x=x^k} \\ &\approx \frac{r_i\big(x^k + \delta \cdot e_j, \overline{y}(b|x^k + \delta \cdot e_j)\big) - r_i\big(x^k, \overline{y}(b, x^k)\big)}{\delta} \,,\end{aligned}$$

wobei $\overline{y}(b|x^k+\delta\cdot e_j)$, $\overline{y}(b, x^k)$ numerische Lösungen mit derselben Schrittweitenfolge sind, e_j der j. normierte Basisvektor ist und schließlich $\delta \approx \sqrt{\text{eps}}$, wobei eps die Maschinengenauigkeit bezeichnet.

(b) Löse linearisierte Gleichung (Variationsgleichung), oft vorteilhafter:

$$v_i' = \frac{\partial f}{\partial y}\big(t, \overline{y}(t|x^k)\big) \cdot v_i \in \mathbb{R}^d \,, \quad v_i(a) = e_i$$

mit e_i dem i. normierten Basisvektor. Damit

$$R^k(b, a) = \big(v_1(b), \ldots, v_d(b)\big)$$

und $E^k = A^k + B^k R^k(b, a)$.

Berechne A^k, B^k, $\frac{\partial f}{\partial y}$ analytisch oder durch numerische Differentiation („innere“). Bei gleichem Aufwand verlässlicher.

Bei schlechten Startwerten: Verwende Newton-Verfahren mit Dämpfung:

$$x^{k+1} = x^k + \lambda_k \Delta x_k \quad (0 < \lambda_k \leq 1) \quad \rightarrow \quad \text{Numerik I} \,.$$

Schießverfahren: Randwertproblem → Folge von Anfangswertproblemen.

Nachteile (des einfachen Schießverfahrens):

(a) Sehr empfindlich gegenüber der Wahl des Startwerts x^0,

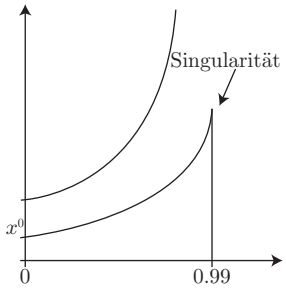
(b) Anfangswertproblem kann schlecht konditioniert sein, obwohl das Randwertproblem gut konditioniert ist,

(c) nichtlinear, oft unmöglich $y(t|x)$ für den gesamten Bereich $[a, b]$ zu berechnen.

Beispiele. (a) $y' = y^2, \quad y(0) = x > 0, \quad y(t|x) = \dfrac{1}{\frac{1}{x} - t}, \quad t < \dfrac{1}{x}$.

Randwertproblem:

$$\begin{cases} y' = y^2 \quad \text{auf } [0, 0.99] \\ y(0) + y(0.99) = 101 \end{cases}$$



Anfangswert der exakten Lösung: $y^*(0) = 1$.
Wählt man Startwert $x^0 \geq \frac{1}{0.99}$, so liegt Singularität in $[0, 0.99]$.

Bei Wahl von $x^0 \approx 0.9$ ist $y(0.99) \approx 10$ weit weg von $y^*(0.99) = 100$. → Schwierigkeiten mit dem Newton-Verfahren.

(b) Anfangswertproblem kann schlecht konditioniert sein, obwohl das Randwertproblem gut konditioniert ist: siehe § 2. Schlecht konditionierte Anfangswertprobleme, die im Newton-Verfahren gebraucht werden, können also große Probleme verursachen.

$$y'' = 100y \ , \quad y(0) = 1 \ , \quad y(3) = \mathrm{e}^{-30} \approx 9.36 \cdot 10^{-14} \ .$$

Allgemeine Lösung dieser Differentialgleichung:

$$y(t) = C_1 \mathrm{e}^{-10t} + C_2 \mathrm{e}^{10t} \ .$$

Lösung des Randwertproblems:

$$y^*(t) = \mathrm{e}^{-10t} \ .$$

$$y(0) = 1 \ , \quad y'(0) = x \ ,$$
$$y(t) = \frac{10 - x}{20} \mathrm{e}^{-10t} + \frac{10 + x}{20} \mathrm{e}^{10t} \qquad (x^* = y^{*\prime}(0) = -10)$$

$$y\left[t \,\middle|\, \begin{pmatrix} y(0) = 1 \\ y'(0) = x + \Delta x \end{pmatrix}\right] - y\left[t \,\middle|\, \begin{pmatrix} y(0) = 1 \\ y'(0) = x \end{pmatrix}\right] = \frac{\Delta x}{20} \left(\mathrm{e}^{10t} - \mathrm{e}^{-10t} \right) \ ,$$

falls $\Delta x / x^* \approx \text{eps}$ und mit $x^* = -10$ sowie ausgewertet an der Stelle $t = 3$:

$$\approx -\frac{\mathrm{e}^{30}}{2} \cdot \text{eps} \approx -5.34 \cdot 10^{12} \cdot \text{eps} \ .$$

§ 4 Mehrzielverfahren

(englisch: multiple shooting)

Idee. Zusätzliche Unterteilung des Intervalls $[a, b]$.

Wähle eine (geeignete) Unterteilung

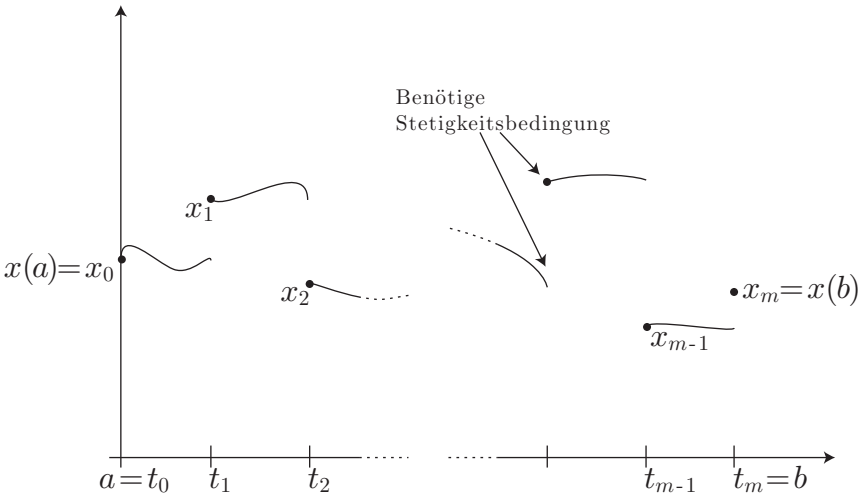
$$a = t_0 < t_1 < \cdots < t_m = b$$

und gib an diesen Punkten die Werte vor:

$$x_0, x_1, \ldots, x_m \in \mathbb{R}^d \ .$$

Löse m Anfangswertprobleme:

$$\begin{cases} y' = f(t,y) \quad \text{für} \quad t \in [t_j, t_{j+1}] \\ y(t_j) = x_j \end{cases}$$



Bezeichnung: $y(t|t_j, x_j)$.
Erhalte Lösung des Randwertproblems, falls

$$\begin{aligned} F_m(x_0, x_m) &:= r(x_0, x_m) = 0 \ , \\ F_j(x_j, x_{j+1}) &:= y(t_{j+1}|t_j, x_j) - x_{j+1} = 0 \ , \quad j = 0, \ldots, m-1 \ . \end{aligned}$$

Stetigkeitsbedingung:

$$F_j(x_j, x_{j+1}) = y(t_{j+1}|t_j, x_j) - x_{j+1} = 0 \ , \quad j = 0, \ldots, m-1$$

Randbedingung:

$$F_m(x_0, x_m) = r(x_0, x_m) = 0$$

Nichtlineares Gleichungssystem:

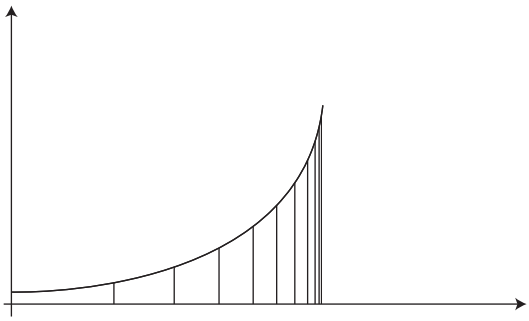
$$F(x) = \begin{pmatrix} F_0(x_0, x_1) \\ \vdots \\ F_{m-1}(x_{m-1}, x_m) \\ F_m(x_0, x_m) \end{pmatrix} = 0$$

Dimension $(m+1)\cdot d$. Lösung mit Newton-Verfahren unter Ausnützung der speziellen Struktur des Gleichungssystems.

Beispiele.

(a) Singularitäten lassen sich durch Einfügen weiterer Zwischenstellen abfangen, z. B. setze neuen Knoten, falls

$$\|y(t|t_j, x_j)\| > 2\|x_j\| + 1 \ .$$



(b) Betrachte wieder

$$y'' = 100y \ , \quad y(0) = 1 \ , \quad y(1) = \mathrm{e}^{-30} \ .$$

Sei $x_j = (y(t_j) = u_j, y'(t_j) = v_j)^T$ das vorgegebenene Wertepaar für $t = t_j$:

$$y\left(t\middle|t_j, \begin{pmatrix} y(t_j) = u_j \\ y'(t_j) = v_j \end{pmatrix}\right) = \frac{10u_j - v_j}{20}\cdot \mathrm{e}^{-10(t-t_j)} + \frac{10u_j + v_j}{20}\cdot \mathrm{e}^{10(t-t_j)}$$

$$y\left(t\middle|t_j, \begin{pmatrix} u_j + \Delta u_j \\ v_j + \Delta v_j \end{pmatrix}\right) - y\left(t\middle|t_j, \begin{pmatrix} u_j \\ v_j \end{pmatrix}\right) =$$

$$= \frac{\Delta u_j}{2}\left(\mathrm{e}^{10(t-t_j)} + \mathrm{e}^{-10(t-t_j)}\right) + \frac{\Delta v_j}{20}\left(\mathrm{e}^{10(t-t_j)} - \mathrm{e}^{-10(t-t_j)}\right)$$

und für $t_{j+1} - t_j = 0.1$:

$$\approx \Delta u_j \frac{\mathrm{e}}{2} + \Delta v_j \frac{\mathrm{e}}{20} \ .$$

→ Keine horrende Fehlerverstärkung, gute Ergebnisse.

Im Newton-Verfahren zu lösen (lasse Iterationsindex weg):

$$F'(x)\Delta x = -F(x) \ .$$

Verwende nun $R_j = R(t_{j+1}, t_j)$, die Resolvente von

$$v' = \frac{\partial f}{\partial y}\big(t, y(t|t_j, x_j)\big)v \ ,$$

sowie

$$A = \frac{\partial r}{\partial y_a}(x_0, x_m) \quad \text{und} \quad B = \frac{\partial r}{\partial y_b}(x_0, x_m) \ .$$

Erhalte damit

$$\begin{matrix} \\ \vdots \\ +BR_{m-1}\cdot \\ +B\cdot \end{matrix} \begin{bmatrix} R_0 & -I & & & & 0 \\ & R_1 & -I & & & \\ & & \ddots & \ddots & & \\ & & & R_{m-2} & -I & \\ & 0 & & & R_{m-1} & -I \\ A & & & & & B \end{bmatrix} \begin{bmatrix} \Delta x_0 \\ \Delta x_1 \\ \vdots \\ \Delta x_{m-2} \\ \Delta x_{m-1} \\ \Delta x_m \end{bmatrix} = - \begin{bmatrix} F_0 \\ F_1 \\ \vdots \\ F_{m-2} \\ F_{m-1} \\ F_m \end{bmatrix} ,$$

alle unbeschrifteten Elemente sind 0.
Lösung des linearen Gleichungssystems durch Block-Gauß-Elimination von unten („Condensing“). Erhalte in 1. Blockreihe:

$$A + B \cdot R_{m-1} \cdots R_0 =: E_m$$

(für exakte Lösung $x_j^* = y^*(t_j) : R_{m-1} \cdots R_0 = R^*(b,a)$). Damit $E_m^* = E^*$ von § 2, invertierbar. Erhalte aus

$$E_m \Delta x_0 = -(F_m + BF_{m-1} + BR_{m-1}F_{m-2} + \cdots + BR_{m-1} \cdots R_1 F_0)$$

Δx_0 mit dem Aufwand von $md^3 + \frac{d^3}{3}$ Operationen.
Rückwärts einsetzen:

$$R_0 \Delta x_0 - \Delta x_1 = -F_0 \ .$$

Allgemein:

$$\Delta x_{j+1} = R_j \Delta x_j + F_j \ , \quad j = 0, \ldots, m-1 \ ,$$

Rekursion md^2 Operationen.

Mögliche Instabilität: Fehler in Δx_0 sei δ_0. Dann ergibt sich der Fehler von Δx_m zu

$$R_{m-1} \cdots R_1 R_0 \delta_0 \overset{\text{„=“ auf exakter}}{\overset{\downarrow}{\approx}} R(b,a)\delta_0 \ .$$

$\|R(b,a)\|$ groß, falls Anfangswertproblem schlecht konditioniert.
Im Fall separierter Randbedingungen

$$r_1\big(y(a)\big) = 0 \in \mathbb{R}^l \ , \quad r_2\big(y(b)\big) = 0 \in \mathbb{R}^{d-l}$$

setze

$$A = \begin{pmatrix} A_1 \\ 0 \end{pmatrix} \ , \qquad B = \begin{pmatrix} 0 \\ B_2 \end{pmatrix} \begin{matrix} \} \ l \\ \} \ d-l \end{matrix}$$

Umordnen des Gleichungssystems, A_1 nach oben:

$$\begin{bmatrix} A_1 & & & & 0 \\ R_0 & -I & & & \\ & R_1 & -I & & \\ & & \ddots & \ddots & \\ & & & R_{m-1} & -I \\ 0 & & & & B_2 \end{bmatrix} .$$

Dies Bandmatrix der Breite $2d + l$. Darauf kann das Gauß-Eliminations-Verfahren mit Spaltenpivotsuche angewendet werden → stabiler! Aufwand: $\mathcal{O}(md^3)$.
Fasse zusammen: Newton-Iteration bei Mehrzielmethode.

(a) Berechnung von F_j und R_j $(j = 0, 1, \ldots, m-1)$:

Löse das Anfangswertproblem

$$\begin{cases} y' = f(t,y) \quad \text{auf } [t_j, t_{j+1}] \ , \\ y(t_j) = x_j \ , \end{cases} \quad \rightarrow \quad F_j = y(t_{j+1}|t_j, x_j) - x_j \ ,$$

und die zugehörige linearisierte Differentialgleichung

$$\begin{aligned} v_i' &= \frac{\partial f}{\partial y}(t, y(t|t_j, x_j))v_i \ , \\ v_i(t_j) &= e_i \ , \quad i = 1, \ldots, d \ , \\ \rightarrow \quad R_j &= (v_1, \ldots, v_d)(t_{j+1}) = \big(v_1(t_{j+1}), \ldots, v_d(t_{j+1})\big) \ , \end{aligned}$$

wobei e_i den i-ten normierten Basisvektor bezeichnet.

Löse dies als ein Anfangswertproblem der Dimension $d + d^2$.

Der Aufwand ist ungefähr wie beim einfachen Schießverfahren, im Allgemeinen entsteht der Hauptaufwand in der Newton-Iteration.

(b) Löse lineares Gleichungssystem in $\mathcal{O}(md^3)$ Operationen, statt in $\mathcal{O}\big((md)^3\big)$ Operationen, wenn man die Struktur nicht berücksichtigt.

§ 5 Kollokationsverfahren

Randwertproblem:

$$\begin{cases} y' = f(t,y) \quad \text{auf } [a,b] \ , \\ r\big(y(a), y(b)\big) = 0 \quad \text{unter Voraussetzung von § 2.} \end{cases}$$

Idee. (a) Suche Näherungslösung in endlichdimensionalem Teilraum von $C[a,b]$.

(b) Verlange, dass die Differentialgleichung in endlich vielen Punkten erfüllt ist.

Unterteile $a = t_0 < t_1 < \cdots < t_m = b$.
Knoten $c_1, \ldots, c_s \in [0,1]$ alle verschieden, $h_j = t_{j+1} - t_j$.

Kollokation: Suche $u : [a,b] \to \mathbb{R}^d$ stetig, sodass $u|_{[t_j, t_{j+1}]}$ Polynom vom Grad $\leq s$ mit

$$\begin{cases} u'(t) = f\big(t, u(t)\big) & \text{für } t = t_j + c_i h_j \ , \\ r\big(u(a), u(b)\big) = 0 & \forall i = 1, \ldots, s \ , \\ & \forall j = 0, \ldots, m-1 \ . \end{cases}$$

Kann aufgefasst werden als Mehrzielmethode, bei der in jedem Teilintervall das Anfangswertproblem $y' = f(t,y)$ auf $[t_j, t_{j+1}]$, $y(t_j) = x_j$ ersetzt wird durch

$$\begin{cases} u'(t) = f\big(t, u(t)\big) & \text{für } t = t_j + c_i h_j \ , \\ u(t_j) = x_j \ , & \\ u \text{ Polynom vom Grad} \le s. \end{cases} \tag{I.2}$$

Im Folgenden sei j fest: Schreibe t_0 statt t_j, h statt h_j, y_0 statt x_j.

Satz 1. *Für genügend kleine Schrittweiten h hat das Kollokationsverfahren* (I.2) *eine eindeutige Lösung u. Der Endwert $u(t_0+h)$ ist gleich dem Ergebnis des impliziten Runge-Kutta-Verfahrens mit Koeffizienten c_i wie gegeben,*

$$b_j = \int_0^1 l_j(x)\,dx \ , \quad a_{ij} = \int_0^{c_i} l_j(x)\,dx \ ,$$

wobei $l_j(x)$ das Lagrange-Polynom zu den (c_i), $i = 1, \dots, s$, ist:

$$l_j(c_i) = \begin{cases} 1 & i = j \ , \\ 0 & i \ne j \ , \end{cases} \quad \deg(l_j) = s-1 \ , \quad l_j(x) = \prod_{\substack{i=1 \\ i \ne j}}^{s} \frac{x - c_i}{c_j - c_i} \ .$$

Erinnerung. Runge-Kutta

$$\begin{cases} y_1 = y_0 + h \sum_{i=1}^{s} b_i Y_i' \ , & Y_i' = f(t_0 + c_i h, Y_i) \ , \\ Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} Y_j' \ , & (i = 1, \dots, s) \ . \end{cases}$$

Setze $Y_i = u(t_0 + c_i h)$, $Y_i' = u'(t_0 + c_i h)$.

Beweis. Nehme zunächst Existenz und Eindeutigkeit des Kollokationspolynoms u an.
Zeige: $u(t_0 + h) = y_1$ aus Runge-Kutta-Verfahren.
Da u' Polynom vom Grad $\le s-1$ ist, gilt

$$u'(t_0 + xh) = \sum_{j=1}^{s} l_j(x) Y_j' \quad \text{(Lagrange Interpolation).}$$

Habe

$$\begin{aligned} Y_i = u(t_0 + c_i h) &= u(t_0) + \int_{t_0}^{t_0 + c_i h} u'(t)\,dt = u(t_0) + h \int_0^{c_i} u'(t_0 + xh)\,dx \\ &= y_0 + h \sum_{j=1}^{s} \int_0^{c_i} l_j(x)\,dx \cdot Y_j' = y_0 + h \sum_{j=1}^{s} a_{ij} Y_j' \ , \end{aligned}$$

ebenso

$$u(t_0 + h) = y_0 + h \sum_{j=1}^{s} \int_0^1 l_j(x)\,dx \cdot Y_j' = y_0 + h \sum_{j=1}^{s} b_j Y_j' = y_1 \ .$$

Umgekehrt: Gehe vom Runge-Kutta-Verfahren aus. Dieses hat für genügend kleine h eine eindeutige Lösung (Satz über implizite Funktionen).
Definiere $u(t)$ als Interpolationspolynom durch (t_0, y_0) und $(t_0 + c_i h, Y_i)$ für $i = 1, \dots, s$. Das so erhaltene u ist dann Kollokationspolynom. □

Satz 2. *Für den Fehler des Kollokationspolynoms $u(t)$ gilt*

$$u(t) = y(t) + \mathcal{O}\left(h^{s+1}\right) \quad \textit{gleichmäßig für } t \in [t_0, t_0 + h]$$

und für die Ableitungen $(k = 1, \dots, s)$

$$u^{(k)}(t) = y^{(k)}(t) + \mathcal{O}\left(h^{s+1-k}\right) \quad \textit{gleichmäßig für } t \in [t_0, t_0 + h] \ .$$

Beweis. Verwende (wieder mit Lagrange-Interpolation)

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \sum_{j=1}^{s} \int_0^{c_i} l_j(x)\,dx \cdot c_j^{k-1} = \int_0^{c_i} \underbrace{\sum_{j=1}^{s} l_j(x) c_j^{k-1}}_{=x^{k-1}}\,dx = \frac{x_k}{k}\Big|_0^{c_i} = \frac{c_i^k}{k}$$

bei der Taylor-Entwicklung von $y(t_0 + c_i h)$ für kleine h, wobei $y(t_0) = y_0$:

$$\begin{aligned}
y(t_0 + c_i h) &= y_0 + h c_i y'(t_0) + \frac{h^2}{2} c_i^2 y''(t_0) + \dots + \frac{c_i^s}{s!} h^s y^{(s)}(t_0) + \mathcal{O}(h^{s+1}) \\
&= y_0 + h \underbrace{\sum_{j=1}^{s} a_{ij}}_{=c_i}\, y'(t_0) + h^2 \underbrace{\sum_{j=1}^{s} a_{ij} c_j}_{=c_i^2/2}\, y''(t_0) + \dots \\
&\qquad + \frac{h^s}{(s-1)!} \underbrace{\sum_{j=1}^{s} a_{ij} c_j^{s-1}}_{=c_i^s/s}\, y^{(s)}(t_0) + \mathcal{O}(h^{s+1}) \\
&= y_0 + h \sum_{j=1}^{s} a_{ij} \Bigg(\underbrace{y'(t_0) + h c_j y''(t_0) + \dots + \frac{h^{(s-1)}}{(s-1)!} c_j^{s-1} y^{(s)}(t_0)}_{=y'(t_0 + c_j h) + \mathcal{O}(h^s)} + \mathcal{O}(h^s) \Bigg) \\
&= y_0 + h \sum_{j=1}^{s} a_{ij} y'(t_0 + c_j h) + \mathcal{O}(h^{s+1}) \ .
\end{aligned}$$

Andererseits habe auch

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} Y_j' \ ,$$

also:

$$y(t_0+c_ih)-Y_i = h\sum_{j=1}^{s} a_{ij}(y'(t_0+c_jh)-Y_j') + \mathcal{O}(h^{s+1})$$

$$\overset{\substack{\text{Differentialgleichung,}\\ \text{Runge-Kutta für } Y_j'}}{\downarrow}{=} \quad h\sum_{j=1}^{s} a_{ij}\overbrace{\Big(f\big(t_0+c_jh,y(t_0+c_jh)\big)-f(t_0+c_jh,Y_j)\Big)}^{\|\%\|\le L\|y(t_0+c_jh)-Y_j\|} + \mathcal{O}(h^{s+1})\ ,$$

wobei L die Lipschitz-Konstante von f ist. Bringe nun die Summe auf die linke Seite. Dann folgt schließlich für hL genügend klein $Y_i - y(t_0+c_ih) = \mathcal{O}(h^{s+1})$ und

$$Y_i' - y'(t_0+c_ih) = f(t_0+c_ih,Y_i) - f\big(t_0+c_ih,y(t_0+c_ih)\big) = \mathcal{O}(h^{s+1})\ .$$

Weiterhin

$$u'(t_0+xh)-y'(t_0+xh) = \sum_{i=1}^{s} l_i(x)\underbrace{\big(Y_i'-y'(t_0+c_ih)\big)}_{\mathcal{O}(h^{s+1})} + \underbrace{\sum_{i=1}^{s} l_i(x)y'(t_0+c_ih) - y'(t_0+xh)}_{\substack{\text{Interpolationsfehler } \mathcal{O}(h^s) \text{ gleichmäßig für } x\in[0,1],\\ \text{sogar mit glatter Funktion } E \text{ (ohne Beweis) } = h^sE(x,h).}}$$

Damit

$$u(t)-y(t) = \int_{t_0}^{t}\underbrace{\big(u'(\tau)-y'(\tau)\big)}_{\mathcal{O}(h^s)}\,d\tau = \mathcal{O}\left(h^{s+1}\right)\ , \qquad \text{gleichmäßig für } t\in[t_0,t_0+h]$$

und durch $(k-1)$-maliges Differenzieren nach x

$$h^{k-1}\cdot\Big(u^{(k)}(t_0+xh)-y^{(k)}(t_0+xh)\Big) = \mathcal{O}(h^s)\ . \qquad \square$$

Bei Verwendung in der Mehrzielmethode interessiert der Fehler am Endpunkt t_0+h (bzw. t_{j+1}, ausgehend von t_j).

Satz 3. *Die Quadraturformel zu den Knoten (c_i), $i=1,\ldots,s$,*

$$\int_0^1 g(x)\,dx \approx \sum_{i=1}^{s} b_ig(c_i)$$

habe die Ordnung p (d. h. exakt für alle Polynome vom Grad $\le p-1$).
Dann hat das Runge-Kutta-Verfahren von Satz 1 die Ordnung p, d. h.

$$u(t_0+h)-y(t_0+h) = \mathcal{O}(h^{p+1})\ .$$

Kurz: Das Kollokationsverfahren hat dieselbe Ordnung wie die zugrunde liegende Quadraturformel.

Günstige Wahl der Knoten c_i: Die Gaußsche Quadraturformel hat die Ordnung $p = 2s$ (verschobene Legendrepolynome).

Beweis. Anfangswertproblem $y'(t) = f\big(t, y(t)\big)$, $y(t_0) = y_0$.
Habe für Kollokationsverfahren $u'(t) = f\big(t, u(t)\big) + d(t)$, $u(t_0) = y_0$ mit $d(t_0 + c_i h) = 0$ für $i = 1, \dots, s$ (Kollokationsbedingung).
Betrachte Homotopie

$$\begin{cases} z' = f(t,z) + \tau d(t)\ , \quad 0 \le \tau \le 1 \text{ Parameter}, \\ \text{Anfangswert } y_0 = z(t_0, \tau)\ . \end{cases} \tag{I.3}$$

Lösung bezeichnet mit $z(t,\tau)$, habe $z(t,0) = y(t)$, $z(t,1) = u(t)$:

$$u(t) - y(t) = z(t,1) - z(t,0) = \int_0^1 \frac{\partial z}{\partial \tau}(t,\tau)\, d\tau\ .$$

Zur Berechnung von $\frac{\partial z}{\partial \tau}$ differenziere (I.3) nach τ:

$$\left(\frac{\partial z}{\partial \tau}\right)' = \underbrace{\frac{\partial f}{\partial y}\big(t, z(t,\tau)\big)}_{=:C_\tau(t)} \cdot \frac{\partial z}{\partial \tau} + d(t)\ , \qquad \text{lineare Differentialgleichung}$$

$$\frac{\partial z}{\partial \tau}(t_0,\tau) = 0\ .$$

Sei $R_\tau(t,s)$ Resolvente hierzu, daraus (siehe Übungsaufgabe 1 b):

$$\frac{\partial z}{\partial \tau}(t,\tau) = \int_{t_0}^t R_\tau(t,s) d(s)\, ds\ .$$

Setze oben ein, vertausche Integrale:

$$\begin{aligned} u(t_0+h) - y(t_0+h) &= \int_{t_0}^{t_0+h} \underbrace{\int_0^1 R_\tau(t_0+h, s)\, d\tau\, d(s)}_{g(s)}\, ds \\ &= h \sum_{i=1}^s b_i \underbrace{g(t_0 + c_i h)}_{=0} + h^{p+1} E(g,h) \end{aligned}$$

mit $\|E(g,h)\| \le C \max_{[t_0, t_0+h]} \|g^{(p)}(t)\|$ wegen $d(t_0 + c_i h) = 0$.
Alle Ableitungen von u sind unabhängig von h beschränkt. Dann sind alle Ableitungen von d unabhängig von h beschränkt (Satz 2). Dann sind auch alle Ableitungen von z, R_τ unabhängig von h beschränkt. Schließlich sind dann alle Ableitungen von g unabhängig von h beschränkt.
Somit: $u(t_0 + h) - y(t_0 + h) = \mathcal{O}(h^{p+1})$. □

Möchte Konvergenz des Kollokationsverfahrens für Randwertprobleme bei feiner werdenden Unterteilungen $a = t_0 < \cdots < t_m = b$:

$$h = \max_j |t_{j+1} - t_j| \to 0 \quad (\text{dann auch } m \to \infty)\ .$$

Brauche dazu Erinnerung:

§ 6 Einschub: Konvergenz des vereinfachten Newton-Verfahrens

Nichtlineares Gleichungssystem $F(x) = 0$, $F : D \subset \mathbb{R}^n \to \mathbb{R}^n$
Vereinfachtes Newton-Verfahren: x^0 gegeben, $J \approx F'(x^0)$

$$x^{k+1} = x^k + \Delta x^k \quad \text{mit } J \cdot \Delta x^k = -F(x^k) \ (k = 0, 1, 2, \dots) \ .$$

Satz 1. *Sei $D \subset \mathbb{R}^n$ offen und $F : D \to \mathbb{R}^n$ stetig differenzierbar. Sei $x^0 \in D$ und gelte*

(a) $\|\Delta x^0\| \le \alpha$ *(Definition von α),*

(b) $\|I - J^{-1}F'(x)\| \le \gamma < 1 \ \forall x \in D$ *(verträgliche Matrixnorm zu $\|\cdot\|$ auf $\mathbb{R}^n$).*

(c) Für $\rho = \alpha/(1-\gamma)$ sei $\overline{B}(x^0, \rho) \subset D$, wobei $\overline{B}(x^0, \rho) := \{x \in \mathbb{R}^n \mid \|x - x^0\| \le \rho\}$, eine abgeschlossene Kugel vom Radius ρ um x^0.

Dann bleibt die Folge (x^k) des vereinfachten Newton-Verfahrens in $\overline{B}(x^0, \rho)$ und konvergiert gegen eine Nullstelle x^* von F. Diese ist die einzige Nullstelle in $\overline{B}(x^0, \rho)$.
Es gilt

$$\|x^{k+1} - x^k\| \le \gamma \cdot \|x^k - x^{k-1}\| \quad \textit{und} \quad \|x^{k+1} - x^*\| \le \frac{\gamma}{1-\gamma} \cdot \|x^{k+1} - x^k\|$$

Beweis. Banachscher Fixpunktsatz angewandt auf (siehe Numerik I)

$$x^{k+1} = \phi(x^k) \quad \text{mit} \quad \phi(x) = x - J^{-1}F(x) \ . \qquad \square$$

§ 7 Konvergenz von Kollokationsverfahren für Randwertprobleme

$$\begin{cases} y' = f(t, y) & \text{auf } [a, b], \\ r\big(y(a), y(b)\big) = 0 \ , & \end{cases}$$

Randwertproblem unter Voraussetzungen von § 2, f, r genügend oft differenzierbar.

Knoten $0 \le c_1 < \dots < c_s \le 1$.
Unterteilung $a = t_0 < t_1 < \dots < t_m = b$.

Kollokation: Suche Näherungslösung u mit

$$\begin{cases} u|_{[t_j, t_{j+1}]} & \text{Polynom vom Grad} \le s & \forall j = 0, \dots, m-1, \\ u'(t) = f\big(t, u(t)\big) & \text{für } t = t_j + c_i(t_{j+1} - t_j) & \forall j \text{ und } \forall i = 1, \dots, s, \\ r\big(u(a), u(b)\big) = 0 \ . & & \end{cases}$$

Fehler $u(t) - y(t) = ?$

Bezeichnung: $h := \max_j |t_{j+1} - t_j|$, die maximale Gitterweite ($\to 0$).
Falls äquidistant: $mh = b - a$, betrachte allgemeine Gitter mit $mh \leq \text{const}$.

Satz 1. *Voraussetzungen wie oben. Die Quadraturformel zu den Knoten c_i habe Ordnung $p \geq 1$. Dann existiert eine lokal eindeutige Kollokationslösung u mit*

$$\|u(t) - y(t)\| \leq C \cdot h^r \quad \textit{mit } r = \min(s+1, p) \textit{ gleichmäßig für } t \in [a,b],$$
$$\|u(t_j) - y(t_j)\| \leq C \cdot h^p \quad \forall j = 0, \ldots, m \ , \quad \textit{„Superkonvergenz“ in den Gitterpunkten.}$$

C unabhängig von Unterteilung mit $mh \leq \text{const}$.

Beweis. Das Kollokationsverfahren ist eine Mehrzielmethode, bei der in jedem Teilintervall die exakte Lösung des Anfangswertproblems durch einen Schritt des Kollokationsverfahrens (Runge-Kutta-Verfahrens) ersetzt wird.
Bezeichne $u(t|t_j, x_j)$ die Kollokationslösung des Anfangswertproblems.

$$\begin{cases} u \text{ Polynom vom Grad} \leq s \ , \\ u'(t) = f\big(t, u(t)\big) \quad \text{für } t = t_j + c_i(t_{j+1} - t_j) \ , \\ u(t_j) = x_j \ . \end{cases}$$

Vergleiche mit der exakten Lösung des Anfangswertproblems $y(t|t_j, x_j)$.

Gleichungen:

Stetigkeitsbedingung ($j = 0, \ldots, m-1$):

$$\tilde{F}_j(x_j, x_{j+1}) := u(t_{j+1}|t_j, x_j) - x_{j+1} = 0 \ .$$

Vergleiche Mehrzielmethode $F_j(x_j, x_{j+1}) = y(t_{j+1}|t_j, x_j) - x_{j+1} = 0$.

Randbedingung:

$$\tilde{F}_m(x_0, x_m) = r(x_0, x_m) = 0 \ .$$

Zu lösen:

$$\tilde{F}(x) = 0 \quad \text{mit} \quad \tilde{F} = (\tilde{F}_0, \ldots, \tilde{F}_m)^T \ , \quad x = (x_1, \ldots, x_m)^T \ .$$

Wende darauf das vereinfachte Newton-Verfahren mit Startvektor x^0 auf der *exakten* Lösung y des Randwertproblems an (nur zur theoretische Untersuchung, nicht für die praktische Berechnung!).
Startvektor:

$$x^0 = (x_0^0, \ldots, x_m^0)^T = \big(y(t_0), \ldots, y(t_m)\big)^T \ .$$

1. Iteration

$$x^1 = x^0 + \Delta x^0 \quad \text{mit} \quad J \Delta x^0 = -\tilde{F}(x^0) \ .$$

Wähle

$$J = F'(x^0) \quad \text{für} \quad F = (F_0, \ldots, F_m)$$

von § 4, Mehrzielmethode (*nicht* $\tilde{F}'(x^0)$).

Wissen aus § 4: J invertierbar. Zeige unten im Hilfssatz 2:

$$\|J^{-1}\|_\infty = \mathcal{O}(m) \ ,$$

$\|\cdot\|_\infty$ ist die von der Maximum-Norm induzierte Matrixnorm.
Die rechte Seite in der Gleichung für Δx^0 lautet $-\tilde{F}(x^0)$.

$$\tilde{F}_j(\underbrace{x_j^0}_{y(t_j)}, \underbrace{x_{j+1}^0}_{y(t_{j+1})}) = u\big(t_{j+1}|t_j, y(t_j)\big) - \underbrace{y(t_{j+1})}_{y\left(t_{j+1}|t_j, y(t_j)\right)} \overset{\text{Satz 3, § 5}}{\downarrow} \mathcal{O}(h^{p+1}) \ , \qquad j = 0, \ldots, m-1,$$

$$\tilde{F}_m(x_0^0, x_m^0) = r\big(y(a), y(b)\big) = 0 \ .$$

Damit:

$$\|\Delta x^0\|_\infty = \|J^{-1}\tilde{F}(x^0)\|_\infty \le \|J^{-1}\|_\infty \cdot \|\tilde{F}(x^0)\|_\infty \le \mathcal{O}(m) \cdot \mathcal{O}(h^{p+1}) \overset{mh \le \text{const}}{\downarrow} \le \mathcal{O}(h^p) \ .$$

Untersuche nun $\|I - J^{-1}\tilde{F}'(x)\|$. Durch Taylorentwicklung folgt

$$\tilde{F}'(x) = \tilde{F}'(x^0) + \mathcal{O}(\|x - x^0\|) \quad \text{für } \|x - x^0\| \le C_0 h^p \text{ und } C_0 \text{ genügend groß.}$$

Habe

$$\tilde{F}'(x^0) = \begin{bmatrix} \tilde{R}_1 & -I & & & \\ & \tilde{R}_2 & -I & 0 & \\ & & \ddots & \ddots & \\ & 0 & & \tilde{R}_{m-1} & -I \\ A & & & & B \end{bmatrix}$$

mit $A = \frac{\partial r}{\partial y_a}\big(y(a), y(b)\big)$, $B = \frac{\partial r}{\partial y_b}\big(y(a), y(b)\big)$ und

$$\tilde{R}_j := \frac{\partial u}{\partial x_j}\big(t_{j+1}|t_j, y(t_j)\big) = ?$$

Erinnerung. $u(t_{j+1}) = y_1$ aus Runge-Kutta-Verfahren der Ordnung p. Schreibe hier t_0 statt t_j:

$$y_1 = \underbrace{y_0}_{\text{hier: } y(t_j)} + h \sum_{i=1}^{s} b_i \underbrace{f(t_0 + c_i h, Y_i)}_{Y_i'} \ ,$$

$$Y_i = y_0 + h \sum_{j=1}^{s} a_{ij} \underbrace{f(t_0 + c_j h, Y_j)}_{Y_j'} \ , \qquad i = 1, \ldots, s.$$

Fasse y_1, Y_i als Funktionen des Anfangswerts y_0 auf, leite nach y_0 ab:

$$\frac{\partial y_1}{\partial y_0} = I + h \sum_{i=1}^{s} b_i \frac{\partial f}{\partial y}(t_0 + c_i h, Y_i) \frac{\partial Y_i}{\partial y_0} \ ,$$

$$\frac{\partial Y_i}{\partial y_0} = I + h \sum_{j=1}^{s} a_{ij} \frac{\partial f}{\partial y}(t_0 + c_j h, Y_j) \frac{\partial Y_j}{\partial y_0} \ , \qquad i = 1, \ldots, s.$$

Habe: $\tilde{R}_j = \partial y_1/\partial y_0$. Dies ist äquivalent zur Anwendung des Runge-Kutta-Verfahrens auf das Anfangswertproblem

$$\begin{cases} y' = f(t,y)\ , \quad \text{Anfangswert } y(t_j)\ , \\ R' = \dfrac{\partial f}{\partial y}(t,y)\cdot R\ , \quad R(t_j) = I \end{cases}$$

mit exakter Lösung $y(t)$, $R(t) = R(t,t_j)$ Resolvente. Runge-Kutta-Verfahren hat Ordnung p, daher $\tilde{R}_j = R_j + \mathcal{O}(h^{p+1})$ mit $R_j = R(t_{j+1},t_j)$ wie bei Mehrzielmethode.
Mit

$$\tilde{F}'(x^0) - \underbrace{F'(x^0)}_{=J} = \begin{bmatrix} \mathcal{O}(h^{p+1}) & & & \\ & \mathcal{O}(h^{p+1}) & & 0 \\ & & \ddots & \\ & 0 & & \mathcal{O}(h^{p+1}) \\ & & & & 0 \end{bmatrix} =: M$$

und wieder mit Hilfssatz 2 ($\|J^{-1}\|_\infty = \mathcal{O}(m)$) sowie $mh \le \text{const}$ habe

$$\begin{aligned} I - J^{-1}\tilde{F}'(x) &= I - J^{-1}\Bigg(\tilde{F}'(x)\overbrace{-\tilde{F}'(x^0) + \underbrace{\tilde{F}'(x^0) - F'(x^0)}_{=M} + \underbrace{F'(x^0)}_{=J}}^{=0}\Bigg) \\ &= I - J^{-1}\big(\underbrace{\tilde{F}'(x) - \tilde{F}'(x^0)}_{\mathcal{O}(\|x-x^0\|)}\big) - \underbrace{J^{-1}M}_{\mathcal{O}(m)\cdot\mathcal{O}(h^{p+1})=\mathcal{O}(h^p)} - I \end{aligned}$$

$$\Rightarrow \quad \|I - J^{-1}\tilde{F}'(x)\| = \mathcal{O}(\|x-x^0\|) + \mathcal{O}(h^p)\ .$$

Damit ist Satz 1 aus § 6 über das vereinfachte Newton-Verfahren anwendbar. Erhalte mit $\alpha,\gamma,\rho = \mathcal{O}(h^p)$, $D = \overline{B}(x^0,\rho)$:
Es existiert genau eine Lösung x^* von $\tilde{F}(x) = 0$ mit $\|x^* - x^0\| \le \rho = \mathcal{O}(h^p)$.
Wegen $x_j^* = u(t_j)$ und $x_j^0 = y(t_j)$ folgt $\|u(t_j) - y(t_j)\| = \mathcal{O}(h^p)$ und weiterhin folgt für $t \in (t_j, t_{j+1})$:

$$\begin{aligned} \|u(t) - y(t)\| &= \|u\big(t|t_j,u(t_j)\big) - y\big(t|t_j,y(t_j)\big)\| \\ &\le \|u\big(t|t_j,u(t_j)\big) - u\big(t|t_j,y(t_j)\big)\| + \|u\big(t|t_j,y(t_j)\big) - y(t|t_j,y(t_j))\| \\ &\le \text{const}\cdot\|u(t_j) - y(t_j)\| + \underbrace{\mathcal{O}(h^{s+1})}_{\text{Satz 2, § 5}} \\ &\le \mathcal{O}(h^p) + \mathcal{O}(h^{s+1}) = \mathcal{O}(h^r) \quad \text{mit} \quad r = \min(s+1,p)\ . \end{aligned}$$ □

Für vollständigen Beweis brauche noch den Hilfssatz, dass $\|J^{-1}\|_\infty = \mathcal{O}(m)$ gilt. Dieser folgt nun noch.
Sei $E(t) = A\cdot R(a,t) + B\cdot R(b,t)$ die invertierbare Sensitivitätsmatrix (siehe § 2) und

$$G(t,s) = \begin{cases} E(t)^{-1}AR(a,s)\ , & s < t\ , \\ -E(t)^{-1}BR(b,s)\ , & s \ge t \end{cases}$$

die Greensche Funktion (siehe Übungsaufgabe 2) des entlang der Lösung $y(t)$ linearisierten Randwertproblems:

$$\begin{cases} v' = \dfrac{\partial f}{\partial y}\big(t, y(t)\big)v \ , \\ Av(a) + Bv(b) = r \ . \end{cases}$$

Hilfssatz 2.

$$J^{-1} = \left[\begin{array}{ccc|c} G_{00} & \cdots & G_{0,m-1} & -E(t_0)^{-1} \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \\ G_{m0} & \cdots & G_{m,m-1} & -E(t_m)^{-1} \end{array}\right]$$

mit $G_{jk} = G(t_j, t_{k+1})$.
Die Einträge von J^{-1} sind somit unabhängig von der Unterteilung des Intervalls beschränkt, damit

$$\|J^{-1}\|_\infty = \max_j \sum_{k=0}^{m} \big|(J^{-1})_{jk}\big| = \mathcal{O}(m) \ .$$

Beweis. Lösung von $J \cdot \Delta x = -F$ ist (Übungsaufgabe 6, folgt aus § 4)

$$\Delta x_j = \sum_{l=0}^{m-1} G_{jl} F_l - E_j^{-1} F_m \ , \qquad\qquad j = 0, \ldots, m,$$

mit

$$\begin{aligned} E_j &= AR_0^{-1} \cdots R_{j-1}^{-1} + BR_{m-1} \cdots R_j \ , \\ G_{jl} &= \begin{cases} E_j^{-1} AR_0^{-1} \cdots R_l^{-1} \ , & l < j \ , \\ -E_j^{-1} BR_{m-1} \cdots R_{l+1} \ , & l \geq j \ , \end{cases} \end{aligned}$$

wobei $R_0^{-1} \cdots R_{j-1}^{-1} := I$ für $j = 1$ und $R_{m-1} \cdots R_j := I$ für $j = m$ gelten soll.
Da $J = F'(x^0)$ mit x^0 auf exakter Lösung $y(t)$, ist hier

$$\begin{aligned} R_{m-1} \cdots R_j &= R(t_m, t_{m-1}) \cdots R(t_{j+1}, t_j) = R(\underbrace{t_m}_{=b}, t_j) = R(b, t_j) \ , \\ R_0^{-1} \cdots R_{j-1}^{-1} &= R(t_1, t_0)^{-1} \cdots R(t_j, t_{j-1})^{-1} = R(t_0, t_1) \cdots R(t_{j-1}, t_j) \\ &= R(\underbrace{t_0}_{=a}, t_j) = R(a, t_j) \ , \end{aligned}$$

damit $E_j = E(t_j)$ und

$$G_{jl} = \left\{ \begin{array}{ll} E(t_j)^{-1} AR(a, t_{l+1}) \ , & l < j \ , \\ -E(t_j)^{-1} BR(b, t_{l+1}) \ , & l \geq j \ , \end{array} \right\} = G(t_j, t_{l+1}) \ . \qquad \square$$

Bemerkungen (zusammenfassend).

- Einfachschießverfahren: $m = 1$.

  Viele Integrationsschritte (z. B. Runge-Kutta) auf $[t_0, t_1] = [a, b]$.

  Sehr empfindlich gegenüber der Wahl der Startwerte.

- Mehrzielverfahren: Typischerweise wird m eher klein gehalten und mithilfe ein Wachstumskriterium an den Lösungsverlauf angepasst.

  Mehrere ($\sim$ viele) Integrationsschritte je Teilintervall.

- Kollokationsverfahren: m groß und meist durch Schrittweitensteuerung, wie beim Runge-Kutta-Verfahren für Anfangswertprobleme, festgelegt.

  Ein Integrationsschritt je Teilintervall.

Programme z. B in `netlib`, `elib`, `nag`.
Bei allen Verfahren sind „gute“ Startwerte für die Newton-Iteration wichtig.
Oft durch Homotopie bezüglich Parametern in den Gleichungen.

Beispiel (Kapillare).

$$\theta_0 = \pi/2 \ , \quad y(r) \equiv 0$$

Nehme dies als Startwert für $\theta_1 = 1.4$.
Löse Randwertproblem für

$$\theta_1 > \theta_2 > \cdots > \theta \ ,$$

wobei θ dem gewünschten θ entspricht.

Betrachte noch wichtige Problemklassen, die auf Randwertprobleme (in Standardform oder anders) führen:

§ 8 Variationsprobleme

Wohlbekannt: Minimierung einer Funktion von einer oder mehreren reellen Veränderlichen $x_1, \ldots, x_n$.
Suche $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, sodass $F(x) = \min!$ ($F : \mathbb{R}^n \to \mathbb{R} \in C^1$).
Notwendige Bedingung: $F'(x) = 0$, d. h.

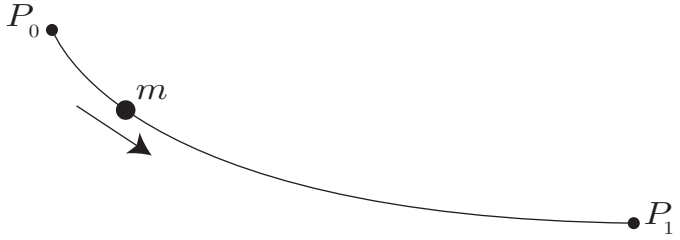
$$\frac{\partial F}{\partial x_i}(x_1, \ldots, x_n) = 0 \ .$$

Jetzt: Minimiere Funktion von Kurven („Funktional“).
Suche $y : [a, b] \to \mathbb{R}^d$ mit vorgegebenen Endpunkten $y(a) = y_a$ und $y(b) = y_b$, sodass

$$\Phi(t) = \int_a^b f\big(t, y(t), y'(t)\big)\, dt \to \min! \tag{I.4}$$

Φ: Teilmenge von $C^1[a, b] \to \mathbb{R}$, $f : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

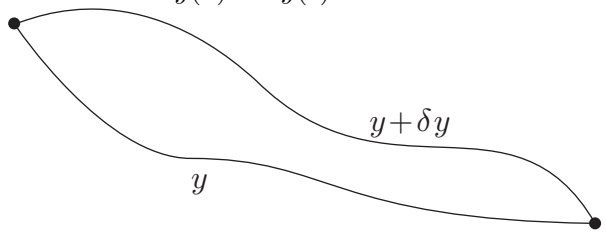
Beispiel (Johannes Bernoulli 1696, Brachystochronen-Problem)**.** Suche die Kurve von P_0 nach P_1, sodass der Massenpunkt, der infolge der Schwerkraft entlang der Kurve gleitet, in kürzester Zeit den Endpunkt P_1 erreicht.



Beispiel. Suche die kürzeste Kurve im $\mathbb{R}^3$ (schwieriger: auf Fläche im $\mathbb{R}^3$), die zwei Punkte verbindet:

$$\min \int_a^b \sqrt{y_1'(t)^2 + y_2'(t)^2 + y_3'(t)^2}\, dt\ , \qquad \text{(Gerade).}$$

Suche eine Bedingung für die Minimumkurve analog $F'(x) = 0$. Sei y Lösung von (I.4). Betrachte nun mit Lagrange (1755) die „Variation“ von y, d. h. die Kurve $y + \delta y$ mit denselben Endpunkten und mit $\delta y(a) = \delta y(b) = 0$.



Betrachte die eindimensionale Minimierung von $F(\epsilon) = \Phi(y + \epsilon\delta y)$, $\epsilon \in \mathbb{R}$ (nahe 0). Da y die Lösung von (I.4) ist, liegt das Minimum bei $\epsilon = 0$, also muss $F'(0) = 0$ sein, d. h.

$$\begin{aligned} 0 &= \frac{d}{d\epsilon}\Big|_{\epsilon=0} \int_a^b f\big(t, y(t) + \epsilon\delta y(t), y'(t) + \epsilon\delta y'(t)\big)\, dt \\ &= \int_a^b \Big\{ \underbrace{\frac{\partial f}{\partial y}\big(t, y(t), y'(t)\big) \cdot \delta y(t)}_{\text{wird nicht partiell integriert}} + \underbrace{\frac{\partial f}{\partial y'}\big(t, y(t), y'(t)\big)}_{\downarrow} \cdot \underbrace{\delta y'(t)}_{\uparrow} \Big\}\, dt \\ &\overset{\substack{\text{partielle}\\ \text{Integration}\\ \downarrow}}{=} \int_a^b \left\{ \frac{\partial f}{\partial y}(t, y, y') - \frac{d}{dt}\frac{\partial f}{\partial y'}(t, y, y') \right\} \delta y(t)\, dt + \underbrace{\frac{\partial f}{\partial y'}(t, y, y')\delta y\Big|_a^b}_{=0} \end{aligned}$$

weil die Endpunkte $y(a)$ und $y(b)$ fest sind: $\delta y(a) = \delta y(b) = 0$. Dies gilt für „beliebige“ Variationen δy.
Vermute, dass $\{\%\} = 0$. Tatsächlich:

Satz 1. *Sei $y \in C^2[a,b]$ Minimum des Variationsproblems* (I.4) *mit zweimal stetig differenzierbaren $f : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.*
Damit gilt

$$\frac{\partial f}{\partial y}\big(t, y(t), y'(t)\big) - \frac{d}{dt}\frac{\partial f}{\partial y'}\big(t, y(t), y'(t)\big) = 0 \quad \forall t \in [a,b] \ ,$$

bezeichnet als die **Euler-Lagrange-Differentialgleichung**. *Sie ist notwendige Bedingung für ein Extremum.*
Kurz:

$$f_y - \frac{d}{dt} f_{y'} = 0 \ .$$

Bemerkung. Die Kettenregel liefert

$$\begin{cases} f_y - f_{y't} - f_{y'y}y' - f_{y'y'}y'' = 0 \ , & \text{Differentialgleichung 2. Ordnung,} \\ y(a) \ , \ y(b) & \text{fest vorgegebene Randbedingungen.} \end{cases}$$

Falls $f_{y'y'}$ invertierbar, erhalte Randwertproblem in *Standardform.*
Numerisch ist es oft günstiger, nicht auszudifferenzieren!

$$\left\{ \begin{array}{l} \left.\begin{array}{l} y' = z \ , \\ p' = f_y(t,y,z) \ , \end{array}\right\} \text{Differentialgleichung,} \\ \left.\begin{array}{l} 0 = p - f_{y'}(t,y,z) \ , \end{array}\right\} \begin{array}{l} \text{nichtlineare Gleichung,} \\ \text{„algebraische“ Gleichung,} \end{array} \end{array} \right\} \begin{array}{l} \text{differentiell-} \\ \text{algebraisches} \\ \text{System,} \end{array}$$

$$y(a) = y_a \ , \quad y(b) = y_b \ .$$

Beweis (des Satzes). Die Voraussetzungen an y und f garantieren, dass obiges Vertauschen von Differentiation und Integral zulässig ist.
Die Behauptung folgt dann mit $\delta y(t) = h(t)e_i$, $i = 1, \ldots, d$, e_i dem $i$ten normierten Basisvektor und $h(t) \in \mathbb{R}$, aus folgendem Hilfssatz: □

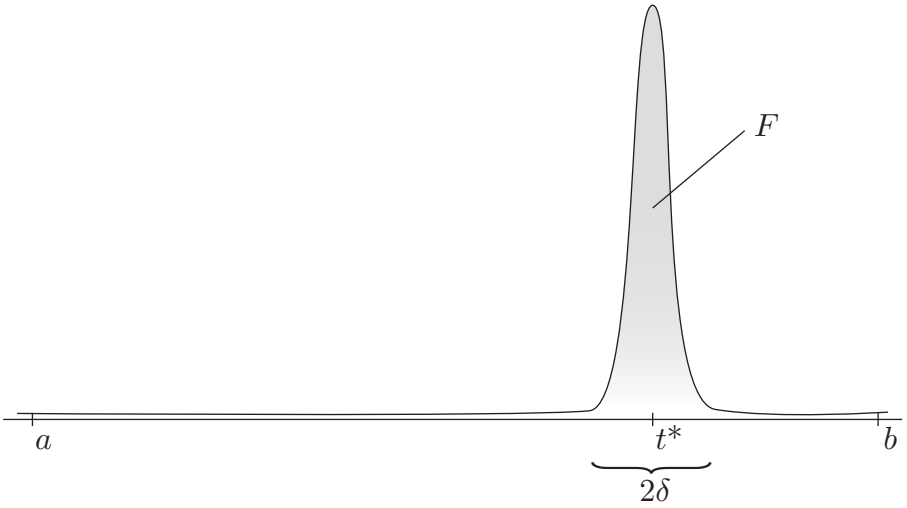
Hilfssatz 2 („Fundamentallemma der Variationsrechung“)**.** *Falls für eine stetige Funktion $g : [a,b] \to \mathbb{R}$ gilt, dass*

$$\int_a^b g(t)h(t)\,dt = 0$$

$\forall$ stetig differenzierbaren Funktionen $h : [a,b] \to \mathbb{R}$ mit $h(a) = h(b) = 0$, dann ist $g(t) = 0$ $\forall t \in [a,b]$.

Beweis. Sei $t^* \in (a,b)$ beliebig. Wähle nun $h = h_\delta \geq 0$ und $h_\delta = 0$ außerhalb $(t^* - \delta, t^* + \delta)$ (h sei im Wesentlichen um t^* konzentriert). Betrachte dazu die Fläche F, für die gilt:

$$F = \int h_\delta(t)\,dt > 0 \ .$$



Habe

$$0 \overset{\text{nach Voraussetzung}}{\downarrow =} \int_a^b g(t)h_\delta(t)\,dt = \int_{t^*-\delta}^{t^*+\delta} g(t)\underbrace{h_\delta(t)}_{\geq 0}\,dt$$

$$\overset{\text{Mittelwertsatz}}{\downarrow =} g(\tau)\underbrace{\int_{t^*-\delta}^{t^*+\delta} h_\delta(t)\,dt}_{>0}$$

für ein $\tau \in [t^*-\delta, t^*+\delta]$. Daraus folgt $g(\tau) = 0$.
Lasse nun $\delta \to 0$. Daraus folgt $\tau \to t^*$ und damit $g(\tau) \to g(t^*)$. Mit der Stetigkeit von g folgt $g(t^*) = 0\ \forall t^* \in (a,b)$. Also ist $g(t) = 0\ \forall t \in [a,b]$. □

§ 9 Erinnerung: Gauß-Newton-Verfahren

- Lineares Gleichungssystem $Ax = b$, $A \in \mathbb{R}^{n\times n}$ invertierbar, $b \in \mathbb{R}^n$.
- Nichtlineares Gleichungssystem $f(x) = 0$.

  Das Newton-Verfahren

  $$x^{k+1} = x^k + \Delta x^k \quad \text{mit } J^k \Delta x^k = -f(x^k)$$

  konvergiert lokal quadratisch: $e^{k+1} \leq C \cdot (e^k)^2$.
- Überbestimmtes lineares Gleichungssystem $Ax \approx b$:

  $$\|Ax - b\|_2 = \min! \ , \quad A \in \mathbb{R}^{m\times n} \ , \quad b \in \mathbb{R}^n \ , \quad m > n$$
  $$\Leftrightarrow \quad A^T Ax = A^T b \qquad \text{(Normalengleichung)}.$$

- Nichtlineares $f : \mathbb{R}^n \to \mathbb{R}^m$, $m > n$,

$$\|f(x)\|_2 = \min!$$

Sei x^* Lösung, x nahe x^*, schreibe $x^* = x + \Delta x$

$$\underset{\text{minimal}}{\|f(x^*)\|} = \|f(x+\Delta x)\| = \Big\| f(x) + \underbrace{J(x)}_{f'(x)} \Delta x + \underbrace{\mathcal{O}(\|\Delta x\|^2)}_{\text{vernachlässige}} \Big\| \ .$$

Löse *lineares* Ausgleichsproblem.

Gauß-Newton-Verfahren x^0 gegeben, iteriere für $k = 0, 1, 2, \ldots$

$$x^{k+1} = x^k + \Delta x^k \quad \text{mit } \big\| J(x^k)\Delta x^k + f(x^k) \big\|_2 = \min!$$

Dies ist eine Folge von linearen Ausgleichsproblemen.

Konvergenz?

Satz 1 (siehe Numerik I)**.** *Für den Fehler $e^k = x^k - x^*$ gilt*

$$e^{k+1} = -(\underbrace{J^T}_{n\times m}\underbrace{J}_{m\times n})^{-1} \underbrace{(J^T)'e^k}_{n\times m} \cdot \underbrace{f}_{m\times 1}\big|_{x=x^*} + \mathcal{O}(\|e^k\|^2) \ .$$

Der erste Term ist linear *in e^k:*

$$e^{k+1} = Be^k + \mathcal{O}(\|e^k\|^2) \ .$$

Das Gauß-Newton-Verfahren konvergiert lokal linear, falls $\|B\| < 1$ erfüllt. Dies ist erfüllt, falls $\|f''\| \cdot \|f\|$ klein.

Beweis. Siehe Numerik I. □

§ 10 Parameteridentifizierung bei Differentialgleichungen

Beispiel (aus der Chemie: Zerfall von Ozon)**.**

$$\begin{aligned} \mathrm{O_3} + \mathrm{O_2} &\underset{p_2}{\overset{p_1}{\rightleftharpoons}} \mathrm{O\bullet} + 2\mathrm{O_2} \ , \\ \mathrm{O_3} + \mathrm{O\bullet} &\underset{p_3}{\to} 2\mathrm{O_2} \ . \end{aligned}$$

Massenwirkungsgesetz liefert Differentialgleichung für Konzentrationen:

$$y_1 = [\mathrm{O\bullet}] \ , \quad y_2 = [\mathrm{O_2}] \ , \quad y_3 = [\mathrm{O_3}] \ , \quad \text{in mol cm}^{-3} \ .$$

$y_i(t)$ bezeichne die Konzentration von O_i zur Zeit t:

$$\begin{cases} y_1' = +p_1y_2y_3 - p_2y_1y_2^2 - p_3y_1y_3 \ , \\ y_2' = +p_1y_2y_3 - p_2y_1y_2^2 + 2p_3y_1y_3 \ , \\ y_3' = -p_1y_2y_3 + p_2y_1y_2^2 - p_3y_1y_3 \ . \end{cases}$$

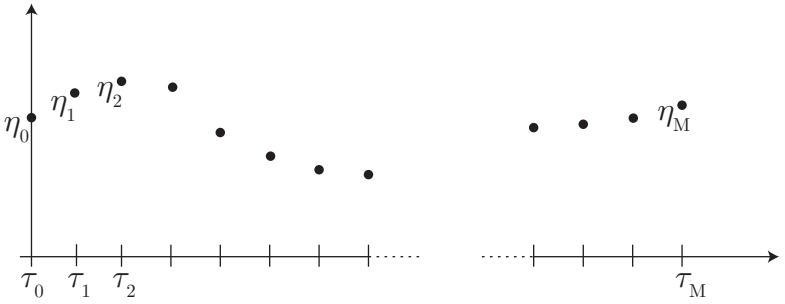
Bemerkung. Gesamtmasse $1 \cdot y_1 + 2 \cdot y_2 + 3 \cdot y_3 = \text{const}$.

„Direkte Simulation“: Bei bekannten Parametern p_i und gegebenem Anfangswert $y_i(t_0)$ löse das Anfangswertproblem numerisch.

„Inverses Problem“: Bestimme aus gemessenen Konzentrationsverläufen die unbekannten Parameter p_i.

Allgemein: Differentialgleichung $y' = f(t, y, p)$ mit unbekannten Parametern

$$p = (p_1, \ldots, p_q)^T \in \mathbb{R}^q \quad (f : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^q \to \mathbb{R}^d) \ .$$



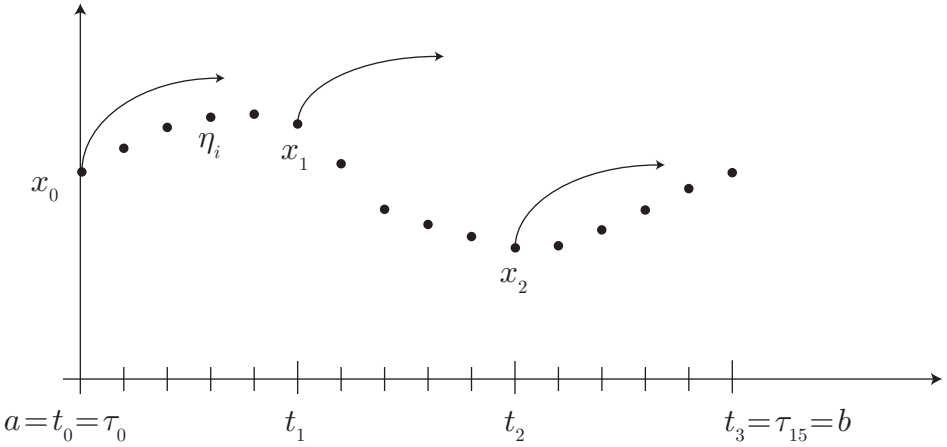
Vorgegebene Messpunkte $(\tau_i, \eta_i)_{i=0,\ldots,M}$,

$$a = \tau_0 < \tau_1 < \cdots < \tau_M = b \ , \quad \eta_i \in \mathbb{R}^d \ .$$

„Inverses Problem“: Bestimme Parameter $p \in \mathbb{R}^q$ und Lösung $y : [a, b] \to \mathbb{R}^d$ der Differentialgleichung $y' = f(t, y, p)$ mit

$$\sum_{i=0}^{M} \|y(\tau_i) - \eta_i\|^2 = \min!$$

Mehrzielmethode: Unterteilung $\{t_0, \ldots, t_m\} \subset \{\tau_0, \ldots, \tau_M\}$, i. A. ist $m \ll M$.
Mit $m = 3$, $M = 15$:



Bezeichnung: $y(\cdot|t,x,p)$ Lösung von $y' = f(t,y,p)$ zum Anfangswert $y(t) = x$.
Habe wieder Stetigkeitsbedingungen:

$$y(t_{j+1}|t_j, x_j, p) - x_{j+1} = 0 \ , \qquad\qquad j = 0, \dots, m-1.$$

Unter diesen Nebenbedingungen suche

$$\sum_{i=0}^{M} \|y(\tau_i|t_j, x_j, p) - \eta_i\|_2^2 = \min! \ ,$$

wobei $j = j(i)$ so, dass $t_j \le \tau_i < t_{j+1}$.

Erhalte:

$$\begin{aligned} F(x,p) = 0 \ , \qquad & F(x,p) = \big(y(t_{j+1}|t_j, x_j, p) - x_{j+1}\big)_{j=0}^{m-1} \ , \\ \|r(x,p)\|_2^2 = \min! \ , \quad & r(x,p) = \big(y(\tau_i|t_j, x_j, p) - \eta_i\big)_{i=0}^{M} \ . \end{aligned}$$

Dies ist ein nichtlineares Gleichungssystem gekoppelt mit einem nichtlinearem Ausgleichsproblem.

Löse numerisch durch Kombination von Newton und Gauß-Newton:

$$z^{k+1} = z^k + \Delta z^k \ , \quad z = (x,p)^T \in \mathbb{R}^{(m+1)d+q} \ ,$$

wobei Δz^k Lösung des *linearen* Problems

$$\left.\begin{aligned} \text{Newton}: \qquad & F'(z^k) \cdot \Delta z^k = -F(z^k) \ , \\ \text{Gauß-Newton}: \quad & \|r'(z^k)\Delta z^k + r(z^k)\|_2^2 = \min! \end{aligned}\right\}$$

und weiter

$$F' = \frac{\partial F}{\partial z} = \left(\frac{\partial F}{\partial x}, \frac{\partial F}{\partial p}\right) \ .$$

Erhalte (lasse den Iterationsindex k weg)

$$\left[\begin{array}{cccc|c} R_0(t_1) & -I & & & P_0(t_1) \\ & R_1(t_2) & -I & 0 & P_1(t_2) \\ & & \ddots & & \vdots \\ & 0 & & \ddots & \vdots \\ & & & R_{m-1}(t_m) \quad -I & P_{m-1}(t_m) \end{array}\right] \begin{pmatrix} \Delta x_0 \\ \vdots \\ \vdots \\ \Delta x_m \\ \Delta p \end{pmatrix} = - \begin{pmatrix} F_0 \\ \vdots \\ \vdots \\ \vdots \\ F_{m-1} \end{pmatrix} \ ,$$

$$\sum_{i=0}^{M} \|R_j(\tau_i)\Delta x_j + P_j(\tau_i)\Delta p + r_i\|_2^2 = \min! \ ,$$

wieder mit $j = j(i)$ so, dass $t_j \le \tau_i < t_{j+1}$. Dabei ist (siehe Übungsaufgabe 9)

$$R_j(t) = \frac{\partial}{\partial x_j} y(t|t_j, x_j, p) \quad \text{die Resolvente zu} \quad v' = C_j(t)v$$
$$\text{mit} \quad C_j(t) = \frac{\partial f}{\partial y}\big(t, y(t|t_j, x_j, p), p\big) \ ,$$

d. h.

$$R_j' = C_j R_j \quad \text{mit} \quad R_j(t_j) = I \ ,$$

und weiterhin ist

$$P_j(t) = \frac{\partial}{\partial p} y(t|t_j, x_j, p) \quad \text{Lösung von} \quad P_j' = C_j P_j + \frac{\partial f}{\partial p}\big(t, y(t|t_j, x_j, p), p\big) \ ,$$
$$P_j(t_j) = 0 \ .$$

Lösung des linearen Teilproblems: Habe für $u = (\Delta x_0, \dots, \Delta x_m, \Delta p)^T$ Problem der Form

$$\begin{cases} \big\|Au - c\big\|_2^2 = \min! \ , \\ \text{unter Nebenbedingung} \quad Bu = d \ , \quad A, B \text{ schwach besetzt.} \end{cases}$$

Setze $v = c - Au$, schreibe um als

$$\frac{1}{2} v^T v = \frac{1}{2} \|v\|_2^2 = \min!$$

unter den Nebenbedingungen

$$Au + v = c \quad \text{und} \quad Bu = d \ .$$

Erhalte mit Lagrange-Multiplikatoren $(\mu, \lambda)^T$ zunächst die Lagrange-Funktion

$$\mathcal{L}(v, u) = \frac{1}{2}\|v\|_2^2 + \lambda(Bu - d) + \mu(Au + v - c)$$

und aus den Ableitungen $\frac{\partial \mathcal{L}}{\partial v} = 0$ und $\frac{\partial \mathcal{L}}{\partial u} = 0$ die folgenden äquivalenten Gleichungssysteme:

$$\begin{cases} \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} A^T & B^T \\ I & 0 \end{pmatrix} \begin{pmatrix} \mu \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \ , \\ \qquad\qquad \begin{pmatrix} A & I \\ B & 0 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} c \\ d \end{pmatrix} \end{cases}$$

bzw.

$$\underbrace{\begin{bmatrix} 0 & 0 & A^T & B^T \\ 0 & I & I & 0 \\ A & I & 0 & 0 \\ B & 0 & 0 & 0 \end{bmatrix}}_{\text{Große schwach besetzte Matrix}} \begin{pmatrix} u \\ v \\ \mu \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ c \\ d \end{pmatrix} \ .$$

Löse mit „sparse solver“, z. B. `netlib`: `MA28`, `MA47`.

Kapitel II

Elliptische partielle Differentialgleichungen: Einführung

§ 1 Lineare partielle Differentialgleichungen 2. Ordnung

§ 1.1 Potentialgleichung als Beispiel einer elliptischen partiellen Differentialgleichung

Erinnerung. Gewöhnliche Differentialgleichung 2. Ordnung

$$\frac{d^2y}{dt^2} = \varphi(t, y) \ ,$$

dazu Anfangsbedingung oder Randbedingung.

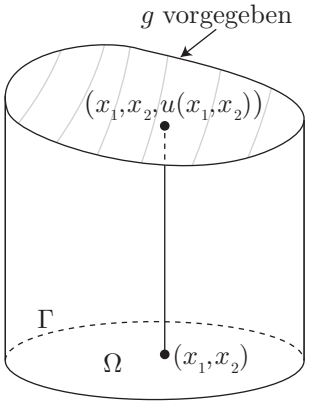
Wichtige Anwendungsklasse: *Variationsprobleme.*
Minimiere Funktion von Kurven:

$$\int_a^b f\big(t, y(t), y'(t)\big)dt = \min!$$

mit festen Endpunkten $y(a)$, $y(b)$.

Jetzt: Minimiere Funktion von *Flächen.*

Beispiel (Minimalflächenproblem)**.**
Suche die Fläche mit dem minimalem Flächeninhalt, die durch eine vorgegebene geschlossene Kurve im $\mathbb{R}^3$ berandet wird (Seifenblase). Die zu minimierende Fläche ist in der Abbildung grau schraffiert.



Punkte auf der Fläche:

$$\big(x_1, x_2, u(x_1, x_2)\big) \quad \text{für} \quad (x_1, x_2) \in \Omega \subset \mathbb{R}^2 \ .$$

Punkte auf der Randkurve:

$$\big(x_1, x_2, g(x_1, x_2)\big) \quad \text{für} \quad (x_1, x_2) \in \partial\Omega =: \Gamma \ .$$

Minimiere den Flächeninhalt

$$\int_\Omega \sqrt{1+\left(\frac{\partial u}{\partial x_1}\right)^2(x_1,x_2)+\left(\frac{\partial u}{\partial x_2}\right)^2(x_1,x_2)}\; d(x_1,x_2)$$

unter allen (stetig differenzierbaren) Funktionen $u:\overline{\Omega}\to\mathbb{R}$, für die gilt

$$u(x_1,x_2)=g(x_1,x_2)\quad \forall (x_1,x_2)\in\Gamma\ .$$

Falls nur kleine Höhenunterschiede, benütze

$$\sqrt{1+u_{x_1}^2+u_{x_2}^2}\approx 1+\frac{1}{2}u_{x_1}^2+\frac{1}{2}u_{x_2}^2+\cdots$$

als Näherung. Minimiere

$$\int_\Omega \left(1+\frac{1}{2}\left(u_{x_1}^2+u_{x_2}^2\right)\right) d(x_1,x_2)$$

unter der Randbedingungen $u=g$ auf Γ. Dies entspricht der Minimierung

$$\int_\Omega \frac{1}{2}\left(\left(\frac{\partial u}{\partial x_1}\right)^2+\left(\frac{\partial u}{\partial x_2}\right)^2\right) d(x_1,x_2)\ ,\quad u(x_1,x_2)=g(x_1,x_2)\quad \forall (x_1,x_2)\in\Gamma\ .$$

Wie in Kapitel I, § 8 betrachte Variation:

$$F(\epsilon)=\int_\Omega \frac{1}{2}\left(\left(\frac{\partial}{\partial x_1}(u+\epsilon v)\right)^2+\left(\frac{\partial}{\partial x_2}(u+\epsilon v)\right)^2\right) dx$$

für $v:\overline{\Omega}\to\mathbb{R}\in C^1$ mit $v(x)=0\ \forall x\in\Gamma$. Für die Minimalkurve u muss gelten $F'(\epsilon=0)=0$. Also:

$$\begin{aligned}0=F'(0)&=\frac{d}{d\epsilon}\Big|_{\epsilon=0}\int_\Omega \frac{1}{2}\left(u_{x_1}^2+2\epsilon u_{x_1}v_{x_1}+\epsilon^2 v_{x_1}^2+u_{x_2}^2+2\epsilon u_{x_2}v_{x_2}+\epsilon^2 v_{x_2}^2\right)dx\\&=\int_\Omega\Big[\underbrace{\frac{\partial u}{\partial x_1}}_{\downarrow}\underbrace{\frac{\partial v}{\partial x_1}}_{\uparrow}+\underbrace{\frac{\partial u}{\partial x_2}}_{\downarrow}\underbrace{\frac{\partial v}{\partial x_2}}_{\uparrow}\Big]dx\ .\end{aligned}$$

Brauche nun die zweidimensionale Version der partiellen Integration.

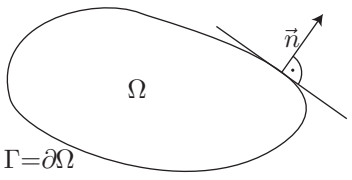
Greensche Formel:

$$\int_\Omega \frac{\partial f}{\partial x_i}(x)g(x)\,dx=\int_\Gamma f(x)g(x)n_i\,d\sigma(x)-\int_\Omega f(x)\frac{\partial g}{\partial x_i}(x)\,dx\ ,\qquad i=1,2,$$

wobei

$$n(x)=\begin{pmatrix}n_1(x)\\n_2(x)\end{pmatrix}$$

der äußere Normaleneinheitsvektor an der Stelle $x\in\Gamma$ ist.



Die Kurve Γ ist parametrisiert durch

$$[0,1] \ni t \mapsto x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} \in \Gamma \ ,$$

$$\int_\Gamma \varphi(x)\, d\sigma(x) = \int_0^1 \varphi\big(x(t)\big) \sqrt{x_1'(t)^2 + x_2'(t)^2}\, dt \ ,$$

$$n = \frac{1}{\sqrt{x_1'^2 + x_2'^2}} \begin{pmatrix} +x_2' \\ -x_1' \end{pmatrix} \ .$$

Damit (beachte $v = 0$ auf Γ):

$$\begin{aligned} 0 &= \int_\Omega \left(\frac{\partial u}{\partial x_1} \frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2} \frac{\partial v}{\partial x_2} \right) dx \\ &\overset{\text{partielle Integration}}{=} \underbrace{\int_\Gamma \left(\frac{\partial u}{\partial x_1} v n_1 + \frac{\partial u}{\partial x_2} v n_2 \right) d\sigma}_{=0} - \int_\Omega \left(\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} \right) v \, dx \end{aligned}$$

$\forall v \in C^1(\overline{\Omega})$ mit $v = 0$ auf Ω.
Schließe mit der zweidimensionalen Version des Fundamentallemmas der Variationsrechnung auf die zugehörige Euler-Lagrange-Gleichung:

$$\begin{cases} \dfrac{\partial^2 u}{\partial x_1^2}(x_1, x_2) + \dfrac{\partial^2 u}{\partial x_2^2} = 0 & \forall (x_1, x_2) \in \Omega \ , \\ \text{Randbedingung:} \quad u(x_1, x_2) = g(x_1, x_2) & \forall (x_1, x_2) \in \Gamma \ . \end{cases}$$

Schreibe kurz mit $\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}$

$$\begin{cases} \Delta u = 0 & \text{in } \Omega \ , \qquad \text{Potentialgleichung} \\ u = g & \text{auf } \Gamma \ , \end{cases}$$

auch für

$$\mathbb{R}^3 \supset \Omega \ni (x_1, x_2, x_3) : \quad \Delta u = \sum_{i=1}^3 \frac{\partial^2 u}{\partial x_i^2}$$

Beispiele.

- Elektrostatik: u Spannung (Potential),
- Temperaturverteilung eines Körper bei gegebener Randtemperatur: u Temperatur,
- Diffusion: Gleichgewichtskonzentration eines diffundierenden Stoffs bei gegebener Randkonzentration.
- Funktionentheorie: $(x_1, x_2) = x_1 + i x_2 \in \mathbb{C}$. Holomorphe Funktion $f = u + iv$, $u = \Re f$. Cauchy-Riemann-Differentialgleichung:

$$\begin{aligned} & \frac{\partial u}{\partial x_1} = \frac{\partial v}{\partial x_2} \ , \quad \frac{\partial u}{\partial x_2} = -\frac{\partial v}{\partial x_1} \\ \Rightarrow \quad & \frac{\partial^2 u}{\partial x_1^2} = \frac{\partial^2 v}{\partial x_1 \partial x_2} = -\frac{\partial^2 u}{\partial x_2^2} \quad \Rightarrow \quad \Delta u = 0 \ . \end{aligned}$$

- Inhomogene Gleichung, wichtig in vielen Problemen der Physik und Chemie:

$$\begin{cases} -\Delta u = f & \text{in } \Omega\ , \qquad \text{Poisson-Gleichung,} \\ u = g & \text{auf } \Gamma\ , \qquad f : \overline{\Omega} \to \mathbb{R}\ , \quad g : \Gamma \to \mathbb{R}\ , \end{cases}$$

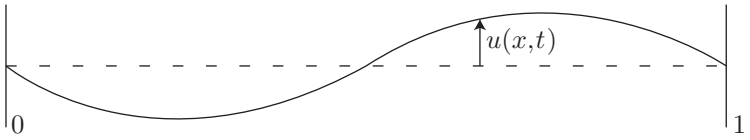
  z. B. Form einer eingespannten Membran unter Wirkung der Schwerkraft im Gleichgewicht (vergleiche Kettenlinie).

- Strömungsprobleme: u Druck einer stationären Strömung.

§ 1.2 Wellengleichung als Beispiel einer hyperbolischen partiellen Differentialgleichung

Schwingende Saite.
Auslenkung:

$u(x,t)\ , \quad x \in [0,1]$ Ort
und $\quad t \in [0,T]$ Zeit.

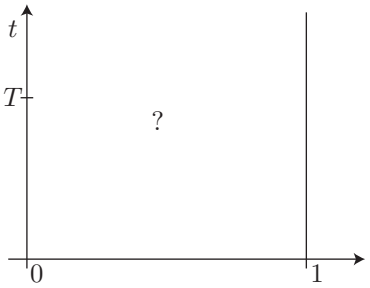


Randbedingung (eingespannt):

$$u(0,t) = u(1,t) = 0 \ \forall t\ .$$

Anfangsbedingung (gegeben):

$$u(x,0)\ , \quad \frac{\partial u}{\partial t}(x,0)\ .$$



Wellengleichung:

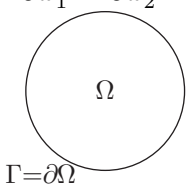
$$\frac{\partial^2 u}{\partial t^2}(x,t) = c^2 \frac{\partial^2 u}{\partial x^2}(x,t)\ .$$

c Konstante in [m/s] (Schallgeschwindigkeit). Für $c = 1$:

$$\frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0\ , \quad \text{vergleiche mit Potentialgleichung:} \quad \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = 0\ .$$

Schwingende Membran. Auslenkung:

$$u(x_1, x_2, t)\ , \quad (x_1, x_2) \in \Omega\ , \quad t \in [0,T]\ .$$



Randbedingung (eingespannt):

$$u(x_1, x_2, t) = 0 \quad \forall (x_1, x_2) \in \Gamma \quad \forall t\ .$$

Anfangsbedingung:

$$u(x_1, x_2, 0)\ , \quad \frac{\partial u}{\partial t}(x_1, x_2, 0) \quad \forall (x_1, x_2) \in \Omega \text{ gegeben.}$$

Wellengleichung:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta u \quad \text{in } \Omega \times [0,T].$$

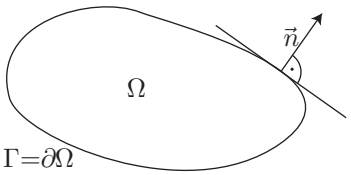
In drei Raumdimensionen: Bewegung in idealem Gas, wobei $u(x_1, x_2, x_3, t)$ den Druck bezeichnet und die Konstante c die Schallgeschwindigkeit.

§ 1.3 Wärmeleitungsgleichung als Beispiel einer parabolischen partiellen Differentialgleichung

$u(x,t)$ Temperatur an der Stelle x zur Zeit t. Im Körper:

$$u(x_1, x_2, x_3) \in \Omega \subset \mathbb{R}^3 \ , \quad t \in [0, T] \ .$$

Randwerte:
Vorgegebene Randtemperatur $u = g$ auf $\partial\Omega \times [0,T]$ oder Körper isoliert, d. h. $\frac{\partial u}{\partial n} = 0$ auf $\partial\Omega \times [0,T]$, wobei



$$\frac{\partial u}{\partial n} := \nabla u \cdot n = \sum_{i=1}^{3} \frac{\partial u}{\partial x_i} n_i \quad \text{Normalenableitung.}$$

Anfangswerte: $u(x,0)$ gegeben $\forall x \in \Omega$.
Wärmeleitungsgleichung:

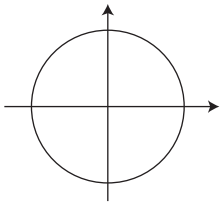
$$\frac{\partial u}{\partial t} = k\Delta u \quad \text{in } \Omega \times (0,T)$$

und Anfangsbedingungen und Randbedingungen wie oben. Auch: Diffusion, u Konzentration.
Eindimensionaler Fall: $u_t = u_{xx}$.

§ 1.4 Typeneinteilung (nach du Bois-Reymond, $\sim$ 1900)

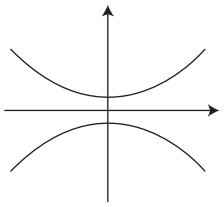
- Poisson-Gleichung (elliptische PDGL)

$$u_{xx} + u_{yy} = f \ ,$$
vergleiche mit:
$$x^2 + y^2 = 1 \ .$$



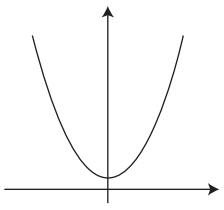
- Wellengleichung (hyperbolische PDGL)

$$u_{tt} - u_{xx} = f \ ,$$
vergleiche mit:
$$t^2 - x^2 = 1 \ .$$



- Wärmeleitungsgleichung (parabolische PDGL)

$$u_t - u_{xx} = f \ ,$$
vergleiche mit:
$$t - x^2 = 1 \ .$$



§ 1.5 Allgemeine lineare partielle Differentialgleichungen 2. Ordnung in n Variablen

Sei $x = (x_1, \ldots, x_n) \in \Omega \subset \mathbb{R}^n$.

$$-\sum_{i,j}^{n} a_{ij}(x)\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{k=1}^{n} b_k(x)\frac{\partial u}{\partial x_k} + c(x)u = f(x) \ , \tag{II.1}$$

wegen

$$\frac{\partial^2 u}{\partial x_i \partial x_j} = \frac{\partial^2 u}{\partial x_j \partial x_i}$$

(für u C^2) kann ohne Einschränkung annehmen, dass $a_{ij} = a_{ji}$.

$$A(x) := \big(a_{ij}(x)\big)_{i,j=1}^{n} \quad \text{symmetrisch} \ .$$

Definition. Die Differentialgleichung (II.1) heißt

- *elliptisch im Punkt x*, falls $A(x)$ positiv definit;
- *elliptisch*, falls sie in jedem Punkt des Gebiets elliptisch ist.

Schreibe dann (II.1) kurz als

$$Lu = f \ ,$$

L heißt elliptischer Differentialoperator 2. Ordnung.

Beispiel (Poisson-Gleichung)**.**

$$\begin{aligned} L = -\Delta \ , \qquad & A(x) = I \ , \\ -\Delta u = f \ , \qquad & \text{die Poisson-Gleichung ist also elliptisch} \ . \end{aligned}$$

Parabolische partielle Differentialgleichung:

$$\frac{\partial u}{\partial t} + Lu = f \ .$$

Hyperbolische partielle Differentialgleichung:

$$\frac{\partial^2 u}{\partial t^2} + Lu = f \ .$$

Bei beiden Gleichungen gelte $u = u(x,t)$ und L elliptisch.

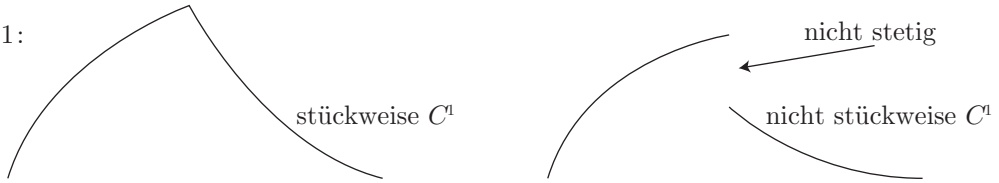
Wann ist das Problem, bestehend aus der partiellen Differentialgleichung, den Rand- und Anfangsbedingungen, *wohlgestellt*?
Existenz, Eindeutigkeit, stetige Abhängigkeit von den Daten (bezüglich welcher Norm?). Diese Fragestellungen gehen auf Hadamard ($\sim$ 1920) zurück.

§ 1.6 Einschub: Gebiete mit stückweise stetig differenzierbarem Rand, Greensche Formel

Definition. Sei $D \subset \mathbb{R}^n$ offen. Eine Funktion $\varphi : \overline{D} \to \mathbb{R}$ heiße *stückweise stetig differenzierbar* (kurz: stw. C^1), falls φ stetig ist und es endlich viele $D_1, \ldots, D_n \subset \mathbb{R}^n$ offen mit $\overline{D} = \bigcup_{i=1}^n \overline{D}_i$ gibt, sodass $\varphi\big|_{D_i}$ stetig differenzierbar und $(\varphi\big|_{D_i})'$ auf $\overline{D}_i$ stetig fortsetzbar ist $(\forall i = 1, \ldots, n)$.

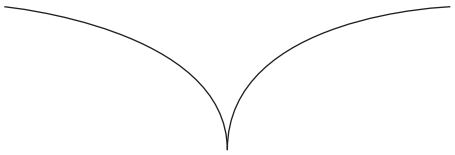
Beispiele. ($n = 1$ und $n = 2$)

$n=1$:



Die linke Funktion ist stückweise C^1, aber ihre Ableitung (rechts) nicht.

$n=1$:



Die Funktion $\sqrt{|x|}$ ist ebenfalls nicht C^1, weil die Ableitung nicht stetig auf $[0, \infty)$ fortsetzbar ist.

$n = 2$: Ein Polyeder ist stückweise C^1.

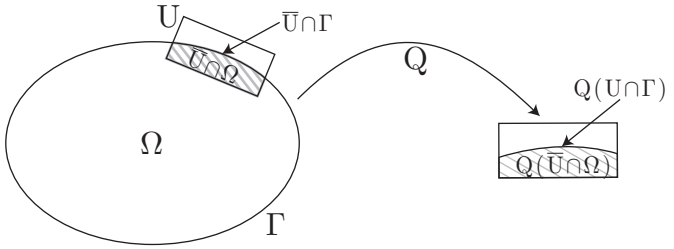
Definition. Sei $\Omega \subset \mathbb{R}^2$ beschränktes Gebiet, $\Gamma := \partial\Omega$. Ω heißt *Gebiet mit stückweise stetig differenzierbarem Rand* (kurz: Ω stw. C^1-Gebiet), falls es endlich viele Rechtecke gibt, die Γ überdecken, sodass gilt: Für jedes solche Rechteck U gibt es eine Drehung Q des $\mathbb{R}^2$, sodass

$$Q(\overline{U} \cap \Gamma) = \text{Graph einer stückweisen } C^1\text{-Funktion } \gamma = \{\big(t, \gamma(t)\big) : t \in [a, b]\}$$

und

$$Q(\overline{U} \cap \Omega) \quad \text{liegt unterhalb von} \quad Q(\overline{U} \cap \Gamma)\ .$$

Anders: Γ ist lokal der Graph einer stückweisen C^1-Funktion und Ω liegt auf einer Seite von Γ.



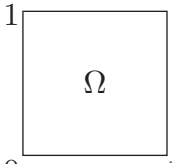
Greensche Formel: $\Omega \subset \mathbb{R}^2$ beschränktes Gebiet, stückweise C^1, $f,\ g : \overline{\Omega} \to \mathbb{R}$ stückweise C^1. Dann gilt

$$\int_\Omega \frac{\partial f}{\partial x_i}(x) \cdot g(x)\, dx = \int_{\partial\Omega} f(x) \cdot g(x) \cdot n_i\, d\sigma - \int_\Omega f(x) \frac{\partial g}{\partial x_i}(x)\, dx\ , \qquad \mathrm{i} = 1,\, 2.$$

Beweis. Analysis II. □

§ 2 Finite Differenzen

Einfachstes und naheliegendes Verfahren zur näherungsweisen Lösung von partiellen Differentialgleichungen. Erläutere an Modellproblem:

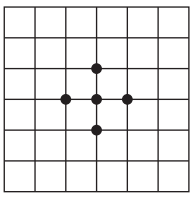


$$\begin{aligned} -\Delta u &= f \quad \text{in} \quad \Omega = (0,1) \times (0,1)\ , \\ u &= g \quad \text{auf} \quad \Gamma = \partial\Omega \quad \text{„}\textit{Dirichlet}\text{-Randbedingungen“}. \end{aligned}$$

Bezeichne Punkte in Ω mit $(x, y) \in \mathbb{R}^2$,

$$-\Delta u = -\frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2}\ .$$

Lege Gitter der Weite $h = 1/(N+1)$ über Ω:



$$\begin{aligned} &x_i = ih\ , \quad y_j = jh\ , \\ &\mathbb{R}_h^2 = \{(x_i, y_j) : i, j \in \mathbb{Z}\}\ , \\ &\Omega_h = \mathbb{R}_h^2 \cap \Omega\ , \quad \text{innere Gitterpunkte}, \\ &\Gamma_h = \mathbb{R}_h^2 \cap \Gamma\ , \quad \text{Gitterpunkte am Rand}, \\ &\overline{\Omega}_h = \Omega_h \cup \Gamma_h\ . \end{aligned}$$

Ersetze Ableitungen durch endlichen Differenzenquotienten (→ Name!):

$$\begin{aligned} -\frac{\partial^2 u}{\partial x^2}(x,y) &\approx \frac{-u(x+h,y) + 2u(x,y) - u(x-h,y)}{h^2}\ , \\ -\frac{\partial^2 u}{\partial y^2}(x,y) &\approx \frac{-u(x,y+h) + 2u(x,y) - u(x,y-h)}{h^2}\ . \end{aligned}$$

Fehler: $\mathcal{O}(h^2)$, falls u C^4 (Taylor), „Abbruchfehler“, englisch „Truncation error“. Somit

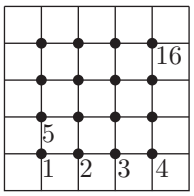
$$\begin{aligned} -\Delta u(x,y) &\approx \frac{1}{h^2}\Big(\underbrace{4u(x,y)}_{\text{mittlerer}} - \underbrace{u(x+h,y)}_{\text{östlicher}} - \underbrace{u(x-h,y)}_{\text{westlicher}} - \underbrace{u(x,y+h)}_{\text{nördlicher}} - \underbrace{u(x,y-h)}_{\text{südlicher}} \Big) \\ &=: -\Delta_h u(x,y)\ . \end{aligned}$$

Schema:

$$\begin{array}{ccccc} & & -1 & & \\ -1 & & 4 & -1 & : \\ & & -1 & & \end{array} \qquad \text{Fünf-Punkte-Stern.}$$

Löse Näherungsproblem auf Gitter: Suche $u_h : \overline{\Gamma}_h \to \mathbb{R}$ mit

$$\begin{cases} -\Delta_h u_h = f_h & \text{in } \Omega_h \quad (f_h = f\big|_{\Omega_h}) \ , \\ u_h = g_h & \text{auf } \Gamma_h \quad (g_h = g\big|_{\Gamma_h}) \ . \end{cases} \tag{II.2}$$



$N = 4$

Nummeriere Punkte von Ω_h durch, erhalte Gitterfunktion

$$u_h : \Omega_h \to \mathbb{R} \mathrel{\hat{=}} \text{Vektor } U \in \mathbb{R}^{N^2} \ ,$$
$$u_h(x_i, y_j) = U_{i+(j-1)N} \ .$$

Aus (II.2) erhalte lineares Gleichungssystem $Au = b$ mit Matrix

$$A = \frac{1}{h^2} \begin{bmatrix} 4 & -1 & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -1 & 4 & -1 & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -1 & 4 & -1 & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & 4 & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ -1 & \cdot & \cdot & \cdot & 4 & -1 & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -1 & \cdot & \cdot & -1 & 4 & -1 & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -1 & \cdot & \cdot & -1 & 4 & -1 & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -1 & \cdot & \cdot & -1 & 4 & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & 4 & -1 & \cdot & \cdot & -1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & -1 & 4 & -1 & \cdot & \cdot & -1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & -1 & 4 & -1 & \cdot & \cdot & -1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & -1 & 4 & \cdot & \cdot & \cdot & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & \cdot & 4 & -1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & -1 & 4 & -1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & -1 & 4 & -1 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -1 & \cdot & \cdot & -1 & 4 \end{bmatrix} \ , \tag{II.3}$$

$A \in \mathbb{R}^{N^2 \times N^2}$.

In der gesamten Matrix wurde zur besseren Übersichtlichkeit $\cdot \equiv 0$ verwendet.

Eigenschaften: Symmetrisch, positiv definit (gilt leider nicht für komplizierte Gebiete Ω). Bandmatrix der Breite $2N + 1$, schwach besetzt.
Die rechte Seite ist

$$b_{i+(j-1)N} = f(x_i, y_j) \ ,$$

falls alle Nachbarpunkte von (x_i, y_j) im Innern von Ω_h liegen. Wenn aber Punkte außerhalb von Ω_h liegen, so gilt

$$\begin{aligned} b_1 &= f(x_1, y_1) + \frac{1}{h^2} g(x_0, y_1) + \frac{1}{h^2} g(x_1, y_0) \ , \\ b_2 &= f(x_2, y_1) + \frac{1}{h^2} g(x_2, y_0) \ , \\ b_3 &= f(x_3, y_1) + \frac{1}{h^2} g(x_3, y_0) \ , \quad \dots \ . \end{aligned}$$

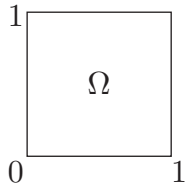
Großes lineares Gleichungssystem zu lösen.

Direkt: Gauß-Elimination, Cholesky-Verfahren. Nütze Bandstruktur, aber innerhalb des Bandes Auffüllen mit Nichtnullelementen („fill-in“).

Iterativ: Konjugierte Gradienten, Mehrgitterverfahren: Siehe weiter unten, kommt später (Aufwand $\sim$ zur Anzahl der Gitterpunkte).

§ 3 Konvergenz des Finite-Differenzen-Verfahrens, Maximumprinzip

Voraussetzungen und Bezeichnungen wie in § 2.



$$\begin{cases} -\Delta_h u_h = f_h & \text{in } \Omega_h \ , \\ u_h = g_h & \text{auf } \Gamma_h \ . \end{cases}$$

Falls exakte Lösung u C^4: Weiß

$$-\Delta_h u = -\Delta u - d_h \quad \text{in } \Omega_h \quad \text{mit} \quad \max_{\Omega_h} |d_h| \le D \cdot h^2 \ , \qquad \text{Abbruchfehler,}$$

und

$$\begin{cases} -\Delta_h u = f_h - d_h & \text{in } \Omega_h \ , \\ u = g_h & \text{auf } \Gamma_h \ . \end{cases}$$

Für Fehler $e_h = u_h - u$:

$$\begin{cases} -\Delta_h e_h = d_h & \text{in } \Omega_h \ , \\ e_h = 0 & \text{auf } \Gamma_h \ . \end{cases}$$

Werde zeigen:

$$\max_{\Omega_h} |e_h| \le \underbrace{C}_{\text{unabhängig von } h} \cdot \max_{\Omega_h} |d_h| \ , \qquad \text{Stabilitäts-Ungleichung.}$$

Damit:

$$\max_{\Omega_h}|e_h| = \mathcal{O}(h^2) \ , \quad \text{falls } u \ C^4 \ .$$

Der Beweis der Stabilitäts-Ungleichung folgt aus dem diskretem Maximumprinzip und steht weiter unten (Satz 4).
Betrachte zunächst das kontinuierliche Maximumprinzip für die Poisson-Gleichung.

Satz 1 (Maximumprinzip)**.** *Sei $\Omega \subset \mathbb{R}^2$ (oder allgemein $\Omega \subset \mathbb{R}^n$, beliebig $n \geq 1$) beschränktes Gebiet. Sei $u \in C^2(\Omega) \cap C(\overline{\Omega})$ und erfülle*

$$\begin{cases} -\Delta u = f & \text{in } \Omega \text{ mit } f \leq 0 \text{ in } \Omega \ , \\ u = g & \text{auf } \Gamma = \partial\Omega \ . \end{cases}$$

Dann nimmt u sein Maximum auf dem Rand an:

$$\max_{\overline{\Omega}} u \leq \max_{\Gamma} g \ .$$

Beweis. (a) *Zeige Behauptung zunächst für $f < 0$.*

Annahme: u nimmt in $(x^*, y^*) \in \Omega$ sein Maximum an. Dann gilt $u_x = 0$ und $u_y = 0$ in (x^*, y^*) und

$$\begin{pmatrix} u_{xx} & u_{xy} \\ u_{yx} & u_{yy} \end{pmatrix} \quad \text{negativ semidefinit.}$$

u_x und u_y haben eine Nullstelle in (x^*, y^*). Diese Nullstelle bleibt auch bei der zweiten partiellen Ableitung nach der jeweils anderen Variablen erhalten, also $u_{xy} = u_{yx} = 0$. Damit gilt weiterhin

$$u_{xx} \leq 0 \ , \quad u_{yy} \leq 0 \quad \text{in } (x^*, y^*)$$

und daraus folgt mit der Differentialgleichung ein Widerspruch:

$$0 \leq -u_{xx} - u_{yy} = f < 0 \quad \Rightarrow \qquad \text{Widerspruch!}$$

(b) *Betrachte nun $f \leq 0$.* Annahme: $\exists (x^*, y^*) \in \Omega$ mit $u(x^*, y^*) > \max_\Gamma g$. Definiere die Hilfsfunktion

$$v(x, y) = (x - x^*)^2 + (y - y^*)^2 \ ,$$

beschränkt auf $\overline{\Omega}$ (beschränktes Gebiet nach Voraussetzung) mit

$$v(x^*, y^*) = 0 \ , \quad -\Delta v(x, y) = -4 \leq 0 \ .$$

Setze nun $w := u + \delta v$. w nimmt für genügend kleines $\delta > 0$ immer noch sein Maximum im Inneren an, in $(x_\delta, y_\delta) \in \Omega$:

$$-\Delta w = -\Delta u - \delta \Delta v = \underbrace{f}_{\leq 0} - 4\delta < 0 \ .$$

Dies ist ein Widerspruch zu Teil (a). □

Satz 2 (diskretes Maximumprinzip)**.** *Voraussetzungen wie in § 2.* $u_h : \overline{\Omega}_h \to \mathbb{R}$ *erfülle*

$$\begin{cases} -\Delta_h u_h = f_h & \text{in } \Omega_h \text{ mit } f_h \le 0 \ , \\ u_h = g_h & \text{auf } \Gamma_h \ . \end{cases}$$

Dann nimmt u_h *sein Maximum auf dem Rand an:*

$$\max_{\Omega_h} u_h \le \max_{\Gamma_h} g_h \ .$$

Beweis. Annahme: u_h nehme sein Maximum in $(x_i, y_j) \in \Omega_h$ an. Setze $u_{ij} = u_h(x_i, y_j)$. Habe dann

$$\begin{aligned} &\left(\frac{1}{h^2}(4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1} \right) = f_{ij} \le 0 \\ \Leftrightarrow \quad & u_{ij} \le \frac{1}{4}\Big(\underbrace{u_{i-1,j}}_{\le u_{ij}} + \underbrace{u_{i+1,j}}_{\le u_{ij}} + \underbrace{u_{i,j-1}}_{\le u_{ij}} + \underbrace{u_{i,j+1}}_{\le u_{ij}} \Big) \ . \end{aligned}$$

Dann nimmt u_h in den Nachbarpunkten ebenfalls sein Maximum an. Durch mehrmalige Anwendung dieses Arguments auf Nachbarpunkte folgt, dass u_h sein Maximum (auch) auf dem Rand annimmt. □

Folgerung (Vergleichsprinzip). Für u_h, $v_h : \overline{\Omega}_h \to \mathbb{R}$ gelte

$$\begin{cases} -\Delta_h u_h \le -\Delta_h v_h & \text{in } \Omega_h \ , \\ u_h \le v_h & \text{auf } \Gamma_h \ . \end{cases}$$

Dann gilt $u_h \le v_h$ in Ω_h.

Beweis.

$$\begin{cases} -\Delta_h(u_h - v_h) \le 0 & \text{in } \Omega_h \ , \\ u_h - v_h \le 0 & \text{auf } \Gamma_h \ . \end{cases}$$

Wende nun mit $f_h = g_h = 0$ das diskrete Maximumprinzip an:

$$\Rightarrow \quad \max_{\Omega_h}(u_h - v_h) \le \max_{\Gamma_h}(u_h - v_h) \le 0 \ . \qquad \square$$

Hilfssatz 3. *Sei* Ω_h *ein Gitter auf dem Einheitsquadrat* $\Omega = (0,1)^2$. $v_h : \overline{\Omega}_h \to \mathbb{R}$ *erfülle*

$$\begin{cases} -\Delta_h v_h = +1 & \text{in } \ \Omega_h \ , \\ v_h = 0 & \text{auf } \ \Gamma_h \ . \end{cases}$$

Dann gilt

$$0 \le v_h \le \frac{1}{2} \quad \text{in } \ \Omega_h \ ,$$

v_h *ist also von* h *unabhängig beschränkt.*

Beweis. Habe

$$\left.\begin{aligned} -\Delta_h(-v_h) &= -1 \quad \text{in } \Omega_h \ , \\ -v_h &= 0 \quad \text{auf } \Gamma_h \ . \end{aligned}\right\}$$

Erhalte daraus mit dem diskreten Maximumprinzip $-v_h \leq 0$ in Ω_h, d. h. $v_h \geq 0$ in Ω_h. Für die obere Schranke definiere die Hilfsfunktion

$$w(x,y) = \frac{1}{4}(2 - x^2 - y^2) \ .$$

Habe $-\Delta w = 1$, $w \geq 0$ in $\overline{\Omega} = [0,1]^2$. Setze $w_h = w\big|_{\overline{\Omega}_h}$. w ist eine quadratische Funktion, daher ist die Differenzenapproximation von Δ exakt. Habe

$$\begin{cases} -\Delta_h w_h = +1 \quad \text{in } \Omega_h \ , \\ w_h \geq 0 \quad \text{auf } \Gamma_h \ , \end{cases}$$

und damit

$$\begin{cases} -\Delta_h v_h = -\Delta_h w_h \quad \text{in } \Omega_h \ , \\ v_h \leq w_h \quad \text{auf } \Gamma_h \ . \end{cases}$$

Da das $=$ in der oberen Zeile auch $\leq$ beinhaltet, ist das Vergleichsprinzip anwendbar und es folgt daraus $v_h \leq w_h \leq \frac{1}{2}$ in Ω_h. □

Als Folgerung erhalte

Satz 4 (Stabilität)**.** *Sei*

$$\begin{cases} -\Delta_h e_h = d_h \quad \textit{in } \Omega_h \ , \\ e_h = 0 \quad \textit{auf } \Gamma_h \ . \end{cases}$$

Dann gilt

$$\max_{\Omega_h} |e_h| \leq \frac{1}{2} \max_{\Omega_h} |d_h| \ .$$

Beweis. Sei $M := \max_{\Omega_h} |d_h|$. Definiere v_h als Lösung von

$$\begin{cases} -\Delta_h v_h = M \quad \text{in } \Omega_h \ , \\ v_h = 0 \quad \text{auf } \Gamma_h \ . \end{cases}$$

Aus dem Vergleichsprinzip folgt $e_h \leq v_h$, mit Hilfssatz 3 folgt weiterhin $v_h \leq M/2$. Habe auch

$$\begin{cases} -\Delta_h(-v_h) = -M \quad \text{in } \Omega_h \\ -v_h = 0 \quad \text{auf } \Gamma_h \ . \end{cases}$$

Mit dem Vergleichsprinzip folgt nun $v_h \leq e_h$ in Ω_h und wieder mit Hilfssatz 3 $-v_h \geq -M/2$. Somit folgt

$$-\frac{M}{2} \leq -v_h \leq e_h \leq v_h \leq \frac{M}{2} \ .$$ □

§ 4 Variationelle Approximation (Ritz-Galerkin)

$$\begin{cases} -\Delta u = f & \text{in } \Omega\ , \quad \Omega \text{ stückweise } C^1\text{-Gebiet in } \mathbb{R}^2, \\ u = 0 & \text{auf } \Gamma = \partial\Omega\ . \end{cases} \tag{II.4}$$

Fasse dies als Euler-Lagrange-Gleichung zu einem Variationsproblem auf:

Satz 1. *Sei $u \in C^2(\Omega) \cap C(\overline{\Omega})$, $u = 0$ auf Γ. Dann sind die folgenden Aussagen äquivalent:*

(i) u ist Lösung von (II.4).

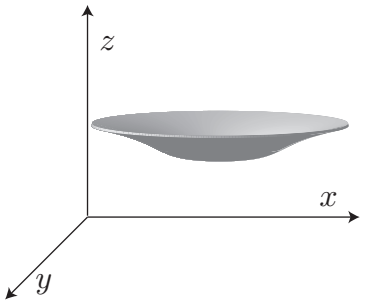
(ii)
$$\int_\Omega \left(\frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} \right) d(x,y) = \int_\Omega f v \, d(x,y)$$
für alle $v : \overline{\Omega} \to \mathbb{R}$ stückweise C^1 mit $v = 0$ auf Γ.

(iii) u ist Lösung des Minimierungsproblems
$$\frac{1}{2}\int_\Omega \left(\left[\frac{\partial v}{\partial x}\right]^2 + \left[\frac{\partial v}{\partial y}\right]^2 \right) d(x,y) - \int_\Omega f v \, d(x,y) = \min!$$
unter allen $v : \overline{\Omega} \to \mathbb{R}$ stückweise C^1 mit $v = 0$ auf Γ.

Bemerkung (Eingespannte Membran im Schwerefeld).

(ii) Prinzip der virtuellen Arbeit

(iii) Minimale Energie



Beweis.

(iii) $\Rightarrow$ Variation $u + \epsilon v$, $\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \% = 0$ (siehe Herleitung der Potentialgleichung) $\Rightarrow$ (ii) $\Rightarrow$ Greensche Formel, Fundamentallemma der Variationsrechnung $\Rightarrow$ (i),

(i) $\Rightarrow$ Greensche Formel $\Rightarrow$ (ii),

(ii) $\Rightarrow$ (iii) folgt gleich mit neuen Bezeichnungen. □

Kurz. Für $u \in C^2(\Omega) \cap \overline{C}(\Omega)$:

$$\begin{cases} -\Delta u = f \quad \text{in } \Omega \ , \\ u = 0 \text{ auf } \Gamma \ , \end{cases} \quad \Leftrightarrow \quad \begin{array}{c} \displaystyle\int_\Omega (u_x v_x + u_y v_y) = \int_\Omega fv \\ \forall v \text{ mit } v = 0 \text{ auf } \Gamma \ , \end{array} \quad \Leftrightarrow \quad \begin{array}{l} u \text{ minimiert} \\ \displaystyle\int_\Omega \frac{1}{2}(v_x^2 + v_y^2) - fv \\ \forall v,\ v\big|_\Gamma = 0 \ . \end{array}$$

Bezeichnung:

$$\begin{aligned} V &= \{v : \overline{\Omega} \to \mathbb{R} \mid v \text{ stückweise } C^1,\ v = 0 \text{ auf } \Gamma\} \ , \\ a(u,v) &= \int_\Omega (u_x v_x + u_y v_y)\, d(x,y) \quad \text{für } u,\ v \in V \ , \\ l(v) &= \int_\Omega fv\, d(x,y) \quad \text{für } v \in V \ , \\ J(v) &= \frac{1}{2}\int_\Omega (v_x^2 + v_y^2)\, d(x,y) - \int_\Omega fv\, d(x,y) = \frac{1}{2}a(v,v) - l(v) \ . \end{aligned}$$

Habe $a : V \times V \to \mathbb{R}$ bilinear, symmetrisch. a ist positiv definit (d. h. $a(v,v) > 0$ $\forall 0 \neq v \in V$), weil wegen

$$a(v,v) = \int_\Omega (v_x^2 + v_y^2)\, d(x,y) \geq 0$$

$a(v,v) = 0$ nur gelten kann für $v_x = v_y = 0$ in Ω und damit $v = \text{const}$. Mit $v = 0$ auf Γ folgt daraus schließlich $v = 0$. Somit ist a Skalarprodukt auf V (*Prä*hilbertraum). Zugehörige Norm:

$$\|v\|_a = \sqrt{a(v,v)} \ , \qquad \text{„Energienorm“.}$$

$l : V \to \mathbb{R}$ linear. Mit diesen Bezeichnungen habe in Satz 1

(ii) $$a(u,v) = l(v)\ \forall v \in V \quad \text{und} \tag{II.5}$$

(iii) $$J(u) = \min_{v \in V} J(v) \quad \text{mit} \quad J(v) = \frac{1}{2}a(v,v) - l(v) \ .$$

Betrachte dies im Falle für allgemeine positiv definite symmetrische Bilinearformen a und Linearformen l auf beliebigem Vektorraum V.
Zeige jetzt (ii) $\Rightarrow$ (iii).
Sei $w \in V$ beliebig, setze $v = w - u$, also $w = u + v$.

$$\begin{aligned} J(u+v) - J(u) &= \frac{1}{2}\Big(a(u+v,u+v) - a(u,u)\Big) - \Big(l(u+v) - l(u)\Big) \\ &\overset{\substack{a \text{ bilinear, symmetrisch} \\ b \text{ linear} \\ \downarrow}}{=} \frac{1}{2}\Big(a(u,u) + 2a(u,v) + a(v,v) - a(u,u)\Big) - \Big(l(u) + l(v) - l(u)\Big) \\ &= \underbrace{a(u,v) - l(v)}_{=0 \text{ wegen (ii)}} + \underbrace{a(v,v)}_{\geq 0} \geq 0 \ . \end{aligned}$$

Damit

$$J(u) \leq J(w)\ \forall w \in V \ .$$

Erinnerung ((iii) $\Rightarrow$ (ii)).

$$0 = \frac{d}{d\epsilon}\Big|_{\epsilon=0} J(u+\epsilon v) = a(u,v) - l(v) \ .$$

Idee. Zur näherungsweisen Lösung der partiellen Differentialgleichung (II.4) gehe von (ii) oder (iii) aus (variationelle Formulierung des Problems).

Wähle einen endlichdimensionalen Unterraum V_N von V, wobei $N = \dim V_N$. Bestimme $u_N \in V_N$, sodass

$$a(u_N, v_N) = l(v_N) \quad \forall v_N \in V_N \ , \qquad \text{Galerkin-Verfahren} \tag{II.6}$$

beziehungsweise (äquivalent)

$$u_N \in V_N : \quad J(u_N) = \min_{v_N \in V_N} J(v_N) \ , \qquad \text{Ritz-Verfahren.}$$

Bemerkung. Die Äquivalenz dieser beiden Verfahren ergibt sich aus (ii) $\Leftrightarrow$ (iii) oben mit V_N statt V.

Satz 2. *Es existiert genau eine Lösung $u_N \in V_N$ des Galerkin-Verfahrens* (II.6).

Beweis. Sei $(\varphi_1, \ldots, \varphi_N)$ Basis von V_N. Gesucht:

$$u_N = \sum_{i=1}^{N} \mu_i \varphi_i \in V_N \ , \quad \text{sodass} \quad a(u_N, v_N) = l(v_n) \quad \forall v_N \quad \text{mit } v_N = \sum_{j=1}^{N} \nu_j \varphi_j \in V_N \ ,$$

d. h. wegen der Linearität von a und l

$$\sum_{i=1}^{N}\sum_{j=1}^{N} \mu_i \nu_j \, a(\varphi_i, \varphi_j) = \sum_{j=1}^{N} \nu_j \, l(\varphi_j) \quad \forall (\nu_j)_{j=1}^{N} \in \mathbb{R}^N \ .$$

In Matrixschreibweise:

$$\nu^T A \mu = \nu^T b \quad \forall \nu \in \mathbb{R}^N$$
$$\text{mit} \quad A = \Big(a(\varphi_i, \varphi_j)\Big)_{i,j=1}^{N} \ , \quad b = \Big(l(\varphi_j)\Big)_{j=1}^{N} \ .$$

Die Matrix A ist symmetrisch (wegen der symmetrischen Bilinearform a) und positiv definit, denn

$$\nu^T A \nu = \sum_i \sum_j \nu_i \nu_j \, a(\varphi_i, \varphi_j) = a\Big(\sum_i \nu_i \varphi_i, \sum_j \nu_j \varphi_j\Big) > 0 \ ,$$

falls $\sum_i \nu_i \varphi_i \neq 0$, d. h. $\nu = (\nu_i)_{i=1}^{N} \neq 0$.
Setze nun $\nu = e_k$, e_k normierter Basisvektor im $\mathbb{R}^N$, und es ergibt sich

$$A\mu = b \ , \qquad \text{lineares Gleichungssystem.}$$

A ist symmetrisch, positiv definit und daher insbesondere invertierbar. Daraus folgt, dass genau eine Lösung μ des linearen Gleichungssystems existiert. Dann ist

$$u_N = \sum_i \mu_i \varphi_i$$

die eindeutige Lösung von (II.6). □

Satz 3 (Céas Lemma)**.** *Falls $u \in V$ Lösung von* (II.5)*, $u_N \in V_N$ Lösung von* (II.6)*, so gilt in der Energienorm ($\|v\|_a = \sqrt{a(v,v)}$)*

$$\|u_N - u\|_a = \min_{v_N \in V_N} \|v_N - u\|_a \ .$$

Bemerkungen. Die Galerkin-Approximation hat also in der Energienorm unter allen Elementen des Approximationsraums V_N den kleinsten Abstand zur exakten Lösung.
Die Fehleruntersuchung des Galerkin-Verfahrens reduziert sich somit auf die Frage, wie gut sich die Lösung in V_N annähern lässt (Approximationsproblem).

Beweis. Da $V_N \leq V$ gilt laut (II.5)

$$a(u, v_N) = l(v_N) \quad \forall v_N \in V_N$$

und laut (II.6)

$$a(u_N, v_N) = l(v_N) \quad \forall v_N \in V_N \ .$$

Subtrahiert man nun diese Gleichungen voneinander, so erhält man

$$a(u_N - u, v_N) = 0 \quad \forall v_N \in V_N$$

bzw. auch (ersetze v_N durch $u_N - v_N \in V_N$)

$$a(u_N - u, \underbrace{u_N - v_N}_{u_N - u + u - v_N}) = 0 \quad \forall v_N \in V_N$$

$$\Leftrightarrow \quad \underbrace{a(u_N - u, u_N - u)}_{= \|u_N - u\|_a^2} = \underbrace{a(u_N - u, v_N - u)}_{\leq \|u_N - u\|_a \cdot \|v_N - u\|_a \text{ (Cauchy-Schwarz)}} \quad \forall v_N \in V_N$$

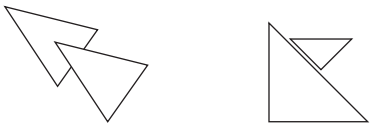
$$\Rightarrow \quad \|u_N - u\|_a \leq \|v_N - u\|_a \quad \forall v_N \in V_N \ .$$ □

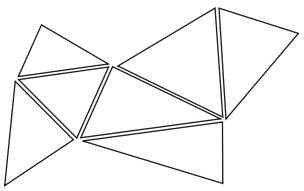
Problem. Wahl von V_N?

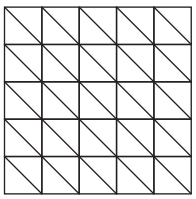
§ 5 Finite Elemente (erste Einführung)

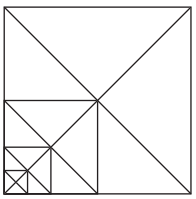
Finite-Elemente-Methode: Galerkin-Verfahren mit spezieller Wahl des Approximationsraums V_N. Ω sei Polygon (oder angenähert durch Polygon).

Triangulierung von Ω: Unterteile Ω in endlich viele Dreiecke („Finite Elemente“), sodass ihre Ecken andere Dreiecke nur wieder in Ecken berühren.

Nicht: 

Ja: 

Beispiele. 



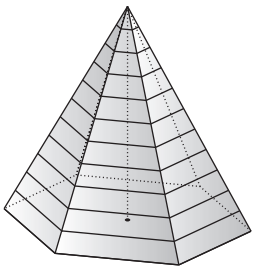
(lokale Verfeinerung möglich)

Notation: Dreiecke K^e, $(e = 1, \dots, E)$,

$$\overline{\Omega} = \bigcup_{e=1}^{E} K^e \ .$$

Ecken der Dreiecke im Inneren von Ω, nicht auf $\partial\Omega$:

$P_i = (x_i, y_i)$, $\quad i = 1, \dots, N$.



Wähle *Basisfunktionen* $\varphi_1, \dots, \varphi_N$ stückweise linear, stetig auf $\overline{\Omega}$ mit

$$\varphi_i(P_j) = \begin{cases} 1 \ , & i = j \ , \\ 0 \ , & \text{sonst} \ . \end{cases}$$

Beachte.

- $\varphi_i = 0$ auf allen Dreiecken, die den Knoten P_i nicht enthalten.
- $\varphi_i = 0$ auf Γ, damit ist erfüllt, dass $\varphi_i \in V$.

Wähle V_N als den von $\varphi_1, \dots, \varphi_N$ aufgespannten Vektorraum, $V_N \leq V$.
Suche Näherungslösung der Form

$$u_N = \sum_{i=1}^{N} \mu_i \varphi_i \ , \quad \text{beachte } u_N(P_j) = \mu_j \ , \quad u_N = 0 \text{ auf } \Gamma \ .$$

Galerkin-Verfahren aus (II.6): u_N erfülle

$$a(u_N, v_N) = l(v_N) \quad \forall v_N \in V_N$$

beziehungsweise äquivalent (siehe Satz 2, § 4) die μ_i erfüllen das lineare Gleichungssystem

$$A\mu = b \quad \text{mit} \quad A = \Big(a(\varphi_i, \varphi_j)\Big)_{i,j=1}^{N} \quad \text{„Steifigkeitsmatrix“, symmetrisch, positiv definit,}$$
$$b = \Big(l(\varphi_i)\Big)_{i=1}^{N} \quad \text{„Lastvektor“.}$$

Habe

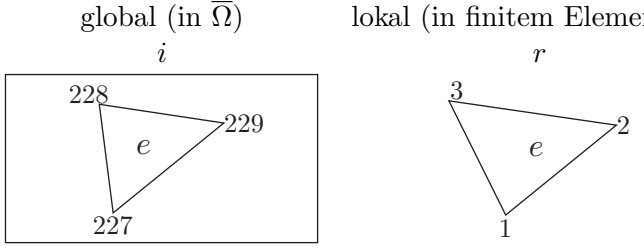
$$a_{ij} = a(\varphi_i, \varphi_j) = \int_\Omega \left(\frac{\partial \varphi_i}{\partial x}\frac{\partial \varphi_j}{\partial x} + \frac{\partial \varphi_i}{\partial y}\frac{\partial \varphi_j}{\partial y} \right) d(x,y)$$
$$= \sum_{\substack{\text{über alle Dreiecke } K^e \\ \text{die sowohl } P_i \\ \text{als auch } P_j \text{ enthalten}}} \int_{K^e} \underbrace{\big(\%\big)}_{\substack{=C^e \\ \text{konstant auf } K^e}} d(x,y) = \sum_{\%} C^e \text{ (Fläche von } K^e) \ .$$

Beachte. $a_{ij} \neq 0$ nur, wenn i, j Knoten eines gemeinsamen Dreiecks sind. Also ist A schwach besetzt.

$$b_i = l(\varphi_i) = \int_\Omega f\varphi_i \, d(x,y) = \sum_{\substack{\text{alle Dreiecke } K^e, \\ \text{die } P_i \text{ enthalten}}} \cdot \underbrace{\int_{K^e} f\varphi_i \, d(x,y)}_{\substack{\text{i. A. näherungsweise berechnet, z. B.} \\ \text{ersetze } f \text{ durch } \sum_j f(P_j)\varphi_j,\ j \text{ über} \\ \text{3 Eckpunkte, dann exakte Rechnung} \\ \text{des Integrals (quadratische Funktion)}}} \ .$$

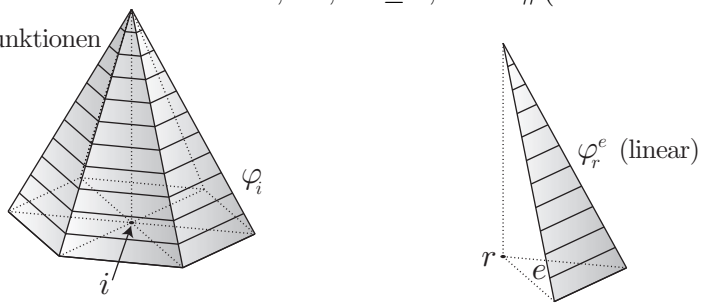
Rechnungsorganisation: elementweises Zusammensetzen

Knotennummerierung



global $\leftarrow$ lokal: $i^e(r)$ = globale Nummer des Knotens mit der lokalen Nummerierung r zu finitem Element e. $r = 1, \ldots, R^e \leq 3$, $R^e = \#$(innere Knoten von K^e).

Basisfunktionen



Einschränkung der globalen Basisfunktion auf ein finites Element = lokale Basisfunktion:

$$\varphi_i|_{K^e} = \varphi_r^e \quad \text{für } i = i^e(r) \ .$$

Steifigkeitsmatrix:

$$a_{ij} = \int_\Omega \left(\frac{\partial \varphi_i}{\partial x} \frac{\partial \varphi_j}{\partial x} + \frac{\partial \varphi_i}{\partial y} \frac{\partial \varphi_j}{\partial y} \right) d(x,y) \ , \quad a^e_{rs} = \int_{K^e} \left(\frac{\partial \varphi^e_r}{\partial x} \frac{\partial \varphi^e_s}{\partial x} + \frac{\partial \varphi^e_r}{\partial y} \frac{\partial \varphi^e_s}{\partial y} \right) d(x,y) \ .$$

Berechnung der Gesamtmatrix aus den Elementarmatrizen:

1. $a_{ij} = 0 \ \forall i,j = 1,\ldots,N$ vorbesetzt.
2. Für $e = 1,\ldots,E$:
   Für $r,s = 1,\ldots,R^e$ (≤ 3):
   Berechne a^e_{rs}.
   Setze $i = i^e(r)$, $j = i^e(s)$ (wegen Symmetrie: nur für $i \leq j$, $r \leq s$).
   Ersetze $a_{ij} := a_{ij} + a^e_{rs}$.

Lastvektor:

$$b_i = \int_\Omega f \varphi_i \, d(x,y) \ , \qquad b^e_r = \int_{K^e} f \varphi^e_r \, d(x,y) \ .$$

Berechnung des Gesamtvektors aus den Elementvektoren:

1. $b_i = 0$ vorbesetzen.
2. Für $e = 1,\ldots,E$:
   Für $r = 1,\ldots,R^e$ (≤ 3):
   Berechne b^e_r.
   Setze $i = i^e(r)$.
   Ersetze $b_i := b_i + b^e_r$.

Bleibt noch das *große* lineare Gleichungssystem $A\mu = b$ zu lösen.

Kapitel III

Variationelle Formulierung elliptischer Randwertprobleme

§ 1 Schwache Lösung, Lax-Milgram-Lemma

Erinnerung.

$$\begin{cases} -\Delta u = f & \text{in } \Omega \ , \\ u = 0 & \text{auf } \Gamma = \partial\Omega \ , \end{cases} \qquad \Omega \text{ stückweise } C^1,\ f : \overline{\Omega} \to \mathbb{R} \text{ stetig.}$$

Falls die „klassische“ Lösung $u \in C^2(\Omega) \cap C(\overline{\Omega})$ existiert, ist u die Lösung eines Variationsproblems. Mit

$$a(u,v) = \int_\Omega (u_x v_x + u_y v_y)\, d(x,y) \ , \qquad l(v) = \int_\Omega f v \, d(x,y)$$

schreibt sich das Variationsproblem zu

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

beziehungsweise äquivalent

$$J(u) = \min_{v \in V} J(v) \ , \qquad \text{wobei} \quad J(v) = \frac{1}{2} a(v,v) - l(v) \ .$$

Der Raum V kann auf verschiedene Weisen gewählt werden:

$$V = \{v \in C^2(\Omega) \cap C(\overline{\Omega}) | v = 0 \text{ auf } \partial\Omega\} \ , \qquad \text{(für klassische Lösungen)}$$

$$\text{oder} \quad V = \{v : \overline{\Omega} \to \mathbb{R} \text{ stückweise } C^1 | v = 0 \text{ auf } \partial\Omega\} \ ,$$

(sinnvoll für finite Elemente, $V_N \leq V$).

Situation: V Vektorraum mit Skalarprodukt $a(\cdot,\cdot)$, also ein Prähilbertraum. Zugehörige Norm

$$\|\cdot\|_a : \|v\|_a = \sqrt{a(v,v)} \ .$$

$l : V \to \mathbb{R}$ stetige Linearform ($|l(v)| \leq C\|v\|_a$) ,

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

Existiert $u \in V$ mit

$$J(u) = \min_{v \in V} J(v) \ ?$$

(Vergleiche: Existiert ein $x \in \mathbb{Q}$ mit $(x^2-2)^2 = \min?$)
Sei $(v_n)_{n\geq 0}$ Folge in V mit

$$J(v_n) \to \inf_{v\in V} J(v) =: C \quad \text{(existiert immer)}.$$

Konvergiert (v_n) gegen ein $u \in V$?
Zeige: (v_n) ist *Cauchy-Folge*:

$$\|v_n - v_m\|_a \to 0 \quad \text{für} \quad n, m \to \infty \ .$$

Weiß:

$$\underbrace{J(v_n) - J(v_m)}_{=:\mathcal{I}} \to 0 \ ,$$

$$\begin{aligned}
\mathcal{I} &= \frac{1}{2}a(v_n, v_n) - l(v_n) - \frac{1}{2}a(v_m, v_m) + l(v_m) \\
&= \frac{1}{4}a(v_n - v_m, v_n - v_m) + a\left(\frac{v_n + v_m}{2}, \frac{v_n + v_m}{2}\right) \\
&\qquad\qquad - a(v_m, v_m) - 2l\left(\frac{v_n + v_m}{2}\right) + 2l(v_m) \\
&= \frac{1}{4}\|v_n - v_m\|_a^2 + 2J\left(\frac{v_n + v_m}{2}\right) - 2J(v_m) \ .
\end{aligned}$$

Habe

$$2J\left(\frac{v+w}{2}\right) \leq J(v) + J(w)$$

wegen

$$\begin{aligned}
&a(v,v) + a(w,w) - 2a\left(\frac{v+w}{2}, \frac{v+w}{2}\right) + 0 = \\
&= \frac{1}{2}a(v,v) + \frac{1}{2}a(w,w) - 2 \cdot 2 \cdot \frac{1}{2} \cdot \frac{1}{2}a(v,w) = \frac{1}{2}a(v-w, v-w) \geq 0 \ .
\end{aligned}$$

Erhalte mit $\inf_{v\in V} J(v) =: C$

$$2C \overset{\text{klar}}{\underset{\downarrow}{\leq}} 2J\left(\frac{v_n + v_m}{2}\right) \leq J(v_n) + J(v_m) \to 2C \ ,$$

somit

$$J\big((v_n + v_m)/2\big) \to C \ .$$

Schließlich bleibt

$$\|v_n - v_m\|_a^2 \to 0 \ ,$$

d. h. (v_n) ist eine Cauchy-Folge.
Falls V Hilbertraum (vollständig: jede Cauchy-Folge konvergiert), dann:
(v_n) konvergiert gegen ein $u \in V$, Lösung des Variationsproblems.

Schwierigkeit: Im obigem Beispiel ist V mit $a(\cdot,\cdot)$ *kein* Hilbertraum!

Idee. Vervollständige V zu Hilbertraum $\overline{V}$ (analog $\mathbb{Q}$ zu $\mathbb{R}$: $(x^2-2)^2$ hat Minimum in $\sqrt{2} \in \mathbb{R}$).

Es gilt

Hilfssatz 1. *Sei V ein Prähilbertraum mit Skalarprodukt $a(\cdot,\cdot)$. Dann existiert ein (bis auf Isometrie eindeutiger) Hilbertraum $\overline{V}$ mit Skalarprodukt $\overline{a}(\cdot,\cdot)$, sodass V dicht in $\overline{V}$ und $\overline{a}|_{V\times V} = a$. (Schreibe wieder a statt $\overline{a}$).*

Beweis. Analog Vervollständigung von $\mathbb{Q}$ zu $\mathbb{R}$.
$\mathbb{R} = (\text{Cauchy-Folgen über } \mathbb{Q})/\sim$ mit Äquivalenzrelation $(x_n) \sim (y_n)$, falls

$$|x_n - y_n| \to 0 \ .$$

$\mathbb{Q} \overset{\text{identifiziert}}{=}$ Äquivalenzklassen konstanter Folgen: $\mathbb{Q} \ni x = \overline{(x,x,x,\dots)}$.

Hier: $\overline{V} = (\text{Cauchy-Folgen über } V)/\sim$ mit Äquivalenzrelation $(v_n) \sim (w_n)$, falls

$$\|v_n - w_n\|_a \to 0 \ .$$

$V \overset{\text{identifiziert}}{=}$ Äquivalenzklassen konstanter Folgen,
Erhalte V dicht in $\overline{V}$. Details nicht vorgeführt. □

Hilfssatz 2 (Fortsetzung durch Dichte)**.** *Sei $\overline{V}$ Hilbertraum, V dichter Teilraum von $\overline{V}$. Dann lässt sich $l : V \to \mathbb{R}$ stetig, linear (wobei statt $\mathbb{R}$ auch ein anderer vollständiger normierter Raum (Hilbertraum) stehen kann) in eindeutiger Weise zu stetigem, linearem $\overline{l} : \overline{V} \to \mathbb{R}$ fortsetzen (d. h. $\overline{l}|_V = l$). Schreibe wieder l statt $\overline{l}$.*

Beweis. Sei (v_n) Cauchy-Folge in $V \subset \overline{V}$,

$$v_n \to v \in \overline{V} \ .$$

Dann ist $l(v_n)$ Cauchy-Folge in $\mathbb{R}$ und es existiert $\lim_{n\to\infty} l(v_n)$ in $\mathbb{R}$.
Definiere

$$\overline{l}(v) := \lim_{n\to\infty} l(v_n) \ .$$

$\overline{l}(v)$ ist wohldefiniert (d. h. es hängt nicht von der Wahl der Cauchy-Folge (v_n) mit $v_n \to v$ ab): Falls auch $(w_n) \to v$, dann gilt

$$w_n - v_n \to 0 \quad \text{in} \quad \overline{V} \ .$$

$l(w_n - v_n) \to 0$ weil l stetig, also

$$\lim_{n\to\infty} l(v_n) = \lim_{n\to\infty} l(w_n) \ .$$

l linear und stetig. □

Oft zweckmäßig, eine zu $\|\cdot\|_a$ auf V äquivalente Norm $\|\cdot\|$ zu wählen: $\exists\gamma,\, C > 0$:

$$\gamma\|v\| \le \|v\|_a \le C\|v\| \ .$$

Beispiel. Betrachte statt der Poisson-Gleichung nun z. B. $7u_{xx}+5u_{yy} = f$ in Ω (ebenfalls elliptisch wie die Poisson-Gleichung). Also nun

$$a(u,v) = \int_\Omega (7u_xv_x + 5u_yv_y)\ d(x,y) \quad \text{statt} \quad a_1(u,v) = \int_\Omega (u_xv_x + u_yv_y)\ d(x,y) \ .$$

Habe $\|\cdot\|_{a,1} \sim \|\cdot\|_a$. Wegen dieser kleinen Unterschiede soll keine neue Norm gewählt werden müssen.

Beachte. V hat bezüglich $\|\cdot\|$ und $\|\cdot\|_a$ dieselbe Vervollständigung $\overline{V}$. Eine Cauchy-Folge bezüglich der einen Norm ist eine Cauchy-Folge bezüglich der anderen Norm.

Situation: $\overline{V}$ Hilbertraum mit Norm $\|\cdot\|$. $a : \overline{V} \times \overline{V} \to \mathbb{R}$ symmetrisch, bilinear mit

(i) $\exists\alpha > 0\ \forall v \in \overline{V} : a(v,v) \ge \alpha\|v\|^2$ und

(ii) $\exists M < \infty\ \forall u, v \in \overline{V} : |a(u,v)| \le M\|u\| \cdot \|v\|$.

Ein solches a heißt *$\overline{V}$-elliptisch.*

Bemerkung. Damit

$$\gamma\|v\| \le \|v\|_a \le C\|v\|$$

mit $\alpha = \gamma^2$, $M = C^2$. Die letzte Gleichheit gilt wegen

$$|a(u,v)| \le \|u\|_a \cdot \|v\|_a \ .$$

Im folgenden Satz schreibe V statt $\overline{V}$.

Satz 3 (Lax-Milgram)**.** *Sei V ein Hilbertraum und $a : V \times V \to \mathbb{R}$ eine V-elliptische Bilinearform, $l : V \to \mathbb{R}$ eine stetige Linearform. Dann hat das Variationsproblem*

$$J(v) := \frac{1}{2}a(v,v) - l(v) = \min!$$

beziehungsweise äquivalent $a(u,v) = l(v)\ \forall v \in V$ genau eine Lösung $u \in V$.

Beweis. Zeige zunächst die Existenz, die aus der Vollständigkeit folgt.
Sei (v_n) Minimalfolge

$$J(v_n) \to \inf_{v\in V} J(v) \ .$$

Gesehen: (v_n) Cauchy-Folge. $\overset{\text{(Vollständigkeit)}}{\Rightarrow}$ (v_n) konvergiert gegen ein $u \in V$

$$J(v_n) \to \inf_{v\in V} J(v)$$

J stetig (weil beschränkt) $\downarrow$

$$J(u) \quad \text{daher} \quad J(u) = \min_{v\in V} J(v) \ .$$

Zeige nun noch die Eindeutigkeit: Sei

$$\left.\begin{array}{ll} a(u_1,v)=l(v) & \forall v\in V \\ a(u_2,v)=l(v) & \forall v\in V \end{array}\right\rangle \quad \Rightarrow \quad a(u_1-u_2,v)=0\ \forall v\in V\ ,$$

insbesondere auch für $v=u_1-u_2$. Also:

$$\|u_1-u_2\|_a^2=0 \quad \Rightarrow \quad u_1=u_2\ . \qquad \square$$

Beispiel.

$$\left\{\begin{array}{rll} -\Delta u=f & \text{in} & \Omega\ , \\ u=0 & \text{auf} & \Gamma\ , \end{array}\right. \qquad \Omega \text{ stückweise } C^1, \quad f:\overline{\Omega}\to\mathbb{R} \text{ stetig.} \tag{III.1}$$

Klassische Lösung $u\in C^2(\Omega)\cap C(\overline{\Omega})$ existiert nicht immer! (ohne Beweis).
Lösung $u\in\overline{V}$ von

$$\frac{1}{2}\int_\Omega (v_x^2+v_y^2)\,d(x,y)-\int_\Omega fv\,d(x,y)=\min!$$

existiert immer (Lax-Milgram), heißt *schwache Lösung* von (III.1).

§ 2 Sobolev-Räume

Wie sehen die Hilberträume $\overline{V}$ aus, die sich aus der Vervollständigung der Räume zulässiger Funktionen bei elliptischen Randwertproblemen ergeben?
Sei $\Omega\subset\mathbb{R}^n$ stückweises C^1-Gebiet, beschränkt (Beweise oft nur für $n=2$).

§ 2.1 Der Raum $L^2(\Omega)$

$L^2(\Omega)$ = Raum der quadratischen Lebesgue-integrierbaren Funktionen auf Ω

$$L^2(\Omega)\ni f:\overline{\Omega}\to\mathbb{R} \quad \text{mit} \quad \int_\Omega f(x)^2\,dx<\infty$$

(Hierbei werden Funktionen f,g identifiziert ($f\sim g$), wenn sie sich nur auf einer Menge vom Maß 0 unterscheiden.)
Beachte $C(\overline{\Omega})\underset{\neq}{\subset} L^2(\Omega)$ z. B. $\log\in L^2(0,1)$.
Es gilt (ohne Beweis): $L^2(\Omega)$ ist ein *Hilbertraum* mit dem Skalarprodukt

$$(f,g)_0=\int_\Omega fg\,dx\ .$$

Zugehörige Norm:

$$\|f\|_0=\sqrt{(f,f)_0}=\left(\int_\Omega f(x)^2\,dx\right)^{\frac{1}{2}}\ .$$

Schreibe manchmal auch $\|f\|_{0,\Omega}$.
$C^\infty(\overline{\Omega})$ ist *dicht* in $L^2(\Omega)$ (bezüglich $\|\cdot\|_0$). Damit auch $C(\overline{\Omega})$.
Kann daher $L^2(\Omega)$ als Vervollständigung von $C^\infty(\overline{\Omega})$ (oder $C(\overline{\Omega})$) bezüglich $\|\cdot\|_0$ auffassen (siehe Hilfssatz 1, § 1).
Schreibe auch $H^0(\Omega) := L^2(\Omega)$: *Sobolev-Raum der Ordnung 0.*
Auf $\Gamma = \partial\Omega$ definiere analog $L^2(\Gamma)$ mit Skalarprodukt

$$(f,g)_{0,\Gamma} = \int_\Gamma f g \, d\sigma \ , \qquad \text{Norm } \|f\|_{0,\Gamma} = \left(\int_\Gamma f^2 \, d\sigma \right)^{\frac{1}{2}} \ .$$

$L^2(\Gamma)$ ist ein Hilbertraum mit der zugehörigen Norm $\|\cdot\|_{0,\Gamma}$.

§ 2.2 Der Raum $H^1(\Omega)$

Betrachte die Vervollständigung von $V = C^1(\overline{\Omega})$ (oder $C^\infty(\overline{\Omega})$) bezüglich

$$\|v\|_1^2 = \int_\Omega v^2 \, dx + \sum_{i=1}^n \int_\Omega \left(\frac{\partial v}{\partial x_i} \right)^2 dx \ .$$

Bezeichnung: $\overline{V} =: H^1(\Omega)$ *Sobolev-Raum der Ordnung 1.*
Wie sehen die Elemente von $H^1(\Omega)$ aus?
Sei (v_k) Cauchy-Folge bezüglich $\|\cdot\|_1$ in $C^\infty(\overline{\Omega})$, d. h.

$$\|v_k - v_l\|_1 \to 0 \quad \text{für } k,l \to \infty \ ,$$

d. h.

$$\|v_k - v_l\|_0 \to 0 \ , \qquad \left\| \frac{\partial v_k}{\partial x_i} - \frac{\partial v_l}{\partial x_i} \right\|_0 \to 0 \qquad \text{für } i = 1,\dots,n.$$

Wegen der Vollständigkeit von $L^2(\Omega)$ bezüglich $\|\cdot\|_0$

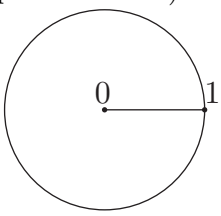
$$\exists v, v^{(1)}, \dots, v^{(n)} \in L^2(\Omega) : \begin{array}{c} \|v_k - v\|_0 \to 0 \\ \left\| \dfrac{\partial v_k}{\partial x_i} - v^{(i)} \right\|_0 \to 0 \end{array} \qquad k \to \infty \ , \qquad i = 1,\dots,n.$$

$v^{(i)}$ ist die verallgemeinerte partielle Ableitung von v, schreibe $v^{(i)} = \partial_i v$ oder auch wieder $\partial v / \partial x_i$.
Damit $H^1(\Omega)$ charakterisiert als

$$H^1(\Omega) = \{ \, v \in L^2(\Omega) \mid v \text{ hat verallgemeinerte partielle Ableitungen } \partial_i v \in L^2(\Omega) \text{ für } i = 1,\dots,n \, \} \ .$$

Man kann zeigen: Für $\Omega \subset \mathbb{R}^1$ ist $H^1(\Omega) \subset C(\overline{\Omega})$. Für $\Omega \subset \mathbb{R}^n$ mit $n \geq 2$ nicht richtig.

Beispiel (Gegenbeispiel für $n = 2$)**.**

$\Omega =$ 

, $\left|\log \underbrace{\sqrt{x_1^2 + x_2^2}}_{=r}\right|^{\frac{1}{4}} \in H^1(\Omega)$,

aber nicht stetig wegen der Singularität bei 0.

Wollen Randwertproblem lösen, aber: Rand hat Maß 0, für $f \in L^2(\Omega)$ ist $f|_\Gamma$ nicht sinnvoll definiert.
Dennoch: Für $v \in H^1(\Omega)$ lässt sich Einschränkung $v|_\Gamma$ (Spur von v auf Γ) wie folgt definieren:

Satz 1 (Spursatz)**.** *Die lineare Abbildung*

$$C^\infty(\overline{\Omega}) \to C(\Gamma) : v \mapsto v|_\Gamma$$

lässt sich in eindeutiger *Weise zu einer* stetigen *linearen Abbildung*

$$H^1(\Omega) \to L^2(\Gamma) : v \mapsto \gamma(v)$$

fortsetzen („Spur von v auf Γ"). Schreibe wieder $\gamma(v) =: v|_\Gamma$.

Einschub: Seien V, W normierte Vektorräume mit Normen $\|\cdot\|_V$, $\|\cdot\|_W$. $L : V \to W$ sei eine lineare Abbildung. Dann sind die folgenden Aussagen äquivalent:

(i) L ist stetig,

(ii) L ist stetig in 0,

(iii) L ist beschränkt: $\exists C < \infty$: $\forall v \in V$: $\|Lv\|_W \le C\|v\|_V$.

Beweis des Einschubs. Siehe Übungsaufgabe 25. □

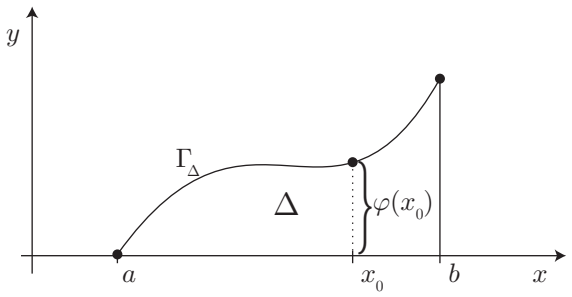
Beweis des Spursatzes.

(a) Zeige die Stetigkeit der linearen Abbildung also durch ihre Beschränktheit:

$$\|v|_\Gamma\|_{0,\Gamma} \le C \cdot \|v\|_{1,\Omega} \ \forall v \in C^\infty(\overline{\Omega}) \tag{III.2}$$

für ein $C < \infty$, das nur von Ω abhängt.
Ω stückweise C^1, beschränkt, kann in endlich viele Elementardreiecke zerlegt werden:

$$\Omega = \bigcup_{i=1}^{m} \Delta_i \ , \quad \Gamma = \bigcup_{i=1}^{m} \Gamma_{\Delta_i} \ .$$



Drehung und Verschiebung ändert

$$\int_{\Gamma_\Delta} v^2 \, d\sigma$$

nicht. Dann kann man ohne Einschränkung annehmen, dass die Strecke $\overline{ab}$ auf der x-Achse ($y = 0$) liegt.

Sei $(x, y) \in \big(x, \varphi(x)\big) \in \Gamma_\Delta$:

$$\begin{aligned}
v(x,y)^2 &= 1v(x,y)^2 - 0v(x,0)^2 = \int_0^1 \frac{\partial}{\partial t}\left(tv(x,ty)^2\right) dt \\
&= \int_0^1 \left(v(x,ty)^2 + t \cdot 2v(x,ty) \cdot \frac{\partial v}{\partial y}(x,ty) \cdot y \right) dt \\
&\overset{2ab \le a^2+b^2}{\le} \int_0^1 \left(v(x,ty)^2 + \left(v(x,ty)^2 + \frac{\partial v}{\partial y}(x,ty)^2 \right) ty \right) dt \\
&\le \int_0^1 \left(v(x,ty)^2 + \frac{\partial v}{\partial y}(x,ty)^2 + \left(v(x,ty)^2 + \frac{\partial v}{\partial y}(x,ty)^2 \right) ty \right) dt \\
&= \int_0^1 \Bigg(\left(v(x,ty)^2 + \frac{\partial v}{\partial y}(x,ty)^2 \right) \underbrace{(1+ty)}_{\le C_1(\Omega)} \Bigg) dt \\
&\le C_1(\Omega) \int_0^1 \left(v(x,ty)^2 + \frac{\partial v}{\partial y}(x,ty)^2 \right) dt \ .
\end{aligned}$$

Damit folgt

$$\begin{aligned}
\int_{\Gamma_\Delta} v^2 \, d\sigma &= \int_a^b v\big(x,\varphi(x)\big)^2 \underbrace{\sqrt{1+\varphi'(x)^2}}_{\le C_2} \, dx \le C_2 \int_a^b v\big(x,\varphi(x)\big)^2 \, dx \\
&\le C_2 C_1 \int_a^b \int_0^1 v\big(x, t\varphi(x)\big)^2 + \frac{\partial v}{\partial y}\big(x, t\varphi(x)\big)^2 \, dt \, dx \\
&= C_2 C_1 \int_\Delta \left(v^2 + \left(\frac{\partial v}{\partial y} \right)^2 \right) d(x,y) \ .
\end{aligned}$$

Nach Summation über Elementardreiecke:

$$\int_\Gamma v^2 \, d\sigma \le C(\Omega) \int_\Omega \left(v^2 + \Big(\underbrace{\frac{\partial v}{\partial x}}_{\substack{\text{durch} \\ \text{Drehungen}}} \Big)^2 + \left(\frac{\partial v}{\partial y} \right)^2 \right) d(x,y) \ .$$

Damit ist (III.2) gezeigt.

(b) *Fortsetzung durch Dichte* wie in Hilfssatz 2, § 1:

Da $C^\infty(\overline{\Omega})$ dicht in $H^1(\Omega)$, existiert zu beliebigem $v \in H^1(\Omega)$ Folge (v_k) in $C^\infty(\overline{\Omega})$ mit

$$\|v_k - v\|_1 \to 0 \ .$$

(v_k) ist Cauchy-Folge in $H^1(\Omega)$:

$$\|v_k - v_l\|_1 \to 0 \quad (k, l \to \infty) \ .$$

Wegen (III.2) gilt

$$\frac{1}{C}\|v_k|_\Gamma - v_l|_\Gamma\|_0 \overset{\text{(III.2)}}{\leq} \|v_k - v_l\|_1 \to 0 \ .$$

Dann folgt, dass $(v_k|_\Gamma)$ eine Cauchy-Folge in $L^2(\Gamma)$ ist:

$$\|v_k|_\Gamma - v_l|_\Gamma\|_{0,\Gamma} \to 0 \ .$$

$L^2(\Gamma)$ vollständig: $\exists w \in L^2(\Gamma)$ mit $\|v_k|_\Gamma - w\|_{0,\Gamma} \to 0$.

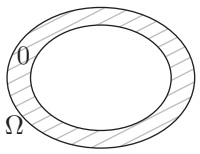
Definiere $\gamma(v) := w$. Zeige wie in Hilfssatz 2, § 1: $\gamma : H^1(\Omega) \to L^2(\Gamma)$ wohldefiniert, linear und stetig (nicht von Folge (v_k) abhängig). □

§ 2.3 Der Raum $H_0^1(\Omega)$

Nach Spursatz ist $v|_\Gamma$ wohldefiniert für $v \in H^1(\Omega)$, setze

$$H_0^1(\Omega) := \{v \in H^1(\Omega) | v|_\Gamma = 0\} \ .$$

Als Kern der Spurabbildung ist $H_0^1(\Omega)$ ein *abgeschlossener* Unterraum von $H^1(\Omega)$ und damit selbst ein Hilbertraum bezüglich $\|\cdot\|_1$.



Es gilt (ohne Beweis):

$$C_0^\infty(\overline{\Omega}) := \{v \in C^\infty(\Omega) | v = 0 \text{ in Umgebung von } \partial\Omega\}$$

liegt *dicht* in $H_0^1(\Omega)$ (bezüglich $\|\cdot\|_1$).
Kann somit $H_0^1(\Omega)$ auffassen als Vervollständigung von $V = C_0^\infty(\overline{\Omega})$, oder auch

$$V = \{v \text{ stückweise } C^1 | v = 0 \text{ auf } \Gamma\} \ , \qquad \text{siehe § 1} \ .$$

In § 1 hatten Vervollständigung bezüglich

$$|v|_1 = \left(\sum_{i=1}^n \int_\Omega \left(\frac{\partial v}{\partial x_i}\right)^2 dx\right)^{\frac{1}{2}} \ ,$$

beachte:

$$\|v\|_1^2 = \|v\|_0^2 + |v|_1^2 \ .$$

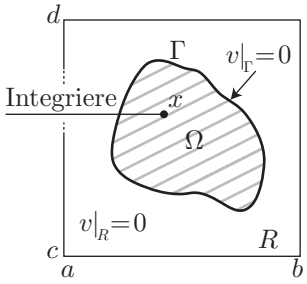
Der folgende Satz zeigt, dass $\|\cdot\|_1$ und $|\cdot|_1$ äquivalente Normen auf $H_0^1(\Omega)$ sind. H_0^1 ist daher der im Beispiel von § 1 betrachtete Hilbertraum $\overline{V}$.

Satz 2 (Poincaré-Ungleichung)**.** *Es existiert eine nur vom Gebiet Ω abhängige Konstante $C = C(\Omega) < \infty$, sodass*

$$\|v\|_0 \le C \cdot |v|_1 \quad \forall v \in H_0^1(\Omega) \ .$$

(Damit $|v|_1^2 \le \|v\|_1^2 = (|v|_1 + \|v\|_0)^2 \le (C^2+1)|v|_1^2$.)

Beweis.



Sei Ω in einem Rechteck $R = (a,b) \times (c,d)$ enthalten.
Zeige die Ungleichung für $v \in C_0^\infty(\overline{\Omega})$ (dicht in $H_0^1(\Omega)$).
Setze v durch 0 auf Rechteck fort: $v \in C_0^\infty(\overline{R})$.
Schreibe mit der Ableitung nach dem ersten Argument $\partial v/\partial x$

$$v(x,y) = \int_a^x 1 \cdot \frac{\partial v}{\partial x}(\xi, y)\, d\xi \quad (a \le x \le b) \ .$$

Dann folgt weiter

$$\begin{aligned} v(x,y)^2 &\overset{\text{Cauchy-Schwarzsche Ungleichung}}{\le} \int_a^x 1^2 d\xi \cdot \int_a^x \left(\frac{\partial v}{\partial x}(\xi,y)\right)^2 d\xi \\ &\le (b-a) \cdot \int_a^b \left(\frac{\partial v}{\partial x}(\xi,y)\right)^2 d\xi \quad \Big| \int_a^b \int_c^d \% \, dx\, dy \end{aligned}$$

Da die rechte Seite der Ungleichung unabhängig von x ist, kann das Integral über x einfach ausgeführt werden ($\int_a^b 1\, dx = (b-a)$):

$$\int_\Omega v^2 \, d(x,y) \le (b-a)^2 \int_c^d \int_a^b \left(\frac{\partial v}{\partial x}\right)^2 dx\, dy = (b-a)^2 \int_\Omega \left(\frac{\partial v}{\partial x}\right)^2 d(x,y) \ ,$$

damit

$$\|v\|_0 \le \underbrace{(b-a)}_{C} |v|_1 \ ,$$

zunächst für $v \in C_0^\infty(\overline{\Omega})$.
Sei nun $v \in H_0^1(\Omega)$. Sei weiterhin (v_n) Folge in $C_0^\infty(\overline{\Omega})$ mit $\|v_n - v\|_1 \to 0$. Dann folgt

$$\begin{aligned} & |v_n - v|_1 \to 0 \quad \text{und} \quad \|v_n - v\|_0 \to 0 \ , \\ \Rightarrow \quad & \|v_n\|_0 \le C \cdot |v_n|_1 \\ & \downarrow \qquad\qquad \downarrow \\ & \|v\|_0 \qquad\quad |v|_1 \ . \end{aligned}$$

□

§ 2.4 Sobolev-Räume höherer Ordnung

Sei $\Omega \subset \mathbb{R}^n$. Vervollständigung von $C^\infty(\overline{\Omega})$ bezüglich

$$\|v\|_m = \left(\sum_{|\alpha| \le m} \int_\Omega (\partial^\alpha v)^2 \, dx \right)^{\frac{1}{2}} .$$

wobei $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}_0^n$ Multi-Index,

$$|\alpha| = \sum_{i=1}^n \alpha_i \ , \quad \partial^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}} \ .$$

Bezeichnung: $H^m(\Omega)$.
Charakterisiert als

$$H^m(\Omega) = \{ v \in L^2(\Omega) | \partial^\alpha v \in L^2(\Omega) \text{ für } |\alpha| \le m \} \ ,$$

$\partial^\alpha v$ ist verallgemeinerte partielle Ableitung.
Schreibe $\|v\|_m^2 = \|v\|_0^2 + |v|_1^2 + |v|_2^2 + \cdots + |v|_m^2$ mit

$$|v|_k^2 = \sum_{|\alpha| = k} \int_\Omega (\partial^\alpha v)^2 \, dx \ .$$

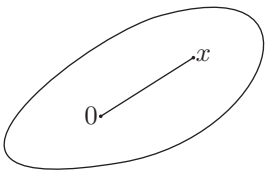
§ 2.5 Sobolev Einbettungssätze

Sei $\Omega \subset \mathbb{R}^n$.

$$\begin{array}{ll} H^m(\Omega) \subset C^{m-1}(\overline{\Omega}) & n = 1 \\ H^m(\Omega) \subset C^{m-2}(\overline{\Omega}) & n = 2, 3 \end{array} \quad \text{(ohne Beweis)}$$

mit stetigen Einbettungen (Inklusionen). (Dabei ist die Norm auf $C^k(\overline{\Omega})$ gegeben durch

$$\max_{|\alpha| \le k} \max_{\Omega} |\partial^\alpha v| \ .)$$



Zeige $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$ stetig für $n = 2, 3$ (dabei bedeutet $\hookrightarrow$ injektive Abbildung).
Genauer: Die Identität $I : C^\infty(\overline{\Omega}) \to C^\infty(\overline{\Omega})$ setzt sich durch die Dichte zu einer stetigen Abbildung $H^2(\Omega) \to C(\overline{\Omega})$ fort.
Für $v \in H^2(\Omega)$ zeige Stetigkeit von v in $x_0 \in \overline{\Omega}$. Sei ohne Einschränkung $x_0 = 0$. Weiterhin (Stetigkeit): lokal ohne Einschränkung $\overline{\Omega}$ konvex.
Für $v \in C^\infty(\overline{\Omega})$ ist

$$v(x) = v(0) + \int_0^1 \underbrace{1}_{\uparrow} \cdot \underbrace{\frac{d}{dt} v(tx)}_{\downarrow} \, dt = v(0) + Dv(x) \cdot x - \int_0^1 t \cdot \underbrace{\frac{d^2}{dt^2} v(tx)}_{D^2 v(tx) \cdot (x,x) = x^T \underbrace{\nabla^2 v(tx)}_{\text{Hesse-Matrix}} x} \, dt \ .$$

Sei B eine Kugel um 0. Habe

$$\int_B \left(\int_0^1 \underbrace{t}_{t^{-\alpha}\cdot t^{1+\alpha}} \cdot \quad D^2 v(tx)\cdot(x,x)\,dt \right)^2 dx$$

$$\overset{\substack{\text{Cauchy-Schwarzsche}\\ \text{Ungleichung}\\ \downarrow}}{\leq} \int_B \underbrace{\int_0^1 t^{-2\alpha}\,dt}_{\substack{=C_0<\infty\\ \text{für } \alpha<\frac{1}{2}}} \cdot \int_0^1 t^{2(1+\alpha)} \overbrace{\underbrace{|x|^4}_{\text{euklidische Norm}}}^{\leq \text{const}} \cdot |\nabla^2 v(\underbrace{tx}_{\xi\in B})|^2\,dt\,dx$$

Setze $\xi = tx$, $d\xi = t^n\,dx$:

$$\leq C \cdot \underbrace{\int_0^1 t^{2(1+\alpha)-n}\,dt}_{=C'<\infty \text{ für } 2(1+\alpha)-n>-1\ ,\ \text{d. h. } 2\alpha>n-3} \cdot \int_B |\nabla^2 v(\xi)|^2\,d\xi \leq \tilde{C}|v|_2^2\ .$$

Beispielsweise $\alpha = 0$ für $n = 1, 2$, $\alpha = \frac{1}{4}$ für $n = 3$.
Erhalte dann für $v(0)$ aus der ursprünglichen Gleichung:

$$v(0) = +v(x) - Dv(x)\cdot x + \int_0^1 t\cdot x^T \nabla^2 v(tx)x\,dt \qquad \Big|^2 \int_B\ ,$$

mit Obigem:

$$v(0)^2 \cdot \int_B dx \leq C\|v\|_2^2\ , \quad \text{für } v\in C^\infty(\overline{\Omega})\ .$$

Daraus folgt $|v(0)| \leq \tilde{C}\|v\|_2^2$.
Damit auch (0 beliebig):

$$\underbrace{\max_{x\in\Omega}|v(x)|}_{=\|v\|_{\tilde{C}(\overline{\Omega})}} \leq \tilde{C}\|v\|_{H^2}\ .$$

$C^\infty(\overline{\Omega})$ dicht in $H^2(\Omega)$: Sei $v\in H^2(\Omega)$, (v_n) in $C^\infty(\overline{\Omega})$ mit $\|v_n - v\|_2 \to 0$.

$$\|v_n - v_m\|_{C(\overline{\Omega})} \leq C\|v_n - v_m\|_2 \to 0\ .$$

$C(\overline{\Omega})$ vollständig:

$$\exists w\in C(\overline{\Omega}) : v_n \to w \text{ in } C(\overline{\Omega})\ , \quad v = w \text{ fast überall (bis auf Nullmenge)}\ , \quad \|w\|_{C(\overline{\Omega})}\ .$$

$$\|v_n\|_{C(\overline{\Omega})} \leq C\cdot \underbrace{\|v_n\|_2}_{\to\|v\|_2=\|w\|_2}\ .$$

Wegen $\|v_n - v\|_0 \to 0$ mit $\|w - v\|_0$ folgt $\int (w-v)^2\,dx = 0$.
$\|w\|_{C(\overline{\Omega})} \leq C\|w\|_2$.
(Wie beim Spursatz): $H^2(\Omega) \hookrightarrow C(\overline{\Omega})$ stetig, linear.

§ 3 Elliptische Randwertprobleme der Ordnung 2

Anwendungen: stationäre Feldprobleme in der Physik (elektrische, magnetische Feldstärke, Temperaturverteilung, stationäre Strömung durch poröse Medien, ...).
In homogenem, isotropen Medium: $-\Delta u = f$, ansonsten Probleme der Gestalt

$$Au = f \quad \text{in } \Omega\ , \quad \Omega \subset \mathbb{R}^n \text{ stückweises } C^1\text{-Gebiet}\ , \quad n = 2, 3\ , \quad f \in L^2(\Omega)$$
$$+\ \text{Randbedingungen auf } \Gamma = \partial\Omega\ ,$$

wobei A elliptischer Differentialoperator 2. Ordnung mit variablen Koeffizienten:

$$Au = -\sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) + a_0 u$$

mit a_{ij}, $a_0 : \Omega \to \mathbb{R}$ *beschränkt*:

$$|a_0(x)|\ , \quad |a_{ij}(x)| \le M\ \forall x \in \Omega\ ,$$

symmetrisch:

$$a_{ij}(x) = a_{ji}(x)\ \forall i, j = 1, \ldots, n \quad \forall x \in \Omega\ ,$$

elliptisch:

$$\begin{aligned} &\exists \alpha_1 > 0 : \forall x \in \Omega \text{ und } \forall \xi \in \mathbb{R}^n : \\ &\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \ge \alpha_1 \sum_{i=1}^{n} \xi_i^2\ , \\ &\exists \underset{\ge}{\alpha_0} 0\ \forall x \in \Omega \quad a_0(x) \ge \alpha_0\ . \end{aligned}$$

§ 3.1 Homogenes Dirichlet-Problem

$$\begin{cases} Au = f & \text{in } \Omega\ , \\ u = 0 & \text{auf } \Gamma\ . \end{cases}$$

Falls die klassische Lösung $u \in C^2(\Omega) \cap C(\overline{\Omega})$ existiert, multipliziere mit $v \in C_0^\infty$, integriere über Ω und erhalte mit der Greenschen Formel

$$\underbrace{\int_\Omega \left(\sum_{i,j=1}^{n} a_{ij}\frac{\partial u}{\partial x_j}\frac{\partial v}{\partial x_i} + a_0 uv\right) dx}_{a(u,v)} = \underbrace{\int_\Omega f \cdot v\, dx}_{l(v)} \quad \forall v \in C_0^\infty(\Omega)\ .$$

Variationelle Formulierung: Lösungsraum $V = H_0^1(\Omega)$ (Vervollständigung von $C_0^\infty(\Omega)$).
Suche $u \in H_0^1(\Omega)$, sodass $a(u,v) = l(v)\ \forall v \in H_0^1(\Omega)$,

Folgerung. Das homogene Dirichlet-Problem hat (in seiner variationellen Formulierung) genau *eine* Lösung $u \in H_0^1(\Omega)$.

Beweis. Weise nach, dass die Voraussetzungen des Satzes von Lax-Milgram erfüllt sind.

- l ist stetige Linearform auf $H^1(\Omega)$ (damit auch auf $H_0^1(\Omega)$), denn

$$|l(v)| = \left|\int_\Omega f v\,dx\right| = |(f,v)_0| \overset{\substack{\text{Cauchy-Schwarzsche}\\ \text{Ungleichung}}}{\underset{\downarrow}{\leq}} \underbrace{\|f\|_0}_{\leq C} \cdot \underbrace{\|v\|_0}_{\leq \|v\|_1} \quad \forall v \in H^1(\Omega)\ .$$

  Laut dem Einschub in § 2 folgt aus der Beschränktheit die Stetigkeit einer linearen Abbildung.

- a ist beschränkte Bilinearform auf $H^1(\Omega) \times H^1(\Omega)$:

$$|a(u,v)| \leq \int_\Omega \left(\sum_{i,j=1}^n \underbrace{|a_{ij}|}_{\leq M} \cdot \left|\frac{\partial u}{\partial x_j}\right| \cdot \left|\frac{\partial v}{\partial x_i}\right| + \underbrace{|a_0|}_{\leq M} \cdot |u| \cdot |v| \right) dx$$

$$\overset{\substack{\text{Cauchy-Schwarzsche}\\ \text{Ungleichung}}}{\underset{\downarrow}{\leq}} M \left(\sum_{i,j=1}^n \underbrace{\left\|\frac{\partial u}{\partial x_j}\right\|_0}_{\leq \|u\|_1} \cdot \underbrace{\left\|\frac{\partial v}{\partial x_i}\right\|_0}_{\leq \|v\|_1} + \underbrace{\|u\|_0}_{\leq \|u\|_1} \cdot \underbrace{\|v\|_0}_{\leq \|v\|_1} \right) \leq M(n^2+1)\|u\|_1 \cdot \|v\|_1\ .$$

  a ist $H_0^1(\Omega)$-elliptisch (im Sinne von § 1):

$$a(v,v) = \int_\Omega \left(\sum_{i,j=1}^n a_{ij} \underbrace{\frac{\partial v}{\partial x_j}}_{\xi_j} \underbrace{\frac{\partial v}{\partial x_i}}_{\xi_i} + a_0 v^2 \right) dx$$

$$\overset{\substack{\text{Elliptizität}\\ \text{von } A}}{\underset{\downarrow}{\geq}} \int_\Omega \left(\alpha_1 \sum_{i=1}^n \left(\frac{\partial v}{\partial x_i}\right)^2 + \alpha_0 v^2 \right) dx$$

$$= \underbrace{\alpha_1}_{>0} |v|_1^2 + \underbrace{\alpha_0}_{\geq 0} \|v\|_0^2 \overset{\substack{\text{Poincaré-Ungleichung}\\ \downarrow}}{\geq} \alpha \|v\|_1^2 \qquad \text{für ein } \alpha > 0\ .$$

Damit sind alle Voraussetzungen des Satzes von Lax-Milgram erfüllt und es folgt die Behauptung. □

Bemerkung. Die Randbedingung ($u = 0$ auf Γ) geht hier in den Lösungsraum $H_0^1(\Omega) = \{v \in H^1(\Omega) : v|_\Gamma = 0\}$ ein („wesentliche Randbedingung“).

§ 3.2 Homogenes Neumann-Problem

Betrachte dasselbe Variationsproblem wie in § 3.1, aber jetzt auf ganz $H^1(\Omega)$:
Suche $u \in H^1(\Omega)$, sodass $a(u,v) = l(v)\ \forall v \in H^1(\Omega)$ mit a, l wie oben. Wegen

$$a(v,v) \geq \alpha_1 |v|_1^2 + \alpha_0 \|v\|_0^2 \geq \min(\alpha_0, \underbrace{\alpha_1}_{>0}) \|v\|_1^2$$

ist a $H^1(\Omega)$-elliptisch, falls $\alpha_0 > 0$.
Die Poincaré-Ungleichung (§ 3, Satz 2) gilt nur für $v \in H_0^1(\Omega)$, nicht aber für $v \in H^1(\Omega$! Deshalb wirklich nur falls $\alpha_0 > 0$.

Folgerung. Es existiert genau eine Lösung $u \in H^1(\Omega)$.

Welche Randbedingungen entsprechen diesem Variationsproblem? Sei u hinreichend regulär, z. B. $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$, verwende die Greensche Formel:

$$\begin{aligned}\int_\Omega f v \, dx &= \int_\Omega \Big(\sum_{i,j=1}^n a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + a_0 uv \Big) \, dx \\ &= \int_\Omega \underbrace{\Big(-\sum_{i,j} \frac{\partial}{\partial x_i} \Big(a_{ij} \frac{\partial u}{\partial x_j} \Big) + a_0 u \Big)}_{=Au} v \, dx + \int_\Gamma \underbrace{\sum_{i,j=1}^n \Big(a_{ij} \frac{\partial u}{\partial x_j} n_i \Big)}_{=: \frac{\partial u}{\partial n_A}} v \, d\sigma\end{aligned}$$

„Konormalen-Ableitung zu A“

mit $n = (n_i)_{i=1}^n$ äußere Normale.
(Falls $A = -\Delta$, ist

$$\frac{\partial u}{\partial n_A} = \sum_{i=1}^n \frac{\partial u}{\partial x_i} n_i = \frac{\partial u}{\partial n}$$

Normalen-Ableitung.)

Folgerung. Sei u Lösung des Variationsproblems und zudem $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$, so gilt

$$\begin{aligned} Au &= f \qquad \text{in } \Omega\ , \\ \frac{\partial u}{\partial n_A} &= 0 \qquad \text{auf } \Gamma \end{aligned} \qquad \text{(im klassischen Sinn).}$$

Beweis. Wähle zuerst v mit $v = 0$ auf Γ mit n-dimensionalem Fundamentallemma:

$$Au = f \quad \text{in } \Omega.$$

Dann v beliebig auf Γ, mit $(n-1)$-dimensionalem Fundamentallemma:

$$\frac{\partial u}{\partial n_A} = 0 \quad \text{auf } \Gamma\ . \qquad \square$$

Bemerkung. Die Randbedingungen treten in der variationellen Formulierung nicht auf. Man nennt sie deshalb *natürliche Randbedingungen.*

$\frac{\partial u}{\partial n_A}$ auf Γ ist für allgemeine $u \in H^1(\Omega)$ (ohne zusätzliche Regularität) nicht definiert.

§ 3.3 Inhomogenes Dirichlet-Problem

Sei $g : \Gamma \to \mathbb{R}$ so, dass $g = u_0|_\Gamma$ für ein $u_0 \in H^1(\Omega)$ (ist z. B. erfüllt, falls g stückweise C^1, ohne Beweis).
Löse

$$\begin{cases} Au = f & \text{in } \Omega \ , \\ u = g & \text{auf } \Gamma \ . \end{cases}$$

Variationelle Formulierung: Suche $u \in H^1(\Omega)$ mit $u - u_0 \in H_0^1(\Omega)$, sodass $a(u,v) = l(v)\ \forall v \in H_0^1(\Omega)$

Folgerung. Es existiert genau eine Lösung $u \in H^1(\Omega)$.

Beweis. Setze $u = u_0 + w$. Äquivalentes Problem für w:

$$\begin{cases} \text{Suche } w \in H_0^1(\Omega), \text{ sodass} \\ a(w,v) = \underbrace{l(v) - a(u_0,v)}_{=:\tilde{l}(v)} \quad \forall v \in H_0^1(\Omega) \ . \end{cases} \tag{III.3}$$

Lineare Abbildung $H_0^1(\Omega) \to \mathbb{R} \ : \ v \mapsto a(u_0,v)$ ist stetig:

$$|a(u_0,v)| \le \underbrace{M(n^2+1)\|u_0\|_1}_{\text{const}} \cdot \|v\|_1 \qquad \text{(früher gezeigt).}$$

Wende Lax-Milgram auf (III.3) an. □

§ 3.4 Inhomogenes Neumann-Problem

Suche $u \in H^1(\Omega)$, sodass

$$a(u,v) = \underbrace{\int_\Omega fv\,dx + \int_\Gamma gv d\sigma}_{=:l(v)} \quad \forall v \in H^1(\Omega)$$

mit $a(\cdot,\cdot)$, f wie zuvor, $g \in L^2(\Gamma)$.
Linearform l ist stetig auf $H^1(\Omega)$, denn:

$$\Big|\int_\Gamma gv\,d\sigma\Big| \overset{\substack{\text{Cauchy-Schwarzsche}\\ \text{Ungleichung}}}{\downarrow \le} \|g\|_{0,\Gamma} \cdot \|v|_\Gamma\|_{0,\Gamma} \overset{\text{Spursatz}}{\downarrow \le} \|g\|_{0,\Gamma} \cdot C \cdot \|v\|_{1,\Omega} \ .$$

Unter Voraussetzung von § 3.2 ($\alpha_0 > 0$) erhalte mit Lax-Milgram:
Folgerung. Es existiert genau eine Lösung $u \in H^1(\Omega)$.
Unter der Regularitätsannahme $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ erhalte wie in § 3.2:

$$Au = f \quad \text{in } \Omega \ , \qquad \frac{\partial u}{\partial n_A} = g \quad \text{auf } \Gamma \ .$$

Bemerkung. Die Randbedingung geht hier in die Linearform l der variationellen Formulierung ein.

Kapitel IV

Finite-Elemente-Approximation

Erinnerung. Elliptisches Randwertproblem in variationeller Formulierung:
Suche $u \in V$, sodass

$$a(u,v) = l(v) \quad \forall v \in V \quad \big(V = H^1(\Omega) \text{ oder } H_0^1(\Omega)\big) \ .$$

Soll mittels Galerkin-Verfahren approximiert werden:
Suche $u_N \in V_N$, sodass

$$a(u_N, v_N) = l(v_N) \quad \forall v_N \in V_N \quad (V_N \leq V \quad N\text{-dimensionaler Unterraum}) \ .$$

Praktische Wahl von V_N?
Finite Elemente: V_N ist der Raum der stückweisen Polynome.

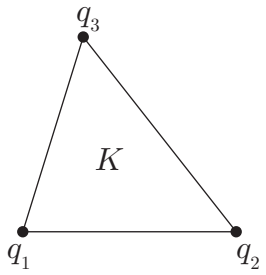
§ 1 Finite Elemente

Kenne bereits spezielle finite Elemente:

Dreiecke mit Knoten (Ecken) q_1, q_2, q_3.

Dazu Basisfunktionen $\varphi_1, \varphi_2, \varphi_3$ linear mit

$$\varphi_r(q_s) = \delta_{rs} \ .$$



Allgemeiner:

Definition. Ein *finites Element* (f. E.) ist eine kompakte, zusammenhängende Teilmenge $K \subset \mathbb{R}^n$, zu der das Folgende gegeben ist:

- Knotenpunkte $q_1, \ldots, q_R \in K$.
- P endlichdimensionaler Vektorraum bestehend aus Polynomfunktionen $p : K \to \mathbb{R}$, sodass
$$\forall c_1, \ldots, c_R \in \mathbb{R} \ \exists_1 p \in P \ : \ p(q_r) = c_r \quad (r = 1, \ldots, R)$$
d. h., $p \in P$ ist durch seine Werte in den Knotenpunkten eindeutig bestimmt.

Anders: Die Abbildung

$$P \cong \mathbb{R}^R \ : \ p \mapsto \big(p(q_r)\big)_{r=1}^R$$

ist ein Isomorphismus.

Notwendig: $\dim P = R = \#(\text{Knotenpunkte})$
Es existieren damit Basisfunktionen $\varphi_1, \ldots, \varphi_R \in P$ mit $\varphi_r(q_s) = \delta_{rs} \quad (r, s = 1, \ldots, R)$.
Jedes Polynom $p \in P$ lässt sich eindeutig darstellen als

$$p = \sum_{r=1}^{R} p(q_r)\varphi_r \ .$$

$$\left(\text{Denn:} \qquad p = \sum_{r=1}^{R} \alpha_r \varphi_r \quad (\exists_1 \alpha_r, \text{ da } \varphi_r \text{ Basis von } P) \ , \right.$$

$$\left. p(q_s) = \sum_{r=1}^{R} \alpha_r \varphi_r(q_s) = \alpha_s \ . \right)$$

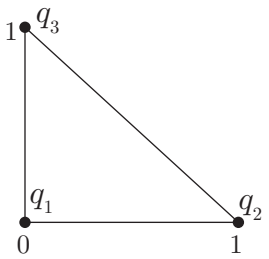
§ 1.1 Wichtige Beispiele in Dimension 2

Dreieckselemente

- Lineare Polynome:

  $P = P_1 = \text{Raum der Polynome vom Grad 1} \ ,$
  $P_1 \ni \alpha + \beta x + \gamma y \ ,$
  $\dim(P_1) = 3 \ ,$

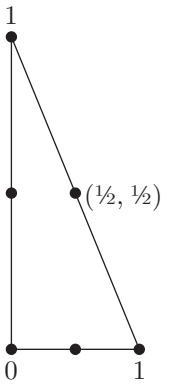
  Basisfunktionen: $x, \ y, \ 1 - x - y$.



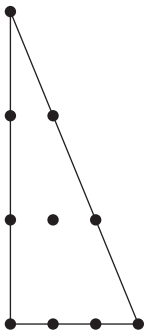
- Quadratische Polynome:

  $P = P_2 = \text{Raum der Polynome vom Grad 2} \ ,$
  $P_2 \ni \alpha + \beta x + \gamma y + \delta x^2 + \epsilon xy + \mu y^2 \ ,$
  $\dim(P_2) = 6 \ ,$

  Basisfunktionen: $2x(x - \frac{1}{2})$, $4xy$, $\ldots$.



- Kubische Polynome:

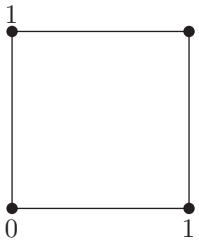


$$P = P_3 = \text{Raum der Polynome vom Grad 3 ,}$$
$$P_3 \ni \alpha + \beta x + \gamma y + \delta x^2 + \epsilon xy + \mu y^2 + \nu x^3 + \eta x^2 y + \lambda x y^2 + \kappa y^3 \ ,$$
$$\dim(P_3) = 10 \ ,$$

Basisfunktionen: ... .

Rechteckselemente

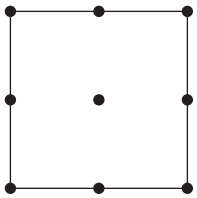
- Bilineare Polynome:



$P = Q_1 =$ Raum der Polynome, die bezüglich jeder einzelnen Variablen x, y linear sind, also
$Q_1 \ni \alpha + \beta x + \gamma y + \delta xy$,
$\dim(Q_1) = 4$,

Basisfunktionen: $xy,\ x(1-y),\ (1-x)y,\ (1-x)(1-y)$ (auf Einheitsquadrat).

- Biquadratische Polynome:



$P = Q_2 =$ Grad $\leq$ 2 bezüglich jeder Variablen, also
$Q_2 \ni \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 xy + \alpha_5 x^2 + \alpha_6 y^2 + \alpha_7 x^2 y + \alpha_8 x y^2 + \alpha_9 x^2 y^2$,
$\dim(Q_2) = 9$,

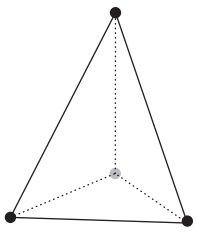
Basisfunktionen: ... .

- Bikubische Polynome (gleiches Vorgehen ...).

§ 1.2 Wichtige Beispiele in Dimension 3

Tetraeder

- Linear:

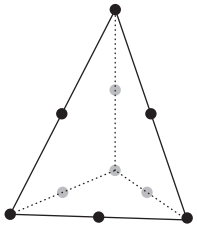


$$P = P_1 \ni \alpha + \beta x + \gamma y + \delta z \ ,$$
$$\dim(P_1) = 4 \ ,$$

Basisfunktionen: ... .

- Quadratisch:

$$P = P_2 \ni \dots \ ,$$
$$\dim(P_2) = 10 \ ,$$

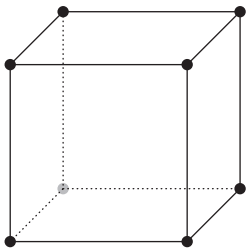


Basisfunktionen: $\dots$.

Quader

- Trilinear:

$$P = Q_1 \ni \alpha + \beta x + \gamma y + \delta z + \\ + \epsilon xy + \mu yz + \xi xz + \eta xyz \ ,$$
$$\dim(Q_1) = 8 \ ,$$



Basisfunktionen: $\dots$.

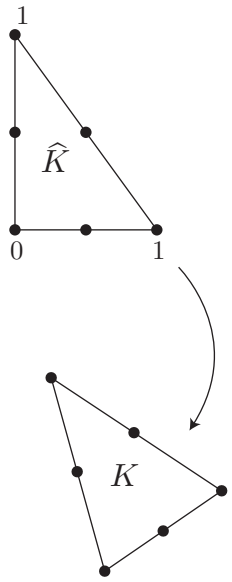
§ 1.3 Transformation von finiten Elementen

Möchte ausgehend vom Referenzelement $\hat{K}$ weitere finite Elemente K erzeugen:

Sei $\hat{K}$ finites Element mit Knoten $\hat{q}_r$ und Basisfunktionen $\hat{\varphi}_r$ und sei

$$F : \hat{K} \to K$$

bijektiv.



Erhalte damit

$$\text{Knoten auf } K\text{:} \quad q_r = F(\hat{q}_r) \ ,$$
$$\text{Basisfunktionen auf } K\text{:} \quad \varphi_r = \hat{\varphi}_r \circ F^{-1} \ .$$

Habe dann

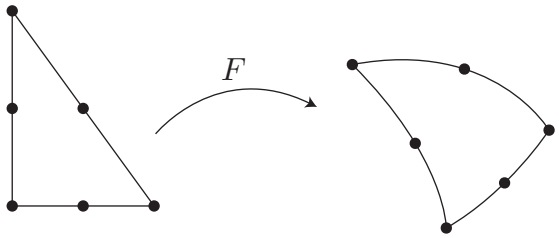
$$\varphi_r(q_s) = \left(\hat{\varphi}_r \circ F^{-1}\right)(F(\hat{q}_s)) = \hat{\varphi}_r(\hat{q}_s) = \delta_{rs} \ .$$

Damit ist K wieder finites Element (falls φ_r Polynome).

Praktische Wahl der Transformation F:

- F *affin*: $F(\hat{x}) = B\hat{x} + b$, B invertierbar.
- F *isoparametrisch*: Gebe Knoten q_r vor, setze

$$F(\hat{x}) := \sum_{r=1}^{R} q_r \hat{\varphi}_r(\hat{x}) \ , \qquad \text{habe } F(\hat{q}_s) = q_s \ .$$



Bijektiv, falls Verzerrungen nicht allzu groß.

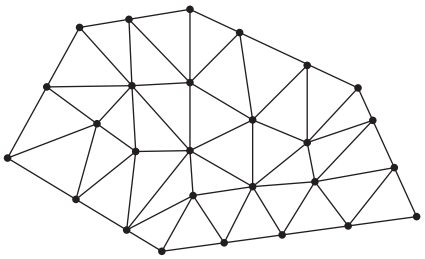
Vorteil: größere geometrische Flexibilität, insbesondere Gebiete Ω mit krummlinigem Rand, rechnerisch ohne Mehraufwand.

§ 2 Zusammensetzen von finiten Elementen, globale Basisfunktionen

Gegeben:

$$\begin{aligned} \text{Finites Element} \quad & K^e \quad (e = 1, \ldots, E) \\ \text{mit Knoten} \quad & Q^e = (q_1^e, \ldots, q_{R^e}^e) \text{ und} \\ \text{Polynomräume} \quad & P^e, \quad \text{Basisfunktionen } \varphi_1^e, \ldots, \varphi_{R^e}^e \, . \end{aligned}$$

Triangulierung von Ω:



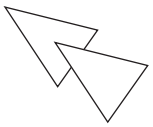
Zerlegung von

$$\overline{\Omega} = \bigcup_{e=1}^{E} K^e \, ,$$

wobei

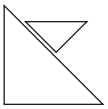
(i) $\overset{\circ}{K}{}^{e_1} \cap \overset{\circ}{K}{}^{e_2} = \emptyset$

Nicht:



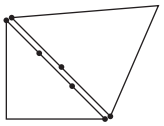
(ii) und jede Seite (Fläche) eines finiten Elements ist entweder Seite eines anderen finiten Elements oder ein Teil von $\partial\Omega$.

Nicht:



Knoten
Kompatibilitätsbedingung: Knoten stimmen auf gemeinsamen Seiten überein.

Nicht:



$Q^{e_1} \cap K' = Q^{e_2} \cap K' \quad \text{für} \quad K' = K^{e_1} \cap K^{e_2}$.

Globale Knotenmenge:

$$Q = \{q_1, \ldots, q_I\} = \bigcup_{e=1}^{E} Q^e \ .$$

Globale Basisfunktionen:

$$\varphi_i : \overline{\Omega} \to \mathbb{R} \ ,$$

definiert durch $\varphi_i\big|_{K^e} = \varphi_r^e$, falls $q_i = q_r^e$ (sonst 0).

Eigenschaften:

$\varphi_i\big|_{K^e}$ Polynom in P^e ,
$\varphi_i(q_j) = \delta_{ij}$,
φ_i haben „kleinen Träger“:
$\varphi_i(x) \neq 0$ nur, falls x und q_i in gemeinsamem finiten Element liegen.

Ohne eine weitere Zusatzannahme sind die φ_i auf gemeinsamen Seiten $K' = K^{e_1} \cap K^{e_2}$ nicht wohldefiniert, möchte aber $\varphi_i : \overline{\Omega} \to \mathbb{R}$ stetig. (Damit $\varphi_i \in H^1(\Omega)$, denn das ist wichtig für die Konvergenz.)

Kompatibilitätsbedingung: Einschränkungen der Polynomräume stimmen auf gemeinsamen Seiten überein, d. h.

$$P' := P^{e_1}\big|_{K'} = P^{e_2}\big|_{K'} \qquad \text{für} \quad K' = K^{e_1} \cap K^{e_2} \ .$$

Es gelte die folgende Interpolationseigenschaft: Polynome in P' sind durch die Werte in den auf K' liegenden Knoten $Q' = Q^e \cap K'$ eindeutig bestimmt.
(Anders: K' ist mit P' und Q' ein finites Element niedrigerer Dimension.)
Damit:

$$\varphi_i : \overline{\Omega} \to \mathbb{R} \quad \text{stetig.}$$

Der Finite-Elemente-Raum:

$$P = \langle \varphi_1, \ldots, \varphi_I \rangle \ ,$$
$$P \ni v = \sum_{i=1}^{I} v(q_i)\varphi_i \qquad \text{stetig, stückweise polynomiell.}$$

§ 3 Aufstellen des Galerkin-Systems

Gegeben: Elliptisches Randwertproblem 2. Ordnung in variationeller Formulierung:

$$a(u, v) = l(v) \quad \forall v \in V \ , \qquad V = H_0^1(\Omega) \text{ oder } H^1(\Omega) \ .$$

wobei

$$a(u,v) = \int_\Omega \left\{ \sum_{i,j=1}^{n} a_{ij} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + a_0 uv \right\} dx$$
$$= \int_\Omega \left\{ \nabla u^T \mathcal{A} \nabla v + a_0 uv \right\} dx$$

mit $\nabla u = (\partial u/\partial x_i, \ldots, \partial u/\partial x_n)^T$ und $\mathcal{A} = (a_{ij})_{i,j=1}^n$,

$$l(v) = \int_\Omega fv\,dx \quad \text{oder} \quad l(v) = \int_\Omega fv\,dx + \int_\Gamma gv\,d\sigma$$

mit Voraussetzungen wie in § 3, Kapitel III.

Gesucht: Galerkin-Approximation zu Approximationsraum $V_N \leq V$:

$$a(u_N, v_N) = l(v_N) \quad \forall v_N \in V_N \ .$$

Wesentliche Randbedingungen:

$$V = H_0^1(\Omega) = \{v \in H^1(\Omega) \ : \ v\big|_\Gamma = 0\} \ , \quad V_N = \langle \varphi_1, \ldots, \varphi_N \rangle \ ,$$

wobei $\varphi_1, \ldots, \varphi_N$ jene Basisfunktionen mit $\varphi_i\big|_\Gamma = 0$.
Dies ist äquivalent zu einem linearem Gleichungssystem $A\mu = b$ mit

$$A = \Big(a(\varphi_j, \varphi_i)\Big)_{i,j=1}^N \ , \qquad \text{„Steifigkeitsmatrix“,}$$
$$b = \Big(l(\varphi_i)\Big)_{i=1}^N \ , \qquad \text{„Lastvektor“,}$$
$$u_N = \sum_{i=1}^N \mu_i \varphi_i \ , \quad \mu_i = u_N(\underbrace{q_i}_{\text{Knoten}}) \ .$$

Das Aufstellen von A und b erfolgt üblicherweise in zwei getrennten Schritten:

(a) A, b für das Problem „ohne“ Randbedingungen (d. h. nur natürliche Randbedingungen).

(b) Berücksichtige wesentliche Randbedingungen.

§ 3.1 Steifigkeitsmatrix

$$a(\varphi_i, \varphi_j) = \int_\Omega \left\{ \nabla \varphi_i^T \mathcal{A} \nabla \varphi_j + a_0 \varphi_i \varphi_j \right\} dx$$

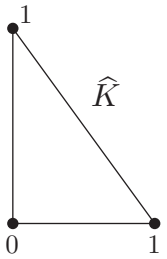
wie in Kapitel II, § 5 berechnet aus Elementarmatrizen

$$a_{rs}^e = \int_{K^e} \left\{ (\nabla \varphi_r^e)^T \mathcal{A} \nabla \varphi_s^e + a_0 \varphi_r^e \varphi_s^e \right\} dx \ .$$

Annahme:
K^e sei auf das Referenzelement $\hat{K}$ rückführbar.

$$F^e : \hat{K} \to K^e$$

bijektiv, lasse im Folgenden Index e weg.



$$\begin{aligned} a_{rs} &= \int_K \left\{ \nabla\varphi_r^T \mathcal{A} \nabla\varphi_s + a_0 \varphi_r \varphi_s \right\} dx \\ &= \int_{\hat{K}} \left\{ \nabla\hat{\varphi}_r^T \hat{\mathcal{A}} \nabla\hat{\varphi}_s + \hat{a}_0 \hat{\varphi}_r \hat{\varphi}_s \right\} |\det DF| d\hat{x} \ , \end{aligned}$$

wobei $\hat{\varphi}_r = \varphi_r \circ F$ Basisfunktion auf $\hat{K}$.

$$\nabla\hat{\varphi}_r = (D\hat{\varphi}_r)^T = \Big(D(\varphi_r \circ F) \Big)^T \overset{\text{Kettenregel}}{\downarrow} = \Big((D\varphi_r) \circ F \cdot \overbrace{DF}^{n\times n} \Big)^T = (DF)^T (\nabla\varphi_r) \circ F \ ,$$

$$\hat{a}_0 = a_0 \circ F \ ,$$

$$\hat{\mathcal{A}} = (DF)^{-1} (\mathcal{A} \circ F)(DF)^{-T} \ ,$$

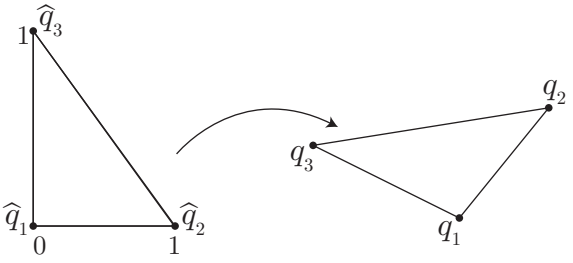
wobei $DF^{-T} = (DF^{-1})^T = (DF^T)^{-1}$ (gilt für alle Matrizen).

Berechnung von F aus Knoten:
F affin:

$$F(\hat{x}) = B\hat{x} + b$$

$$DF = B \ , \qquad b = q_1 \ , \quad q_i \in \mathbb{R}^2 \ ,$$

$$B = (q_2 - q_1, q_3 - q_1) \in \mathbb{R}^{2\times 2} \ .$$



F isoparametrisch:

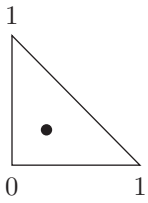
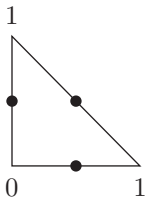
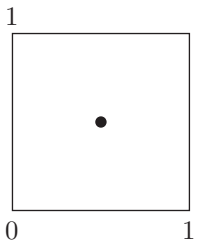
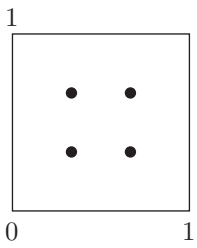
$$F(\hat{x}) = \sum_{r=1}^{R} q_r \hat{\varphi}_r(\hat{x}) \ ,$$

$$DF = \sum_{r=1}^{R} q_r \underbrace{D\hat{\varphi}_r(\hat{x})}_{\text{ohnehin berechnet}} \ .$$

Berechnung des Integrals im Allgemeinen näherungsweise durch Quadraturformel (oder exakt, falls $\mathcal{A}$ und a_0 konstant, z. B. bei der Poisson-Gleichung):

$$\int_{\hat{K}} \phi \, d\hat{x} \approx \sum_{m=1}^{M} w_m \phi(\hat{x}_m) \ .$$

Beispiele.

$\hat{K}$	$\hat{x}_m$	w_m	exakt für Polynome in:
	$\left(\frac{1}{3},\frac{1}{3}\right)$	$\frac{1}{2}$	P_1 (lineare)
	$\left(\frac{1}{2},0\right),\left(0,\frac{1}{2}\right),\left(\frac{1}{2},\frac{1}{2}\right)$	$\frac{1}{6}$	P_2 (quadratische)
	$\left(\frac{1}{2},\frac{1}{2}\right)$	1	Q_1
	Gauß-Knoten $\left(\frac{1}{2}\pm\frac{1}{2\sqrt{3}},\frac{1}{2}\pm\frac{1}{2\sqrt{3}}\right)$	$\frac{1}{4}$	Q_3

§ 3.2 Berechnung des Lastvektors b

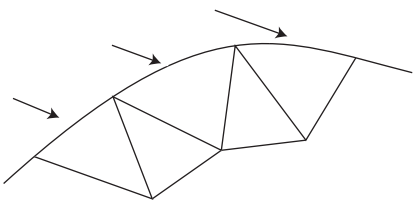
$$b_i = \int_\Omega f\varphi_i\,dx \quad \text{aus Elementvektoren:} \quad \int_K f\varphi_r\,dx = \int_{\hat{K}} \hat{f}\hat{\varphi}_r|\det DF|\,d\hat{x}$$

mit $\hat{f} = f \circ F$, $\hat{\varphi}_r = \varphi_r \circ F$ mit Quadraturformel wie bei Elementmatrizen. Bei

$$b_i = \int_\Omega f\varphi_i\,dx + \int_\Gamma g\varphi_i\,d\sigma$$

ist zusätzlich noch ein Integral über den Rand zu berechnen.

In 2 Dimensionen ($n = 2$):
Kurvenintegral über Finite-Elemente-Kanten.



Rückführbar auf Integrale

$$\int_0^1 \psi\,dx\ .$$

Berechenbar z. B. mit Gauß-Quadraturformel

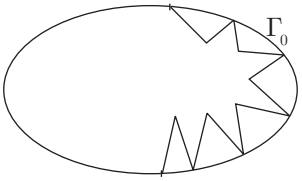
$$\int_0^1 \psi \, dx \approx \psi\left(\frac{1}{2}\right) : \qquad P_1,$$

$$\text{oder:} \qquad \approx \frac{1}{2}\psi\left(\frac{1}{2} - \frac{1}{2\sqrt{3}}\right) + \frac{1}{2}\psi\left(\frac{1}{2} + \frac{1}{2\sqrt{3}}\right) : \qquad P_3.$$

(Genau für Polynome P_1 beziehungsweise P_3.)

§ 3.3 Berücksichtigung von Dirichlet-Randbedingungen (auf $\Gamma_0 \subset \Gamma$)

Bezeichnung: Knoten $q_1, \ldots, q_I$ wie bisher.
Seien $\varphi_1, \ldots, \varphi_N$ jene Basisfunktionen, die auf Γ_0 verschwinden: $\varphi_i\big|_{\Gamma_0} = 0$, $V_N = \langle \varphi_1, \ldots, \varphi_N \rangle$.
Sei $J_0 = \{N+1, \ldots, I\}$, wobei $j \in J_0 \Leftrightarrow q_j \in \Gamma_0$.
Außerdem sei die Voraussetzung erfüllt, dass Γ_0 aus vollständigen Seiten von finiten Elementen besteht.



Dirichlet-Randbedingungen:

$$u = g \quad \text{auf} \quad \Gamma_0 \subset \Gamma \ .$$

Lösbar, falls $g = u_0\big|_{\Gamma_0}$ für ein $u_0 \in H^1(\Omega)$ (Kapitel III, § 3).
Setze als Approximation

$$\tilde{u}_0 = \sum_{j \in J_0} g(q_j)\varphi_j \ .$$

Damit:

$$\tilde{u}_0 \in H^1(\Omega) \ , \quad \tilde{u}_0(q_j) = g(q_j) \quad \forall j \in J_0 \ .$$

(Dies gilt auf den Knoten! Im Allgemeinen gilt nämlich $\tilde{u}_0\big|_{\Gamma_0} \neq g$.)

$$\tilde{u}_0(q_i) = 0 \quad \forall i \notin J_0 \ .$$

Approximiertes Problem: Suche u_N mit $u_N - \tilde{u}_0 \in V_N$, sodass

$$a(u_N, v_N) = l(v_N) \quad \forall v_N \in V_N$$

beziehungsweise äquivalent: Suche $w_n = u_N - \tilde{u}_0 \in V_N$ mit

$$a(w_N, v_N) = l(v_N) - a(\tilde{u}_0, v_N) \quad \forall v_N \in V_N$$

beziehungsweise löse das lineare Gleichungssystem

$$A\mu = b - \left(\sum_{j \in J_0} g(q_j) a(\varphi_j, \varphi_i) \right)_{i=1}^{N} \ , \quad \mu \in \mathbb{R}^N \ ,$$

$$u_N = w_N + \tilde{u}_0 = \sum_{i=1}^{N} \mu_i \varphi_i + \sum_{j=N+1}^{I} g(q_j)\varphi_j \ .$$

§ 4 Fehlerabschätzung und Konvergenz: Vorbemerkungen

Gegeben: Elliptisches Randwertproblem 2. Ordnung in variationeller Formulierung

$$a(u,v) = l(v) \quad \forall v \in V \qquad \Big(H_0^1(\Omega) \subset V \subset H^1(\Omega)\Big)$$

werde mit Finiter-Elemente-Methode „approximiert“.

Fehlerquellen:

- Galerkin-Ansatz (V ersetzt durch endlichendimensionalen Unterraum V_N)
- numerische Integration in Steifigkeitsmatrix, Lastvektor
- Approximation des Gebietsrandes (bei nichtpolygonalen Gebieten)
- Lösen des linearen Gleichungssystems
- Rundungsfehler

Konzentrieren uns hier auf den *Galerkin-Fehler:* Sei $V_N = \langle \varphi_1, \ldots, \varphi_N \rangle \leq V$ Finite-Elemente-Raum und sei h der maximale Durchmesser eines finiten Elements der Triangulierung von Ω. (Der Durchmesser ist der maximale Abstand zweier Punkte.) Betrachte nun Familien von Triangulierungen mit $h \to 0$, schreibe im Folgenden V_h statt V_N.

Galerkin: Suche $u_h \in V_h$, sodass

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h \ .$$

Erhoffe Konvergenz

$$u_h \to u \quad \text{für } h \to 0 \ .$$

Weiß nach Céas Lemma:

$$\|u_h - u\|_a = \min_{v_h \in V_h} \|v_h - u\|_a \quad \Big(\|v\|_a = \sqrt{a(v,v)}\Big) \ .$$

Weiß: $\|\cdot\|_a$ äquivalent zur Sobolevnorm $\|\cdot\|_1 \quad \Big(\alpha\|v\|_1^2 \leq a(v,v) \leq M\|v\|_1^2\Big)$, daher

$$\|u_h - u\|_1 \leq C \cdot \min_{v_h \in V_h} \|v_h - u\|_1 \ , \qquad C = \sqrt{\frac{M}{\alpha}} \ .$$

Wähle speziell die Interpolation

$$v_h = \Pi_h u := \sum_{i=1}^{N} u(q_i)\varphi_i \in V_h$$

in den Knoten q_i. (Sie ist wohldefiniert, falls u stetig ist, z. B falls $u \in H^2(\Omega)$, $n \leq 3$ nach dem Sobolev Einbettungssatz.)

$$\|u_h - u\|_1 \leq C \cdot \|\Pi_h u - u\|_1 \ , \qquad \text{Interpolationsfehler.}$$

Rückführung auf einzelne finite Elemente:

$$\begin{aligned}\|\Pi_h u - u\|_{1,\Omega}^2 &= \int_\Omega (\Pi_h u - u)^2 \, dx + \sum_{i=1}^n \int_\Omega \left(\frac{\partial}{\partial x_i} (\Pi_h u - u) \right)^2 dx \\ &= \sum_{e=1}^E \|\Pi_{K^e} u - u\|_{1,K^e}^2 = \sum_{e=1}^E \int_{K^e} \%\end{aligned}$$

mit

$$\Pi_{K^e} u = \Pi_h u\big|_{K^e} = \sum_{i=1}^n u(q_i)\varphi_i\big|_{K^e} = \sum_{r=1}^R u(q_r^e) \cdot \underbrace{\varphi_r^e}_{\text{lokale Basisfunktion}} \ .$$

Untersuche daher Interpolationsfehler auf dem finiten Element K ($\operatorname{diam} K \leq h$):

$$\|\Pi_k u - u\|_{1,K} \quad \overset{\substack{\text{werde zeigen} \\ \text{(unter zusätzlichen Voraussetzungen)} \\ \downarrow}}{\leq} \quad C_k h^k \cdot |u|_{k+1,K} \ ,$$

falls der Polynomraum des finiten Elements alle Polynome vom Grad $\leq k$ enthält und $u \in H^{k+1}(\Omega)$.

Erinnerung.

$$|u|_{k+1,K}^2 = \int_K \sum_{|\alpha|=k+1} (\partial^\alpha u)^2 \, dx \ .$$

Zusammensetzen:

$$\|u_h - u\|_{1,\Omega} \leq C \cdot C_k \cdot h^k |u|_{k+1,\Omega} \ , \qquad \text{falls } u \in H^{k+1}(\Omega).$$

Dabei ist h der maximale Durchmesser der finiten Elemente.
Zunächst für $k = 1$:

§ 5 Fehlerabschätzungen für lineare finite Elemente

K Dreieck 

P_1: $\|\Pi_k u - u\|_1 \leq ?$

Hilfssatz 1. *$K \subset \mathbb{R}^n$ kompakt, konvex, mit Durchmesser $\leq h$. (Der Durchmesser ist wieder der maximale Abstand zweier Punkte.)*
Brauche Variante der Poincaré-Ungleichung (vergleiche Kapitel III, § 2). Bezeichne

$$Mv := \frac{\int_K v \, dx}{\int_K 1 \, dx}$$

den Mittelwert einer Funktion v auf K. Dann gilt

$$\|v - Mv\|_{0,K} \le C(n) \cdot h \cdot |v|_{1,K} \quad \forall v \in H^1(K) \ ,$$

das heißt

$$\left(\int_K (v - Mv)^2 \, dx \right)^{\frac{1}{2}} \le C(n) \cdot h \left(\int_K \sum_{i=1}^{n} \left(\frac{\partial v}{\partial x_i} \right)^2 \, dx \right)^{\frac{1}{2}} \ .$$

Beweis. Sei $V = \int_K 1 \, dx$ das Volumen von K. Dann gilt:

$$\int_K (v(y) - Mv)^2 \, dy = \int_K \underbrace{\left(\int_K 1 \cdot \Big(v(y) - v(x) \Big) \, dx \right)^2}_{\left(v(y)V - VMv \right)^2} \frac{1}{V^2} \, dy$$

$$\overset{\substack{\text{Cauchy-Schwarzsche} \\ \text{Ungleichung}}}{\downarrow} \le \quad \frac{1}{V} \int_K \int_K \Big[v(y) - v(x) \Big]^2 \, dx \, dy$$

Mit $v(y) - v(x) = \int_0^1 Dv\big(ty + (1-t)x\big) \cdot (y - x) \, dt$ und $|y - x| \le h$, wobei $|\cdot|$ die euklidische Norm bezeichnet, folgt weiterhin

$$\overset{\substack{\text{Cauchy-Schwarzsche} \\ \text{Ungleichung}}}{\downarrow} \le \quad \frac{h^2}{V} \int_K \int_K \int_0^1 \underbrace{\big| Dv\big(ty + (1-t)x\big) \big|^2}_{\text{euklidische Norm im } \mathbb{R}^n} \, dt \, dx \, dy$$

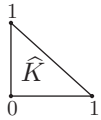
Teile das innere Integral auf: $\int_0^1 \% = \int_0^{\frac{1}{2}} \% + \int_{\frac{1}{2}}^1 \%$. Damit:

$$\overset{\substack{\text{Symmetrie} \\ x \leftrightarrow y}}{\downarrow} \le \frac{2h^2}{V} \int_K \int_K \int_0^{\frac{1}{2}} \big| Dv\big(\underbrace{ty + (1-t)x}_{\xi \in K} \big) \big|^2 \, dt \, dx \, dy$$

Transformationssatz mit $d\xi = (1-t)^n \, dx$:

$$\begin{aligned} &\le \frac{2h^2}{V} \int_K \int_0^{\frac{1}{2}} \int_K |Dv(\xi)|^2 \cdot (1-t)^{-n} \, d\xi \, dt \, dy \\ &= \frac{2h^2}{V} \quad \underbrace{\int_0^{\frac{1}{2}} (1-t)^{-n} \, dt}_{-\frac{(1-t)^{-n+1}}{-n+1}\Big|_0^{\frac{1}{2}} = \frac{2^{n-1}-1}{n-1}} \quad \cdot \underbrace{\int_K dy}_{V} \cdot \underbrace{\int_K |Dv(\xi)|^2 \, d\xi}_{|v|_{1,K}^2} \\ &= \underbrace{\frac{2^n - 2}{n-1}}_{C(n)^2} h^2 |v|_{1,K}^2 \ . \end{aligned}$$

□

Sei $\hat{K} = $ 

das Referenz-Dreieck und $\hat{\Pi}v = \Pi_{\hat{K}} v$ lineare Interpolation in den Ecken.

Hilfssatz 2.

$$\left|v - \hat{\Pi}v\right|_1 \leq \hat{C} \cdot |v|_2 \quad \forall v \in H^2(\hat{K}) \ .$$

Beweis.

$$\hat{\Pi}v(x_1, x_2) = x_1 v(1,0) + x_2 v(0,1) + (1 - x_1 - x_2) v(0,0) \ ,$$
$$\frac{\partial}{\partial x_1}\hat{\Pi}v = v(1,0) - v(0,0) = \int_0^1 \frac{\partial v}{\partial x_1}(t,0)\, dt =: L\left(\frac{\partial v}{\partial x_1}\right) \ , \quad \frac{\partial v}{\partial x_1} \in H^1(\hat{K}) \ ,$$

weil $v \in H^2(\hat{K})$. Da wegen

$$L : H^1(\hat{K}) \xrightarrow[\substack{\text{Spurabbildung}\\ \text{linear, stetig}}]{\cdot|_{\partial\hat{K}}} \underbrace{L^2(\partial\hat{K}) \quad \to \quad \overbrace{\mathbb{R}}^{c \mapsto \text{konstante Funktion mit Wert } c}}_{\substack{w \mapsto \int_0^1 1 \cdot w(t,0)\, dt\\ \text{stetig, beschränkt mit Cauchy-Schwarz}}} \hookrightarrow \quad L^2(\hat{K})$$

L linear, stetig, folgt

$$\begin{aligned}
\int_{\hat{K}} \left(\frac{\partial v}{\partial x_1} - \frac{\partial}{\partial x_1}\hat{\Pi}v\right)^2 dx &= \left\| \frac{\partial v}{\partial x_1} - L\left(\frac{\partial v}{\partial x_1}\right)\right\|_0^2 \\
&\overset{\substack{Lc = c\\ \forall c \in \mathbb{R}}}{=} \left\| \left(\frac{\partial v}{\partial x_1} - c\right) - L\left(\frac{\partial v}{\partial x_1} - c\right)\right\|_0^2 \\
&= \left\| \underbrace{(I - L)}_{\substack{\text{beschränkt:} \quad \|I-L\| \leq l\\ \text{Operator-Norm von } H^1 \to L^2}} \cdot \left(\frac{\partial v}{\partial x_1} - c\right)\right\|_0^2 \\
&\leq l^2 \cdot \left\| \frac{\partial v}{\partial x_1} - c\right\|_1^2 \quad \forall c \in \mathbb{R} \ .
\end{aligned}$$

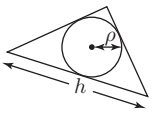
Wähle speziell $c = M(\partial v/\partial x_1)$, verwende Hilfssatz 1:

$$\left\| \frac{\partial v}{\partial x_1} - c\right\|_1^2 = \underbrace{\left\| \frac{\partial v}{\partial x_1} - c\right\|_0^2}_{\substack{\text{Hilfssatz 1}\\ \leq\\ 4 \cdot \left|\frac{\partial v}{\partial x_1}\right|_1^2}} + \left|\frac{\partial v}{\partial x_1}\right|_1^2 \leq 5\left|\frac{\partial v}{\partial x_1}\right|_1^2 \leq 5|v|_2^2 \ ,$$

ebenso für x_2. Erhalte

$$\left|v - \hat{\Pi}v\right|_1^2 = \int_{\hat{K}} \sum_{i=1}^{2} \left(\frac{\partial v}{\partial x_i} - \frac{\partial}{\partial x_i}\hat{\Pi}v\right)^2 dx \leq 5\left(\left|\frac{\partial v}{\partial x_1}\right|_1^2 + \left|\frac{\partial v}{\partial x_2}\right|_1^2\right) = \underbrace{\hat{C}^2}_{5} \cdot |v|_2^2 \ . \qquad \square$$

Hilfssatz 3. *Sei K ein beliebiges Dreieck mit Durchmesser h und Inkreisradius ρ. Sei Π lineare Interpolation in den Ecken von K. Dann gilt*



$$|v - \Pi v|_{1,K} \le C \cdot \frac{h^2}{\rho} \cdot |v|_{2,K} \quad \forall v \in H^2(K) \ .$$

Bemerkung. Falls noch $h/\rho \le \text{const}$ gilt, so folgt:

Zu vermeiden:

$$|v - \Pi v|_{1,K} \le C \cdot h \cdot |v|_{2,K} \ .$$

Beweis. Zurückzuführen auf das Referenz-Dreieck $\hat{K}$ mittels $F : \hat{K} \to K$ affin, bijektiv:

$$F(\hat{x}) = B\hat{x} + b \ .$$

(a)

$$|v - \Pi v|_{1,K}^2 = \int_K \underbrace{\big|\overbrace{Dv}^{\left(\frac{\partial v}{\partial x_1}, \frac{\partial v}{\partial x_2}\right)} - D\Pi v\big|}_{\text{Euklidische Norm}}^2 dx = \int_{\hat{K}} \left| D\hat{v} \cdot B^{-1} - D\hat{\Pi}\hat{v} \cdot B^{-1} \right|^2 \cdot |\det B| \, d\hat{x}$$

wobei $\hat{v} = v \circ F$, $D\hat{v} = (Dv \circ F) \cdot \overbrace{DF}^{B}$ und $\hat{\Pi}\hat{v} = (\Pi v) \circ F$. Weiterhin

$$\le \int_{\hat{K}} |D\hat{v} - D\hat{\Pi}\hat{v}|^2 \cdot \|B^{-1}\|^2 \cdot |\det B| \, d\hat{x}$$

$$\le \int_{\hat{K}} |D\hat{v} - D\hat{\Pi}\hat{v}|^2 \, d\hat{x} \cdot \|B^{-1}\|^2 \cdot |\det B|$$

$$\overset{\substack{\text{Hilfssatz 2} \\ \downarrow}}{\le} \hat{C}^2 \cdot \int_{\hat{K}} \underbrace{|D^2\hat{v}|^2}_{\sum_{|\alpha|=2}(\partial^\alpha v)^2} d\hat{x} \cdot \|B^{-1}\|^2 \cdot |\det B|$$

$$\le \hat{C}^2 \int_K |D^2 v|^2 \, dx \cdot \|B\|^4 \cdot \|B^{-1}\|^2 = \Big(\hat{C} \cdot |v|_{2,K} \cdot \underbrace{\|B\|^2}_{\le h^2/\hat{\rho}^2} \cdot \underbrace{\|B^{-1}\|}_{\le \hat{h}/\rho} \Big)^2 \ .$$

(b) Zeige noch $\|B\| \le h/\hat{\rho}$ und durch Rollentausch von K und $\hat{K}$ auch $\|B^{-1}\| \le \hat{h}/\rho$. Sei $\hat{x} \in \mathbb{R}^2$,

$$|\hat{x}| = \hat{\rho} \quad \overset{\substack{\text{Definition des} \\ \text{Inkreisradius} \\ \downarrow}}{\Rightarrow} \quad \exists \hat{y}, \hat{z} \in \hat{K} : \hat{x} = \hat{y} - \hat{z} \ .$$

$$B\hat{x} = B\hat{y} - B\hat{z} = \underbrace{F(\hat{y})}_{\in K} - \underbrace{F(\hat{z})}_{\in K} \in K \quad \Rightarrow \quad \|B\hat{x}\| \le h \ .$$

Damit

$$\|B\| = \max_{\|\hat{x}\|=\hat{\rho}} \frac{\|B\hat{x}\|}{\hat{\rho}} \le \frac{h}{\hat{\rho}} \ . \qquad \square$$

Hilfssatz 4. *Unter den Voraussetzungen von Hilfssatz 3 ist*

$$\|v - \Pi v\|_{0,K} \le Ch^2 |v|_{2,K} \quad \forall v \in H^2(K) \ .$$

Beweis. Übungsaufgabe 34. □

Hinweis (für Übungsaufgabe 34). Zeige die Behauptung zunächst für das Referenz-Dreieck $\hat{K}$, verwende dazu

$$v(x) - v(0) = \int_0^1 1 \cdot \frac{d}{dt} v(tx)\, dt = Dv(x) \cdot x - \int_0^1 t \cdot \frac{d^2}{dt^2} v(tx)\, dt$$

und dieselbe Formel auch für $\hat{\Pi} v$. Beachte dann: $v(0) = \hat{\Pi} v(0)$. Erhalte damit:

$$\|v - \hat{\Pi} v\|_{0,\hat{K}}^2 \le C_1^2 \cdot |v - \hat{\Pi} v|_{1,\hat{K}}^2 + C_2^2 \cdot |v|_{2,\hat{K}}^2 \overset{\text{Hilfssatz 2}}{\downarrow} \le C^2 |v|_{2,\hat{K}}^2 \ .$$

Daraus folgt die Behauptung für das allgemeine Dreieck wie in Hilfssatz 3.
Aus den Hilfssätzen 3 und 4 folgt weiterhin:

$$\|v - \Pi v\|_{1,K} \le C \cdot h \cdot |v|_{2,K} \quad \forall v \in H^2(K) \quad (h \le \text{const}) \ , \quad \text{falls} \quad h/\rho \le \text{const} \ .$$

Erhalte aus den Überlegungen in § 4

$$\|v - \Pi_h v\|_{1,\Omega} \le C \cdot h \cdot |v|_{2,\Omega} \quad \forall v \in H^2(\Omega)$$

und damit

Satz 5. *Es gelte:*

(i) Die Lösung u des elliptischen Randwertproblems ist in $H^2(\Omega)$.

(ii) Alle Dreiecke aller Triangulierungen haben $h/\rho \le$ const.

Dann gilt für die Finite-Elemente-Methode mit linearen finiten Elementen

$$\|u_h - u\|_{1,\Omega} \le C_1 \cdot h \cdot |u|_{2,\Omega} \ .$$

Bemerkung (H^2-Regularität, ohne Beweis). Falls Ω konvex ist oder einen C^2-Rand hat und nur Dirichlet- oder nur Neumann-Randbedingungen auftreten (also keine gemischten Randbedingungen), so ist das Randwertproblem *H^2-regulär*, d. h. die Lösung u von

$$a(u, v) = \int_\Omega f v\, dx \quad \forall v \in V$$

liegt in $H^2(\Omega) \cap V$ für jedes $f \in L^2(\Omega)$ und $\|u\|_2 \le C_2 \|f\|_0$ (C_2 unabhängig von f).
Wissen aus Hilfssatz 4

$$\|\Pi_h u - u\|_0 \le \tilde{C} h^2 |u|_2 \ ,$$

habe auch

Satz 6. *Voraussetzungen wie in Satz 5, Randwertproblem sei H^2-regulär. Dann gilt*

$$\|u_h - u\|_0 \le C h^2 |u|_2 \ .$$

Beweis (mit Nitsche-Trick). Betrachte das Randwertproblem für φ:

$$a(v,\varphi) = \int_\Omega (u_h - u)v\,dx \quad \forall v \in V \ .$$

Es ist auch H^2-regulär, also $\varphi \in H^2(\Omega) \cap V$ und $\|\varphi\|_2 \le C \cdot \|u_h - u\|_0$.
Insbesondere für $v = u_h - u \in V$: $a(u_h - u, \varphi) = \|u_h - u\|_0^2$.
Habe

$$\left.\begin{aligned} a(u,v) &= l(v) \quad \forall v \in V \\ a(u_h, v_h) &= l(v_h) \ \forall v_h \in V_h \end{aligned}\right\}_{V_h \le V} \quad \Rightarrow \quad a(u_h - u, v_h) = 0 \quad \forall v_h \in V_h \ .$$

Somit:

$$\underbrace{a(u_h - u, \varphi - v_h)}_{\substack{\text{Cauchy-Schwarzsche Ungleichung} \\ \downarrow \\ \le M \cdot \|u_h - u\|_1 \cdot \|\varphi - v_h\|_1}} = \|u_h - u\|_0^2 \quad \forall v_h \in V_h \ .$$

Wähle $v_h = \Pi_h \varphi$, erhalte

$$\|\varphi - \Pi_h\varphi\|_1 \overset{\text{Hilfssätze 3, 4}}{\downarrow} \le Ch|\varphi|_2 \overset{H^2\text{-regulär}}{\downarrow} \le CC_2 h \|u_h - u\|_0 \ .$$

Damit

$$\|u_h - u\|_0^2 \le M \cdot \underbrace{\|u_h - u\|_1}_{\substack{\text{Satz 5} \\ \le C_1 h|u|_2}} \cdot CC_2 h \cdot \|u_h - u\|_0 \le \underbrace{\tilde{C}}_{\substack{\text{Produkt aus} \\ \text{allen Konstanten}}} h^2 \cdot |u|_2 \cdot \|u_h - u\|_0 \ .$$

Kürze noch auf beiden Seiten der Ungleichung mit $\|u_h - u\|_0$. □

Für finite Elemente höherer Ordnung brauche:

§ 6 Kompakte Einbettungen, Satz von Rellich

Seien V, W Hilberträume (oder auch Banachräume). $T : V \to W$ linear, stetig. T heißt *kompakt*, wenn für jede beschränkte Folge (v_n) in V die Bildfolge (Tv_n) eine in W konvergente Teilfolge hat.

Beispiel. $T : V \to W$ mit $\operatorname{Im} T$ endlichdimensional ist kompakt. Denn sei (v_n) beschränkt in V und T linear, stetig. Dann ist (Tv_n) eine beschränkte Folge in $\operatorname{Im} T \subset W$. Da $\operatorname{Im} T \cong \mathbb{R}^n$ endlichdimensional hat (Tv_n) nach dem Satz von Bolzano-Weierstraß eine konvergente Teilfolge.

Hilfssatz 1. *Seien V, W Banachräume, T_n, $T : V \to W$ linear, stetig ($n = 0, 1, 2, \dots$). Weiterhin sei T_n kompakt (z. B.* $\dim \operatorname{Im} T_n < \infty$*).*

$$\|T_n - T\| = \sup_{v \in V \setminus \{0\}} \frac{\|(T_n - T)v\|_W}{\|v\|_V} \xrightarrow{n \to \infty} 0 \quad \Rightarrow \quad T \text{ kompakt.}$$

Beweis. Sei (v_n) eine beschränkte Folge in V.

$$\begin{array}{lll}
(T_1 v_n) & \text{hat konvergente Teilfolge in } W\text{:} & (T_1 v_n^1)\ , \quad \|T_1 v_n^1 - w_1\| \le \frac{1}{n}\ , \\
(T_2 v_n^1) & \text{hat konvergente Teilfolge in } W\text{:} & (T_2 v_n^2)\ , \quad \|T_2 v_n^2 - w_2\| \le \frac{1}{n}\ , \\
\vdots & & \vdots \qquad\qquad \vdots \\
(T_k v_n^{k-1}) & \text{hat konvergente Teilfolge in } W\text{:} & (T_k v_n^k)\ , \quad \|T_k v_n^k - w_k\| \le \frac{1}{n}\ .
\end{array}$$

Betrachte die Diagonalfolge $\tilde{v}_n := v_n^n$.
Damit gilt für $n,\, m \to \infty$ und $n \ge m$:

$$\|T\tilde{v}_n - T\tilde{v}_m\| \le \underbrace{\|T\tilde{v}_n - T_m\tilde{v}_n\|}_{\le \underbrace{\|T - T_m\|}_{\to 0} \cdot \underbrace{\|\tilde{v}_n\|}_{\text{beschränkt}}} + \underbrace{\|T_m\tilde{v}_n - w_m\|}_{\le \frac{1}{n}} + \underbrace{\|w_m - T_m\tilde{v}_m\|}_{\le \frac{1}{m}} + \underbrace{\|T_m\tilde{v}_m - T\tilde{v}_m\|}_{\|T_m - T\| \cdot \|\tilde{v}_m\| \to 0}\ .$$

Daraus folgt: $(T\tilde{v}_n)$ ist eine Cauchy-Folge, also konvergent. Somit ist T kompakt. □

Satz 2. *Sei Ω ein beschränktes Gebiet in $\mathbb{R}^n$, stückweise C^1. Dann ist die Einbettung $H^1(\Omega) \hookrightarrow L^2(\Omega)$ kompakt, d. h. jede beschränkte Folge in $H^1(\Omega)$ (bezüglich $\|\cdot\|_1$) hat eine konvergente Teilfolge in $L^2(\Omega)$ (bezüglich $\|\cdot\|_0$).*

Beweis. Unterteile Ω in endlich viele Elementardreiecke Δ_j mit Durchmesser $\le 1/n$.
Definiere $T_n : H^1(\Omega) \to L^2(\Omega)$ durch $T_n v|_{\Delta_j} = M_{\Delta_j} v\ \forall j$, wobei $M_{\Delta_j} v$ den Mittelwert von v auf Δ_j darstellt.

$$T = \mathrm{inj} : H^1(\Omega) \hookrightarrow L^2(\Omega)\ .$$

Habe: T_n kompakt, da $\dim \operatorname{Im} T_n < \infty$.
$\|T_n - T\| \to 0$ für $n \to \infty$, denn:

$$\begin{aligned}
\|T_n v - Tv\|_{0,\Omega}^2 &= \sum_j \|M_{\Delta_j} v - v\|_{0,\Delta_j}^2 \overset{\text{Hilfssatz 1, § 5}}{\downarrow} \le \sum_j \left(C \frac{1}{n} |v|_{1,\Delta_j} \right)^2 \\
&= \frac{C^2}{n^2} \underbrace{\sum_j |v|_{1,\Delta_j}^2}_{|v|_{1,\Omega}^2} \le \frac{C^2}{n^2} \|v\|_{1,\Omega}^2\ , \quad \text{damit } \|T_n - T\| \le \frac{C}{n} \to 0\ .
\end{aligned}$$

Mit Hilfssatz 1: $T = \mathrm{inj} : H^1(\Omega) \hookrightarrow L^2(\Omega)$ kompakt. □

Satz 3 (Satz von Rellich, $\sim$ 1940)**.** *Sei Ω ein beschränktes stückweises C^1-Gebiet in $\mathbb{R}^n$. Dann ist die Einbettung $H^{m+1}(\Omega) \hookrightarrow H^m(\Omega)$ für jedes $m = 0,\, 1,\, 2,\, \ldots$ kompakt.*

Beweis. Für $m = 0$ folgt die Aussage aus Satz 2.

$$H^{m+1}(\Omega) = \left\{ v : \partial^\alpha v \in L^2(\Omega) \quad \text{für } |\alpha| \le m+1 \right\} = \left\{ v : \partial^\beta v \in H^1(\Omega) \quad \text{für } |\beta| \le m \right\}\ .$$

Sei $\{\beta \in \mathbb{N}_0^n : |\beta| \le m\} = \{\beta_1, \ldots, \beta_M\}$ die Menge aller Multiindizes der Dimension n mit $|\beta| \le m$. Die Mächtigkeit dieser Menge sei M.
Sei (v_n) eine beschränkte Folge in H^{m+1}.

$$
\begin{array}{lll}
(\partial^{\beta_1} v_n) & \text{beschränkte Folge in } H^1(\Omega) \overset{\text{Satz } 2}{\Rightarrow} \exists \text{ konvergente Teilfolge } & (\partial^{\beta_1} v_n^1) \text{ in } L^2(\Omega), \\
(\partial^{\beta_2} v_n^1) & \text{beschränkte Folge in } H^1(\Omega) \overset{\text{Satz } 2}{\Rightarrow} \exists \text{ konvergente Teilfolge } & (\partial^{\beta_2} v_n^2) \text{ in } L^2(\Omega), \\
\vdots & \vdots & \vdots \\
(\partial^{\beta_M} v_n^{M-1}) & \text{beschränkte Folge in } H^1(\Omega) \overset{\text{Satz } 2}{\Rightarrow} \exists \text{ konvergente Teilfolge } & (\partial^{\beta_M} v_n^M) \text{ in } L^2(\Omega).
\end{array}
$$

Somit: Es existiert eine Teilfolge $(\tilde{v}_n) := (v_n^M)$ mit $(\partial^\beta \tilde{v}_n)$ konvergent in $L^2(\Omega)$ $\forall |\beta| \le m$.
Dann ist $(\tilde{v}_n)$ konvergent in $H^m(\Omega)$. □

§ 7 Approximationssätze für Polynominterpolation

Sei $K \subset \mathbb{R}^n$, $n = 2$: Polygon, $n = 3$: Polyeder.
Sei weiterhin P_k der Raum der Polynome vom Grad $\le k$, also in Multiindexnotation:

$$p(x) = \sum_{|\alpha| \le k} c_\alpha x^\alpha \ , \quad k \ge 1 \ , \quad \text{mit Multiindex } \alpha \text{ und mit } x^\alpha = x_1^{\alpha_1} \dots x_n^{\alpha_n} \ .$$

Satz 1. *Es existiert eine Konstante C (nur abhängig von K, k), sodass*

$$\inf_{p \in P_k} \|v - p\|_{k+1} \le C \cdot |v|_{k+1} \quad \forall v \in H^{k+1}(K) \ .$$

Beweis. (a) Sei $R = \dim P_k$ und $x_1, \dots, x_R \in K$ so, dass

$$P_k \overset{\cong}{\longrightarrow} \mathbb{R}^R : p \mapsto \Big(p(x_r)\Big)_{r=1}^R \ .$$

Zeigen in (b) *verallgemeinerte Poincaré-Ungleichung*:

$$\|v\|_{k+1} \le C\left(|v|_{k+1} + \sum_{r=1}^R |v(x_r)|\right) \quad \forall v \in H^{k+1}(K) \ , \tag{IV.1}$$

$v(x_r)$ wohldefiniert, da wegen Sobolev-Einbettung

$$H^{k+1}(K) \subset H^2(K) \subset C(K) \quad \text{für } K \subset \mathbb{R}^n \ , \quad n = 2, 3 \ .$$

Damit: Zu gegebenem $v \in H^{k+1}(K)$ wähle Interpolationspolynom $p \in P_k$ mit

$$p(x_r) = v(x_r) \ , \quad r = 1, \dots, R \ .$$

Mit (IV.1) folgt

$$\|v - p\|_{k+1} \le C|v - p|_{k+1} + 0 \overset{p \in P_k}{=} C|v|_{k+1} \ .$$

Dieser kurze Beweis sagt leider nichts über die Größe von C aus.

(b) Zeigen (IV.1) indirekt: Angenommen, es existiert eine Folge (w_n) in $H^{k+1}(K)$ mit

$$\|w_n\|_{k+1} \ge n \cdot \Big(|w_n|_{k+1} + \sum_{r=1}^R |w_n(x_r)|\Big) \ , \quad \forall n \in \mathbb{N} \ .$$

Setze $v_n = w_n / (\|w_n\|_{k+1})$, habe also:

(i) $\|v_n\|_{k+1} = 1$ und
(ii) $|v_n|_{k+1} + \sum_{r=1}^{R} |v_n(x_r)| \le \frac{1}{n} \to 0$ für $n \to \infty$.

Nach dem Satz von Rellich ist $H^{k+1}(K) \hookrightarrow H^k(K)$ kompakt und jede beschränkte Folge in H^{k+1} hat eine in H^k konvergente Teilfolge. Damit existiert eine Teilfolge $(\tilde{v}_n)$ von (v_n), die in $H^k(K)$ (bezüglich $\|\cdot\|_k$) konvergent ist.

Aus (ii) folgt $|\tilde{v}_n|_{k+1} \to 0$, da beide Summanden unabhängig voneinander gegen Null gehen müssen. Somit ist $(\tilde{v}_n)$ eine Cauchy-Folge in $H^{k+1}(K)$, denn es gilt ja

$$\|\cdot\|_{k+1}^2 = \|\cdot\|_k^2 + |\cdot|_{k+1}^2 \ .$$

Damit ist $(\tilde{v}_n)$ konvergent in $H^{k+1}(K)$.

Der Grenzwert $v \in H^{k+1}(K)$ erfüllt $|v|_{k+1} = 0$, d. h. $\partial^\alpha v = 0 \ \forall \alpha$, $|\alpha| = k+1$. Damit ist v ein Polynom vom Grad $\le k$. Außerdem folgt aus (ii) $|v_n(x_r)| \to 0$, daher $v(x_r) = 0$. Somit $v \in P_k$, $v(x_r) = 0 \ \forall r$ und daraus folgt $v = 0$.

Dies steht im Widerspruch zu (i): $\|v\|_{k+1} = \lim_n \|v_n\|_{k+1} = 1$. □

Als Folgerung erhalte auf dem Referenzelement $\hat{K}$

Satz 2 (Bramble-Hilbert-Lemma, $\sim$ 1970)**.** *Sei $\Pi : H^{k+1}(\hat{K}) \to H^m(\hat{K})$ linear stetig, $m \le k+1$. Es gelte $\Pi p = p \ \forall p \in P_k$ (z. B. Finite-Elemente-Interpolation).*
Dann existiert eine Konstante $\hat{C}$ (abhängig von $\hat{K}$, k und m) so, dass

$$\|v - \Pi v\|_m \le \hat{C} \cdot |v|_{k+1} \quad \forall v \in H^{k+1}(\hat{K}) \ .$$

Beweis. Für ein beliebiges Polynom $p \in P_k$ gilt

$$\|v - \Pi v\|_m = \|v - p - \Pi(v-p)\|_m = \|(I - \Pi)(v-p)\|_m$$

$$\overset{I-\Pi \,:\, H^{k+1} \to H^m \text{ stetig}}{\downarrow}$$

$$\le \underbrace{\|I - \Pi\|_{H^{k+1} \to H^m}}_{=:C_1} \cdot \|v - p\|_{k+1} \ ,$$

also noch mit Satz 1 $\qquad \|v - \Pi v\|_m \le C_1 \inf_{p \in P_k} \|v - p\|_{k+1} \le C_1 \cdot C \cdot |v|_{k+1} \ .$ □

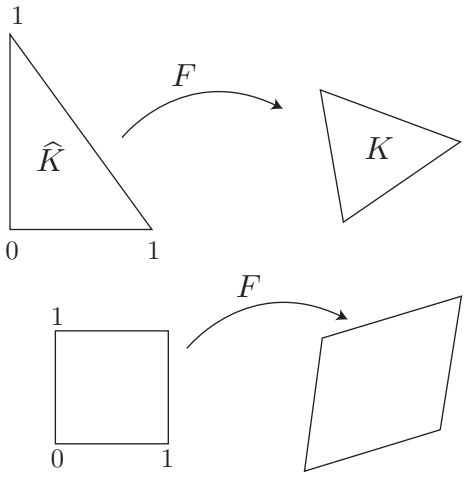
Sei K ein aus dem Referenzelement $\hat{K}$ affin erzeugtes finites Element mit Polynomraum P und den Knoten $q_1, \ldots, q_R$.

$$F(\hat{x}) = B\hat{x} + b$$

(bzw. Tetraeder, Quader in Dimension 3).

Sei h Durchmesser von K, ρ Inkreisradius von K.

Interpolation $\Pi v(x) = \sum_{r=1}^{R} v(q_r)\varphi_r(x)$.



Satz 3. *Situation wie oben, $P_k \subset P$, $m \leq k+1$. Dann gilt*

$$|v - \Pi v|_{m,K} \leq C \cdot \frac{h^{k+1}}{\rho^m} |v|_{k+1,K}$$

mit C unabhängig von K.

Beweis. Wie in Hilfssatz 3, § 5 durch Transformation auf das Referenzelement. Wende dort Satz 2 an. □

Falls $h/\rho \leq$ const für alle finiten Elemente, erhalte daraus für den Interpolationsfehler auf dem Gebiet Ω

$$\begin{aligned} \|v - \Pi v\|_{1,\Omega} &\leq C \cdot h^k \cdot |v|_{k+1,\Omega} \\ \|v - \Pi v\|_{0,\Omega} &\leq C \cdot h^{k+1} \cdot |v|_{k+1,\Omega} \end{aligned} \qquad \forall v \in H^{k+1}(\Omega)$$

und daraus wie in § 4 und § 5

Satz 4. *Es gelte:*

(i) Die Lösung u des elliptischen Randwertproblems sei in $H^{k+1}(\Omega)$.

(ii) Für alle finiten Elemente sei $P_k \subset P$ und $h/\rho \leq$ const. Dann gilt für den Fehler der Finiten-Elemente-Methode

$$\|u_h - u\|_{1,\Omega} \leq C \cdot h^k |u|_{k+1,\Omega} \ .$$

Falls das Problem H^2-regulär ist, erhalte wieder (mit Nitsche-Trick)

$$\|u_h - u\|_{0,\Omega} \leq C h^{k+1} |u|_{k+1,\Omega} \ . \qquad \square$$

Kapitel V

Mehrgitterverfahren

Die Finite-Elemente-Methode führt auf ein lineares Gleichungssystem:

$$A\mu = b \quad \text{in } \mathbb{R}^N \ .$$

N ist sehr groß:

- 2-dimensionales Randwertproblem: $N \sim 10^4$,
- 3-dimensionales Randwertproblem: $N \sim 10^6$.

A ist symmetrisch, positiv definit und schwach besetzt.

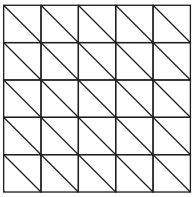
Direktes (Eliminations-) Verfahren: Gauß (LR), Cholesky-Zerlegung (LL^T).
„Fill in“: Auffüllen der Faktoren L, R mit Nichtnullelementen: Aufwand $\mathcal{O}(N^3)$.
→ Zu aufwändig (insbesondere für 3-dimensionales Randwertproblem).

Iterative Verfahren:

- Konjugierte Gradienten (hier nicht)
- Mehrgitterverfahren (Fedorenko 1961, theoretische Untersuchungen. A. Brandt 1972, Hackbusch 1976)

  Aufwand zur Lösung des linearen Gleichungssystems bis zu einer Genauigkeit, die größer als der Fehler der finiten-Elemente-Methode ist: $\mathcal{O}(N) \approx 10\text{–}20N$.

Betrachte Modellproblem:
Poisson-Gleichung auf Einheitsquadrat:



$$\begin{cases} -\Delta u = f & \text{in } \Omega = (0,1) \times (0,1) \\ u = 0 & \text{auf } \Gamma = \partial\Gamma \ . \end{cases}$$

Diskretisierung mit linearen finiten Elementen oder finiten Differenzen:

$$A\mu = b \ , \qquad \begin{array}{ccc} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{array} \qquad \text{Fünf-Punkte-Stern}$$

und mit A wie in (II.3).

Brauche als Vorbereitung:

§ 1 Klassische Iterationsverfahren (Gauß-Seidel, Jacobi)

Betrachte

$$Au = b \ ,$$

zerlege nun $A = M - N$, sodass das Gleichungssystem mit M leicht zu lösen ist, z. B. $M =$ Diagonalmatrix mit Diagonalelementen von A,

$$A = D - (L + U) = \diagdown - (\diagdown\!\!\!\diagdown + \diagdown\!\!\!\diagdown) \ , \qquad \text{Jacobi-Verfahren,}$$

oder $M =$ untere Dreiecksmatrix mit Elementen von A

$$A = (D - L) - U(\diagdown + \diagdown\!\!\!\diagdown) - \diagdown\!\!\!\diagdown \ . \qquad \text{Gauß-Seidel-Verfahren,}$$

Iteriere: Startwert u^0 (gegeben).

$$Au = b \quad \Leftrightarrow \quad Mu = Nu + b \ .$$

Löse für $k = 0, 1, 2, \ldots$

$$Mu^{k+1} = Nu^k + b$$

bis sich (hoffentlich) Konvergenz einstellt.
Klar: Falls $u^k \to u$, dann $Mu = Nu + b \Leftrightarrow Au = b$, u Lösung des Gleichungssystems.
Sei also weiterhin

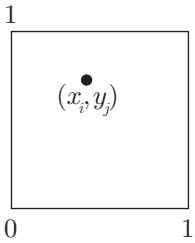
$$A = D - L - U = \diagdown - \diagdown\!\!\!\diagdown - \diagdown\!\!\!\diagdown \ .$$

§ 1.1 Jacobi-Verfahren (Gesamtschrittverfahren)

$M = D$:

Beispiel (Poisson-Gleichung)**.**

$$\begin{matrix} & -1 & \\ -1 & 4 & -1 \\ & -1 & \end{matrix} \ , \qquad D = \frac{4}{h^2} I \ .$$



$$u_{ij} = 0 \quad \text{für } (x_i, y_j) \in \partial\Omega \ , \quad i, j = 1, \ldots, n = \frac{1}{h} - 1 \ ,$$

wobei $h = 1/(n+1)$.
Für $i, j = 1, \ldots, n$:

$$u_{ij}^{k+1} = \frac{1}{4}\left(u_{i-1,j}^k + u_{i+1,j}^k + u_{i,j-1}^k + u_{i,j+1}^k\right) + \frac{h^2}{4} b_{ij} \ .$$

§ 1.2 Gauß-Seidel (Einzelschrittverfahren)

$M = D - L$:

Beispiel (Poisson-Gleichung)**.** Für $i, j = 1, \ldots, n$:

$$u_{ij}^{k+1} = \frac{1}{4}\left(u_{i-1,j}^{k+1} + u_{i+1,j}^{k} + u_{i,j-1}^{k+1} + u_{i,j+1}^{k}\right) + \frac{h^2}{4} b_{ij} \ .$$

§ 1.3 Konvergenzverhalten von Iterationsverfahren

$$A = M - N \ , \qquad Mu^{k+1} = Nu^k + b \ ,$$
$$Mu = Nu + b \ .$$

Daraus erfolgt durch Subtraktion der Fehler $e^k = u^k - u$:

$$Me^{k+1} = Ne^k$$

bzw.

$$e^{k+1} = \underbrace{M^{-1}N}_{\substack{=:K \\ \text{Iterationsmatrix}}} \cdot e^k = Ke^k \ .$$

Falls K diagonalisierbar:

$$Kv_i = \underbrace{\lambda_i}_{\text{Eigenwert}} \underbrace{v_i}_{\text{Eigenvektor}} \ .$$

Zerlege $e^k = \sum_{i=1}^N \varepsilon_i^k v_i$,

$$\underbrace{e^{k+1}}_{=\sum_{i=1}^N \varepsilon_i^{k+1} v_i} = Ke^k = K\sum_{i=1}^N \varepsilon_i^k v_i = \sum_{i=1}^N \varepsilon_i^k \lambda_i v_i$$
$$\Rightarrow \quad \varepsilon_i^{k+1} = \lambda_i \varepsilon_i^k \ , \qquad i = 1, \ldots, N.$$

Somit: $e^k \to 0$ für $k \to \infty$ (für beliebige e^0) genau dann, wenn $|\lambda_i| < 1 \ \forall i$. Konvergiert umso schneller, je kleiner $\max_i |\lambda_i|$. (Je feiner das Gitter, umso langsamere Konvergenz.) Dies gilt auch, wenn K nicht diagonalisierbar ist. Verwende dann die Jordan-Zerlegung.

§ 1.4 Jacobi-Verfahren

$$A = D - (L + U) \ , \quad \text{also} \quad Du^{k+1} = (L + U)u^k + b \ ,$$
die Iterationsmatrix ergibt sich also zu: $J = D^{-1}(L + U)$.

Beispiel. Poisson-Gleichung auf Quadrat (siehe auch in (II.3) für A speziell mit $n = 4$):

$$A = \frac{1}{h^2}\begin{bmatrix} C & -I & & \\ -I & C & \ddots & \\ & \ddots & \ddots & -I \\ & & -I & C \end{bmatrix} \in \mathbb{R}^{N\times N}, \quad \text{wobei} \quad C = \begin{bmatrix} 4 & -1 & & \\ -1 & 4 & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{bmatrix} \in \mathbb{R}^{n\times n},$$

und $h = 1/n + 1$ und $N = n^2$. Mit $D = 4 \cdot I/h^2 \Leftrightarrow D^{-1} = h^2 \cdot I/4$ folgt weiterhin

$$J = \frac{1}{4}\begin{bmatrix} 0 & 1 & & 1 & & \\ 1 & 0 & \ddots & & \ddots & \\ & \ddots & \ddots & 0 & & 1 \\ 1 & & 0 & & \ddots & \\ & \ddots & & \ddots & & \\ & & 1 & & & 0 \end{bmatrix} \in \mathbb{R}^{N\times N}.$$

Bemerkung. Diese Matrix (abgesehen von Vorfaktoren) entsteht auch bei Triangulierung des Einheitsquadrats in regelmäßige Dreiecke im Finiten-Elemente-Verfahren für die Poisson-Gleichung.

Wie lauten die Eigenwerte von J?

Beachte.

$$4J = I \otimes B + B \otimes I \quad \text{mit} \quad B = \begin{bmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix},$$

wobei das Kronecker- bzw. Tensorprodukt wie folgt definiert ist:

$$\underbrace{X}_{m\times n} \otimes \underbrace{Y}_{p\times q} = \begin{bmatrix} X_{11}Y & X_{12}Y & \cdots & X_{1n}Y \\ X_{21}Y & \ddots & & X_{2n}Y \\ \vdots & & & \vdots \\ X_{m1}Y & X_{m2}Y & \cdots & X_{mn}Y \end{bmatrix} \in \mathbb{R}^{(mp)\times(nq)}.$$

Dann folgt

$$J = \frac{1}{4}\begin{bmatrix} B & I & & \\ I & B & \ddots & \\ & \ddots & \ddots & I \\ & & I & B \end{bmatrix}.$$

Rechenregel: $(X \otimes Y)(v \otimes w) = (Xv) \otimes (Yw)$

Bemerkung. Falls v Eigenvektor von X zum Eigenwert λ ist sowie w Eigenvektor von Y zum Eigenwert μ, so folgt

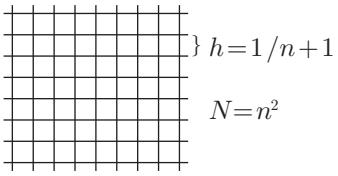
$$\begin{aligned}&(X \otimes Y)(v \otimes w) = \lambda\mu(v \otimes w)\\ \Rightarrow\quad &(v \otimes w) \quad \text{ist Eigenvektor von } X \otimes Y \text{ zum Eigenwert } \lambda\mu.\end{aligned}$$

Hier: Falls w_j, w_k Eigenvektor von B zu Eigenwert μ_j, μ_k, dann ist $w_j \otimes w_k$ Eigenvektor von $I \otimes B + B \otimes I$ zum Eigenwert $\mu_j + \mu_k$, denn es gilt

$$(I \otimes B + B \otimes I)(w_j \otimes w_k) = (w_j \otimes \underbrace{Bw_k}_{\mu_k w_k} + \underbrace{Bw_j}_{\mu_j w_j} \otimes w_k) = (\mu_j + \mu_k)(w_j \otimes w_k)\ .$$

Die Eigenwerte von

$$B = \begin{bmatrix} 0 & 1 & & \\ 1 & 0 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 0 \end{bmatrix} \in \mathbb{R}^{n\times n} \qquad \text{(V.1)}$$



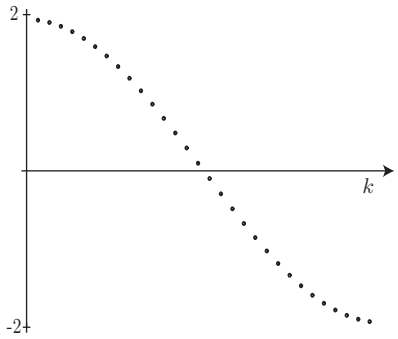
Eigenwerte:

sind

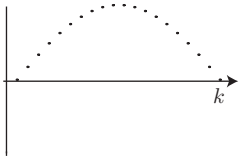
$$\mu_k = 2\cos\frac{k\pi}{n+1}\ , \qquad k = 1, \ldots, n,$$

mit den zugehörigen Eigenvektoren

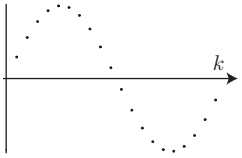
$$w_k = \left(\sin\frac{jk\pi}{n+1}\right)_{j=1}^{n}\ .$$



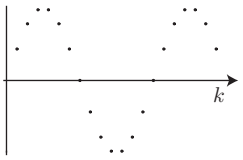
Eigenvektoren: $k = 1$, Gitterweite von $\sin(\pi x)$, $0 \le x \le 1$:



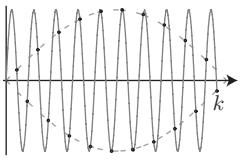
$k = 2$, $\sin(2\pi x)$:



$k = 3, \ldots,$



$\ldots, k = n$:



Je kleiner k, um so weniger Oszillationen, um so „glatter“.
Fasse zusammen:

Hilfssatz 1. *Die Eigenwerte von J sind*

$$\frac{1}{2}\left(\cos\frac{k\pi}{n+1}+\cos\frac{l\pi}{n+1}\right) \quad \text{für } k,\, l = 1,\, \ldots,\, n,$$

die zugehörigen Eigenvektoren sind $w_k \otimes w_l$ mit $w_k = \left(\sin\frac{jk\pi}{n+1}\right)_{j=1}^{n}$.
Betragsgrößter Eigenwert von J ist

$$\frac{1}{2}\left(\cos\frac{1\cdot\pi}{n+1}+\cos\frac{1\cdot\pi}{n+1}\right) = \cos\frac{\pi}{n+1} \approx 1-\frac{1}{2}\left(\frac{\pi}{n+1}\right)^2 .$$

Er liegt nahe bei 1, falls n groß. Daraus folgt eine sehr langsame Konvergenz des Jacobi-Verfahrens!
Brauche $\approx n^2/10$ Iterationen, um den Fehler auf $1/10$ zu reduzieren (siehe Übungen).

§ 1.5 Gedämpftes Jacobi-Verfahren

Sei $0 < \omega < 1$ Parameter. Weiterhin

$$A = M - N = \frac{1}{\omega}D - \left[\left(\frac{1}{\omega}-1\right)D + L + U\right] .$$

Dann ergibt sich für die Iterationsvorschrift

$$\underbrace{\frac{1}{\omega}D}_{M}\, u^{k+1} = \underbrace{\left[\left(\frac{1}{\omega}-1\right)D + (L+U)\right]}_{N} u^k + b$$

und für die Iterationsmatrix $J_\omega = (1-\omega)I + \omega J$. Ihre Eigenwerte sind

$$1-\omega+\frac{\omega}{2}\left(\cos\frac{k\pi}{n+1}+\cos\frac{l\pi}{n+1}\right) , \qquad k,l = 1,\ldots,n,$$

mit den Eigenvektoren $w_k \otimes w_l$ mit $\omega_k = \left(\sin\frac{jk\pi}{n+1}\right)_{j=1}^{n}$.
Betrachte speziell $\omega = 1/2$:

$$\begin{aligned} k,l \approx n \;&\Rightarrow\; \text{Eigenwert} \approx \frac{1}{2}+\frac{1}{4}(-1-1) = 0 \\ k,l \approx 0 \;&\Rightarrow\; \text{Eigenwert} \approx \frac{1}{2}+\frac{1}{4}(+1+1) = 1 . \end{aligned}$$

Die Fehlerkomponenten zu *hochfrequenten* Eigenvektoren werden also im gedämpften Jacobi-Verfahren stark reduziert, wogegen jene zu *„glatten“* Eigenvektoren kaum reduziert werden. Diese können jedoch auch auf einem gröberem Gitter gut approximiert werden. Auf dieser Beobachtung basiert die Idee des *Mehrgitterverfahrens*.

§ 1.6 Gauß-Seidel-Verfahren

$$\underbrace{(D-L)}_{M} u^{k+1} = \underbrace{U}_{N} u^k + b \ .$$

Wie sehen die Eigenwerte der Iterationsmatrix $G = (D-L)^{-1}U$ aus? Füge I ein:

$$G = (D-L)^{-1} \underbrace{(D^{-1})^{-1}D^{-1}}_{=I} U = (I - \underbrace{D^{-1}L}_{\text{◺}})^{-1} \underbrace{D^{-1}U}_{\text{◹}} = (I-\overline{L})^{-1}\overline{U} \ ,$$

wobei $D^{-1}L =: \overline{L}$ die obere und $D^{-1}U =: \overline{U}$ die untere Hälfte von $J = D^{-1}(L+U) = \overline{L}+\overline{U}$ ist.

Erinnerung. $J = I \otimes B + B \otimes I$ mit B wie in (V.1).

Die Eigenwerte von

$$\begin{bmatrix} 0 & \frac{1}{\mu} & & \\ \mu & 0 & \ddots & \\ & \ddots & \ddots & \frac{1}{\mu} \\ & & \mu & 0 \end{bmatrix} = TBT^{-1} \ , \qquad \text{wobei} \quad T = \begin{bmatrix} 1 & & & & \\ & \mu & & & \\ & & \mu^2 & & \\ & & & \ddots & \\ & & & & \mu^{n-1} \end{bmatrix} \ ,$$

sind unabhängig von μ, wie man direkt nachrechnet. (Die zugehörigen Eigenvektoren sind Tw_j.) Dann sind auch die Eigenwerte von

$$\begin{aligned} \mu\overline{L} + \frac{1}{\mu}\overline{U} &= (I + TBT^{-1}) \otimes (TBT^{-1} + I) = (TIT^{-1} + TBT^{-1}) \otimes (TBT^{-1} + TIT^{-1}) \\ &= (T \otimes T)(I \otimes B + B \otimes I)(T^{-1} \otimes T^{-1}) \end{aligned} \tag{V.2}$$

unabhängig von μ. Damit:

$$\begin{aligned} \det(\lambda I - G) &= \det\left(\lambda I - (I-\overline{L})^{-1}\overline{U}\right) = \det\left((I-\overline{L})^{-1} \cdot \left(\lambda(I-\overline{L}) - \overline{U}\right)\right) \\ &= \underbrace{\det(I-\overline{L})^{-1}}_{1} \cdot \underbrace{\det(\sqrt{\lambda}I)}_{(\sqrt{\lambda})^n} \cdot \underbrace{\det\left(\sqrt{\lambda}I - (\sqrt{\lambda}\overline{L} + \frac{1}{\sqrt{\lambda}}\overline{U})\right)}_{\overset{(V.2)}{=}\det(\sqrt{\lambda}I - \underbrace{(\overline{L}+\overline{U})}_{J})} \\ &= \lambda^{\frac{n}{2}} \det(\sqrt{\lambda}I - J) \ . \end{aligned}$$

Eigenvektoren: Für J gilt $Jw = (\overline{L}+\overline{U})w = \mu w$ wie zuvor. Sei noch $\overline{T} := T \otimes T$. Dann:

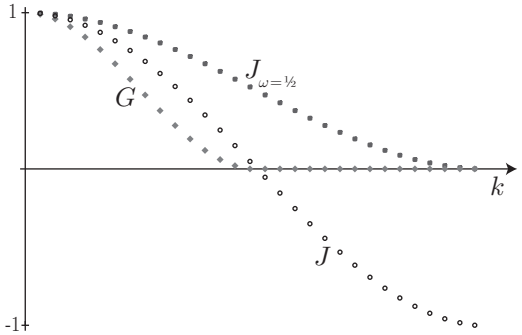
$$(\mu\overline{L} + \frac{1}{\mu}\overline{U}) \underbrace{\overline{T}w}_{v} = \mu \underbrace{\overline{T}w}_{v} \quad \Leftrightarrow \quad (\mu^2\overline{L} + \overline{U})v = \mu^2 v$$

Multipliziere die Gleichung nun von links mit $(I-\overline{L})^{-1}$:

$$\begin{aligned} \Leftrightarrow \quad (I-\overline{L})^{-1}\mu^2 v \cdot \overline{L} + (I-\overline{L})^{-1}\overline{U}v &= (I-\overline{L})^{-1}\mu^2 v \cdot I \\ \Leftrightarrow \quad (I-\overline{L})^{-1}\overline{U}v = (I-\overline{L})^{-1}\mu^2 v(I-\overline{L}) &= \mu^2 v \ , \\ v = v_k \otimes v_l \ , \quad v_k = Tw_k \ . & \end{aligned}$$

Damit wurde gezeigt:

Hilfssatz 2. *Die Hälfte der Eigenwerte von G sind Quadrate der Eigenwerte von J, die anderen sind 0.*



Das Gauß-Seidel-Verfahren konvergiert doppelt so schnell wie das Jacobi-Verfahren (aber immer noch zu langsam).

Beachte (wie beim gedämpften Jacobi-Verfahren). Es werden nur Fehlerkomponenten zu „glatten“ (niedrigfrequenten) Eigenvektoren schlecht reduziert.

§ 2 Zweigitterverfahren

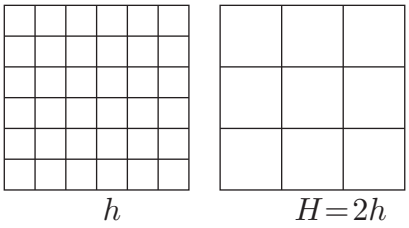
Elliptisches Randwertproblem: $a(u,v) = l(v)\ \forall v \in V$,

Finites Element: $a(u_h, v_h) = l(v_h)\ \forall v_h \in V_h \leq V \quad \Rightarrow \quad A_h\mu = b_h$.

Starte mit u_h^0.
Führe einige wenige (1–3) Iterationen des *gedämpften* Jacobi- oder des Gauß-Seidel-Verfahrens durch.
Erhalte $\overline{u}_h = \sum_{i=1}^{N} \overline{\mu}_i \varphi_i$ bzw. Koeffizientenvektor $\overline{\mu}$.

$$\left.\begin{aligned} A_h\mu_h &= b_h \\ A_h\overline{\mu}_h &= b_h + \overbrace{\quad}^{\text{„Defekt“}} \\ & \qquad \delta_h \end{aligned}\right\} \quad \Rightarrow \quad A_h(\overline{\mu}_h - \mu_h) =: A_h\varepsilon_h = \delta_h \ , \qquad \text{Fehlergleichung.}$$

Idee. Da im Fehler $e_h = \overline{u}_h - u_h$ nur noch „glatte“ Fehlerkomponenten groß sind, kann die Fehlergleichung gut auf einem gröberen Gitter approximiert werden:



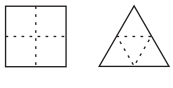
Fehlergleichung: $a(e_h, v_h) = a(\overline{u}_h, v_h) - l(v_h)\ \forall v_h \in V_h$ für $e_h = \overline{u}_h - u_h$.
Löse stattdessen auf gröberem Gitter (Galerkin-Approximation der Fehlergleichung auf $V_H \leq V_h$): Suche $e_H \in V_H$ mit

$$a(e_H, v_H) = a(\overline{u}_h, v_H) - l(v_H) \quad \forall v_H \in V_H \ ,$$

d. h. löse $A_H\varepsilon_H = \delta_H$. Dies ist ein Gleichungssystem niedrigerer Dimension!

$$e_H = \sum_{i=1}^{N_H} \varepsilon_i \varphi_i^H \ , \quad \varepsilon_H = (\varepsilon_i)_{i=1}^{N_H} \ .$$

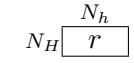
Korrektur: $u_h^{(1)} = \overline{u}_h - e_H$.

Rechnerische Durchführung:
Seien $\varphi_j^h \in V_h$ Basisfunktionen des Feingitterraums und seien $\varphi_j^H \in V_H$ Basisfunktionen des Grobgitterraums.
Voraussetzung: $V_H \leq V_h (\leq V)$, d. h. *echte* Vergröberung der ursprünglichen Triangulierung.  , nicht:  .

Somit $\varphi_i^H \in V_h = \langle \varphi_j^h \rangle$,

$$\varphi_i^H = \sum_j r_{ij} \varphi_j^h \quad \text{mit} \quad r_{ij} = \varphi_i^H(q_j^h) =: p_{ji} \ , \quad (r = p^T) \ .$$

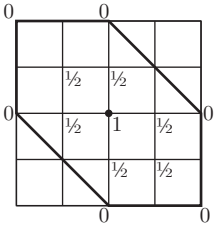
Habe

$r = (r_{ij}) \in \mathbb{R}^{N_H \times N_h}$ „Restriktion“: vom feinen auf das grobe Gitter, 

$p = (p_{ij}) \in \mathbb{R}^{N_h \times N_H}$ „Prolongation“: vom groben auf das feine Gitter. 

Die Prolongation wird auch als Interpolation bezeichnet,

$r = p^T$ ist schwach besetzt .



Grobgittergleichung: (Galerkin-Verfahren mit V_H als Approximationsraum an V_h)

$$a(e_H, v_H) = a(\overline{u}_h, v_H) - l(v_H) \ \forall v_H \in V_H \ .$$

Schreibe einzelne Terme in Basisdarstellung:

$$v_H = \sum_{i=1}^{N_H} \underbrace{v_i}_{\substack{\text{Knotenwerte auf} \\ \text{grobem Gitter}}} \qquad \varphi_i^H = \sum_{i=1}^{N_H} \sum_{j=1}^{N_h} v_i r_{ij} \varphi_j^h = \sum_j \Bigg(\underbrace{\sum_i p_{ji} v_i}_{(pv)_j} \Bigg) \varphi_j^h \ ,$$

$(pv)_j$ sind die Knotenwerte auf dem feinen Gitter.

$$e_H = \sum_k \varepsilon_k \varphi_k^H = \sum_k \sum_l \varepsilon_k r_{kl} \varphi_l^h \quad \text{mit } \varepsilon_k = e_H(q_k^H),\ q_k^H \text{ Knoten des groben Gitters.}$$
$$\overline{u}_h = \sum_l \overline{\mu}_l \varphi_l^h \ , \quad \overline{\mu}_l = \overline{u}_h(q_l^h) \ .$$

Für die Steifigkeitsmatrizen gilt:

$$a(e_H, v_H) = \sum_k \sum_i \varepsilon_k v_i \underbrace{a(\varphi_k^H, \varphi_i^H)}_{\substack{\text{Steifigkeitsmatrix} \\ \text{für grobes Gitter: } A_H}} = v^T A_H \varepsilon \ ,$$

$$(A_H)_{ki} = a(\varphi_k^H, \varphi_i^H) = \sum_l \sum_j r_{kl} r_{ij} \underbrace{a(\varphi_l^h, \varphi_j^h)}_{\substack{\text{Steifigkeitsmatrix} \\ \text{für feines Gitter}}} = (r A_h r^T)_{ki} \ .$$

Somit: $A_H = r A_h p$ Steifigkeitsmatrix für grobes Gitter.
Berechnung des Lastvektors:

$$l(v_H) = \underbrace{\sum_i v_i \underbrace{l(\varphi_i^H)}_{b_i^H}}_{=v^T b_H} = \sum_i \sum_j v_i r_{ij} \underbrace{l(\varphi_j^h)}_{b_j^h} = v^T r b_h \ .$$

Somit: $b_H = r b_h$, Lastvektor für grobes Gitter,

$$a(\overline{u}_h, v_H) = \sum_l \sum_i \sum_j \overline{\mu}_l v_i r_{ij} a(\varphi_l^h, \varphi_j^h) = v^T r A_h \overline{\mu} \ .$$

Damit Grobgittergleichung: $v^T A_H \varepsilon = v^T r A_h \overline{\mu} - v^T r b_h \ \forall v \in \mathbb{R}^{N_H}$.
Daraus folgt ein Gleichungssystem für $\varepsilon = (\varepsilon_k)_{k=1}^{N_H}$:

$$A_H \varepsilon = r(A_h \overline{\mu} - b_h) \ ,$$

wobei A_H die Grobgittersteifigkeitsmatrix ist und $r(b_h - A_h \overline{\mu})$ der auf das grobe Gitter restringierte Defekt. Dies ist ein Gleichungssystem kleinerer Dimension (N_H statt N_h).
Korrektur:

$$\begin{aligned} u_h^{(1)} &= \overline{u}_h - e_H \\ &= \sum_l \overline{\mu}_l \varphi_l^h - \sum_k \sum_l \varepsilon_k \underbrace{r_{kl}}_{p_{lk}} \varphi_l^h = \sum_l (\overline{\mu}_l - \sum_k p_{lk} \varepsilon_k) \varphi_l^h \ . \end{aligned}$$

Also: $u_h^{(1)} = \sum_l \mu_l^{(1)} \varphi_l^h$ mit $\mu^{(1)} = \overline{\mu} - p\varepsilon$.
Fasse zusammen:

Zweigitter-Algorithmus

1) Glättungsiterationen (gedämpftes Jacobi, Gauß-Seidel): $\mu_h^{(0)} \to \overline{\mu}_h$.
2) Berechne Defekt: $\delta_h = A_h \overline{\mu}_h - b_h$
3) Löse Grobgittergleichung $A_H \varepsilon_H = r \delta_h$ ($A_H = r A_h p$)
4) Korrektur $\mu_h^{(1)} = \overline{\mu}_h - p \varepsilon_H$

Aufwand (mit Ausnahme der Grobgittergleichung): $\mathcal{O}(N_h)$.
Konvergenzrate (Fehlerreduktion je Iterationsschritt) $< 1/10$ unabhängig von h.

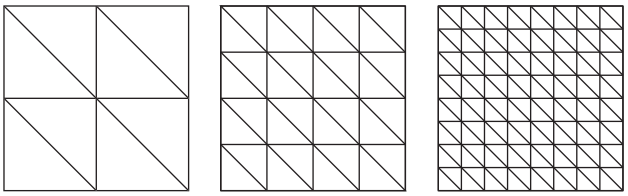
§ 3 Mehrgitterverfahren

Idee. Löse die Grobgittergleichung nicht exakt, sondern näherungsweise durch ein oder zwei Iterationen des Zweigitterverfahrens zu Gittern H und $2H$. Ebenso wieder auf Gitter $2H$ usw. rekursiv, bis auf dem größtem Gitter das lineare Gleichungssystem so klein ist, dass es gut direkt zu lösen ist (im Extremfall: eine skalare lineare Gleichung).

Triangulierungshierarchie:

$$V_{h_0} \leq V_{h_1} \leq V_{h_2} \leq \cdots \leq V_{h_L} \ ,$$
$$h_0 > h_1 > h_2 > \cdots > h_L \ ,$$

also eine geschachtelte Triangulierung; auch eine lokale Verfeinerung ist möglich.



Mehrgitterverfahren zur Lösung von $A_{h_l}\mu_{h_l} = b_{h_l}$, kurz

$$A_l\mu_l = b_l \qquad (A_{l-1} = rA_lp) \qquad \text{(eigentlich auch } r \text{ und } p \text{ mit Indizes).}$$

$u_l^{(1)} = \mathrm{MGV}(l, b_l, u_l^{(0)})$ berechnet durch:

1) Glättungsschritt (ein bis zwei Gauß-Seidel- oder gedämpfte Jacobi-Iterationen):

$$u_l^{(0)} \to \overline{u}_l \ .$$

2) $\delta_l = A_l\overline{u}_l - b_l$ Defekt.

3) Restriktion $\delta_{l-1} = r\delta_l$.

4) Falls $l > 1$: $e_{l-1} = \mathrm{MGV}(l-1, \delta_{l-1}, 0)$ (V-Zyklus, ein MGV) bzw. eventuell

$$\begin{cases} e_{l-1}^{(1)} = \mathrm{MGV}(l-1, \delta_{l-1}, 0) \\ e_{l-1} = \mathrm{MGV}(l-1, \delta_{l-1}, e_{l-1}^{(1)}) \end{cases} \qquad \text{(W-Zyklus, zwei MGV).}$$

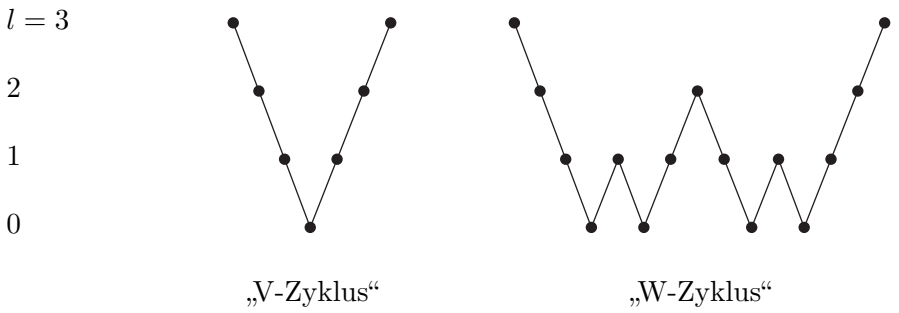
Falls $l = 1$: Löse $A_0\varepsilon_0 = \delta_0$, $e_0 = \sum_{i=1}^{N_0} \varepsilon_{0,i}\varphi_i^0$.

5) Korrektur: $\overline{u}_l^{(1)} = \overline{u}_l - pe_{l-1}$.

6) Nachglätten (ein bis zwei Gauß-Seidel- oder gedämpfte Jacobi-Iterationen):

$$\overline{u}_l^{(1)} \to u_l^{(1)} \ .$$

Schema Eine MGV-Iteration in 4) Zwei MGV-Iterationen in 4)

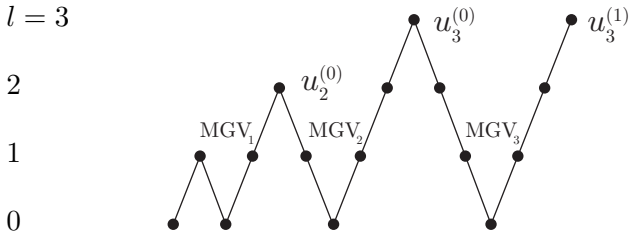


„V-Zyklus“ „W-Zyklus“

Konstruktion von Startwerten durch *geschachtelte Iteration*:
Löse $A_0 u_0 = b_0 \quad (A_{l-1} = r A_l p,\ b_{l-1} = r b_l)$.

Setze
$$
\begin{aligned}
& u_1^{(0)} = p u_0, \\
& u_1^{(1)} = \mathrm{MGV}(1, b_1, u_1^{(0)}), \\
& u_2^{(0)} = p u_1^{(1)}, \\
& u_2^{(1)} = \mathrm{MGV}(2, b_2, u_2^{(0)}), \\
& u_3^{(0)} = p u_2^{(1)}, \\
& \quad \vdots \\
& u_L^{(0)} = p u_{L-1}^{(1)} \qquad \text{„Startwert“}, \\
& u_L^{(1)} = \mathrm{MGV}(L, b_L, u_L^{(0)}) \qquad \text{„Endergebnis“}.
\end{aligned}
$$

Typisch: Fehler $u_L^{(1)} - u_L <$ Finiter-Elemente-Fehler $u_L - u$.
Schema (für V-Zyklus im Mehrgitterverfahren):



(Aufwand: $\mathcal{O}(N)$).

Konvergenzuntersuchungen der Mehrgitterverfahren: Finden sich in der Literatur, z. B. in [Braess92] oder [Hackbu93].

Rechenaufwand: MGV(l), $N_l := \dim V_l$.

1)	Gauß-Seidel (glätten)	cN_l Operationen	(Modellproblem $c = 5$)
2)	Defekt	cN_l	
3)	Restriktion	cN_l	
4)	MGV($l-1$)		
5)	Korrektur	cN_l	$C = 5c(= 25)$
6)	Nachglätten	cN_l	

Aufwand:

$$\begin{aligned} M_l &= CN_l + M_{l-1} \quad \text{(V-Zyklus)} , \\ M_l &= CN_l + 2M_{l-1} \quad \text{(W-Zyklus)} . \end{aligned}$$

$$M_L = \sum_{l=1}^{L} CN_l + \underbrace{M_0}_{\substack{\text{Lösung der f. E.-Gleichung} \\ \text{auf gröbstem Gitter}}} \quad (N_l = 4N_{l-1} \text{ in 2D; } N_l = 8N_{l-1} \text{ im 3D Fall})$$

$$\begin{aligned} &\leq CN_L\left(1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \cdots\right) + M_0 \overset{\substack{\text{geometrische} \\ \text{Reihe} \\ \downarrow}}{=} CN_L \frac{1}{1 - \frac{1}{4}} + M_0 \\ &= \frac{4}{3}CN_L + \underbrace{M_0}_{\text{vernachlässigbar}} , \\ \Rightarrow \quad M_L &\leq \frac{4}{3}CN_L . \end{aligned}$$

Aufwand für die geschachtelte Iteration:

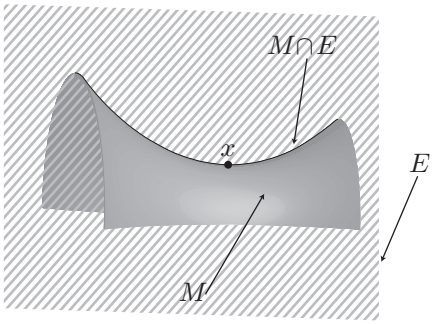
$$\begin{aligned} A_L &= M_0 + (CN_1 + M_1) + (CN_2 + \underbrace{M_2}_{\leq \frac{4}{3}CN_2 + M_0}) + \cdots + (CN_L + \underbrace{M_L}_{\leq \frac{4}{3}CN_L + M_0}) \\ &\leq \underbrace{(L+1)M_0}_{\text{zu vernachlässigen}} + \underbrace{\frac{4}{3}CN_L + \frac{4}{3} \cdot \frac{4}{3}CN_L}_{= \frac{28}{9}CN_L} \approx \underbrace{3C}_{\approx 75} N_L . \end{aligned}$$

Mit diesem Aufwand wird u_L mit einem Fehler kleiner dem Finiten-Elemente-Fehler berechnet. Dieser Aufwand ist im Allgemeinen geringer als der für das Erstellen des linearen Gleichungssystems (Steifigkeitsmatrix, ...).

Kapitel VI

Sattelpunktmethode

Minimierung unter Nebenbedingungen:

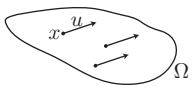


x: Minimum in $M \cap E$.

§ 1 Stokes-System

Strömung einer viskosen inkompressiblen Flüssigkeit.

Navier-Stokes-Gleichungen
$x = (x_1, x_2, x_3)^T \in \Omega \subset \mathbb{R}^3$ Gebiet,
$u = (u_1, u_2, u_3)^T \in \mathbb{R}^3$ Geschwindigkeit, $u = u(x,t)$.



$$\begin{cases} \dfrac{\partial u}{\partial t} = \underbrace{\mu}_{\text{Viskosität}} \Delta u - \displaystyle\sum_{i=1}^{3} u_i \frac{\partial u}{\partial x_i} - \operatorname{grad} \underbrace{p}_{\text{Druck}} + \underbrace{f}_{\text{äußere Kraft}} & \text{in } \Omega\ , \\ \operatorname{div} u = 0 & \text{in } \Omega\ . \end{cases}$$

Navier-Stokes

Randbedingung: $u = 0$ auf $\partial\Omega$, Anfangsbedingung $u = u_0$ zu $t = 0$.
Vereinfachungen. Stationäres Geschwindigkeitsfeld: $u = u(x)$. Setze $\mu = 1$.
Langsam, also ist die Geschwindigkeit u klein: Vernachlässige daher $\sum_i u_i \partial u / \partial x_i$.

Stokes-Gleichungen

$$\begin{cases} -\Delta u + \operatorname{grad} p = f \quad \text{in } \Omega \subset \mathbb{R}^d\ , \qquad d = 2, 3\ , \\ \operatorname{div} u = \displaystyle\sum_{i=1}^{d} \frac{\partial u_i}{\partial x_i} = 0 \quad \text{in } \Omega\ . \end{cases}$$

Randbedingung: $u = 0$ auf $\partial\Omega$. *Gegeben:* $f = (f_1, f_2, f_3)^T$.
Gesucht: $u = (u_1, u_2, u_3)^T$, $p \in \mathbb{R}$. p wird über die Normierung festgelegt: $\int_\Omega p\, dx = 0$.

Variationelle Formulierung: Multipliziere die obere Gleichung mit $v \in C_0^\infty(\Omega)^d$ und multipliziere die untere ($\operatorname{div} u = 0$) mit $q \in C_0^\infty(\Omega)$. Integriere über Ω, verwende partielle Integration (Greensche Formel). Verwende das euklidische Skalarprodukt in $\mathbb{R}^d$:

$$-\int_\Omega \sum_{i=1}^d \Delta u_i v_i \, dx + \int_\Omega \sum_{i=1}^d \frac{\partial p}{\partial x_i} v_i \, dx = \int_\Omega \sum_{i=1}^d f_i v_i \, dx \ ,$$

erhalte mit partieller Integration beider Summanden auf der linken Seite

$$\int_\Omega \sum_{i=1}^d \sum_{j=1}^d \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j} \, dx - \int_\Omega p \underbrace{\sum_{i=1}^d \frac{\partial v_i}{\partial x_i}}_{\operatorname{div} v} \, dx = \int_\Omega \sum_{i=1}^d f_i v_i \, dx \ . \tag{VI.1}$$

Betrachte $V = \{v \in H_0^1(\Omega)^d : \operatorname{div} v = 0\}$ ist abgeschlossener Unterraum von $H_0^1(\Omega)^d$. Daher: V ist Hilbertraum mit Norm (Poincaré-Ungleichung)

$$|v|_{1,\Omega} = \left(\int_\Omega \sum_{i,j=1}^d \left(\frac{\partial v_i}{\partial x_j} \right)^2 dx \right)^{\frac{1}{2}} \ .$$

Betrachte Bilinearform auf $V \times V$:

$$a(u,v) = \int_\Omega \sum_{i,j=1}^d \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j} \, dx \ .$$

Linearform auf V:

$$l(v) = \int_\Omega \sum_{i=1}^d f_i v_i \, dx \ .$$

Wegen $a(v,v) = |v|_1^2$ ist a V-elliptisch; mit Lax-Milgram erhalte

Satz 1. *Es existiert genau ein $u \in V$ mit $a(u,v) = l(v)$ $\forall v \in V$.*

Galerkin: Brauche endlichdimensionalen Unterraum $V_h \leq V$. *Aber:* $\operatorname{div} v_h = 0$, $v_h \in H_0^1(\Omega)^d$, wobei die v_h stückweise Polynome sein sollen, ist nicht handhabbar.

Ausweg?

Beachte. u ist Lösung eines Minimierungsproblems unter Nebenbedingungen auf V. Mit $J(v) = \frac{1}{2} a(v,v) - l(v)$ gilt

$$J(u) = \min_{v \in V} \{ J(v) : \overbrace{v \in H_0^1(\Omega)^d \ , \ \operatorname{div} v = 0}^{v \in V} \} \ .$$

Setze nun für eine übersichtliche Notation $X := H_0^1(\Omega)^d$.

Sei $M = \{q \in L^2(\Omega) : \int_\Omega q \, dx = 0\}$ (Lagrange-Multiplikatoren) und b eine Bilinearform auf $X \times M$:

$$b(v,q) = -\int_\Omega \operatorname{div} v \cdot q \, dx$$

Sei weiterhin a Bilinearform auf $X \times X$ und sei l Linearform auf X.

Variationelle Formulierung: Schreibe (VI.1) in dieser Notation:

$$\begin{aligned} a(u,v) + b(v,p) &= l(v) \quad \forall v \in X \ , \\ b(u,q) &= 0 \qquad \forall q \in M \ . \end{aligned} \tag{VI.2}$$

Satz 2. *Es existiert genau eine Lösung $(u,p) \in X \times M$ von* (VI.2).

Beweis von Satz 2. Ohne Beweis (schwierig: Existenz von p). □

Beachte. u ist Lösung von Satz 1, denn $b(v,q) = 0 \ \forall v \in V$.

Zusammenfassung.

$$u = \begin{pmatrix} u_1(x_1,x_2,x_3) \\ u_2(x_1,x_2,x_3) \\ u_3(x_1,x_2,x_3) \end{pmatrix} \ , \quad p = p(x_1,x_2,x_3) \ .$$

$$\begin{aligned} -\Delta u + \operatorname{grad} p &= f \quad \text{in } \Omega \quad | \cdot v \quad \Big| \int_\Omega \\ \operatorname{div} u &= 0 \quad \text{in } \Omega \quad | \cdot q \quad \Big| \end{aligned}$$

Randbedingung: $u = 0$ auf Γ .

$$a(u,v) = \int_\Omega \sum_{i,j=1}^{3} \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j} \, dx \ , \qquad \int_\Omega \operatorname{grad} p \cdot v \, dx = -\int_\Omega p \operatorname{div} v \, dx = b(v,p) \ ,$$

$$b(u,q) = -\int_\Omega \operatorname{div} u \cdot q \, dx \ , \qquad l(v) = \int_\Omega f v \, dx \ .$$

$$a : \Big(H_0^1(\Omega)\Big)^3 \times \Big(H_0^1(\Omega)\Big)^3 \to \mathbb{R} \ , \qquad l : H_0^1(\Omega)^3 \to \mathbb{R} \quad \text{linear} \ ,$$

$$b : \underbrace{\Big(H_0^1(\Omega)\Big)^3}_{=X} \times \underbrace{L_0^2(\Omega)}_{=\{q \in L^2(\Omega) \,\big|\, \int_\Omega q \, dx = 0\} =: M} \to \mathbb{R} \ ,$$

wobei X und M Hilberträume sind und a eine X-elliptische Bilinearform.

Satz 3. *Sei $u \in \big(C^2(\Omega) \cap C(\overline{\Omega})\big)^d$, $u = 0$ auf $\partial\Omega$ und sei $p \in C^1(\Omega) \cap C(\overline{\Omega})$, $\int_\Omega p \, dx = 0$. Es gilt: (u,p) ist genau dann klassische Lösung der Stokes-Gleichung, wenn (u,p) Lösung der variationellen Formulierung* (VI.2) *ist.*

Beweis. Übung, mit partieller Integration, wie bei der Poisson-Gleichung. □

§ 2 Gemischte finite Elemente

Seien $X_h \leq X$ und $M_h \leq M$ endlichdimensionale Finite-Elemente-Räume.

Approximiere (VI.2) durch: Suche $(u_h, p_h) \in X_h \times M_h$ mit

$$\begin{aligned} a(u_h, v_h) + b(v_h, p_h) &= l(v_h) \quad \forall v_h \in X_h \\ b(u_h, q_h) &= 0 \qquad \forall q_h \in M_h \ . \end{aligned} \tag{VI.3}$$

Rechnerisch: Sei $(\varphi_1, \ldots, \varphi_n)$ Basis von X_h und $(\psi_1, \ldots, \psi_m)$ Basis von M_h.

$$u_h = \sum_{i=1}^{n} u_i \varphi_i \ , \qquad p_h = \sum_{k=1}^{m} p_k \psi_k \ ,$$

und analog:

$$\begin{aligned} v_h &= \sum_{i=1}^{n} v_i \varphi_i \ , & q_h &= \sum_{k=1}^{m} q_k \psi_k \ , \\ \boldsymbol{u} &= (u_i)_{i=1}^{n} \ , & \boldsymbol{p} &= (p_k)_{k=1}^{m} \ , \\ A &= \big(\underbrace{a(\varphi_i, \varphi_j)}_{A_{ji}} \big)_{i,j=1}^{n} \ , & B &= \big(\underbrace{b(\varphi_j, \psi_k)}_{B_{kj}} \big)_{\substack{j=1,\ldots,n \\ k=1,\ldots,m}} \ , \quad \boldsymbol{f} = \big(l(\varphi_i) \big)_{i=1}^{n} \ . \end{aligned}$$

Beachte. $b(v_h, p_h) = \sum_{j=1}^{n} \sum_{k=1}^{m} p_k \underbrace{b(\varphi_j, \psi_k)}_{B_{kj}} v_j = \boldsymbol{p}^T B \boldsymbol{v} = \boldsymbol{v}^T B^T \boldsymbol{p}$ und $a(u_h, v_h) = \boldsymbol{v}^t A \boldsymbol{u}$.

Erhalte aus (VI.2)

$$\begin{aligned} \boldsymbol{v}^T A \boldsymbol{u} + \boldsymbol{v}^T B^T \boldsymbol{p} &= \boldsymbol{v}^T \boldsymbol{f} \quad \forall v \in \mathbb{R}^n \ , \\ \boldsymbol{q}^T B \boldsymbol{u} &= 0 \qquad \forall q \in \mathbb{R}^m \ . \end{aligned}$$

Dann ergibt sich ein lineares Gleichungssystem:

$$\begin{aligned} A\boldsymbol{u} + B^T \boldsymbol{p} &= \boldsymbol{f} \\ B\boldsymbol{u} &= 0 \end{aligned} \quad \text{bzw.} \quad \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{p} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f} \\ 0 \end{pmatrix} \ .$$

Lösbarkeit? Stabilität (der Diskretisierung)?

$$\|u_h\|_1^2 + \|p_h\|_0^2 \leq \underbrace{C}_{\text{unabhängig von } h!} \cdot \|f_h\|_0^2 \ .$$

Fehler $\|u_h - u\|_1$, $\|p_h - p\|_0$?
Klar: Kann X_h und M_h nicht beliebig unabhängig voneinander wählen.

Einschub: Singulärwertzerlegung

Satz. *Zu jeder Matrix $B \in \mathbb{R}^{m \times n}$ existieren orthogonale Matrizen $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$, so dass $B = U \Sigma V^T$ mit*

$$\begin{aligned} \Sigma &= \operatorname{diag}(\sigma_1, \ldots, \sigma_l) \in \mathbb{R}^{m \times n} \ , \\ l &= \min(m, n) \quad \textit{und} \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_l \geq 0 \ . \end{aligned}$$

Die σ_j heißen Singulärwerte *von B. Die Matrix Σ ist nur diagonal für $m = n$ (da $\Sigma \in \mathbb{R}^{m \times n}$!), im nichtdiagonalen Fall wird sie mit 0-Zeilen oder 0-Spalten aufgefüllt.*

Bemerkung. $\underbrace{BB^T}_{m\times m} = U\Sigma\underbrace{V^TV}_{=I}\underbrace{\Sigma^T}_{=\Sigma}U^T = U\begin{pmatrix}\sigma_1^2 & & \\ & \ddots & \\ & & \sigma_m^2\end{pmatrix}U^T$, falls $m \le n$; d. h. die σ_j sind die Wurzeln der Eigenwerte von BB^T. Falls $m \ge n$, so sind die σ_j die Wurzeln der Eigenwerte von B^TB.

Hilfssatz 1. *Situation wie oben. Schreibe $U = (u_1, \dots, u_m)$, $V = (v_1, \dots, v_n)$ mit $u_1, \dots, u_m \in \mathbb{R}^m$ sowie $v_1, \dots, v_n \in \mathbb{R}^n$ (Orthonormalbasis).*
Falls $\sigma_1 \ge \cdots \ge \sigma_r > \sigma_{r+1} = \cdots = \sigma_l = 0$, $l = \min(m,n)$, so gilt

$$\operatorname{Rang} B = r\ , \quad \operatorname{Ker} B = \langle v_{r+1}, \dots, v_n\rangle\ , \quad \operatorname{Im} B = \langle u_1, \dots, u_r\rangle\ ,$$
$$\|B\|_2 = \sigma_1 = \sup_{v\neq 0}\frac{\|Bv\|_2}{\|v\|_2}\ , \quad \|B\|_F := \sqrt{\textstyle\sum_{i=1}^m\sum_{j=1}^n b_{ij}^2} = \|\Sigma\|_F = \sqrt{\sigma_1^2+\cdots+\sigma_r^2}\ .$$

Beweis. Übung. □

Bemerkung (Niedrigrangapproximation). Anwendung bei Suchmaschinen, z. B. Google (riesige Matrix, Zeilen = Begriffe, Einträge = Ort des Dokuments).
Gegeben: Matrix $B \in \mathbb{R}^{m\times n}$ (zu groß).
Suche Matrix X vom Rang k mit $\|B - X\|_F$ minimal!
Erhalte $X = \sum_{j=1}^k \sigma_j u_j v_j^T$ (wie im Hilfssatz).

Beweis des Satzes. Seien $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ mit $\|x\| = 1$ und $\|y\| = 1$ (euklidische Norm) und $Bx = \sigma y$ mit $\sigma = \|B\| = \max_{\|v\|=1}\|Bv\|$.
Sei V_1 orthogonale $n\times n$-Matrix mit x in erster Spalte.
Sei U_1 orthogonale $m\times m$-Matrix mit y in erster Spalte. Damit:

$$B_1 := U_1^TBV_1 = \left(\begin{array}{c|c} \sigma & \\ 0 & \\ \vdots & \ast \\ 0 & \end{array}\right) = \begin{bmatrix} \sigma & \omega^T \\ \underbrace{0}_{1} & \underbrace{\hat{B}_1}_{n-1}\end{bmatrix}\begin{matrix}\}\ 1\\ \}\ m-1\end{matrix}\ ,$$

denn für das $(1,1)$-Element gilt $y^TBx = y^T\sigma y = \sigma y^Ty = \sigma\|y\|^2 = \sigma$. Alle anderen Elemente der ersten Spalte $(k,1)_{k=2,\dots,m}$ müssen verschwinden, da alle Zeilen außer der ersten orthogonal zu y sind (da U_1 orthogonal!).
Zeige weiterhin: $\omega = 0$.

$$\left\|B_1\begin{pmatrix}\sigma\\ \omega\end{pmatrix}\right\| = \left\|\begin{pmatrix}\sigma^2+\omega^T\omega\\ \ast\end{pmatrix}\right\| \ge \sigma^2+\omega^T\omega = \sqrt{\sigma^2+\omega^T\omega}\cdot\sqrt{\sigma^2+\omega^T\omega}\ ,$$

aber auch

$$\left\|B_1\begin{pmatrix}\sigma\\ \omega\end{pmatrix}\right\| \le \|B\|\cdot\left\|\begin{pmatrix}\sigma\\ \omega\end{pmatrix}\right\| = \sigma\sqrt{\sigma^2+\omega^T\omega}\ .$$

Es folgt $\sqrt{\sigma^2+\omega^T\omega} \le \sigma$. Damit muss $\omega = 0$ gelten.
Wende dasselbe Argument auf $\hat{B}_1$ an. Per Induktion folgt schließlich die Behauptung. □

§ 3 Die inf-sup-Bedingung

Seien X, M Hilberträume. Sei weiterhin

$$\begin{aligned} a: &\quad X \times X \to \mathbb{R} \quad \text{eine symmetrische, beschränkte, positiv definite Bilinearform,} \\ b: &\quad X \times M \to \mathbb{R} \quad \text{eine beschränkte Bilinearform,} \\ f: &\quad X \to \mathbb{R} \quad \text{linear und beschränkt,} \\ g: &\quad M \to \mathbb{R} \quad \text{linear und beschränkt,} \\ V &= \{\, v \in X : b(v,q) = 0 \ \forall q \in M \,\} \ . \end{aligned}$$

Suche $(u,p) \in X \times M$ mit

$$\begin{aligned} a(u,v) + b(v,p) &= f(v) \quad \forall v \in X \ , \\ b(u,q) &= g(q) \quad \forall q \in M \ . \end{aligned} \tag{VI.4}$$

Satz 1. (VI.4) *hat genau dann eine eindeutige Lösung* $(u,p) \in X \times M$*, wenn* $\alpha > 0$ *und* $\beta > 0$ *existieren mit*

(a) $a(v,v) \geq \alpha \|v\|_X^2 \ \forall v \in V$ *(d. h.* $a|_{V \times V}$ *V-elliptisch) und*

(b) $\inf\limits_{0 \neq q \in M} \sup\limits_{0 \neq v \in X} \dfrac{b(v,q)}{\|v\|_X \cdot \|q\|_M} \geq \beta$ *(Babuška-Brezzi-Bedingung).*

Zu (b) äquivalent:

$$\begin{aligned} & \forall 0 \neq q \in M : \sup_{0 \neq v \in X} \frac{b(v,q)}{\|v\|_X \cdot \|q\|_M} \geq \beta \\ \Leftrightarrow \quad & \forall 0 \neq q \in M \ \exists 0 \neq v \in X : \frac{b(v,q)}{\|v\|_X \cdot \|q\|_M} \geq \beta \ . \end{aligned}$$

Beweis nur für endlichdimensionalen *Fall, wie für finite Elemente benötigt.*
Sei $(\varphi_1, \ldots, \varphi_n)$ Orthonormalbasis von X und $(\psi_1, \ldots, \psi_n)$ Orthonormalbasis von M.

$$\begin{aligned} X \ni v &= \sum_{i=1}^{n} v_i \varphi_i \ , \quad \boldsymbol{v} = (v_i)_{i=1}^n \in \mathbb{R}^n \ , \\ M \ni q &= \sum_{k=1}^{m} q_k \psi_k \ , \quad \boldsymbol{q} = (q_k)_{k=1}^m \in \mathbb{R}^m \ . \end{aligned}$$

Da (φ_i) und (ψ_k) Orthonormalbasen sind, gilt $\|v\|_X = \|\boldsymbol{v}\|$ und $\|q\|_M = \|\boldsymbol{q}\|$, wobei $\|\cdot\|$ die euklidische Norm bezeichnet.
(VI.4) ist äquivalent zu

$$\begin{aligned} A\boldsymbol{u} + B^T \boldsymbol{p} &= \boldsymbol{f} \ , \quad \boldsymbol{f} = (f_i) \text{ mit } f_i = f(\varphi_i) \ , \\ B\boldsymbol{u} &= \boldsymbol{g} \ , \quad \boldsymbol{g} = (g_k) \text{ mit } g_k = g(\psi_k) \ . \end{aligned}$$

Sei ohne Einschränkung $m \le n$ (für $m > n$: überbestimmt). Lasse im Folgenden die Fettschreibweise (Beispiele: $\boldsymbol{u}$, $\boldsymbol{p}$, $\boldsymbol{f}$, $\boldsymbol{g}$) weg.

$$A = \Big(a(\varphi_i, \varphi_j)\Big)_{i,j=1}^{n} , \quad B = \Big(b(\varphi_i, \psi_k)\Big)_{\substack{i=1,\dots,n \\ k=1,\dots,m}} .$$

Singulärwertzerlegung von B: $\underset{m\times n}{\boxed{B}} = \underset{m\times m}{U}\,\underset{m\times n}{\Sigma}\,\underset{n\times n}{V^T}$ mit orthogonalen Matrizen U und V.

$$b(v,q) = q^T B v = \underbrace{(U^T q)^T}_{=:\hat{q}^T}\,\Sigma\,\underbrace{(V^T v)}_{=:\hat{v}} = \sum_{k=1}^{m} \sigma_k \hat{q}_k \hat{v}_k ,$$

wobei

$$\Sigma = \left(\begin{array}{cccc|cccc} \sigma_1 & 0 & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & & \vdots & \vdots & \ddots & & \vdots \\ \vdots & & \ddots & 0 & \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_m & 0 & \cdots & \cdots & 0 \end{array}\right) = \Big(\Sigma_1 \Big| 0\Big) \Big\}\, m ,$$

wobei $\sigma_1 \ge \cdots \ge \sigma_m \ge 0$. Weiter:

$$\sup_{\|v\|_X=1} b(v,q) = \sup_{\|\hat{v}\|=1} \sum_{k=1}^{m} \sigma_k \hat{q}_k \hat{v}_k = \max_{k=1,\dots,m} |\sigma_k \hat{q}_k| ,$$

$$\inf_{\|q\|_M=1} \sup_{\|v\|_X=1} b(v,q) = \inf_{\|\hat{q}\|=1} \max_{k=1,\dots,m} |\sigma_k \hat{q}_k| \ge \beta \quad \Leftrightarrow \quad \sigma_m > 0 \quad \Leftrightarrow \quad \Sigma_1 \text{ invertierbar}$$

$$\overset{\text{aus erstem Term}}{\overset{\downarrow}{=}} \quad \inf_{M \ni q \neq 0} \sup_{X \ni v \neq 0} \frac{b(v,q)}{\|v\|_X \cdot \|q\|_M} .$$

Somit:

$$\underset{\text{bzw. (b)}}{\text{inf-sup-Bedingung}} \quad \Leftrightarrow \quad \sigma_m > 0 \quad \Leftrightarrow \quad \Sigma_1 \text{ invertierbar} \quad \Leftrightarrow \quad \frac{1}{\sigma_m} = \big\|\Sigma_1^{-1}\big\| \le \frac{1}{\beta} .$$

Spalte auf:

$$\hat{u} = V^T u = \begin{pmatrix} \hat{u}_1 \\ \hat{u}_0 \end{pmatrix} \begin{matrix} \} & m \\ \} & n-m \end{matrix} , \qquad \hat{p} = U^T p .$$

$$V^T A V = \begin{pmatrix} A_{11} & A_{10} \\ A_{01} & A_{00} \end{pmatrix} \begin{matrix} \} & m \\ \} & n-m \end{matrix} , \qquad V^T f = \hat{f} = \begin{pmatrix} \hat{f}_1 \\ \hat{f}_0 \end{pmatrix} , \qquad \hat{g} = U^T g .$$

Damit:

$$\text{(VI.4)} \qquad \Leftrightarrow \qquad \left\{ \begin{array}{ll} A_{11}\hat{u}_1 + A_{10}\hat{u}_0 + \Sigma_1 \hat{p} = \hat{f}_1 , \\ A_{01}\hat{u}_1 + A_{00}\hat{u}_0 \qquad\quad = \hat{f}_0 , \\ \underbrace{\Sigma_1}_{\text{invertierbar}} \hat{u}_1 \qquad\qquad\quad = \hat{g} . \end{array} \right. \qquad \text{(VI.5)}$$

Berechne $\hat{u}_1$ in der untersten Gleichung und setze es in die oberen Gleichungen ein. Berechne dann mithilfe der mittleren Gleichung $\hat{u}_0$ und setze es oben ein. Berechne schließlich mit der obersten Gleichung $\hat{p}$. Also müssen nur Σ_1 und A_{00} invertierbar sein, um (VI.5) lösen zu können.
A positiv semidefinit $\Rightarrow$ A_{00} positiv semidefinit.
Daher: A_{00} invertierbar $\Leftrightarrow$ A_{00} positiv definit. Also:

$$\text{(a)} \quad \Leftrightarrow \quad \lambda_{\min}(A_{00}) \geq \alpha \quad \Leftrightarrow \quad \|A_{00}^{-1}\| \leq \frac{1}{\alpha} \ .$$

Somit:

$$\begin{aligned} \text{(VI.4) eindeutig lösbar} \quad &\Leftrightarrow \quad \text{(VI.5) eindeutig lösbar} \\ &\Leftrightarrow \quad \Sigma_1 \text{ und } A_{00} \text{ invertierbar} \\ &\Leftrightarrow \quad \exists \beta > 0,\ \alpha > 0 \text{ mit (b), (a).} \end{aligned}$$ □

Beachte.

$$\|A_{kl}\| \leq \|A\| = \sup_{\|v\|_X = \|w\|_X = 1} |a(v,w)| \leq K \ .$$

Erhalte damit aus (VI.5):

$$\begin{aligned} &\|u_1\| \leq \frac{1}{\sigma_m}\|g\| \ , && \|\Sigma_1^{-1}\| = \frac{1}{\sigma_m} \ , \\ &\|u_0\| \leq \frac{1}{\alpha}\Big(\underbrace{\|f_0\|}_{\leq \|f\|} + K \underbrace{\frac{1}{\sigma_m}}_{\leq \frac{1}{\beta}^{\ddagger}} \|g\| \Big) \ , && \|A_{00}^{-1}\| \leq \frac{1}{\alpha} \ , \\ &\|p\| \leq \frac{1}{\sigma_m}\Big(\underbrace{\|f_1\|}_{\leq \|f\|} + K\|u\| \Big) \ . \end{aligned}$$

Weiterhin gilt noch $\|u\| \leq \|u_1\| + \|u_0\|$.

Folgerung. (a), (b) $\Rightarrow$ $\|u\|_X + \|p\|_M \leq C\,(\|f\|_{X'} + \|g\|_{M'})$, wobei C nur von α, $\sigma_m = \beta^{\ddagger}$, K abhängt und

$$\begin{aligned} \|f\|_{X'} &= \sup_{\|v\|_X = 1} |f(v)| = \sup_{\|\boldsymbol{v}\| = 1} |\sum_{i=1}^{n} f_i v_i| = \|\boldsymbol{f}\| \ , \\ \|g\|_{M'} &= \sup_{\|q\|_M = 1} |g(q)| = \|\boldsymbol{g}\| \ . \end{aligned}$$

Bemerkung. (b) $\Leftrightarrow \forall q \in M\ \exists v \in X$ mit $b(v,q) \geq \beta \|q\| \cdot \|v\|$.

[‡]Ersetze σ_m durch β, da gilt: $\beta := \inf \sup \frac{b(v,q)}{\|v\| \cdot \|q\|} \leq \sigma_m$, $\frac{1}{\sigma_m} \leq \frac{1}{\beta}$.

§ 4 Ein Approximationssatz für gemischte finite Elemente

Seien X_h, M_h endlichdimensionale Unterräume der Hilberträume X, M.

$$\begin{cases} a(u_h, v_h) + b(v_h, p_h) = f(v_h) & \forall v_h \in X_h \ , \\ b(u_h, q_h) = g(q_h) & \forall q_h \in M_h \ . \end{cases} \tag{VI.6}$$

Satz 1. *Für X_h, M_h seien mit α, β unabhängig von h die Babuška-Brezzi-Bedingungen erfüllt. Dann hat* (VI.6) *eine eindeutige Lösung $(u_h, p_h) \in X_h \times M_h$ und*

$$\|u_h - u\|_X + \|p_h - p\|_M \le c\Big(\inf_{v_h \in X_h} \|v_h - u\|_X + \inf_{q_h \in M_h} \|q_h - p\|_M \Big)$$

mit c unabhängig von h (vergleiche Céas Lemma).

Beweis. $a(u, v) + b(v, p) = f(v)\ \forall v \in X$, insbesondere $\forall v \in X_h \subset X$, und es gilt auch

$$a(u_h, v_h) + b(v_h, p_h) = f(v_h) \quad \forall v_h \in X_h \ .$$

Daher für beliebige $\tilde{v}_h \in X_h$, $\tilde{q}_h \in M_h$:

$$a(u_h - \tilde{v}_h, v_h) + b(v_h, p_h - \tilde{q}_h) = a(u - \tilde{v}_h, v_h) + b(v_h, p - \tilde{q}_h) \quad \forall v_h \in X_h \ .$$

Ebenso

$$b(u_h - \tilde{v}_h, q_h) = b(u - \tilde{v}_h, q_h) \quad \forall q_h \in M_h \ .$$

Mit Folgerung aus § 3 (in X_h, M_h):

$$\|u_h - \tilde{v}_h\| + \|p_h - \tilde{q}_h\| \le C\Big[\sup_{\|v_h\|_X = 1} |a(u - \tilde{v}_h, v_h) + b(v_h, p - \tilde{q}_h)| + \sup_{\|q_h\|_M = 1} |b(u - \tilde{v}_h, q_h)| \Big]$$

$$\overset{\substack{a,\ b \text{ beschränkt} \\ \downarrow}}{\le} \hat{C}\Big[\|u - \tilde{v}_h\|_X + \|p - \tilde{q}_h\|_M \Big] \ .$$

Erhalte die Behauptung mit der Dreiecksungleichung, angewendet auf die linke Seite:

$$\|u_h - u\| \le \|u_h - \tilde{v}_h\| + \|\tilde{v}_h - u\| \ , \quad \|p_h - p\| \le \|p_h - \tilde{q}_h\| + \|\tilde{q}_h - p\| \ . \qquad \square$$

Kriterium für die Babuška-Brezzi-Bedingung in X_h, M_h

Hilfssatz 2. *Auf X, M gelte die inf-sup-Bedingung mit Konstante β:*

$$\inf_{q \in M} \sup_{v \in X} \frac{b(v, q)}{\|v\|_X \|q\|_M} \ge \beta > 0 \ .$$

Falls es eine lineare Abbildung $\Pi_h : X \to X_h$ gibt mit

$$\begin{cases} b(v - \Pi_h v, q_h) = 0 & \forall q_h \in M_h \ , \\ \|\Pi_h v\|_X \le c_0 \|v\|_X & \forall v \in X \ , \end{cases}$$

so ist

$$\inf_{q_h \in M_h} \sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_X \|q_h\|_M} \ge \frac{\beta}{c_0} \ .$$

Beweis. Inf-sup-Bedingung auf X, M,

$$\forall q \in M \; \exists v \in X : \frac{b(v,q)}{\|v\|\|q\|} \geq \beta \; ,$$

insbesondere $\forall q_h \in M_h \subset M$.
Habe $\Pi_h v \neq 0$, denn sonst wäre

$$b(v, q_h) = b(v - \Pi_h v, q_h) = 0$$

im Widerspruch zur inf-sup-Bedingung.
Damit:

$$\frac{b(\Pi_h v, q_h)}{\|\Pi_h v\| \cdot \|q_h\|} = \frac{b(v, q_h)}{\|\Pi_h v\| \cdot \|q_h\|} \geq \frac{1}{c_0} \frac{b(v, q_h)}{\|v\| \cdot \|q_h\|} \geq \frac{\beta}{c_0} \; . \qquad \square$$

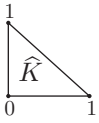
§ 5 Finite Elemente für das Stokes-Problem

Zweidimensionaler Fall

1. Versuch  

	Geschwindigkeit zwei Komponenten $u_h \in X_h$	Druck skalar $p_h \in M_h$	
	$(P_1)^2$	P_1	instabil!
ebenso	$(P_2)^2$	P_2	instabil!
2. Versuch	$(P_1)^2$	P_0	instabil!
3. Versuch	$(P_2)^2$	P_0	stabil ✓

Problem mit $(P_1)^2$, P_1: zu wenig Geschwindigkeiten in X_h.
(Brauche $\forall q_h \in M_h \; \exists v_h \in X_h$: $\dfrac{b(v_h, q_h)}{\|\%\| \cdot \|\%\|} \geq \beta$.)

Idee. Füge zum Finiten-Elemente-Raum der Geschwindigkeiten „bubble functions“ hinzu: Die Bubble-Funktionen seien stückweise kubisch und verschwinden auf allen Dreiecksseiten.
Für das Referenzelement 

lautet die bubble function speziell

$$\hat{b}(\hat{x}_1, \hat{x}_2) = \hat{x}_1 \hat{x}_2 (1 - \hat{x}_1 - \hat{x}_2) \; .$$

Damit ergibt sie sich auf einem allgemeinen Dreieckselement K mit einer affinen Transformation $F : \hat{K} \to K$ zu

$$b_K(x) = \hat{b}(\hat{x}) \quad \text{für} \quad x = F(\hat{x}) \; ,$$

$$X_h = \left\{ v_h \in \left(H_0^1(\Omega) \right)^2 : v_h\big|_K \in (P_1 \oplus \mathbb{R} b_K)^2 \; \forall K \in T_h \right\} \; ,$$

$$M_h = \left\{ q_h \in L^2(\Omega) : q_h\big|_K \in P_1 \; \forall K \in T_h \; , \quad q_h \text{ stetig} \; , \quad \int_\Omega q_h \, dx = 0 \right\} \; .$$

Das sogenannte *Mini-Element* ist stabil! Es erfüllt die Brezzi-Bedingungen gleichmäßig in h.

Satz 1. *Sei T_h eine Familie von Triangulierungen mit $h_K/\rho_K \leq \text{const}\ \forall K \in T_h\ \forall h$, wobei h_K den Durchmesser und ρ_K den Inkreisradius des Dreiecks K bezeichnet.*
Dann erfüllen die Räume X_h, M_h des Mini-Elements die Babuška-Brezzi-Bedingungen für das Stokes-Problem gleichmäßig in h. Es ist

$$\|u_h - u\|_1 + \|p_h - p\|_0 \leq ch(|u|_2 + |p|_1) \ .$$

Idee. Wähle im Hilfssatz von § 4 Π_h als lineare Interpolation plus eine Korrektur mit Bubble-Funktionen so, dass

$$b(v, q_h) = b(\Pi_h v, q_h) \quad \forall q_h \in M_h\ \forall v \in \left(H_0^1(\Omega)\right)^2 \ .$$

Technische Schwierigkeit: $H_0^1(\Omega) \not\subset C(\Omega)$ für $\Omega \subset \mathbb{R}^2$. Die Knotenwerte $v(q_i)$ sind nicht definiert! Verwende stattdessen lokale Mittelung: Für $v \in H_0^1(\Omega)$ setze

$$R_h v = \sum_{i=1}^{N} \underbrace{\frac{\int_\Omega v\varphi_i\,dx}{\int_\Omega \varphi_i\,dx}}_{\text{statt } v(q_i)} \varphi_i$$

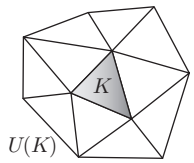
und für $v = (v_1, v_2) \in \left(H_0^1(\Omega)\right)^2$ komponentenweise.

Hilfssatz 2.

$$\begin{aligned} |R_h v|_1 &\leq C|v|_1 \\ \|R_h v - v\|_0 &\leq Ch|v|_1 \end{aligned} \qquad \forall v \in H_0^1(\Omega) \ .$$

Beweis. Unter (schwachen) zusätzlichen Voraussetzungen an die Triangulierung:
Zu $K \in T_h$ sei

$$U(K) = \bigcup \left\{ K^1 \in T_h : K \cap K^1 \neq \emptyset \right\} \ .$$



Voraussetzungen:
$U(K)$ sei konvex $\forall K \in T_h$ (für Hilfssatz 1, § 5, Kapitel IV).

$$\forall K^1 \in U(K) : c_1 h_K \leq h_{K^1} \leq C_1 h_K \quad \text{(mit Konstanten } c_1,\ C_1 > 0).$$

(Die Triangulierung ist *lokal quasiuniform.*)

Damit ist der Beweis des Hilfssatzes mit früheren Ergebnissen möglich.

Beachte. Für alle konstanten Funktionen $c \in \mathbb{R}$ gilt $R_h c = c$.

Habe $\forall K \in T_h$ und $\forall c \in \mathbb{R}$

$$|R_h v|_{1,K} = |R_h v - c|_{1,K} = |R_h(v - c)|_{1,K}$$

und es folgt mit Hilfssatz 2 („Inverse Abschätzung“), § 5 aus Kapitel V (in diesem Semester nicht in der Vorlesung vorgeführt) weiterhin

$$\leq \frac{c}{h_K}\|R_h(v-c)\|_{0,K} \quad \overset{\substack{R_h \text{ beschränkt} \\ \text{(Cauchy-Schwarzsche Ungleichung)} \\ \downarrow}}{\leq} \quad \frac{c'}{h_K}\|v - c\|_{0,U(K)} \ . \qquad \square$$

Wähle c als Mittelwert von v auf $U(K)$, verwende Hilfssatz 1, § 5, Kapitel IV:

$$\|v-c\|_{0,U(K)} \le \sqrt{2}\cdot \underbrace{\operatorname{diam}\big(U(K)\big)}_{C'' h_K} |v|_{1,U(K)} \ ,$$

somit

$$|R_h v|_{1,K} \le \tilde{C}|v|_{1,U(K)}$$

und damit (da jedes Dreieck K nur in endlich vielen $U(K)$ enthalten ist)

$$|R_h v|_{1,\Omega} \le C|v|_{1,\Omega} \ .$$

Zudem:

$$\|R_h v - v\|_{0,K} = \|R_h(v-c)-(v-c)\|_{0,K} \le \tilde{C}\|v-c\|_{0,U(K)} \overset{\substack{c \text{ ist Mittelwert}\\ \downarrow}}{\le} \hat{C} h_K |v|_{1,U(K)} \ .$$

Damit:

$$\|R_h v - v\|_{0,\Omega} \le C|v|_{1,\Omega} \ .$$

Beweis des Satzes. Zeige, dass die Bedingungen des Hilfssatzes aus § 4 erfüllt sind. Brauche dazu $\Pi_h : X \to X_h$, wobei $X = (H_0^1(\Omega))^2$, mit

(i) $|\Pi_h v|_1 \le C|v|_1 \ \forall v \in X$ und

(ii) $b(v-\Pi_h v, q_h) = 0 \ \forall v \in X$ und $\forall q_h \in M_h \subset H^1(\Omega)$.

(ii) bedeutet

$$-\int_\Omega \operatorname{div}(v-\Pi_h v) q_h \, dx \overset{!}{=} 0$$

für alle stetigen, stückweise linearen Funktionen q_h und mit der Greenschen Formel (Randterme verschwinden)

$$= \int_\Omega (v-\Pi_h v) \underbrace{\operatorname{grad} q_h}_{\text{stückweise konstant}} dx \overset{!}{=} 0 \ .$$

Brauche also

$$\int_K (v-\Pi_h v)\, dx = 0 \quad \forall K \in T_h \quad \forall v \in X \ . \tag{VI.7}$$

Definiere $\Pi_h v \in X_h$ stückweise durch

$$\Pi_h v\big|_K = \underbrace{R_h v\big|_K}_{\in P_1} + \alpha_K \underbrace{b_K^{\ddagger}}_{\text{bubble}} \quad \text{für ein } \alpha_K \in \mathbb{R}.$$

Klar: $\Pi_h v \in X_h$ (stetig in Ω, weil $R_h v$ stetig und $b_K = 0$ auf ∂K).
Wähle α_K so, dass (VI.7) gilt:

$$\alpha_K \underbrace{\int_K b_K(x)\, dx}_{\ge \gamma h_K^2 \ (\gamma>0)} \overset{!}{=} \int_K (v - R_h v)\, dx \overset{\substack{\text{Cauchy-Schwarzsche}\\ \text{Ungleichung}\\ \downarrow}}{\le} \|v-R_h v\|_{0,K} \cdot \underbrace{\left(\int_K 1\, dx\right)^{\frac{1}{2}}}_{\le C'' h_K} \ .$$

[‡]Die bubble function b_K ist nicht zu verwechseln mit der Bilinearform b!

Beachte. $b_K = \mathcal{O}(1)$, $\dfrac{\partial b_K}{\partial x} = \mathcal{O}(h_K^{-1})$.
Zeige noch (i): $|\Pi_h v|_1 \le C|v|_1 \ \forall v \in X = (H_0^1(\Omega))^2$.

$$|\Pi_h v|_1 \le \overbrace{|R_h v|_1}^{\le C|v|_1 \text{ nach Hilfssatz}} + \Big|\sum_K \alpha_k b_K\Big|_1$$

$$\Big|\sum_K \alpha_K b_K\Big|_1^2 = \sum_{K\in T_h} \int_K \alpha_K^2 \underbrace{\Big[\Big(\frac{\partial b_K}{\partial x_1}\Big)^2 + \Big(\frac{\partial b_K}{\partial x_2}\Big)^2\Big]}_{=\mathcal{O}(h_K^{-2})}$$

$$\overset{\substack{\text{mit obiger}\\ \text{Ungleichung für } \alpha_K}}{\downarrow}$$

$$\le \sum_K \underbrace{\|v - R_h v\|_{0,K}^2}_{\le (Ch_K|v|_1)^2} \cdot \frac{1}{\gamma^2} h_K^{-2} \cdot C' h_K^{-2} \cdot \underbrace{\int_K 1\, dx}_{\le C'' h_K^2} \le \tilde{C}|v|_1^2 \ .$$

Damit: $|\Pi_h v|_1 \le C|v|_1$.
Voraussetzungen des Hilfssatzes aus § 4 erfüllt. Damit ist die Brezzi-Bedingung gleichmäßig in h erfüllt.
Nach Hilfssatz, § 4:

$$\|u_h - u\|_1 + \|p_h - p\|_0 \le C\Big\{ \inf_{v_h \in X_h} \|v_h - u\|_1 + \inf_{q_h \in M_h} \|q_h - p\|_0 \Big\}$$

$$\overset{\substack{(P_1)^2 \subset X_h,\\ P_0 \subset M_h}}{\downarrow}$$

$$\le C\left(C_1 h|u|_2 + C_2 h|p|_1\right) \ . \qquad \square$$

Die zweite Ungleichung folgt aus den Abschätzungen in § 5, Kapitel IV.

Bemerkung. Es existiert eine Vielzahl von stabilen finiten Elementen für das Stokes-Problem und verwandte Sattelpunktprobleme („mixed finite elements“).

Literatur: [BrezFor91], [GirRav86].

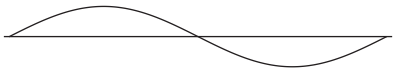
Kapitel VII

Eigenwertprobleme

Suche Eigenwert $\lambda \in \mathbb{C}$ und Eigenfunktionen $u : \Omega \to \mathbb{R}$ mit

$$\begin{cases} -\Delta u = \lambda u & \text{in } \Omega \ , \\ u = 0 & \text{auf } \partial\Omega \ , \end{cases} \qquad \text{(VII.1)}$$

$\Omega \subset \mathbb{R}^d$ Gebiet.



Beispiel (Eigenfrequenzen einer Trommel)**.** $u(x,t)$ Auslenkung am Ort x zur Zeit t. Wellengleichung:

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} &= \Delta u \quad \text{für } x \in \Omega \ , \qquad t \in \mathbb{R} \ , \\ u &= 0 \qquad \text{für } x \in \partial\Omega \ . \end{aligned}$$

Suche Lösungen der Form (stationäre Schwingung):

$$u(x,t) = w(x)\varphi(t) \ .$$

Einsetzen in die Wellengleichung

$$w(x)\varphi''(t) = \Delta w(x)\varphi(t) \ ,$$

d. h.

$$\frac{\varphi''(t)}{\varphi(t)} = \frac{\Delta w(x)}{w(x)} = -\lambda \ , \quad \lambda = \text{const} \ , \ \text{unabhängig von } x, t.$$

Löse (VII.1) für w und $\varphi'' + \lambda\varphi = 0$.
Habe $\lambda > 0$ (wird später gezeigt), erhalte mit $\lambda = \omega^2$

$$u(x,t) = \overbrace{w(x)}^{\text{Eigenfunktion aus (VII.1)}} \underbrace{\left(C_1 \cos \omega t + C_2 \sin \overbrace{\omega}^{\omega \text{ Eigenfrequenz}} t\right)}_{A_1 e^{i\omega t} + A_2 e^{-i\omega t}} \ .$$

Fragen. Existenz, Eindeutigkeit von Eigenpaaren (Eigenwerte und Eigenfunktionen), Approximation, Berechnung?

§ 1 Spektraltheorie kompakter Operatoren

Sei H ein Hilbertraum mit Skalarprodukt $(\cdot,\cdot)$ und Norm $|\cdot|$, $|v| = \sqrt{(v,v)}$.
Im Folgenden oft: H separabel, d. h. H hat abzählbare dichte Teilmenge oder äquivalent (ohne Beweis).
H hat eine *Hilbertbasis* $(e_n)_{n\geq 0}$:

- e_n orthonormal: $(e_n, e_m) = \begin{cases} 1\ , & n = m, \\ 0\ , & \text{sonst.} \end{cases}$
- $v = \sum_{n=0}^{\infty} (v, e_n) e_n \ \forall v \in H.$
- $|v|^2 = \sum_{n=0}^{\infty} |(v, e_n)|^2$, (anders: $H \cong l^2$, $v \mapsto \Big((v, e_n)\Big)_{n\geq 0}$).

$T : H \to H$ *kompakt*, falls für jede beschränkte Folge (v_n) in H die Bildfolge (Tv_n) eine konvergente Teilfolge hat.
$T : H \to H$ linear.
Spektrum:

$$\sigma(T) = \{\, \lambda \in \mathbb{C} : (T - \lambda I) \text{ nicht bijektiv} \,\}\ .$$

Punktspektrum:

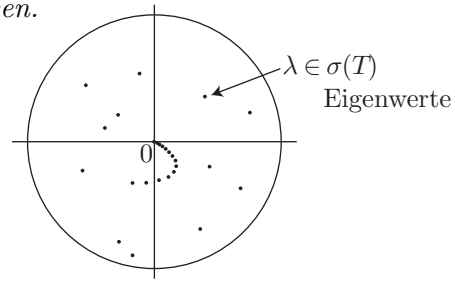
$$\begin{aligned} \sigma_p(T) &= \{\, \lambda \in \mathbb{C} : (T - \lambda I) \text{ nicht injektiv} \,\} \\ &= \{\, \lambda \in \mathbb{C} : \exists 0 \neq v \in H : \underbrace{(T - \lambda I)v}_{Tv = \lambda v} = 0 \,\} \\ &= \{\, \lambda \in \mathbb{C} : \lambda \text{ Eigenwert von } T \,\}\ . \end{aligned}$$

Satz 1 (Fredholm (1900), Schmidt (1906), Riesz (1913))**.** *Sei H Hilbertraum, $T : H \to H$ kompakt, $\lambda \in \mathbb{C}$, $\lambda \neq 0$. Dann:*

(a) Entweder hat $(\lambda I - T)$ eine beschränkte Inverse oder λ ist ein Eigenwert von T. (Mit anderen Worten: $(\lambda I - T)$ surjektiv $\Leftrightarrow$ $(\lambda I - T)$ injektiv.)

(b) $\dim \ker(\lambda I - T) < \infty$, $\operatorname{Im}(\lambda I - T)$ *abgeschlossen.*
$\dim \ker(\lambda I - T) = \dim\big(H / \operatorname{Im}(\lambda I - T)\big)$.

(c) Das Spektrum $\sigma(T)$ besteht aus höchstens abzählbar vielen Elementen, die sich nur in 0 häufen können.



(d) Falls $\dim H = \infty$, *so ist* $0 \in \sigma(T)$ *(0 muss kein Eigenwert sein).*

§ 2 Spektralzerlegung symmetrischer kompakter Operatoren

Motivation: Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch. Dann hat der $\mathbb{R}^n$ eine Orthonormalbasis aus Eigenvektoren von A. Zeige analoges Resultat für symmetrische kompakte Operatoren auf Hilberträumen.

Situation: H Hilbertraum mit Skalarprodukt $(\cdot,\cdot)$, Norm $\|\cdot\|$. $T : H \to H$ sei linear, kompakt und *symmetrisch*, d. h. $(Tu, v) = (u, Tv)\ \forall u,\ v \in H$.
Klar: Alle Eigenwerte von T reell, denn: Falls $Tu = \lambda u$ mit Eigenvektor $u \neq 0$, so ist

$$\lambda(u,u) = (Tu,u) \overset{\substack{T \text{ symmetrisch}\\ \downarrow}}{=} (u,Tu) = (u,\lambda u) = \overline{\lambda}(u,u)\ .$$

Charakterisiere im folgenden Hilfssatz den größten und den kleinsten Eigenwert (sofern nicht 0) durch den *Rayleigh-Quotienten*:

$$R(u) = R_T(u) = \frac{(Tu,u)}{(u,u)}\ , \qquad (0 \neq u \in H)\ .$$

Hilfssatz 1. *Falls $\Lambda := \sup_{u\neq 0} R(u) \neq 0$, so ist Λ der größte Eigenwert von T. (Entsprechend: Ist $\Lambda := \inf_{u\neq 0} R(u) \neq 0$, so ist Λ kleinster Eigenwert.)*

Beweis. Betrachte symmetrische, positiv semidefinite Bilinearform a:

$$a(u,v) = (\Lambda u - Tu, v)\ .$$

Mit der Cauchy-Schwarzschen Ungleichung ergibt sich

$$|a(u,v)| \leq \sqrt{a(u,u)}\sqrt{a(v,v)} \quad \forall u,v \in H\ ,$$

d. h.

$$|(\Lambda u - Tu, v)| \leq \sqrt{(\Lambda u - Tu, u)}\sqrt{(\Lambda v - Tv, v)}\ .$$

Für den zweiten Faktor der rechten Seite erhalte weiterhin

$$\sqrt{(\Lambda v - Tv, v)} \overset{\substack{\text{Cauchy-Schwarzsche}\\ \text{Ungleichung}\\ \downarrow}}{\leq} \sqrt{\|(\Lambda I - T)v\| \cdot \|v\|} \overset{\substack{\text{Verträglichkeit von}\\ \text{Vektor- und Matrixnorm}\\ \downarrow}}{\leq} \sqrt{\|\Lambda I - T\| \cdot \|v\|} \cdot \sqrt{\|v\|} \overset{\substack{T \text{ beschränkt}\\ \downarrow}}{\leq} C\|v\|\ ,$$

oben eingesetzt:

$$|(\Lambda u - Tu, v)| \leq C\sqrt{(\Lambda u - Tu, u)} \cdot \|v\|\ .$$

Für $v = \Lambda u - Tu$ erhalte dann

$$\begin{aligned} & \|\Lambda u - Tu\|^2 \leq C\sqrt{(\Lambda u - Tu, u)} \cdot \|\Lambda u - Tu\| \\ \Leftrightarrow \qquad & \|\Lambda u - Tu\| \ \leq C\sqrt{(\Lambda u - Tu, u)}\ . \end{aligned} \tag{VII.2}$$

Sei (u_n) Folge in H mit $\|u_n\| = 1$, sodass $(Tu_n, u_n) = R(u_n) \to \Lambda$. Da T kompakt hat (Tu_n) eine konvergente Teilfolge, ohne Einschränkung $Tu_n \to w$.

Habe

$$\|\Lambda u_n - w\| \leq \|\Lambda u_n - Tu_n\| + \underbrace{\|Tu_n - w\|}_{\to 0}$$
$$\overset{\text{(VII.2)}}{\downarrow}$$
$$\leq C\sqrt{(\Lambda u_n - Tu_n, u_n)} = C\sqrt{\Lambda - R(u_n)} \to 0 \ .$$

Damit $u_n \to w/\Lambda =: u$, $\|u\| = \lim\|u_n\| = 1$, da $\|u_n\| = 1$.

$$\left.\begin{array}{l} Tu_n \to Tu \\ \downarrow \\ w = \Lambda u \end{array}\right\rangle \begin{array}{l} Tu = \Lambda u \ , \quad u \neq 0 \ , \\ \text{also } \Lambda \text{ Eigenwert.} \end{array}$$

Falls λ weiterer Eigenwert zu Eigenvektor v ist, so ist $\lambda = R(v) \leq \Lambda$. □

Folgerung. Situation wie oben. Falls das Spektrum von T nur aus 0 besteht, so ist $T = 0$.

Beweis. Nach Hilfssatz 1 ist $R(u) = 0 \ \forall 0 \neq u \in H$, d. h. $(Tu, u) = 0 \ \forall u \in H$. Dann gilt

$$2(Tu, v) \overset{\text{Parallelogrammidentität}}{=} \underbrace{\Big(T(u+v), (u+v)\Big)}_{0} - \underbrace{(Tu, u)}_{0} - \underbrace{(Tv, v)}_{0} = 0$$
$$\Rightarrow \quad Tu = 0 \ \forall u \in H \ , \quad \text{d. h. } T = 0.$$ □

Satz 2. *Sei H ein separabler Hilbertraum (d. h. H hat eine abzählbare Teilmenge, die dicht in H liegt) und sei T ein symmetrischer, kompakter linearer Operator auf H. Dann hat H eine Hilbertbasis aus Eigenvektoren von T.*

Beweis. Sei $(\lambda_n)_{n\geq 1}$ Folge der von 0 verschiedenen Eigenwerte von T, $\lambda_0 := 0$.
Setze $E_n := \ker(\lambda_n I - T)$, $n \geq 0$.
Weiß $0 \leq \dim E_0 \leq \infty$ und $0 < \dim E_n < \infty$ für $n \geq 1$ (aus Satz 1).

(a) Die Räume E_n *sind paarweise orthogonal*: Sei $u \in E_m$, $v \in E_n$, $m \neq n$. $Tu = \lambda_m u$, $Tv = \lambda_n v$.

$$\lambda_m(u, v) = (Tu, v) \overset{\text{symmetrisch}}{=} (u, Tv) = \lambda_n(u, v) \quad \Rightarrow \quad (u, v) = 0 \ .$$

(b) Sei F der von den $(E_n)_{n\geq 0}$ erzeugte Vektorraum.

$$F \ni v = \sum_{\text{endlich}} v_n \ , \quad v_n \in E_n \ .$$

Zeige: F *dicht in* H. Wegen $H = \overline{F} \oplus F^\perp$ zeige $F^\perp = 0$.

Klar: $T(F) \subset F$ (wegen $Tv = \lambda_n v$ für $v \in E_n$).

Auch: $T(F^\perp) \subset F^\perp$, denn:

$$u \in F^\perp \ , \quad v \in F \quad \Rightarrow \quad (Tu, v) \overset{\text{symmetrisch}}{=} (u, Tv) \overset{\substack{u \in F^\perp \\ Tv \in F}}{=} 0 \ .$$

(i) $T_0 := T\big|_{F^\perp}$ ist symmetrisch, kompakt.

(ii) T_0 hat das Spektrum $\sigma(T_0) = \{0\}$, denn: Falls $\lambda \in \sigma(T_0)$ und $\lambda \neq 0$, so ist λ Eigenwert von T_0, d. h. $\exists 0 \neq u \in F^\perp$ mit $T_0 u = \lambda u = Tu$. Dann ist $0 \neq \lambda$ Eigenwert von T zum Eigenvektor u. Dann gilt aber $0 \neq u \in F$: Dies steht im Widerspruch zu $u \in F^\perp$.

Aus (i) und (ii) mit der Folgerung aus Hilfssatz 1: $T_0 = T\big|_{F^\perp} = 0$. Damit $F^\perp \subset \ker T = E_0 \subset F \Rightarrow F^\perp = 0$.

(c) $E_0 = \ker T$ ist ein abgeschlossener Unterraum von H, und damit ist E_0 selbst wieder ein separabler Hilbertraum.

Wähle in E_0 Hilbertbasis, in E_n $(n \geq 1)$ Orthonormalbasis aus Eigenvektoren von $T\big|_{E_n}$ $(\dim E_n < \infty)$.

Die Vereinigung dieser Basen ist Hilbertbasis von H aus Eigenvektoren von T. □

§ 3 Elliptische Eigenwertprobleme

Beispiel (Standardproblem)**.** Laplace-Operator auf $\Omega \subset \mathbb{R}^d$ mit Dirichlet-Randbedingungen:

$$-\Delta u = \lambda u \quad \text{in } \Omega\ , \quad u = 0 \quad \text{auf } \partial\Omega\ .$$

Variationelle Formulierung (multipliziere mit v, integriere über Ω):

$$\int_\Omega \sum_{i=1}^{d} \frac{\partial u}{\partial x_i} \cdot \frac{\partial v}{\partial x_i}\, dx = \lambda \int_\Omega uv\, dx \quad \forall v \in H_0^1(\Omega)\ .$$

Kurz:

$$a(u,v) = \lambda \underbrace{(u,v)_0}_{\text{Skalarprodukt zu } H = L^2(\Omega),\ V \subset H} \quad \forall v \in V = H_0^1(\Omega)\ .$$

Allgemeine Voraussetzungen: Sei V Hilbertraum, $a : V \times V \to \mathbb{R}$ eine symmetrische, V-elliptische Bilinearform (Skalarprodukt auf V). Sei H ein separabler Hilbertraum mit Skalarprodukt $(\cdot,\cdot)$, wobei $V \subset H$ dicht in H. Die Einbettung $V \hookrightarrow H$ ist *kompakt.*

Erinnerung. $H_0^1(\Omega) \subset L^2(\Omega)$ kompakt (Satz von Rellich, IV, § 6).

Bezeichnungen: $\|\cdot\| = \sqrt{a(\cdot,\cdot)}$ Norm auf V, $|\cdot| = \sqrt{(\cdot,\cdot)}$ Norm auf H.
Habe $|v| \leq C \cdot \|v\|$ $\forall v \in V$ (aus der Poincaré-Ungleichung).
Weiß (Lax-Milgram): Zu einem beliebigen $f \in H$ existiert genau ein $u = Sf \in H$ mit

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

Die Linearform $l(v) = (f,v)$ ist beschränkt durch

$$|l(v)| = |(f,v)| \leq |f| \cdot |v| \leq C|f| \cdot \|v\| \quad \forall v \in V\ .$$

Habe auch $\|u\| = \|Sf\| \leq \tilde{C}|f| \ \forall f \in H$. Erhalte damit beschränkte lineare Abbildung

$$S : H \to V : f \mapsto u = Sf \ .$$

Setze $T = S\big|_V : V \xhookrightarrow[\text{kompakt}]{} H \xrightarrow[\text{beschränkt}]{S} V$. Damit $T : V \to V$ kompakt.
Wegen $u = Tf$ gilt für jedes $f \in V$: $a(u, v) = a(Tf, v) = (f, v) \ \forall v \in V$.

Hilfssatz. *Sei $T : V \to V$ definiert durch*

$$a(Tw, v) = (w, v) \quad \forall v \in V \ , \quad w \in V \ .$$

Dann ist T ein kompakter linearer Operator auf V, symmetrisch bezüglich des Skalarprodukts $a(\cdot, \cdot)$ auf V:

$$a(Tw, v) = a(w, Tv) \quad \forall w, v \in V \ .$$

Weiterhin ist T positiv definit:

$$a(Tv, v) > 0 \quad \forall 0 \neq v \in V \ .$$

Beweis. T ist kompakt, da $V \xhookrightarrow{\text{kompakt}} H \xrightarrow{\text{stetig}} V$.
T ist symmetrisch und positiv definit, da

$$a(Tu, v) = (u, v) \overset{(\cdot,\cdot) \text{ symmetrisch}}{\downarrow =} (v, u) = a(Tv, u) \overset{a(\cdot,\cdot) \text{ symmetrisch}}{\downarrow =} a(u, Tv) \ ,$$
$$a(Tv, v) = (v, v) > 0 \quad \forall v \neq 0 \ . \qquad \square$$

Damit § 2 auf T anwendbar.
Sei μ Eigenwert von T zum Eigenvektor u:

$$(u, v) = a(Tu, v) = a(\mu u, v) = \mu a(u, v) \quad \forall v \in V \ ,$$

d. h. $a(u, v) = \lambda(u, v) \ \forall v \in V$ mit $\lambda = 1/\mu$ ($\mu \neq 0$, da T positiv definit).
u ist dann Eigenfunktion zu $a(\cdot, \cdot)$ zum Eigenwert λ.

Satz 1. *Voraussetzungen wie oben (insbesondere $V \hookrightarrow H$ kompakt). Dann gibt es zum Eigenwertproblem*

$$a(u, v) = \lambda(u, v) \quad \forall v \in V$$

eine Folge von Eigenwerten $0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \cdots \to \infty$ und eine Hilbertbasis von H aus Eigenvektoren $w_k \in V$. Die Folge $v_k = w_k/\sqrt{\lambda_k}$ bildet eine Hilbertbasis von V bezüglich des Skalarprodukts $a(\cdot, \cdot)$.

Beweis. Nach dem Hilfssatz ist T ein symmetrischer, kompakter linearer Operator auf V.
Nach Satz 2 aus § 2 hat V eine Hilbertbasis (v_k) aus Eigenvektoren von T zu Eigenwerten (μ_k) mit $\mu_k \to 0$.
Nach dem Hilfssatz ist T positiv definit: $\mu_k > 0$, also $0 < \cdots \leq \mu_3 \leq \mu_2 \leq \mu_1$. Ohne Einschränkung $Tv_k = \mu_k v_k$, daraus

$$a(Tv_k, v) = a(\mu_k v_k, v) = \mu_k a(v_k, v) \overset{\text{Definition von } T}{\downarrow =} (v_k, v) \quad \forall v \in V \ .$$

Also:

$$a(v_k, v) = \frac{1}{\mu_k}(v_k, v) \quad \forall v \in V \ , \quad \frac{1}{\mu_k} := \lambda_k \to +\infty \ .$$

Habe Hilbertbasis (v_k) aus Eigenvektoren in V.
Zeige nun noch, dass die $w_k := \sqrt{\lambda_k} v_k$ eine Hilbertbasis von H bilden:

$$(w_k, w_l) = \sqrt{\lambda_k}\sqrt{\lambda_l}(v_k, v_l) = \sqrt{\lambda_k}\sqrt{\lambda_l}\frac{1}{\lambda_k}a(v_k, v_l) = \begin{cases} 1 \ , & k = l \ , \\ 0 \ , & k \neq l \ . \end{cases}$$

Der von den w_k erzeugte Vektorraum ist dicht in H, denn: Falls für $f \in H$

$$(f, w_k) = 0 \quad \forall k \quad \Rightarrow \quad (f, v_k) = 0 \quad \forall k \quad \Rightarrow \quad (f, v) = 0 \quad \forall v \in V \ ,$$

d. h. $f \in V^\perp$. Aber: V dicht in H (d. h. $V^\perp = \{0\}$). Also ist $f = 0$. □

Da (w_k) Hilbertbasis von H, habe für jedes $f \in H$

$$f = \sum_{k \geq 1} (f, w_k) w_k \ , \quad (f, f) = |f|^2 = \sum_{k \geq 1} (f, w_k)^2 \ .$$

Da $(v_k) = (w_k / \sqrt{\lambda_k})$ Hilbertbasis von V, habe für jedes $v \in V$

$$v = \sum_{k \geq 1} a(v, v_k) v_k = \sum_{k \geq 1} \frac{1}{\lambda_k} a(v, w_k) w_k \ ,$$
$$a(v, v) = \sum_{k \geq 1} a(v, v_k)^2 = \sum_{k \geq 1} \frac{1}{\lambda_k} a(v, w_k)^2 = \sum_{k \geq 1} \lambda_k (v, w_k)^2$$

Rayleigh-Quotient $R(v) = a(v, v)/(v, v)$. Habe

$$R(w_k) = \frac{a(w_k, w_k)}{(w_k, w_k)} = a(w_k, w_k) = \lambda_k (w_k, w_k) = \lambda_k \quad \forall k \geq 1 \ .$$

Für allgemeines $v = \sum_{k \geq 1} \alpha_k w_k$, wobei $\alpha_k = (v, w_k)$, habe

$$R(v) = \frac{\sum_{k \geq 1} \lambda_k \alpha_k^2}{\sum_{k \geq 1} \alpha_k^2} \overset{\substack{\lambda_1 \text{ kleinster} \\ \text{Eigenwert}}}{\geq} \frac{\sum_{k \geq 1} \lambda_1 \alpha_k^2}{\sum_{k \geq 1} \alpha_k^2} = \lambda_1 \ , \quad \text{damit} \quad \lambda_1 = \min_{0 \neq v \in V} R(v) \ .$$

Bezeichne $V_m := \langle w_1, \ldots, w_m \rangle$ (Aufspann)

$$\begin{aligned} V_m^\perp &= \big\{ v \in V \mid a(v, w) = 0 \quad \forall w \in V_m \big\} \qquad (a\text{-orthogonaler Unterraum}) \\ &= \big\{ v \in V \mid a(v, w_k) = \lambda_k (v, w_k) = 0 \ , \ k = 1, \ldots, m \big\} \\ &= \big\{ v \in V \mid (v, w_k) = 0 \ , \ k = 1, \ldots, m \big\} \ . \end{aligned}$$

Für $v = \sum_{k \geq 1} \alpha_k w_k \in V_{m-1}^\perp$ ist $\alpha_1 = \cdots = \alpha_{m-1} = 0$, also

$$R(v) = \frac{\sum_{k \geq m} \lambda_k \alpha_k^2}{\sum_{k \geq m} \alpha_k^2} \overset{\substack{\lambda_m \text{ kleinster Eigenwert} \\ \downarrow}}{\geq} \lambda_m$$

und somit

$$\lambda_m = \min_{0\neq v\in V_{m-1}^{\perp}} R(v)\ . \tag{VII.3}$$

Wichtigere Charakterisierung durch

Satz 2 (Minimax-Prinzip von Courant-Fischer)**.** *Unter den Voraussetzungen von Satz 1 ist*

$$\lambda_m = \min_{\substack{E_m\ m\text{-dimensionaler}\\ \text{Unterraum von } V}} \max_{0\neq v\in E_m} R(v)\ .$$

(Hier wird die Bestimmung der Eigenräume vermieden.)

Beweis. Betrachte zuerst $R(v)$ für $v \in V_m$:

$$v=\sum_{k=1}^{m}\alpha_k w_k\ :\quad R(v)=\frac{\sum_{k=1}^{m}\lambda_k\alpha_k^2}{\sum_{k=1}^{m}\alpha_k^2}\quad \overset{\lambda_m \text{ größter Eigenwert}}{\overset{\downarrow}{\leq}}\quad \lambda_m\ ,\qquad \text{also}\quad \max_{0\neq v\in V_m} R(v)=\lambda_m\ .$$

Sei nun E_m beliebiger m-dimensionaler Unterraum von V. Zeige:

$$\lambda_m \leq \max_{0\neq v\in E_m} R(v)\ .$$

Habe $E_m \cap V_{m-1}^{\perp} \neq 0$, denn:
Sei $(\varphi_1,\ldots,\varphi_m)$ Basis von E_m, $v=\sum_{i=1}^{m}\nu_i\varphi_i \in E_m$.

$$\begin{aligned} v\in V_{m-1}^{\perp} \quad &\Leftrightarrow \quad (v,w_k)=0\ , && k=1,\ldots,m-1,\\ &\Leftrightarrow \quad \sum_{i=1}^{m}(\varphi_i,w_k)\nu_i=0\ , && k=1,\ldots,m-1. \end{aligned}$$

Dies ist ein homogenes lineares Gleichungssystem mit $m-1$ Gleichungen in m Unbekannten. Es hat daher eine nichttriviale Lösung $\nu=(\nu_i)_{i=1}^{m}\neq 0$.
Sei $0\neq v\in E_m\cap V_{m-1}^{\perp}$, mit (VII.3):

$$\lambda_m=\min_{0\neq w\in V_{m-1}^{\perp}} R(w)\leq R(v)\leq \max_{0\neq u\in E_m} R(u)\ .$$

Damit $\lambda_m \leq \max_{0\neq v\in E_m} R(v)$. Also auch

$$\left.\begin{aligned} \Rightarrow \quad \inf_{\substack{E_m\ m\text{-dimensionaler}\\ \text{Unterraum von } V}} \max_{0\neq v\in E_m} R(v)\geq\lambda_m\ ,\\ \text{andererseits}\quad \max_{0\neq v\in V_m} R(v)=\lambda_m\ . \end{aligned}\right\}\ \Rightarrow\quad \text{Behauptung.}\qquad \square$$

§ 4 Galerkin-Approximation des Eigenwertproblems

(Rayleigh-Ritz-Approximation)

Situation wie in § 3.
(P) Suche $\lambda>0$, $0\neq u\in V$ mit $a(u,v)=\lambda(u,v)\ \forall v\in V$.
Wähle endlichdimensionalen Unterraum $V_h\subset V$ (z. B. Finiten-Elemente-Raum).
(P_h) Suche $\lambda_h>0$, $0\neq u_h\in V_h$ mit $a(u_h,v_h)=\lambda_h(u_h,v_h)\ \forall v_h\in V_h$

Weiß aus § 3 mit V_h statt V und H:
Es existieren Eigenwerte $0 < \lambda_{1,h} \le \lambda_{2,h} \le \cdots \le \lambda_{N,h}$ ($\dim V_h = N$).
V_h hat eine Orthonormalbasis aus Eigenvektoren $(v_{k,h})$ bezüglich $a(\cdot,\cdot)$ und aus Eigenvektoren $(w_{k,h}) = (\sqrt{\lambda_{k,h}} v_{k,h})$ bezüglich $(\cdot,\cdot)$. Aus Minimax-Prinzip erhalte

$$\begin{aligned}\lambda_{m,h} &= \min_{\substack{E_m \ m\text{-dimensionaler} \\ \text{Unterraum von } V_h}} \ \max_{0 \neq v \in E_m} R(v) \\ &\ge \min_{\substack{E_m \ m\text{-dimensionaler} \\ \text{Unterraum von } V}} \ \max_{0 \neq v \in E_m} R(v) = \lambda_m \ .\end{aligned}$$

Also: $\lambda_{m,h} \ge \lambda_m \ \forall m = 1, \ldots, N$.

Berechnung? Sei $(\varphi_1, \ldots, \varphi_N)$ Basis von V_h. Suche $\lambda > 0$, $u_h = \sum_{i=1}^N \mu_i \varphi_i$, sodass

$$a\Big(\sum_{i=1}^N \mu_i \varphi_i, \sum_{j=1}^n \nu_j \varphi_j\Big) = \lambda\Big(\sum_{i=1}^N \mu_i \varphi_i, \sum_{j=1}^N \nu_j \varphi_j\Big) \quad \forall v_h = \sum_{j=1}^N \nu_j \varphi_j \in V_h \ ,$$

d. h. $\nu^T A \mu = \lambda \nu^T M \mu \ \forall \nu \in \mathbb{R}^N$, wobei $\nu = (\nu_j)_{j=1}^N$, $\mu = (\mu_i)_{i=1}^N$ und

$$A = \Big(a(\varphi_i, \varphi_j)\Big)_{i,j=1}^N \qquad \text{Steifigkeitsmatrix,}$$

$$M = \Big((\varphi_i, \varphi_j)\Big)_{i,j=1}^N \qquad \text{Massematrix.}$$

Suche $\lambda > 0$ und $0 \neq \mu \in \mathbb{R}^N$ mit

$$A\mu = \lambda \cdot M\mu \ , \qquad \text{verallgemeinertes Eigenwertproblem im } \mathbb{R}^N.$$

Sei $M = LL^T$ Cholesky-Zerlegung. Schreibe äquivalent

$$\underbrace{L^{-1} A (L^T)^{-1}}_{\tilde{A}} \underbrace{L^T \mu}_{\xi} = \lambda \cdot \underbrace{L^T \mu}_{\xi} \ .$$

Dies ist ein gewöhnliches Eigenwertproblem im $\mathbb{R}^N$: Suche $\lambda > 0$, $0 \neq \xi \in \mathbb{R}^N$ mit $\tilde{A}\xi = \lambda \xi$. (QR-Algorithmus: iteratives Verfahren mit Aufwand $\sim \mathcal{O}(N^3)$.)
Weiß: $\lambda_m \le \lambda_{m,h}$.
Möchte: $\lambda_{m,h} \le \lambda_m + \varepsilon(h)$, wobei $\varepsilon(h) \to 0$ für $h \to 0$.

§ 5 Konvergenz der Eigenwertapproximation

Situation wie zuvor: $V_h \subset V \underset{\text{kompakt}}{\subset} H$, $\dim V_h \to \infty$ für $h \to 0$ (z. B. $V = H_0^1$, $H = L^2$).

$$\|v\| = \|v\|_a = \sqrt{a(v,v)} \qquad \text{Energienorm auf } V,$$

$$|v| = \sqrt{(v,v)} \qquad \text{Norm auf } H.$$

Approximationsannahme:

$$\forall u \in V \quad \lim_{h\to 0} \inf_{v_h \in V_h} \|v_h - u\| = 0 \ . \tag{VII.4}$$

(Elemente in V sind beliebig genau durch Elemente in V_h approximierbar, wenn $h \to 0$.)
Bei finiten Elementen: $\|\cdot\| \sim \|\cdot\|_1$ H^1-Norm, $V = H_0^1(\Omega)$ (oder $H^1(\Omega)$).
Gezeigt (in Kapitel IV): Für $u \in H^2(\Omega) \cap H_0^1(\Omega)$:

$$\inf_{v_h \in V_h} \|v_h - u\|_1 \le Ch|u|_2 \ .$$

$H^2(\Omega) \cap H_0^1(\Omega)$ dicht in $H_0^1(\Omega)$: Es existiert eine Folge (u_n) in $H^2(\Omega)$: $\|u_n - u\|_1 \le \frac{1}{n}$.

Interpolationsfehler: $\quad \|\Pi_h u_n - u_n\|_1 \le Ch|u_n|_2 \le \dfrac{1}{n}$ für $h \le h_n$ (genügend klein),

$$\forall n \ \exists h_n > 0 \quad \forall h \le h_n : \quad \|\underbrace{\Pi_h u_n}_{\in V_h} + (u_n - u_n) - u\| \le \frac{2}{n} \to 0 \ .$$

Satz 1. *Voraussetzungen wie in § 3, § 4. Die Approximationsannahme* (VII.4) *sei erfüllt. Dann:*

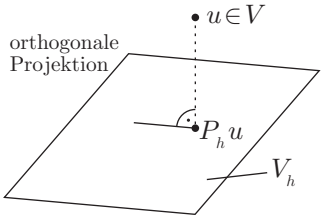
$$\forall m \ge 1 \quad \exists h_m > 0 \quad \forall h \le h_m : \qquad 0 \le \lambda_{m,h} - \lambda_m \le 4\frac{\lambda_m}{\lambda_1}\varepsilon_{m,h}^2 \ ,$$

wobei $\varepsilon_{m,h}^2 = \sum_{i=1}^{m} \inf_{v_h \in V_h} \|w_i - v_h\|^2 \to 0$ *für* $h \to 0$ *nach* (VII.4)*,* w_i *Eigenfunktionen mit* $|w_i| = 1$.

Beweis. In zwei Schritten.

Erinnerung. Ritz-Projektion $P_h : V \to V_h$, orthogonale Projektion bezüglich des Skalarprodukts $a(\cdot,\cdot)$, d. h. $P_h u \in V_h$ definiert durch $a(P_h u, v_h) = a(u, v_h)$ $\forall v_h \in V_h$.
Bemerkung. Für $a(u,v) = l(v)$ $\forall v \in V$, also u Lösung eines elliptischen Randwertproblems, ist

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h$$



Finite-Elemente-Approximation, hier $P_h u = u_h$.

Habe $\|u - P_h u\| = \inf_{v_h \in V_h} \|u - v_h\|$ (siehe Céa).

Hilfssatz 2. $\lambda_{m,h} \le \dfrac{\lambda_m}{\sigma_{m,h}^2}$, *wobei* $\sigma_{m,h} = \inf\limits_{\substack{v \in V_m \\ |v|=1}} |P_h v|$, *falls* $\sigma_{m,h} > 0$, $V_m = \langle w_1, \ldots, w_m \rangle$.

Beweis (von Hilfssatz 2). Minimax-Prinzip:

$$\lambda_{m,h} = \min_{\substack{E_m \ m\text{-dimensionaler} \\ \text{Unterraum von } V_h}} \max_{0 \ne v \in E_m} R(v) \ .$$

Wähle $E_m = P_h V_m \subset V_h$:

$$\lambda_{m,h} \le \max_{v \in E_m} R(v) \ .$$

Zeige zunächst: $\dim E_m = m$.
Indirekt: Nehme an, dass $\dim E_m < m$ (weiß $\dim V_m = m$). Dann ist $P_h V_m \to E_m$ keine Bijektion, d. h. es existiert $0 \neq v \in V_m$ mit $P_h v = 0$. Da aber $\sigma_{m,h} = 0$ ausgeschlossen, ist dies ein Widerspruch. Es muss also $\dim E_m = m$ gelten.

$$\lambda_{m,h} \leq \max_{0\neq v_h\in E_m} \frac{a(v_h,v_h)}{(v_h,v_h)} \overset{\substack{v_h = P_h v \\ \text{mit } v\in V_m}}{\downarrow =} \max_{0\neq v\in V_m} \frac{a(P_h v, P_h v)}{(P_h v, P_h v)} \overset{\substack{P_h v \text{ } a\text{-orthogonale} \\ \text{Projektion}}}{\downarrow \leq} \max_{0\neq v\in V_m} \frac{a(v,v)}{(P_h v, P_h v)}$$
$$\leq \underbrace{\max_{0\neq v\in V_m} \frac{a(v,v)}{(v,v)}}_{\lambda_m} \cdot \max_{0\neq v\in V_m} \frac{(v,v)}{(P_h v, P_h v)} = \lambda_m \cdot \underbrace{\frac{1}{\inf_{0\neq v\in V_m} \frac{|P_h v|^2}{|v|^2}}}_{1/\sigma_{m,h}^2} = \frac{\lambda_m}{\sigma_{m,h}^2} \ . \qquad \square$$

Hilfssatz 3. $\sigma_{m,h}^2 \geq 1 - \frac{2}{\lambda_1} \sum_{i=1}^m \|w_i - P_h w_i\|^2 = 1 - \frac{2}{\lambda_1} \varepsilon_{m,h}^2$.

Beweis (von Hilfssatz 3). Sei $v \in V_m$, $|v| = 1$. Habe $v = \sum_{i=1}^m \alpha_i w_i$, $1 = |v|^2 = \sum_{i=1}^m \alpha_i^2$:

$$1 - |P_h v|^2 = |v|^2 - |P_h v|^2 = (v - P_h v, v + P_h v) = -\underbrace{|v - P_h v|^2}_{\geq 0} + 2(v - P_h v, v)$$

und damit

$$|P_h v|^2 \geq 1 - 2(v - P_h v, v) \ . \tag{VII.5}$$

Habe

$$(v - P_h v, v) = \sum_{i=1}^m \alpha_i (v - P_h v, w_i) \overset{w_i \text{ Eigenfunktionen}}{\downarrow =} \sum_{i=1}^m \frac{\alpha_i}{\lambda_i} a(v - P_h v, w_i)$$
$$\overset{\substack{a(v-P_h v, v_h) = 0 \ \forall v_h \in V_h, \\ \text{wähle } v_h = -P_h w_i}}{\downarrow =} \sum_{i=1}^m \frac{\alpha_i}{\lambda_i} a(v - P_h v, w_i - P_h w_i) \ ,$$

damit:

$$(v - P_h v, v) \overset{\text{Cauchy-Schwarz}}{\downarrow \leq} \frac{1}{\lambda_1} \sum_{i=1}^m |\alpha_i| \cdot \|v - P_h v\| \cdot \|w_i - P_h w_i\|$$
$$\overset{\substack{\text{Cauchy-Schwarz} \\ \text{auf } \mathbb{R}^m}}{\downarrow \leq} \frac{1}{\lambda_1} \underbrace{\|v - P_h v\|}_{\text{(Berechnung folgt)}} \underbrace{\sqrt{\sum_{i=1}^m \alpha_i^2}}_{1} \cdot \sqrt{\sum_{i=1}^m \|w_i - P_h w_i\|^2} \ ,$$

$$\|v - P_h v\| = \left\| \sum_{i=1}^m \alpha_i (w_i - P_h w_i) \right\| \leq \underbrace{\sqrt{\sum_{i=1}^m \alpha_i^2}}_{1} \sqrt{\sum_{i=1}^m \|w_i - P_h w_i\|^2} \ .$$

Damit:

$$(v - P_h v, v) \leq \frac{1}{\lambda_1} \sum_{i=1}^m \|w_i - P_h w_i\|^2 = \frac{\varepsilon_{m,h}^2}{\lambda_1} \ .$$

Durch Einsetzen in (VII.5) folgt die Behauptung. □

Beweis von Satz 1.

$$\lambda_{m,h} \overset{\text{Hilfssatz 1}}{\underset{\downarrow}{\leq}} \frac{\lambda_m}{\sigma_{m,h}^2} \overset{\text{Hilfsatz 2}}{\underset{\downarrow}{\leq}} \frac{\lambda_m}{1-2\frac{\varepsilon_{m,h}^2}{\lambda_1}} \overset{\text{geometrische Reihe}}{\underset{\downarrow}{\leq}} \lambda_m\left(1+\frac{4}{\lambda_1}\varepsilon_{m,h}^2\right) .$$

Dabei galt für die geometrische Reihe

$$\frac{1}{1-x} = 1 + x + x^2 + \cdots \leq 1 + 2x \quad \text{für} \quad x \leq \frac{1}{2} .$$

Für h genügend klein ist $2\dfrac{\varepsilon_{m,h}^2}{\lambda_1} \leq \dfrac{1}{2}$ (wegen $\varepsilon_{m,h} \to 0$ für $h \to 0$ nach (VII.4)). □

Anwendung des Satzes auf finite Elemente: $V = H_0^1(\Omega)$, a V-elliptische Bilinearform. V_h Finiter-Elemente-Raum mit

$$\inf_{v_h \in V_h} \|u - v_h\|_{1,\Omega} \leq C_1 h^k |u|_{k+1,\Omega} \quad \forall u \in H^{k+1}(\Omega) . \tag{VII.6}$$

Siehe Kapitel IV, § 7, Satz 3: Gilt, falls $h/\rho \leq \text{const}$, $P_k \subset P$.

Satz 4. *Falls für die ersten m Eigenfunktionen $w_1, \ldots, w_m \in H^{k+1}(\Omega)$ gilt und* (VII.6) *erfüllt ist, so ist*

$$0 \leq \lambda_{m,h} - \lambda_m \leq C h^{2k} \lambda_m \sum_{i=1}^{m} |w_i|_{k+1,\Omega}^2 . \qquad \text{(Doppelte Konvergenzordnung)}$$

Beweis. $\varepsilon_{m,h}^2 = \displaystyle\sum_{i=1}^{m} \|w_i - P_h w_i\|_a^2$,

$$\|w - P_h w\|_a^2 \overset{\substack{P_h\ a\text{-orthogonale} \\ \text{Projektion (Céa)}}}{\underset{\downarrow}{=}} \inf_{v_h \in V_h} \|w - v_h\|_a^2$$

$$\overset{\substack{\text{Normäquivalenz} \\ \text{zwischen 1- und Energie-Norm}}}{\underset{\downarrow}{\leq}} M \cdot \inf_{v_h \in V_h} \|w - v_h\|_{1,\Omega}^2 \overset{\text{(VII.6)}}{\underset{\downarrow}{\leq}} M \cdot C_1^2 \cdot h^{2k} |w|_{k+1,\Omega}^2 .$$

Einsetzen in Satz 1 und es folgt die Behauptung. □

§ 6 Konvergenz der Eigenvektoren

Situation wie gehabt, λ_m sei (algebraisch) einfacher Eigenwert: $\lambda_i \neq \lambda_m$ $\forall i \neq m$. Eigenfunktionen w_m, $|w_m| = 1$. Eigenfunktion $w_{m,h}$ zu $\lambda_{m,h}$, $|w_{m,h}| = 1$. Setze

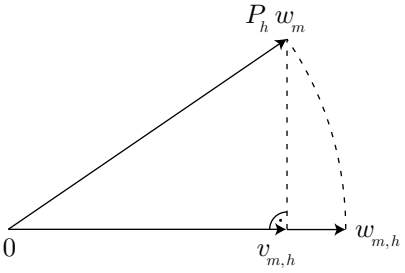
$$\rho_{m,h} := \max_{i \neq m} \frac{\lambda_m}{|\lambda_{i,h} - \lambda_m|} \quad \xrightarrow{h \to 0} \quad \rho_m = \max_{i \neq m} \frac{\lambda_m}{|\lambda_i - \lambda_m|} .$$

Satz 1. *Voraussetzungen wie in § 5. Sei λ_m einfacher Eigenwert. Dann gibt es für genügend kleine $h > 0$ Eigenfunktionen $v_{m,h}$ mit*

$$|v_{m,h} - w_m| \le (1 + \rho_{m,h})|w_m - P_h w_m| \ .$$

Beweis.

Sei $w_{m,h}$ Eigenfunktion zu $\lambda_{m,h}$ mit $|w_{m,h}| = 1$. Betrachte Orthogonalprojektion bezüglich $(\cdot,\cdot)$ von $P_h w_m$ auf $\mathbb{R} \cdot w_{m,h}$: $v_{m,h} = (P_h w_m, w_{m,h}) w_{m,h}$. Habe



$$P_h w_m - v_{m,h} = \sum_{\substack{i=1 \\ i \ne m}}^{N} (P_h w_m, w_{i,h}) w_{i,h} \ ,$$

da $(w_{i,h})$ Orthonormalbasis von V_h bezüglich $(\cdot,\cdot)$.

$$|P_h w_m - v_{m,h}|^2 = \sum_{i \ne m} (P_h w_m, w_{i,h})^2 \ , \tag{VII.7}$$

$$(\underbrace{P_h w_m}_{\in V_h}, w_{i,h}) \overset{\substack{w_{i,h} \text{ Eigenfunktion} \\ \text{in } V_h \\ \downarrow}}{=} \frac{1}{\lambda_{i,h}} a(P_h w_m, w_{i,h}) \overset{\substack{P_h \ a\text{-orthogonale} \\ \text{Projektion} \\ \downarrow}}{=} \frac{1}{\lambda_{i,h}} a(w_m, \underbrace{w_{i,h}}_{\in V}) \overset{\substack{w_m \text{ Eigenfunktion} \\ \text{in } V \\ \downarrow}}{=} \frac{\lambda_m}{\lambda_{i,h}} (w_m, w_{i,h}) \ .$$

Daraus:

$$\begin{aligned} (\lambda_{i,h} - \lambda_m)(P_h w_m, w_{i,h}) &= \lambda_m (w_m - P_h w_m, w_{i,h}) \\ \Rightarrow \quad (P_h w_m, w_{i,h}) &\le \rho_{m,h} |(w_m - P_h w_m, w_{i,h})| \ . \end{aligned}$$

In (VII.7) einsetzen:

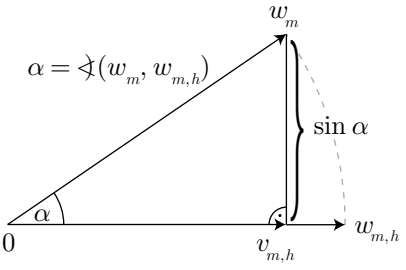
$$\begin{aligned} |P_h w_m - v_{m,h}|^2 &\le \rho_{m,h}^2 \sum_{i \ne m} (w_m - P_h w_m, w_{i,h})^2 \le \rho_{m,h}^2 \sum_{i=1}^{N} (w_m - P_h w_m, w_{i,h})^2 \\ &= \rho_{m,h}^2 |w_m - P_h w_m|^2 \\ \Leftrightarrow \quad |P_h w_m - v_{m,h}| &\le \rho_{m,h} |w_m - P_h w_m| \ . \end{aligned}$$

Dreiecksungleichung:

$$|w_m - v_{m,h}| \le |w_m - P_h w_m| + |P_h w_m - v_{m,h}| \le (1 + \rho_{m,h}) |w_m - P_h w_m| \ . \qquad \square$$

In Satz 1 gezeigt:

$$\sin\alpha = |w_m - v_{m,h}| \le (1 + \rho_{m,h}) |w_m - P_h w_m| \ .$$



Klar: Falls $\sin\alpha \le \frac{1}{2}$, ist

$$|w_{m,h} - w_m| \le 2 |v_{m,h} - w_m| \ .$$

Hilfssatz 2. *Situation wie zuvor. Dann gilt*

$$\|w_{m,h} - w_m\|^2 = \lambda_m |w_{m,h} - w_m|^2 + (\lambda_{m,h} - \lambda_m) \ .$$

Beweis.

$$\|w_{m,h} - w_m\|^2 = \underbrace{\|w_{m,h}\|^2}_{\substack{a(w_{m,h},w_{m,h}) \\ \lambda_{m,h}(w_{m,h},w_{m,h}) \\ |w_{m,h}|^2 \\ 1}} + \underbrace{\|w_m\|^2}_{\substack{a(w_m,w_m) \\ \lambda_m (w_m,w_m) \\ |w_m|^2 \\ 1}} - 2a(w_{m,h}, w_m)$$

$$= \lambda_{m,h} + \lambda_m - 2\lambda_m (w_{m,h}, w_m) \ .$$

Andererseits:

$$|w_m - w_{m,h}|^2 = \underbrace{|w_{m,h}|^2}_{1} + \underbrace{|w_m|^2}_{1} - 2(w_{m,h}, w_m) = 2 - 2(w_{m,h}, w_m)$$

Multipliziere die Gleichung mit λ_m. Dann folgt die Behauptung. □

Anwendung auf finite Elemente $(V = H_0^1(\Omega),\ H = L^2(\Omega).)$
V_h sei Finiter-Elemente-Raum mit

$$\inf_{v_h \in V_h} \left\{ \|v_h - v\|_{0,\Omega} + h|v_h - v|_{1,\Omega} \right\} \le C \cdot h^{k+1} |v|_{k+1,\Omega} \quad \forall v \in H^{k+1}(\Omega) \ . \tag{VII.8}$$

(Kapitel IV: Gezeigt, falls Finiter-Elemente-Raum Polynome vom Grad k enthält).

Satz 3. *λ_m sei einfacher Eigenwert. Falls die ersten m Eigenfunktionen $w_1, \ldots, w_m \in H^{k+1}(\Omega)$ und falls* (VII.8) *gilt, ist*

$$\|w_{m,h} - w_m\|_{1,\Omega} \le C_m h^k \ .$$

Falls zusätzlich das elliptische Randwertproblem $a(u,v) = (f,v)$, $f \in L^2$, H^2-regulär ist, so gilt

$$\|w_{m,h} - w_m\|_{0,\Omega} \le \tilde{C}_m h^{k+1} \ .$$

(C_m, $\tilde{C}_m$ unabhängig von h.)

Beweis (nach Hilfssatz 2).

$$\|w_{m,h} - w_m\|_1^2 \le C_m \|w_{m,h} - w_m\|_0^2 + (\lambda_{m,h} - \lambda_m) \ ,$$

nach Satz 1 und Bemerkung danach:

$$\begin{aligned} \|w_{m,h} - w_m\|_0 &\le C'_m \|w_m - P_h w_m\|_0 \\ &\le C'_m \|w_m - P_h w_m\|_1 \le C''_m \inf_{v_h \in V_h} \|w_m - v_h\|_1 \overset{\text{(VII.8)}}{\downarrow} \le C'''_m h^k \ . \end{aligned}$$

Nach Satz 2, § 5: $\lambda_{m,h} - \lambda_m \le C'''_m h^{2k}$. Dann gilt $\|w_{m,h} - w_m\|_1 \le C^{IV}_m h^k$.
Nach Satz 1:

$$\|w_{m,h} - w_m\|_0 \overset{\text{Satz 1}}{\downarrow} \le C'_m \|w_m - P_h w_m\|_0 \overset{\substack{\text{Nitsche-Trick} \\ H^2\text{-Regularität}}}{\downarrow} \le C^V_m h \|w_m - P_h w_m\|_1 \le \tilde{C}_m h^{k+1} \ .$$ □

Literaturverzeichnis

[StoeBul05] Josef Stoer und Roland Bulirsch, *Numerische Mathematik 2*, 5. Auflage, Springer, 2005.

[AschMat88] Uri M. Ascher, Robert M. M. Mattheij und Robert D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall, 1988.

[Braess92] Dietrich Braess, *Finite Elemente*, 1. Auflage, Springer, 1992.

[Hackbu93] Wolfgang Hackbusch, *Iterative Lösung großer schwachbesetzter Gleichungssysteme*, 1. Auflage, Teubner, 1993.

[BrezFor91] Franco Brezzi und Michel Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, 1991.

[GirRav86] Vivette Girault und Pierre-Arnaud Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer, 1986.

[HankBou06] Martin Hanke-Bourgeois, *Grundlagen der Numerischen Mathematik und des Wissenschaftlichen Rechnens*, 2. Auflage, Teubner, 2006.

[GroRoos94] Christian Großmann, Hans-Görg Roos, *Numerik partieller Differentialgleichungen*, 2. Auflage, Teubner, 1994.