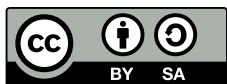


# Numerik partieller Differentialgleichungen

Oliver Sander

29. August 2020

Getippt und gesetzt von Johannes R. Stojanow



Oliver Sander, 2017

Copyright 2017 by Oliver Sander. This work is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-sa/4.0/>.

# Inhaltsverzeichnis

<b>1</b>	<b>Partielle Differentialgleichungen</b>	<b>1</b>
1.1	Beispiel: Die eindimensionale Wellengleichung . . . . .	1
1.2	Beispiel: Erhaltungsgleichungen . . . . .	3
<b>2</b>	<b>Differenzenverfahren für das Poisson-Problem</b>	<b>7</b>
2.1	Algebraische Darstellung . . . . .	8
2.2	Fourier-Analyse . . . . .	9
2.3	Diskretisierungsfehleranalyse . . . . .	11
<b>3</b>	<b>Finite Elemente</b>	<b>15</b>
3.1	Die Variationsformulierung . . . . .	16
3.2	Sobolev-Räume . . . . .	17
3.2.1	Schwache Ableitung . . . . .	18
3.2.2	Der Spursatz . . . . .	19
3.3	Wohlgestelltheit . . . . .	20
3.4	Finite Elemente . . . . .	24
3.5	Diskretisierungsfehler . . . . .	26
3.6	Quasi-Uniformität . . . . .	28
3.7	Approximationsfehlerabschätzungen in der $H^1$ -Norm . . . . .	28
3.8	Diskretisierungsfehler in der $L^2$ -Norm . . . . .	30
3.9	Crouzeix-Raviart-Element (Poisson-Gl.) . . . . .	31
3.10	Die Stokes-Gleichung . . . . .	34
3.10.1	Schwache Formulierung . . . . .	35
3.10.2	Exkurs: Sattelpunktprobleme . . . . .	35
3.10.3	Gemischte FE-Methode . . . . .	38
3.10.4	Diskretisierungsfehler . . . . .	39
<b>4</b>	<b>Konstruktion von adaptiven Gittern</b>	<b>43</b>
4.1	Residuenfehlerschätzer . . . . .	43
4.1.1	Effizienz des Residuenschätzers . . . . .	43
<b>5</b>	<b>Finite-Volumen-Verfahren</b>	<b>47</b>

# 1 Partielle Differentialgleichungen

Differentialgleichung:

- Gleichung für unbekannte Funktion  $u$ .
- Die Gleichungen nutzt Ableitungen von  $u$ .

Bereits bekannt sind gewöhnliche Differentialgleichungen:

- Es kommen nur Ableitungen bezüglich einer Variablen vor.
- Kanonische Form:  $u'(t) = f(t, u(t))$ .

Partielle Differentialgleichung

- Es kommen Ableitungen bzgl. mehrerer Variablen vor.
- Beschreiben häufig Prozesse, die vom Ort und von der Zeit abhängen., z.B.:
  - Wärmeausbreitung, Stofftransport,
  - Spannungsverteilung in deformierbaren Körpern,
  - Raumkrümmung in der Kosmologie,
  - Dichtfelder in der Quantenmechanik.

Erstaunlich ist, dass sich Theorie und Numerik partieller Differentialgleichungen stark von denjenigen gewöhnlicher Differentialgleichungen unterscheiden.

## 1.1 Beispiel: Die eindimensionale Wellengleichung

Wir betrachten ein elastisches eingespanntes Band und wollen das Schwingungsverhalten beschreiben. Wir nehmen vereinfachend an:

- Keine Gravitation,
- nur Longitudinalschwingungen.

Wir suchen eine Funktion  $u: [0, \infty) \times [0, \ell] \rightarrow \mathbb{R}$ .

- $u(t, x)$  ist die longitudinale Auslenkung zur Zeit  $t$  am Ort  $x$ .

Vermutung: die gesuchte Funktion  $u$  löst eine partielle Differentialgleichung. Wir approximieren das Band durch ein Masse-Feder-System.

## 1 Partielle Differentialgleichungen

- Teilchen  $i$  habe Masse  $m_i$  und Auslenkung  $u_i$ .
- Auf Teilchen  $i$  wirken die folgenden Kräfte:
  - Zugkraft der linken Feder:  $\kappa \cdot (u_i - u_{i-1})$  mit  $\kappa$  Federkonstante (Gesetz von HOOKE),
  - Zugkraft der rechten Feder:  $-\kappa \cdot (u_{i+1} - u_i)$ ,
  - Massenträgheit:  $0.5 \cdot m\ddot{u}$ .

Physikerin des Vertrauens: Für jedes innere Teilchen  $i$  gilt das Kräftegleichgewicht:

$$\frac{1}{2}m\ddot{u}_i + \kappa(u_i - u_{i-1}) + \kappa(u_i - u_{i+1}) = 0. \quad (*)$$

Die Federkonstante  $\kappa$  hängt von der Länge  $h$  der Feder ab:

$$\kappa = \frac{EA}{h}$$

mit  $A$  Querschnittsfläche,  $E$  Elastizitätsmodul,  $h$  Länge in der Ruhelage. Weiterhin stellen wir uns die Masse  $m_i$  gleichmäßig über das  $i$ -te Teilchen und die links danebenliegende Feder vor. Wir nehmen an, die Masse  $m_i$  werde skaliert mit der Federlänge

$$m_i = \rho h$$

mit  $\rho$  Massendichte (Masse pro Länge). Die Gleichung  $(*)$  des Kräftegleichgewichts wird zu

$$\begin{aligned} 0 &= \frac{1}{2}\rho h\ddot{u}_i + \frac{EA}{h}(-u_{i+1} + 2u_i - u_{i-1}) \\ &= \frac{1}{2}\rho\ddot{u}_i + \frac{EA}{h^2}(-u_{i+1} + 2u_i - u_{i-1}). \end{aligned}$$

Für den Grenzübergang  $h \rightarrow 0$  benötigen wir das folgende Lemma.

**Lemma 1.1.** Sei  $f: \mathbb{R} \rightarrow \mathbb{R}$  hinreichend glatt. Dann gilt

$$\frac{\partial^2 f}{\partial x^2}(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} - \frac{h^2}{12}f^{(4)}(\xi).$$

Mit diesem Lemma wird die Gleichung des Kräftegleichgewichts zur partiellen Differentialgleichung

$$\frac{1}{2}\rho \frac{\partial^2 u}{\partial t^2} - EA \frac{\partial^2 u}{\partial x^2} = 0.$$

Diese Gleichung nennt man „eindimensionale linearen Wellengleichung“. Lösungen sind z.B.:

$$\begin{aligned} u(t, x) &= \sin \sqrt{EA}t \cdot \cos \sqrt{\rho}x, \\ u(t, x) &= \sin(\sqrt{\rho}x + \sqrt{EA}t). \end{aligned}$$

Sie enthält zweite partielle Ableitungen nach  $t$  und  $x$ . In höheren Dimensionen, z.B.  $\mathbb{R}^3$ , führen ähnliche Überlegungen auf die dreidimensionale Wellengleichung

$$\frac{1}{2}\rho \frac{\partial^2 u}{\partial t^2} - EA \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) = 0.$$

Sie beschreibt Druckwellen in zwei- und dreidimensionalen Objekten. Zur Abkürzung definiert man den Laplace-Operator

$$\Delta u := \sum_{i=1}^d \frac{\partial^2 u}{\partial x_i^2}.$$

Mit ihm wird die Wellengleichung zu

$$\frac{\partial^2 u}{\partial t^2} = \Delta u.$$

## 1.2 Beispiel: Erhaltungsgleichungen

Wir modellieren die Bewegung eines Gases in einem langen, dünnen Rohr.

- Das Rohr ist so dünn, dass alle physikalischen Größen als konstant auf jeder Querschnittsfläche angenommen werden können.
- Wir betrachten die Massendichte  $\rho(x, t)$  des Gases im Ort  $x \in [0, \ell]$  zur Zeit  $t \in [0, \infty)$  und
- die Gesamtmasse im Rohrabschnitt zwischen zwei Punkten  $x_1, x_2$  mit  $x_1 \leq x_2$ :

$$M(x_1, x_2, t) := \int_{x_1}^{x_2} \rho(x, t) dx.$$

Weitere Annahmen:

- Das Rohr ist gasundurchlässig.
- Gas wird weder erzeugt noch vernichtet.

Die Gasmasse kann sich nur ändern, wenn Gas an den Rohrenden ein- oder austritt.

- Sei  $v(x, t)$  die Geschwindigkeit des Gases am Ort  $x$  zur Zeit  $t$ .

**Definition 1.2.** Der Gasfluss im Punkt  $x$  zur Zeit  $t$  ist

$$f(x, t) := \rho(x, t) \cdot v(x, t).$$

Massenerhaltung: Betrachte zwei Punkte  $x_1 \leq x_2$  und zwei Zeitpunkte  $t_1 \leq t_2$ .

## 1 Partielle Differentialgleichungen

- Massenänderung in  $[x_1, x_2]$  von  $t_1$  nach  $t_2$ :

$$\int_{x_1}^{x_2} \rho(x, t_2) dx - \int_{x_1}^{x_2} \rho(x, t_1) dx.$$

- Fluss durch den Rand von  $[x_1, x_2]$  in dieser Zeit:

$$\int_{t_1}^{t_2} f(x_1, t) dt - \int_{t_1}^{t_2} f(x_2, t) dt.$$

Massenerhaltungsgesetz in einer Integralform:

$$\begin{aligned} \int_{x_1}^{x_2} [\rho(x, t_2) - \rho(x, t_1)] dx &= \int_{t_1}^{t_2} [f(x_1, t) - f(x_2, t)] dt \\ &= \int_{t_1}^{t_2} [\rho(x_1, t)v(x_1, t) - \rho(x_2, t)v(x_2, t)] dt. \end{aligned}$$

Diese Massenerhaltungsgleichung wurde hier für Gase vorgestellt, die Argumentation gilt aber auch in anderen Situationen:

- Teilchen in einer Flüssigkeit,
- Bakterien in einer Lösung,
- Verkehrsflüsse, etc.

Jetzt wollen wir daraus eine Differentialgleichung ableiten. Zusatzannahme:  $\rho$  und  $v$  sind stetig differenzierbar. Der Hauptsatz der Analysis liefert

$$\begin{aligned} \rho(x, t_2) - \rho(x, t_1) &= \int_{t_1}^{t_2} \frac{\partial}{\partial t} \rho(x, t) dt \\ \rho(x_2, t)v(x_2, t) - \rho(x_1, t)v(x_1, t) &= \int_{x_1}^{x_2} \frac{\partial}{\partial t} (\rho(x, t)v(x, t)) dx \end{aligned}$$

Wir setzen ein:

$$\int_{t_1}^{t_2} \int_{x_1}^{x_2} \left[ \frac{\partial}{\partial t} \rho(x, t) + \frac{\partial}{\partial x} (\rho(x, t) \cdot v(x, t)) \right] dx dt = 0$$

für alle  $x_1 \leq x_2$  und  $t_1 \leq t_2$ . Daraus folgt, dass der Integrand punktweise = 0 ist, also

$$\frac{\partial}{\partial t} \rho(x, t) + \frac{\partial}{\partial x} (\rho \cdot v)(x, t) = 0 \quad \forall x \in [0, \ell], t \in [0, \infty).$$

Dies ist eine Partielle Differentialgleichung (Erhaltungsgleichung in Differentialform)! Wir haben nun eine Gleichung für zwei Unbekannte. Wir führen daher zusätzlich ein konstitutives Gesetz ein, welches eine Abhängigkeit von  $v$  als Funktion von  $\rho$  ausdrückt. Z.B.:

- $v$  ist unabhängig von  $\rho$ ,  $x$ ,  $t$ . Also

$$\frac{\partial \rho}{\partial t} + v \frac{\partial \rho}{\partial x} = 0 \quad \forall x, t.$$

Diese Gleichung heißt lineare Transportgleichung/Advektionsgleichung. Wenn man Startwerte  $\rho(x, 0) = \rho_0(x)$  für alle  $x \in \mathbb{R}$  vorgibt, dann erhält man als Lösung  $\rho(x, t) = \rho_0(x - vt)$ . Die Anfangsdichte  $\rho_0$  wird mit Geschwindigkeit  $v$  weitertransportiert.

- $v(x, t) = \frac{1}{2}\rho(x, t)$  führt auf die Burgers-Gleichung

$$\frac{\partial \rho}{\partial t} + \frac{1}{2} \frac{\partial}{\partial x} \rho^2 = 0.$$

Sie ist eine einfache nichtlineare Transportgleichung und interessant, denn Lösungen dieser Gleichung können Schocks entwickeln. Das bedeutet, dass  $\rho$  für endliche  $t$  unstetig wird. Die Gleichung wird dann undefiniert.





## 2 Differenzenverfahren für das Poisson-Problem

Als erstes betrachten wir *Differenzenverfahren*, auch: *Verfahren der finiten Differenzen*.

- Diese haben signifikante Nachteile.
- Sie sind aber sehr einfach.
- Deshalb werden sie dennoch verwendet.

Wir betrachten das Poisson-Problem: Sei  $\Omega$  ein Gebiet (offen, beschränkt, zusammenhängend) in  $\mathbb{R}^d$ . Finde  $u \in C^2(\Omega) \cap C(\bar{\Omega})$ , so dass

$$-\Delta u(x) := -\left(\frac{\partial^2 u}{\partial x_1^2} + \dots + \frac{\partial^2 u}{\partial x_d^2}\right) = f(x) \quad \forall x \in \Omega$$

mit den Randbedingungen

$$u(x) = g(x) \quad \forall x \in \partial\Omega.$$

Es existiert eine eindeutige Lösung, wenn  $f, g, \partial\Omega$  hinreichend glatt sind. Im Folgenden sei  $\Omega = (0, 1)^2$ . Diskretisierung: Wir ersetzen die Differentialgleichung durch eine endlich-dimensionale Approximation.

- Diesen Prozess nennt man *Diskretisierung*.
- Wir brauchen dafür zunächst ein Gitter. Bei FD-Methoden ist dies eine regelmäßige Anordnung von Punkten in  $\Omega$ .
- Definiere (mit Schrittweite  $h = \frac{1}{n-1}$ ,  $n \in \mathbb{N}$ )

$$\Omega_h := \{(ih, jh) : 1 \leq i, j \leq n-1\},$$

$$\bar{\Omega}_h := \{(ih, jh) : 0 \leq i, j \leq n\}.$$

Ab jetzt interessieren wir uns nur noch für Werte von  $u$  an den Gitterpunkten.

- Wie schon bei der Herleitung der Wellengleichung zeigt eine Taylor-Entwicklung:

$$v''(x) = \frac{v(x-h) - 2v(x) + v(x+h)}{h^2} + \mathcal{O}(h^2).$$

## 2 Differenzenverfahren für das Poisson-Problem

- Ersetze jetzt  $\Delta u$  durch die Differenzenformel:

$$\Delta_h u(x, y) := \frac{1}{h^2} [u(x-h, y) + u(x+h, y) + u(x, y+h) + u(x, y-h) - 4u(x, y)]$$

für alle  $x, y \in \Omega_h$ .

Solche Differenzenformeln schreibt man häufig auch als sogenannte *Differenzensterne*. Unser Verfahren besitzt folgende Darstellung:

$$[-\Delta]_{(x,y)} = \frac{1}{h^2} \begin{bmatrix} & -1 & & \\ -1 & 4 & -1 & \\ & -1 & & \end{bmatrix}_{(x,y)}$$

Sei  $\ell^2(\Omega_h)$  bzw.  $\ell^2(\bar{\Omega})$  die Menge aller Gitterfunktionen auf  $\Omega_h$  bzw.  $\bar{\Omega}_h$ . D.h.  $v_h \in \ell(\Omega_h)$  heißt  $v_h: \Omega_h \rightarrow \mathbb{R}$ . Man erhält das *diskrete Poisson-Problem*: Finde  $u_h \in \ell^2(\bar{\Omega})$ , so dass

$$\begin{aligned} -(\Delta_h u_h)(x, y) &= f(x, y) \quad \forall (x, y) \in \Omega_h \\ u_h(x, y) &= g(x, y) \quad \forall (x, y) \in \bar{\Omega}_h \setminus \Omega_h. \end{aligned}$$

Dieses Problem führt auf ein lineares Gleichungssystem. Jedem Punkt aus  $\bar{\Omega}_h$  wird eine Gleichung zugeordnet. Wir haben zwei Möglichkeiten zum Fortfahren:

1. Unbekannte sind die Werte von  $u_h$  auf  $\bar{\Omega}_h$ .
2. Eliminiere (durch Einsetzen) die Randbedingung  $\rightarrow$  Unbekannte sind die Werte von  $u_h$  auf  $\Omega_h$ .

### 2.1 Algebraische Darstellung

- Wähle eine Nummerierung der Gitterpunkte in  $\Omega_h$ .
- Das diskrete Poisson-Problem ist äquivalent zu einem LGS  $Ax = b$ , wobei  $x \in \mathbb{R}^{(n-1)^2}$  und  $A \in \mathbb{R}^{(n-1)^2 \times (n-1)^2}$  definiert ist durch

$$A = \frac{1}{h^2} \begin{pmatrix} T & -I & & & 0 \\ -I & T & -I & & \\ & -I & \ddots & \ddots & \\ & & \ddots & \ddots & -I \\ 0 & & & -I & T & -I \\ & & & -I & T & \end{pmatrix} \quad \text{mit } T = \begin{pmatrix} 4 & -1 & & & \\ -1 & 4 & -1 & & \\ & -1 & \ddots & & \\ & & & -1 & 4 \\ & & & & -1 & 4 \end{pmatrix},$$

sodass  $T \in \mathbb{R}^{(n-1) \times (n-1)}$ .

Ist das lineare Gleichungssystem  $Ax = b$  eindeutig lösbar? Wie gut approximiert  $u_h$  die Funktion  $u$ . Hier soll die Fourier-Analyse zu Hilfe genommen werden.

## 2.2 Fourier-Analyse

Die Matrix  $A$  ist symmetrisch, es existieren also nur reelle Eigenwerte und damit eine Basis aus orthogonalen Eigenvektoren. Diese kann man explizit angeben.

**Bemerkung 2.1.** Betrachte den Ursprung von  $A$ , den Laplace-Operator. Erfüllt ein Vektor  $v$  die Gleichung  $Bv = \lambda v$  mit einer Matrix  $B$  und zugehörigem Eigenwert  $\lambda$  von  $B$ , so heißt  $v$  Eigenvektor von  $B$  zum Eigenwert  $\lambda$ . Erfüllt nun eine Funktion  $u$  die Gleichung  $\Delta u = \lambda u$ , so nennt man  $u$  Eigenfunktion von  $\Delta$  zum Eigenwert  $\lambda$ .

Betrachte für  $\nu, \mu \in \mathbb{N}$  die Funktionen

$$e^{\nu\mu}(x, y) := \sin(\nu\pi x) \cdot \sin(\mu\pi y), \quad (x, y) \in \Omega.$$

Diese Funktion/-en ist/sind Eigenfunktion/-en von  $\Delta$  auf  $\Omega$ . Zudem gilt  $e^{\nu\mu}$  ist Null auf  $\partial\Omega$  für alle  $\nu, \mu \in \mathbb{N}$ . Es gilt

$$-\Delta e^{\nu\mu} = [(\nu\pi)^2 + (\mu\pi)^2] e^{\nu\mu}.$$

Betrachte die Restriktionen auf das Gitter  $\bar{\Omega}_h$ :

$$e_h^{\nu\mu}(x, y) := e^{\nu\mu}(x, y) \quad \forall (x, y) \in \bar{\Omega}_h, \quad 1 \leq \nu, \mu \leq n-1.$$

**Lemma 2.2.** Die  $e^{\nu\mu}$  bilden eine Orthogonalbasis von  $\ell^2(\Omega)$ , d.h.

$$\sum_{\xi \in \Omega_h} e_h^{\nu\mu}(\xi) \cdot e_h^{\nu'\mu'}(\xi) = \begin{cases} \frac{4}{h^2} & , (\nu, \mu) = (\nu', \mu') \\ 0 & , \text{sonst.} \end{cases}$$

*Beweis.* Siehe Bemerkung 8.52 bei Dahmen, Reusken. □

Sind die diskretisierten Eigenfunktionen  $e_h^{\nu\mu}$  auch Eigenfunktionen des diskretisierten Laplace-Operators  $\Delta_h$ ? Die Antwort ist: Ja! Die  $e_h^{\nu\mu}$  sind Eigenvektoren von  $\Delta_h$ : für alle  $\xi \in \Omega_h$  und  $1 \leq \nu, \mu \leq n-1$ :

$$-\Delta_h e_h^{\nu\mu}(\xi) = \lambda_{\nu\mu} e_h^{\nu\mu}(\xi) \text{ mit} \\ \lambda_{\nu\mu} = \left[ \sin^2\left(\frac{1}{2}\nu\pi h\right) + \sin^2\left(\frac{1}{2}\mu\pi h\right) \right] \cdot \frac{4}{h^2}.$$

Wähle Nummerierung der Gitterknoten in  $\Omega_h$ .

- Gitterfunktion  $e_h^{\nu\mu} \in \ell^2(\Omega_h)$  kann als Vektor  $\bar{e}_h^{\nu\mu} \in \mathbb{R}^{(n-1)^2}$  dargestellt werden.
- $A$  ist Matrixdarstellung des Operators  $\Delta_h: \ell_0^2(\bar{\Omega}_h) \rightarrow \ell^2(\Omega_h)$ .
- Die  $\bar{e}_h^{\nu\mu}$  sind Eigenvektoren von  $A$  zu den Eigenwerten  $\lambda_{\nu\mu}$ .

Da alle  $\lambda_{\nu\mu} > 0$  folgt:

## 2 Differenzenverfahren für das Poisson-Problem

**Lemma 2.3.** Die Matrix  $A$  des diskretisierten Poisson-Problems ist symmetrisch und positiv definit.

Die zweite wichtige Folgerung betrifft die Kondition von  $A$ .

**Lemma 2.4.** Für die Kondition von  $A$  bezüglich der  $\ell^2$ -Norm gilt

$$\kappa_2(A) = \frac{\cos^2\left(\frac{1}{2}\pi h\right)}{\sin^2\left(\frac{1}{2}\pi h\right)} = \frac{4}{(\pi h)^2} \left(1 + \mathcal{O}(h^2)\right).$$

Wir begegnen einem großen Problem. Für  $h \rightarrow 0$  (immer feineres Gitter, also immer bessere Approximation) wächst die Kondition der Matrix wie  $h^{-2}$ . Dies führt auf eine Langsamkeit eines Löser.

*Beweis.*  $A$  ist symmetrisch und positiv definit. Deshalb ist

$$\kappa_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

An der expliziten Darstellung der Eigenwerte sieht man

$$\begin{aligned} \kappa_2(A) &= \frac{\lambda_{n-1,n-1}}{\lambda_{1,1}} = \frac{\frac{4}{h^2} \cdot 2 \cdot \sin^2\left(\frac{1}{2}(n-1)\pi h\right)}{\frac{4}{h^2} \cdot 2 \cdot \sin^2\left(\frac{1}{2}\pi h\right)} \\ &= \frac{\sin^2\left(\frac{1}{2}\pi - \frac{1}{2}\pi h\right)}{\sin^2\left(\frac{1}{2}\pi h\right)} = \frac{\cos^2\left(\frac{1}{2}\pi h\right)}{\sin^2\left(\frac{1}{2}\pi h\right)} \\ &= \cot^2\left(\frac{1}{2}\pi h\right). \end{aligned}$$

Die Laurent-Reihe von  $\cot$  ist

$$\cot x = \frac{1}{x} - \frac{1}{3}x + \mathcal{O}(x^3) \quad 0 \leq |x| \leq \pi.$$

Deshalb gilt für die Kondition

$$\begin{aligned} \kappa_2(A) &= \left[ \frac{1}{\frac{1}{2}\pi h} - \frac{1}{3} \cdot \frac{1}{2}\pi h + \mathcal{O}(h^3) \right]^2 \\ &= \left( \frac{2}{\pi h} \right)^2 - \frac{2}{3} + \mathcal{O}(h^2) + \left( \frac{\pi h}{6} \right)^2 + \mathcal{O}(h^4) + \mathcal{O}(h^6) \\ &= \left( \frac{2}{\pi h} \right)^2 \left[ 1 - \frac{2}{3} \left( \frac{\pi h}{2} \right)^2 + \mathcal{O}(h^4) \right] \\ &= \left( \frac{2}{\pi h} \right)^2 \left[ 1 + \mathcal{O}(h^2) \right]. \end{aligned} \quad \square$$

## 2.3 Diskretisierungsfehleranalyse

Wir suchen eine Funktion  $u \in C^2(\Omega) \cap C(\bar{\Omega})$ , welche das Poisson-Problem löst. Dazu berechnen wir eine Funktion  $u_h \in \ell^2(\bar{\Omega}_h)$ , welche  $u$  auf dem Gitter  $\Omega_h$  approximiert. Wie gut ist diese Approximation?

**Definition 2.5.** Der Diskretisierungsfehler ist

$$e_h := u|_{\bar{\Omega}_h} - u_h.$$

Sei  $\ell_0^2(\bar{\Omega}_h)$  die Menge aller Gitterfunktionen, die auf  $\bar{\Omega}_h \setminus \Omega_h$  den Wert 0 annehmen. Da  $u_h$  die korrekten Randwerte hat, gilt  $e_h \in \ell_0^2(\bar{\Omega}_h)$ . Betrachte  $\Delta_h$  als Operator  $\ell_0^2(\bar{\Omega}_h) \rightarrow \ell^2(\Omega_h)$ . Die folgende Teilung ist fundamental für die Theorie des Diskretisierungsfehlers von FD-Methoden:

$$\begin{aligned} \|e_h\| &= \|\Delta_h^{-1} \Delta_h e_h\| = \|\Delta_h^{-1} \Delta_h (u|_{\bar{\Omega}_h} - u_h)\| \\ &= \|\Delta_h^{-1} (\Delta_h u|_{\bar{\Omega}_h} + f|_{\Omega_h})\| \\ &\leq \|\Delta_h^{-1}\| \cdot \|\Delta_h u|_{\Omega_h} + f|_{\Omega_h}\|. \end{aligned}$$

Der Ausdruck  $\|\Delta_h u|_{\Omega_h} + f|_{\Omega_h}\|$  heißt *Konsistenzfehler*. Er misst den *Defekt*, d.h. den Fehler, der entsteht, wenn man die exakte Lösung  $u$  in die diskrete Gleichung  $-\Delta u_h = f$  einsetzt. Die Zahl  $\|\Delta_h^{-1}\|$  bezeichnet ein Maß für die *Stabilität*.

- Ein Verfahren heißt stabil, wenn  $\|\Delta_h^{-1}\|$  beschränkt und unabhängig von  $h$  ist.

**Bemerkung 2.6.** Finde ein  $u$  mit  $-\Delta u = f$ ,  $u = g$  auf  $\partial\Omega$ . Sei  $\tilde{u}: \Omega \rightarrow \mathbb{R}$  mit  $\tilde{u} = g$  auf  $\partial\Omega$ . Dann ist  $u = \tilde{u} + w$  mit  $w = 0$  auf  $\partial\Omega$ .

$$-\Delta u = f \iff -\Delta(\tilde{u} + w) = f \iff -\Delta w = \underbrace{f + \Delta\tilde{u}}_{=: \tilde{f}}.$$

Wir erhalten ein Problem gleicher Art, aber mit Randbedingungen  $w = 0$  auf  $\partial\Omega$ .

**Satz 2.7** (Hauptsatz der Numerik). *Aus Konsistenz  $\|\Delta_h u|_{\Omega_h} + f|_{\Omega_h}\| \xrightarrow{h \rightarrow 0} 0$  und Stabilität  $\|\Delta_h^{-1}\| \leq C$  ( $h \rightarrow 0$ ) folgt Konvergenz.*

Unsere Aufgabe ist es, Konsistenz und Stabilität zu untersuchen.

- Wir wählen  $\|\cdot\| = \|\cdot\|_\infty$ .
- Stärkste Forderung: Konvergenz in der  $\|\cdot\|_\infty$ -Norm.

## 2 Differenzenverfahren für das Poisson-Problem

**Lemma 2.8** (Konsistenz in  $\|\cdot\|_\infty$ , DR, Lemma 12.11). Sei  $u \in C^4(\bar{\Omega})$  und

$$C := \max \left\{ \left\| \frac{\partial^4 u}{\partial x^4} \right\|_{\infty, \bar{\Omega}}, \left\| \frac{\partial^4 u}{\partial y^4} \right\|_{\infty, \bar{\Omega}} \right\}.$$

Dann gilt

$$\left\| -\Delta_h u|_{\bar{\Omega}_h} - f|_{\Omega_h} \right\|_\infty \leq \frac{1}{6} Ch^2.$$

*Beweis.* Approximationsgüte unserer FD-Formel war

$$v''(x) = \frac{v(x+h) - 2v(x) + v(x-h)}{h^2} - \frac{h^2}{12} v^{(4)}(\zeta)$$

für ein  $\zeta \in [x-h, x+h]$ . Sei jetzt  $(x, y) \in \Omega_h$ .

$$\begin{aligned} (-\Delta_h u)(x, y) &= \frac{1}{h^2} \begin{bmatrix} -1 & 2 & -1 \end{bmatrix}_{(x,y)} u + \frac{1}{h^2} \begin{bmatrix} -1 \\ 2 \\ -1 \end{bmatrix}_{(x,y)} u \\ &= -\frac{\partial^2 u}{\partial x^2}(x, y) - \frac{h^2}{12} \frac{\partial^4 u}{\partial x^4}(\eta, y) - \frac{\partial^2 u}{\partial y^2}(x, y) - \frac{h^2}{12} \frac{\partial^4 u}{\partial y^4}(x, \tilde{\eta}), \quad \eta, \tilde{\eta} \in [0, 1] \\ &= -\Delta u(x, y) - \frac{h^2}{12} \left[ \frac{\partial^4 u}{\partial x^4}(\eta, y) + \frac{\partial^4 u}{\partial y^4}(x, \tilde{\eta}) \right] \\ &= f(x, y) - \frac{h^2}{12} \left[ \frac{\partial^4 u}{\partial x^4}(\eta, y) + \frac{\partial^4 u}{\partial y^4}(x, \tilde{\eta}) \right] \end{aligned}$$

Also

$$\begin{aligned} \left\| -\Delta_h u|_{\bar{\Omega}_h} - f|_{\Omega_h} \right\|_\infty &= \max_{(x,y) \in \Omega_h} \left| -(\Delta_h u|_{\bar{\Omega}_h})(x, y) - f(x, y) \right| \\ &= \max_{(x,y) \in \Omega_h} \left| \frac{h^2}{12} \left[ \frac{\partial^4 u}{\partial x^4}(\eta, y) + \frac{\partial^4 u}{\partial y^4}(x, \tilde{\eta}) \right] \right| \\ &\leq \frac{h^2}{12} \left[ \left\| \frac{\partial^4 u}{\partial x^4} \right\|_{\infty, \bar{\Omega}_h} + \left\| \frac{\partial^4 u}{\partial y^4} \right\|_{\infty, \bar{\Omega}_h} \right] \leq \frac{1}{6} Ch^2. \quad \square \end{aligned}$$

Wir müssen jetzt noch Stabilität des Verfahrens zeigen, d.h.  $\|\Delta_h^{-1}\|$  bleibt beschränkt für  $h \rightarrow 0$ . Dies ist einfacher, wenn man die algebraische Darstellung  $A^{-1}$  statt  $\Delta_h^{-1}$  betrachtet. Insbesondere gilt  $\|\Delta_h^{-1}\|_\infty = \|A^{-1}\|_\infty$  (Operatornorm = Matrixnorm).

**Definition 2.9.** Eine Matrix  $M \in \mathbb{R}^{n \times n}$  heißt *reduzibel*, wenn man die Zeilen und Spalten so permutieren kann, dass die folgende Blockgestalt

$$\begin{pmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{pmatrix}$$

mit quadratischem  $M_{11}$  entsteht. Eine Matrix heißt diagonaldominant, wenn

$$\sum_{j \neq i} |m_{ij}| \leq |m_{ii}| \quad \forall i = 1, \dots, n$$

mit strikter Ungleichheit für mind. einen Index  $i$ .

Matrizen, die durch die Diskretisierung partieller Differentialgleichungen entstehen, sind fast immer irreduzibel, aber nur manchmal auch diagonaldominant. Wir benutzen die folgende wichtige Eigenschaft.

**Satz 2.10.** *Sei  $A \in \mathbb{R}^{n \times n}$  irreduzibel und diagonaldominant mit  $a_{ij} \leq 0$  für alle  $i \neq j$ . Dann ist  $A$  invertierbar und  $(A^{-1})_{ij} \geq 0$  für alle  $i, j$ .*

*Beweis.* Hackbusch, „Iterative Solution of Large Sparse Systems of Equations“, Springer, 1994. □

Diese Eigenschaft ist eng mit diskreten Maximumprinzipien verwandt. Wenn  $f(x) \leq 0$  für alle  $x \in \Omega$ , dann ist auch  $u \leq 0$  für alle  $x \in \Omega$ . Eine FD-Diskretisierung führt auf ein Gleichungssystem  $Ax = b$ . Wenn  $\forall x \in \Omega_h: b \leq 0$ , dann  $\forall x \in \Omega_0: x \leq 0$ ,  $x = A^{-1}b$ . Gilt insbesondere dann, wenn  $(A^{-1})_{ij} \geq 0$  für alle  $i, j$ .

**Satz 2.11** (DR, Lemma 12.21). *Sei  $A$  die Matrix des FD-diskretisierten Poisson-Problems. Dann gilt die Ungleichung*

$$\|A^{-1}\|_{\infty} \leq \frac{1}{8}.$$

*Beweis.*  $A$  ist irreduzibel und diagonaldominant und  $a_{ij} \leq 0$  für alle  $i \neq j$ . Dann existiert  $A^{-1}$  und hat nur nicht-negative Einträge. Für solche Matrizen kann man  $\|A^{-1}\|_{\infty}$  als Vektornorm schreiben:

$$\begin{aligned} \|A^{-1}\|_{\infty} &= \max_{1 \leq i \leq n} \sum_{j=1}^n |(A^{-1})_{ij}| \\ &= \max_{1 \leq i \leq n} \sum_{j=1}^n (A^{-1})_{ij} = \max_{1 \leq i \leq n} \left( A^{-1} \underbrace{\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}}_{=: \zeta} \right)_i = \|A^{-1}\zeta\|_{\infty} \end{aligned}$$

Betrachte die Funktion  $u: \Omega = (0, 1)^2 \rightarrow \mathbb{R}$ ,  $u(x, y) := x(1-x) + y(1-y)$ . Es gilt  $-\Delta u = 4$ .  $u$  löst das Poisson-Problem mit  $f(x, y) = 4$  und Randwerten  $g(x, y) = x(1-x) + y(1-y)$  für alle  $(x, y) \in \partial\Omega$ . Sei  $b \in \mathbb{R}^n$  die rechte Seite des FD-diskretisierten Poisson-Problems. Nach Konstruktion gilt  $b \geq 4\zeta$  (komponentenweise). Sei  $u_h$  die diskrete Lösung. Dann ist

$$\frac{1}{4}u_h = \frac{1}{4}A^{-1}b \geq \frac{1}{4}A^{-1}(4\zeta) \geq 0.$$



## 2 Differenzenverfahren für das Poisson-Problem

Für dieses spezielle Beispiel ist die FD-Lösung exakt:

$$\begin{aligned} \|e_h\| &= \|u|_{\Omega_h} - u_h\| \leq \|\Delta_h^{-1}\| \cdot \|-\Delta_h u|_{\overline{\Omega}_h} - f|_{\Omega_h}\| \\ &\leq \|\Delta_h^{-1}\| \cdot \frac{1}{6} h^2 \underbrace{\max \left\{ \left\| \frac{\partial^4 u}{\partial x^4} \right\|_{\infty} + \left\| \frac{\partial^4 u}{\partial y^4} \right\|_{\infty} \right\}}_{=0} \end{aligned}$$

nach Konstruktion von  $u$ . Damit ist

$$\begin{aligned} \|A^{-1}\|_{\infty} &= \|A^{-1}\zeta\|_{\infty} \leq \frac{1}{4} \|u_h\|_{\infty} = \frac{1}{4} \max_{(x,y) \in \Omega_h} |u_h(x,y)| = \frac{1}{4} \max_{(x,y) \in \Omega_h} |u(x,y)| \\ &\leq \frac{1}{4} \max_{(x,y) \in \Omega} |u(x,y)|. \end{aligned}$$

Das Maximum von  $u$  ist in  $(\frac{1}{2}, \frac{1}{2})$ , also

$$\|A^{-1}\|_{\infty} \leq \frac{1}{4} u\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{8}. \quad \square$$

Zusammengefasst erhält man folgendes Konvergenzresultat:

**Satz 2.12** (Konvergenz von FD für das Poisson-Problem). Sei  $u \in C^4(\overline{\Omega})$  und

$$C := \max \left\{ \left\| \frac{\partial^4 u}{\partial x^4} \right\|_{\infty} + \left\| \frac{\partial^4 u}{\partial y^4} \right\|_{\infty} \right\}.$$

Dann gilt

$$\|e_h\|_{\infty} \leq \frac{1}{48} \cdot Ch^2.$$

Die Forderung  $u \in C^4(\overline{\Omega})$  ist sehr stark und häufig nicht erfüllt. Sie ist einer der Hauptgründe, sich nach Alternativen zur FD-Methode umzusehen.

### 3 Finite Elemente

FD-Verfahren haben zwei prinzipielle Schwächen:

- Konstruktion von Differenzensternen wird schwierig, wenn das Gebiet  $\Omega$  nicht nur achsenparallele Kanten hat.
- Für die Diskretisierungsfehlerabschätzungen mussten wir  $u \in C^4(\bar{\Omega})$  annehmen. Diese Annahme gilt häufig nicht.

Es kommt noch schlimmer: Man kann relativ leicht Poisson-Probleme konstruieren, die nicht mal in  $C^2(\Omega) \cap C(\bar{\Omega})$  eine Lösung haben. In so einem Fall ist das klassische/starke Poisson-Problem

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

nicht wohlgestellt. Wir betrachten jetzt ein etwas allgemeineres Modellproblem: Sei  $\Omega \subset \mathbb{R}^d$  ein Gebiet. Finde ein  $u: \Omega \rightarrow \mathbb{R}$ , so dass

$$-\operatorname{div}(C\nabla u + bu) + ru = f \text{ auf } \Omega$$

und bestimmte Randbedingungen gelten. Dabei sind  $C$  eine  $\mathbb{R}^{d \times d}$ -Matrix,  $b \in \mathbb{R}^d$ ,  $r \in \mathbb{R}$  und  $f: \Omega \rightarrow \mathbb{R}$  eine gegebene Funktion. Diese Gleichung umfasst diverse interessante Fälle.

#### Beispiel 3.1.

- $C = \text{Id}$ ,  $b = 0$ ,  $r = 0$ . Die Gleichung wird zu  $-\Delta u = f$ .
- $C = 0$ ,  $r = 0$  führt zu stationären, linearen Transportgleichung  $b\nabla u = f$ .

Weiterhin betrachten wir allgemeinere Randbedingungen:

- Sei  $\partial\Omega$  in zwei disjunkte Teile  $\Gamma_D$  und  $\Gamma_N$  zerlegt, sodass  $\partial\Omega = \overline{\Gamma_D} \cap \overline{\Gamma_N}$ .

Wir fordern

- $u = g$  auf  $\Gamma_D$  („Dirichlet“-Randbedingung).
- $(C\nabla u + bu)n = j$  auf  $\Gamma_N$  („Neumann“-Randbedingung).

Für die Poisson-Gleichung ist die letzte Bedingung gerade

$$\nabla u \cdot n = \frac{\partial u}{\partial n} = j,$$

die Normalenableitung wird vorgegeben. Für die stationäre Transportgleichung:

$$u\langle b, n \rangle = j.$$

### 3.1 Die Variationsformulierung

Wir wollen jetzt eine Umformulierung der Gleichung herleiten, die auch für Funktionen  $u$  Sinn hat, die nicht in  $C^2(\Omega) \cap C(\bar{\Omega})$  sind. Sei  $C_D^\infty$  die Menge aller unendlich oft differenzierbaren Funktionen, die auf  $\Gamma_D$  Null sind<sup>1</sup>. Sei  $\phi \in C_D^\infty$  beliebig. Wir multiplizieren die PDG mit  $\phi$  und integrieren über  $\Omega$ :

$$\int_{\Omega} -\operatorname{div}(C\nabla u + bu)\phi \, dx + \int_{\Omega} ru\phi \, dx = \int_{\Omega} f\phi \, dx.$$

Wir benutzen die Greensche Formel

$$\int_{\Omega} v \frac{\partial w}{\partial x_i} \, dx = - \int_{\Omega} \frac{\partial v}{\partial x_i} w \, dx + \int_{\partial\Omega} vw\nu_i \, dS$$

und erhalten

$$\int_{\Omega} (C\nabla u + bu) \cdot \nabla\phi \, dx + \int_{\Omega} ru\phi \, dx = \int_{\partial\Omega} (C\nabla u + bu)\nu\phi \, dS + \int_{\Omega} f\phi \, dS.$$

Die Funktion  $\phi$  heißt Testfunktion. Der Rand  $\partial\Omega$  besteht aus  $\Gamma_N$  und  $\Gamma_D$ :

$$\int_{\partial\Omega} (C\nabla u + bu)\nu\phi \, dS = \underbrace{\int_{\Gamma_N} (C\nabla u + bu)\nu\phi \, dS}_{=\int_{\Gamma_N} j\phi \, dS} + \underbrace{\int_{\Gamma_D} (C\nabla u + bu)\nu\phi \, dS}_{=0}.$$

Schwache Form: Finde ein  $u: \Omega \rightarrow \mathbb{R}$  mit  $u|_{\Gamma_D} = g$  und

$$\int_{\Omega} (C\nabla u + bu) \cdot \nabla\phi \, dx + \int_{\Omega} ru\phi \, dx = \int_{\Gamma_N} j \cdot \phi \, dS + \int_{\Omega} f\phi \, dS$$

für alle  $\phi \in C_D^\infty$ . Kurzschreibweise: Für zwei Funktionen  $v, w: \Omega \rightarrow \mathbb{R}$  sei

$$a(v, w) := \int_{\Omega} (C\nabla v + bv)\nabla w \, dx + \int_{\Omega} rvw \, dx$$

und

$$\ell(v) := \int_{\Gamma_N} jv \, dS + \int_{\Omega} fv \, dx.$$

Finde also ein  $u: \Omega \rightarrow \mathbb{R}$  mit  $u|_{\Gamma_D} = g$ , sodass

$$a(u, \phi) = \ell(\phi) \quad \forall \phi \in C_D^\infty.$$

Die Menge aller  $u$  mit  $u|_{\Gamma_D} = g$  ist ein affiner Raum. Wir hätten lieber  $u$  aus einem Vektorraum. Sei  $u_g$  eine Funktion mit  $u_g|_{\Gamma_D} = g$ . Dann ist  $u = \tilde{u} + u_g$  mit  $\tilde{u}|_{\Gamma_D} = 0$ . Sei  $\phi \in C_D^\infty$ .

$$a(u, \phi) = a(\tilde{u} + u_g, \phi) = a(\tilde{u}, \phi) + a(u_g, \phi) = \ell(\phi)$$

$$a(\tilde{u}, \phi) = \ell(\phi) - a(u_g, \phi) =: \tilde{\ell}(\phi).$$

<sup>1</sup>In der Literatur findet man häufig noch die Bedingung „mit kompaktem Träger“, aber unser  $\Omega$  ist ohnehin immer beschränkt.

Finde  $\tilde{u}: \Omega \rightarrow \mathbb{R}$  mit  $\tilde{u}|_{\Gamma_D} = 0$ , so dass

$$\forall \phi \in C_D^\infty: a(\tilde{u}, \phi) = \tilde{\ell}(\phi).$$

Wir betrachten deshalb jetzt nur noch dieses Problem.

## 3.2 Sobolev-Räume

In welchem Funktionenraum suchen wir nach  $u$ ?

- Im größten, in dem die Variationsgleichung sinnvoll definiert werden kann.

Wir brauchen dazu von  $u$

1. erste Ableitungen,
2. Integrierbarkeit und
3. Restriktion auf  $\Gamma_D$ .

Sei  $\Omega \subset \mathbb{R}^d$  offen mit stückweise glattem Rand. Der Lebesgue-Raum  $L^p(\Omega)$  für  $1 \leq p < \infty$  ist

$$L^p(\Omega) := \left\{ v: v \rightarrow \mathbb{R} \mid \int_{\Omega} |v|^p dx < \infty \right\}.$$

Genauer:  $L^p$  sei ein Raum von Äquivalenzklassen solcher Funktionen. Zwei Funktionen in  $L^p$  seien äquivalent, falls sie sich nur auf einer Menge vom Maß = 0 unterscheiden. Daraus folgt, dass  $L^p$ -Funktionen nicht punktweise ausgerechnet werden können.

**Lemma 3.2.** Die Funktion  $\|v\|_{L^p} := \sqrt[p]{\int_{\Omega} |v|^p dx}$  ist eine Norm auf  $L^p$ . Damit ist  $L^p$  ein Banach-Raum.

**Lemma 3.3.** Die Bilinearform

$$\langle v, w \rangle_{L^2} := \int_{\Omega} vw dx$$

ist ein Skalarprodukt auf  $L^2$ . Damit bildet  $L^2$  ein Hilbertraum.

Die Eigenschaft, dass man  $L^p$ -Funktionen nicht punktweise auswerten sondern nur integrieren kann, wirkt zunächst komisch. Tatsächlich kommen in unserer Variationsformulierung wirklich nur Integrale von Funktionen vor. Andererseits brauchen wir zusätzlich auch Ableitungen von Funktionen und  $L^p$ -Funktionen sind im Allgemeinen nicht ableitbar.

### 3.2.1 Schwache Ableitung

Seien  $\Omega = (a, b) \subset \mathbb{R}$ ,  $u \in C^1(\Omega)$  und  $v := u'$ . Sei  $C_0^\infty(\Omega)$  der Raum aller beliebig oft stetig differenzierbaren Funktionen auf  $(a, b)$  mit kompakten Träger und  $\phi \in C_0^\infty$ .  $C_0^\infty$  heißt *Raum der Testfunktionen*.

$$\int_{\Omega} \phi u' dx = \underbrace{\int_{\partial\Omega} \phi v dx}_{=0} - \int_{\Omega} \phi' u dx.$$

Also

$$\int_{\Omega} \phi v dx = - \int_{\Omega} \phi' u dx$$

für alle Testfunktionen  $\phi \in C_0^\infty$ .

**Definition 3.4.** Die Funktion  $u \in L^2(\Omega)$  besitzt die schwache Ableitung  $v$ , falls  $v \in L^2(\Omega)$  und

$$\int_{\Omega} \phi v dx = - \int_{\Omega} \phi' u dx$$

für alle Testfunktionen  $\phi \in C_0^\infty$ .

**Beispiel 3.5.** Die Funktion  $x \mapsto |x|$  hat die schwache Ableitung

$$v(x) = \begin{cases} 1 & , x > 0 \\ -1 & , x < 0 \end{cases}.$$

Beweis: Sei  $v \in C_0^\infty$ . Dann

$$\begin{aligned} \int_{\Omega} \phi v &= \int_{-\infty}^0 \phi(-1) dx + \int_0^{\infty} \phi(1) dx \\ &= - \int_{-\infty}^0 \phi x dx - \int_0^{\infty} \phi(-x) dx = - \int_{-\infty}^{\infty} \phi |x| dx. \end{aligned}$$

Was ist  $v(0)$ ? Der Wert  $v(0)$  ist undefiniert. Es handelt sich hierbei um eine Lebesgue-Funktion, deren Punktwerte nicht ausgerechnet werden können. Nur Integrale sind wohldefiniert.

Für höhere schwache Ableitungen verwenden wir die Multiindex-Notation:

- Sei  $\alpha \in \mathbb{N}_0^d$  ein Multiindex.
- Wir definieren

$$\partial^\alpha u := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \frac{\partial^{\alpha_2}}{\partial x_2^{\alpha_2}} \cdots \frac{\partial^{\alpha_d}}{\partial x_d^{\alpha_d}}, \quad |\alpha| := \sum_{i=1}^d \alpha_i.$$

**Definition 3.6.**  $u \in L^2(\Omega)$  besitzt die schwache Ableitung  $v = \partial^\alpha u$ , falls  $v \in L^2(\Omega)$  und

$$\int_{\Omega} \phi v = (-1)^{|\alpha|} \int_{\Omega} \partial^\alpha \phi u \quad \forall \phi \in C_0^\infty.$$

**Definition 3.7** (Sobolev-Räume). Für  $m \in \mathbb{N}$  bezeichne  $H^m(\Omega)$  die Menge aller Funktionen  $u \in L^2(\Omega)$ , die schwache Ableitungen  $\partial^\alpha u$  für alle  $|\alpha| \leq m$ .

Mit dem Skalarprodukt

$$\langle v, w \rangle_{H^m} := \sum_{|\alpha| \leq m} \int_{\Omega} \partial^\alpha v \partial^\alpha w$$

ist  $H^m(\Omega)$  ein Hilbertraum. Die dazugehörige Norm ist

$$\|v\|_{H^m} := \sqrt{\sum_{|\alpha| \leq m} \int_{\Omega} |\partial^\alpha v|^2}.$$

Die Sobolev-Halbnorm (Semi-Norm) ist definiert durch

$$|v|_{H^m} := \sqrt{\sum_{|\alpha|=m} \int_{\Omega} |\partial^\alpha v|^2}.$$

Die Funktion  $a: u \mapsto a(u, \phi) \in \mathbb{R}$  ist auf  $H^1(\Omega)$  wohldefiniert.

### 3.2.2 Der Spursatz

Problem: Bedingung  $u|_{\Gamma_D} = 0$ . Jedoch sind Elemente von  $H^1 \subset L^2$  Äquivalenzklassen von Funktionen. Zwei Funktionen sind äquivalent, wenn sie sich nur auf Lebesgue-Nullmengen unterscheiden. Die Menge  $\Gamma_D$  hat Maß = 0. Zur Rettung gibt es den Spursatz.

**Satz 3.8.** Sei  $\Omega$  beschränkt und habe  $\Omega$  einen glatten Rand. Dann existiert eine lineare, beschränkte Abbildung

$$\text{tr}: H^1(\Omega) \rightarrow L^2(\Gamma_D),$$

so dass  $\text{tr } v = v|_{\Gamma_D}$  für alle  $v \in C^1(\overline{\Omega})$ .

**Bemerkung 3.9.**  $\text{tr}$  ist nicht surjektiv. Das Bild  $\text{tr}(H^1(\Omega))$  nennt man  $H^{\frac{1}{2}}(\Gamma_D)$ . Es gilt

$$L^2(\Gamma_D) \supsetneq H^{\frac{1}{2}}(\Gamma_D) \supsetneq H^1(\Gamma_D).$$

Sei  $H_D^1$  der Teilraum aller Funktionen  $u \in H^1(\Omega)$ , für den  $\text{tr } u = 0$ . Finde nun  $u \in H_D^1(\Omega)$ , so dass

$$a(u, \phi) = \ell(\phi) \quad \forall \phi \in C_D^\infty$$

für irgendeine Bilinearform  $a$  und Linearform  $\ell$ . Häufig ersetzt man  $C_D^\infty$  durch  $H_D^1$ . Und das geht so:

- Annahme:  $a(\cdot, \cdot)$  und  $\ell(\cdot)$  seien stetig.
- Sei  $\{\phi_n\}$  eine  $\|\cdot\|_{H^1}$ -konvergente Folge in  $C_D^\infty$  mit Grenzwert  $\phi \in \overline{C_D^\infty}$ .
- Für jedes  $\phi_n$  gilt  $a(u, \phi_n) = \ell(\phi_n)$ .

### 3 Finite Elemente

- Da  $a$  und  $\ell$  stetig sind, gilt auch  $a(u, \phi) = \ell(\phi)$ .

**Lemma 3.10.**  $H_D^1$  ist der Abschluss von  $C_D^\infty$  bezüglich  $\|\cdot\|_{H^1}$ .

Deshalb äquivalent: Finde  $u \in H_D^1(\Omega)$ , sodass

$$a(u, \phi) = \ell(\phi) \quad \forall \phi \in H_D^1(\Omega).$$

**Beispiel 3.11.**  $a(v, w) := \int_\Omega C \nabla v \cdot \nabla w$  ist stetig in  $H^1$ . Beweis:

$$\begin{aligned} |a(v, w)| &\leq c \|v\|_{H^1} \|w\|_{H^1} \\ |a(v, w)| &= \left| \sum_{i,j} C_{ij} \int \partial^i v \partial^j w \right| \leq K_1 \left| \sum_{i,j} \int \partial^i v \partial^j w \right| \\ &\leq K_1 \sum_{i,j} \int |\partial^i v \partial^j w| \\ &\leq K_1 \sum_{i,j} \left[ \int (\partial^i v)^2 \cdot \int (\partial^j w)^2 \right]^{\frac{1}{2}} \\ &\leq K_2 |v|_{H^1} \cdot |w|_{H^1} \leq K_2 \|v\|_{H^1} \cdot \|w\|_{H^1}. \end{aligned}$$

### 3.3 Wohlgestelltheit

Seien  $U, V$  Banach-Räume. Seien  $a(\cdot, \cdot): U \times V \rightarrow \mathbb{R}$  bilinear und  $\ell \in V'$ . Finde  $u \in U$ , sodass  $a(u, v) = \ell(v)$  für alle  $v \in V$ . Wohlgestelltheit (im Sinne von Hadamard) heißt:

- Es existiert eine Lösung  $u$ .
- Diese ist eindeutig.
- $u$  hängt stetig von  $\ell$  ab.

Der Bilinearform  $a$  wird ein linearer Operator

$$L: U \rightarrow V'$$

zugeordnet, und zwar durch

$$\langle Lu, v \rangle_{V' \times V} = a(u, v) \quad \forall v \in V.$$

Das Variationsproblem kann man damit schreiben als: Finde  $u \in U$ , sodass  $Lu = \ell$ .

- Diese Gleichung hat eine Lösung, wenn  $L$  surjektiv ist.
- Sie hat höchstens eine Lösung, wenn  $L$  injektiv ist.
- Lösung hängen stetig von  $\ell$  ab, wenn  $L^{-1}$  stetig ist ( $u = L^{-1}\ell$ ,  $\|u\| \leq c\|\ell\|$ ).

**Definition 3.12.** Eine lineare, bijektive Abbildung  $L: U \rightarrow V'$  (oder allgemeiner:  $L: X \rightarrow Y$ ) heißt Isomorphismus, wenn  $L$  und  $L^{-1}$  stetig sind.

**Satz 3.13** (Banach, Nečas, Babuška). *Seien  $U, V$  Banachräume und  $V$  reflexiv. Eine lineare Abbildung  $L: U \rightarrow V'$  ist genau dann ein Isomorphismus, wenn die zugehörige Bilinearform  $a: U \times V \rightarrow \mathbb{R}$  die folgenden Eigenschaften erfüllt:*

1. *Stetigkeit:*

$$\exists C > 0 \forall u \in U \forall v \in V: |a(u, v)| \leq C \|u\|_U \|v\|_V.$$

2. *inf-sup-Bedingung:*

$$\begin{aligned} \exists \alpha > 0: \inf_{u \in U} \sup_{v \in V} \frac{a(u, v)}{\|u\|_U \|v\|_V} &\geq \alpha \\ \iff \exists \alpha > 0 \forall u \in U: \sup_{v \in V} \frac{a(u, v)}{\|v\|_V} &\geq \alpha \|u\|_U. \end{aligned}$$

3. *Nicht-Degeneriertheit:*

$$\forall v \in V \setminus \{0\} \exists u \in U: a(u, v) \neq 0.$$

**Definition 3.14** (Polare). Sei  $X$  ein Banachraum und  $W \subseteq X$  ein abgeschlossener Unterraum. Dann heißt die Menge

$$W^\circ := \{\ell \in X' \mid \ell(w) = 0 \forall w \in W\} \subseteq X'$$

Polare von  $W$ .

**Satz 3.15** (Satz vom abgeschlossenen Bild). *Seien  $X, Y$  Banachräume,  $Y$  reflexiv und  $L: X \rightarrow Y$  eine stetige, lineare Abbildung. Dann sind äquivalent:*

1. *Das Bild  $L(X)$  ist abgeschlossen in  $Y$ .*
2. *Es gilt  $L(X) = (\ker L)^\circ$ .*

*Beweis von BNB.* Wir zeigen nur „ $\implies$ “.

1. Aus Stetigkeit von  $a(\cdot, \cdot)$  folgt Stetigkeit von  $L$ . Sei dazu  $u \in U$ .

$$\begin{aligned} \|Lu\|_{V'} &= \sup_{\|v\|_V=1} \langle Lu, v \rangle_{V', V} = \sup_{\|v\|_V} a(u, v) \\ &\leq C \cdot \sup_{\|v\|_V} \|u\|_U \cdot \|v\|_V = C \|u\|_U. \quad \checkmark \end{aligned}$$

2. Aus der inf-sup-Bedingung folgt Injektivität von  $L$ . Seien dazu  $u_1, u_2 \in U$  mit  $Lu_1 = Lu_2$ . Nach Definition von  $L$  folgt

$$a(u_1, v) = a(u_2, v) \quad \forall v \in V.$$

Wende die inf-sup-Bedingung auf  $u := u_1 - u_2$  an:

$$\alpha \|u_1 - u_2\|_U \leq \sup_{v \in V} \frac{a(u_1 - u_2, v)}{\|v\|_V} = 0.$$

Es folgt  $u_1 = u_2$ .  $\checkmark$



### 3 Finite Elemente

3.  $L^{-1}$  ist stetig auf dem Bild von  $L$ . Sei  $\ell \in L(U)$ .  $L$  ist injektiv, daher existiert genau ein  $u \in U$  mit  $u = L^{-1}\ell$  ( $u$  löst  $Lu = \ell$  bzw.  $a(u, v) = \ell(v)$ ). Wir wollen:  $\|u\|_U = \|L^{-1}\ell\| \leq C\|\ell\|_{V'}$ . Mit der inf-sup-Bedingung folgt

$$\alpha\|u\|_U \leq \sup_{v \in V} \frac{a(u, v)}{\|v\|_V} = \sup_{v \in V} \frac{\ell(v)}{\|v\|_V} = \|\ell\|_{V'}. \quad \checkmark$$

4.  $L$  ist surjektiv. Die Abbildungen  $L$  und  $L^{-1}$  sind stetig, daher ist  $L(U)$  abgeschlossen in  $V'$ . Argument: Da  $L^{-1}$  stetig ist, existiert ein  $C > 0$  mit  $\|Lu\|_{V'} \geq C\|u\|_U$ . Sei  $(v_n)_n$  eine Folge in  $L(U)$  mit  $v_n \rightarrow v$ . Wir zeigen, dass  $v \in L(U)$ . Es gilt mit Injektivität von  $L$

$$\|v_n - v_m\| = \|LL^{-1}v_n - LL^{-1}v_m\| \geq c\|L^{-1}v_n - L^{-1}v_m\|.$$

Also ist  $L^{-1}v_n$  Cauchy-Folge in  $U$ .  $U$  ist jedoch ein Banachraum, damit existiert der Grenzwert  $z \in U$ .  $L$  ist stetig, deshalb ist

$$L(z) = L\left(\lim_{n \rightarrow \infty} L^{-1}v_n\right) = \lim_{n \rightarrow \infty} v_n = v,$$

also  $U \ni L(z) = v \implies L(U)$  ist abgeschlossen. Mit dem Satz vom abgeschlossenen Bild gilt

$$L(U) = (\ker L')^\circ.$$

Es ist  $L': V'' \rightarrow U'$  die Dualabbildung von  $L$ , bzw. mit Reflexivität von  $V: L': V \rightarrow U'$ . Dann ist

$$\begin{aligned} \ker L' &= \{v \in V \mid L'(v) = 0 \in U'\} \\ &= \{v \in V \mid \forall u \in U: L'(v)(u) = 0\} \\ &= \{v \in V \mid \forall u \in U: \langle L'v, u \rangle = 0\} \\ &= \{v \in V \mid \forall u \in U: a(u, v) = 0\}. \end{aligned}$$

Also  $L(U) = (\ker L')^\circ = \{v \in V \mid a(u, v) = 0 \forall u \in U\}^\circ$ . Wegen Voraussetzung 3 folgt

$$\{v \in V \mid \forall u \in U: a(u, v) = 0\}^\circ = \{0\}^\circ = V'.$$

Also ist  $L$  surjektiv.  $\checkmark$  □

Sei nun  $U = V$ .

**Definition 3.16.** Sei  $a: V \times V \rightarrow \mathbb{R}$  eine Bilinearform.  $a$  heißt *koerziv* oder *V-elliptisch*, wenn sie stetig ist und ein  $\alpha > 0$  existiert, sodass

$$\forall v \in V: \alpha\|v\|^2 \leq a(v, v).$$

**Korollar 3.17** (Lemma von Lax-Milgram). Sei  $a: V \times V \rightarrow \mathbb{R}$  eine koerzive Bilinearform und  $\ell \in V'$ . Dann hat die Variationsgleichung

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

eine eindeutige Lösung  $u \in V$  und es gilt

$$\|u\|_V \leq \frac{1}{\alpha}\|\ell\|_{V'}.$$

*Beweis.*

1. Stetigkeit. Ist klar. ✓
2. inf-sup-Bedingung. Sei  $u \neq 0$  beliebig. Dann gilt

$$\sup_{v \in V} \frac{a(u, v)}{\|v\|_V} \geq \frac{a(u, u)}{\|u\|_V} \geq \alpha \|u\|_V,$$

da  $a$  koerziv ist. ✓

3.  $\forall v \in V \setminus \{0\} \exists u \in V : a(u, v) \neq 0$ . Wähle dazu  $u = v$ . ✓ □

**Bemerkung 3.18.** Die endlich-dimensionale Version von BNB und Lax-Milgram wäre:

BNB Das LGS  $Ax = b$  ist eindeutig lösbar genau dann, wenn  $A$  invertierbar ist.

L-M  $Ax = b$  ist eindeutig lösbar, wenn  $A$  positiv definit.

**Beispiel 3.19.** Sei  $V := H^1(\Omega)$ ,  $a(v, w) := \int C \nabla v \nabla w + r v w$ ,  $C$  positiv definit und  $r > 0$ . Dann ist  $a(\cdot, \cdot)$   $H^1$ -elliptisch.

*Beweis.* Sei  $v \in H^1$ .

$$\begin{aligned} a(v, v) &= \int C \nabla v \nabla v + r v^2 \\ &\geq \gamma \int_{\Omega} |\nabla v|^2 + r \int_{\Omega} v^2 \\ &\geq \min\{\gamma, r\} \left( \int |\nabla v|^2 + \int v^2 \right) \\ &= \min\{\gamma, r\} \cdot \|v\|_{H^1}^2. \end{aligned} \quad \square$$

**Beispiel 3.20.**  $a(v, w) := \int \nabla v \nabla w$ .  $a(v, v) = \int |\nabla v|^2 \geq |v|_{H^1}^2$  (nicht  $H^1$ -elliptisch). Wenn  $\Gamma_D$  positives  $(d-1)$ -dimensionales Maß hat, dann ist  $a(\cdot, \cdot)$   $H_D^1$ -elliptisch.

*Beweis.* Sei  $v \in H_0^1(\Omega) \subset H^1(\Omega)$ .  $a(v, v) = |v|_{H^1}^2$ . Wir brauchen  $|v|_{H^1} \geq \alpha \|v\|_{H^1}$ .

**Lemma 3.21** (Poincaré-Friedrich-Ungleichung). Sei  $\Omega$  in einem  $d$ -dimensionalen Würfel der Kantenlänge  $s$  enthalten. Dann gilt

$$\|v\|_{L^2} \leq s |v|_{H^1}$$

für alle  $v \in H_D^1$ .

Es gilt also

$$\begin{aligned} a(v, v) &= |v|_{H^1}^2 = \frac{1}{2} |v|_{H^1}^2 + \frac{1}{2} |v|_{H^1}^2 \\ &\geq \frac{s^2}{2} \|v\|_{L^2}^2 + \frac{1}{2} |v|_{H^1}^2 \geq \alpha \|v\|_{H^1}^2. \end{aligned}$$

□

### 3.4 Finite Elemente

Seien  $U, V$  Banachräume,  $a: U \times V \rightarrow \mathbb{R}$  bilinear und  $\ell \in V'$ . Wir suchen  $u \in U$ , sodass

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

Idee der FE-Elemente (Galerkin-Approximation):

- Ersetze  $U, V$  durch endlich-dimensionale Räume  $U_h, V_h$ .

Bis auf weiteres betrachten wir *konforme* Approximationen, d.h.  $U_h \subset U$  und  $V_h \subset V$ .  
Diskretes Problem: Finde  $u_h \in U_h$ , sodass

$$\forall v_h \in V_h: a(u_h, v_h) = \ell(v_h).$$

Existieren Lösungen zu diesem Problem und sind sie eindeutig?

**Lemma 3.22.** Falls  $U = V$ ,  $U_h = V_h$  und  $a(\cdot, \cdot)$   $V$ -elliptisch ( $\exists \alpha > 0 \forall v \in V: a(v, v) \geq \alpha \|v\|_V^2$ ) ist, dann hat das diskrete Problem eine eindeutige Lösung.

*Beweis.*

- $V_h$  ist endlich-dimensional, also abgeschlossen im Banachraum  $V$ .
- Daher ist  $V_h$  selbst ein Banachraum.
- $a$  ist  $V$ -elliptisch und  $V_h \subset V$ , also ist  $a$   $V_h$ -elliptisch.
- Die Behauptung folgt aus dem Lemma von Lax-Milgram. □

FE-Methoden sind eine bestimmte Art, die Räume  $U_h, V_h$  zu wählen.

**Definition 3.23** (Gitter). Das Gebiet  $\Omega$  habe stückweise affinen Rand. Eine Menge  $\mathcal{T}$  von abgeschlossenen konvexen Polytopen  $T$  heißt *Gitter*, wenn gilt

1. (*Partitionierung*)  $\bar{\Omega} = \bigcup_{T \in \mathcal{T}} T$ .
2. (*Konformität*) Der Schnitt zweier Polytope aus  $\mathcal{T}$  ist eine gemeinsame Seitenfläche mit Dimension  $\leq d - 1$  oder leer.

Die  $T \in \mathcal{T}$  heißen *Elemente* des Gitters. Meistens verwendet man in

2D: Dreiecke, Vierecke,

3D: Tetraeder, Hexaeder.

FE-Räume  $U_h, V_h$  sind Räume von Funktionen  $\Omega \rightarrow \mathbb{R}^m$ , die auf jedem Element  $T \in \mathcal{T}$  polynomiell sind und gewissen Stetigkeitsbedingungen genügen.

**Beispiel 3.24.**

1. Lagrange-Elemente.

- Das Gitter  $\mathcal{T}$  bestehe nur aus Simplizes.
- $\Pi_p$  sei der Raum aller Polynome in  $d$  Variablen vom Grad höchstens  $p$ .
- Der Lagrange-Raum  $p$ -ter Ordnung ist

$$V_h^p = \{v_h \in C(\Omega) : v_h|_T \in \Pi_p \ \forall T \in \mathcal{T}\}.$$

- $V_h^p$  ist keine Teilmenge von  $C^1(\Omega)$ . Aber  $V_h^p \subset H^1(\Omega)$ .

## 2. Raviart-Thomas-Element.

- Raum von vektorwertigen Funktionen. Sei  $d = 2$ .

$$RT = \left\{ v \in L^2(\Omega) \mid \forall T \in \mathcal{T} : v_T = \begin{pmatrix} a_T \\ b_T \end{pmatrix} + c_T \begin{pmatrix} x \\ y \end{pmatrix}, \ a_T, b_T, c_T \in \mathbb{R}, \right\}$$

wobei zusätzlich  $v \cdot n$  stetig an den Elementgrenzen ist.  $n$  ist der äußere Einheitsnormalenvektor.

- $RT$  ist keine Teilmenge von  $C(\Omega)$ . Auch nicht von  $H^1(\Omega)$ .
- $RT \subset H(\text{div}) = \{v \in L^2 \mid \text{div } v \in L^2\} \supset (H^1)^2$ .

## 3. Crouzeix-Raviart-Element.

- Sei  $d = 2$ .

$$CR = \left\{ v \in L^2 \mid \forall T \in \mathcal{T} : v_T \in \Pi_1 \right\},$$

wobei zusätzlich  $v$  stetig an allen Kantenmitten ist.

- $CR$  ist kein Teilraum von  $H^1$ .
- Nichtkonformes Element.

Sei weiterhin  $U = V$ ,  $U_h = V_h$ . FE-Problem: Finde  $u_h \in V_h$ , sodass

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

- Wähle eine Basis  $\{\phi_i\}_{i=1}^n$  von  $V_h$ .
- Dann ist (I) äquivalent zu: Finde  $u_h \in V_h$ , sodass

$$a(u_h, \phi_i) = \ell(\phi_i) \quad \forall i = 1, \dots, n.$$

- Seien  $\bar{u}_i$  für  $i = 1, \dots, n$  die Koeffizienten von  $u_h$  bzgl.  $\{\phi_i\}_i$ .

$$\sum_{i=1}^n \bar{u}_i a(\phi_i, \phi_j) = \ell(\phi_j) \quad \forall j = 1, \dots, n.$$

- Definiere die *Steifigkeitsmatrix*  $A \in \mathbb{R}^{n \times n}$  durch  $A_{ij} = a(\phi_i, \phi_j)$ .
- Definiere den *Lastvektor*  $b \in \mathbb{R}^n$  durch  $b_i = \ell(\phi_i)$ .

### 3 Finite Elemente

- Dann löst  $\bar{u} \in \mathbb{R}^n$  das lineare Gleichungssystem  $A^T \bar{u} = b$ .

Wahl von  $\{\phi_i\}$ : Wähle die Basis  $\{\phi_i\}$  so, dass  $A$  dünnbesetzt ist.

**Beispiel 3.25** (Knotenbasis).

- Betrachte den Lagrange-Raum  $V_h^p$ .
- Uniform verteilte Lagrange-Punkte  $x_i$ .
- Knotenbasis

$$\phi_i \in V_h, \quad \phi_i(x_j) = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

## 3.5 Diskretisierungsfehler

Wie gut wird  $u$  durch  $u_h$  approximiert?

- Wie groß ist  $\|u - u_h\|$ ?
- Welche Norm verwendet man?
- Wie verhält sich  $u - u_h$  für  $h \rightarrow 0$ ?

**Satz 3.26.** *Sei eine Familie von quasi-uniformen Triangulierungen  $\mathcal{T}_h$  mit Gitterweite  $h$  gegeben. Ist  $u$   $p+1$ -regulär, d.h.  $u \in H^{p+1}$ , dann gilt für die FE-Lösung  $u_h \in V_h^p$  durch Lagrange-Elemente  $p$ -ter Ordnung*

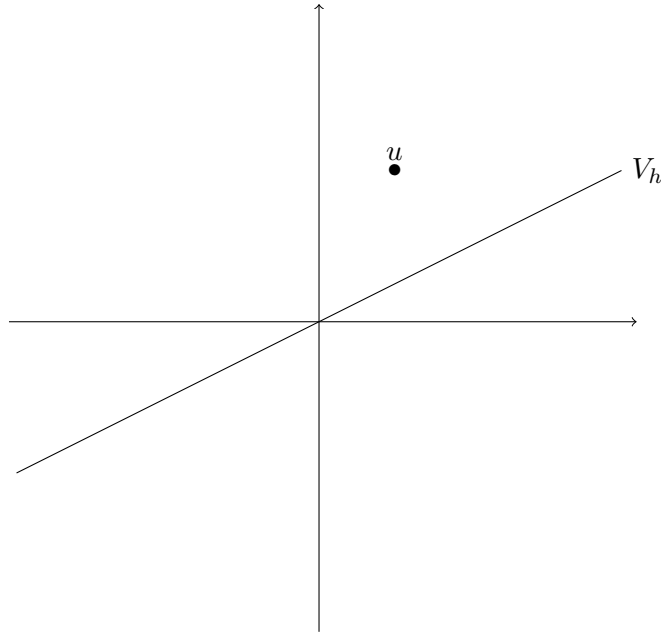
$$\|u - u_h\|_{H^1} < ch^p |u|_{p+1}.$$

Für den Beweis benötigen wir das folgende Lemma.

**Lemma 3.27** (Céa). *Die Bilinearform  $a(\cdot, \cdot)$  sei  $V$ -elliptisch mit Konstante  $\alpha$  und mit Stetigkeitskonstante  $C$ , also  $a(v, v) \geq \alpha \|v\|^2$ ,  $a(v, w) \leq C \|v\| \cdot \|w\|$ . Seien  $u$ ,  $u_h$  die Lösungen von  $a(\cdot, v) = \ell(v)$  in  $V$  bzw.  $V_h$ . Dann ist*

$$\|u - u_h\|_{H^1} \leq C \alpha^{-1} \cdot \inf_{v_h \in V_h} \|u - v_h\|_{H^1}$$

Der Ausdruck  $\inf_{v_h \in V_h} \|u - v_h\|_{H^1}$  ist der kleinstmögliche Fehler.



*Beweis des Satzes in zwei Schritten.*

1. „Bestapproximation“, Anwendung des Céa-Lemmas.
2. Konstruiere eine Funktion  $\tilde{u}_h$  mit  $\|u - \tilde{u}_h\|_{H^1} \leq ch^p |u|_{p+1}$ . Einfache Variante:  $\tilde{u}_h := Iu$ , die punktweise an den Lagrange-Knoten interpolierende Funktion.  $\square$

*Beweis des Céa-Lemmas.*

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

Wegen  $V_h \subset V$  kann subtrahiert werden:

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

Diese Eigenschaft heißt *Galerkin-Orthogonalität*. Sei  $v_h \in V_h$  eine beliebige feste Funktion.

$$a(u - u_h, v_h - u_h) = 0.$$

$v_h - u_h \in V_h$ . Wir schätzen ab:

$$\begin{aligned} a\|u - u_h\|^2 &\leq a(u - u_h, u - u_h) \\ &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) = a(u - u_h, u - v_h) \\ &\leq C\|u - u_h\|_{H^1}\|u - v_h\|_{H^1} \end{aligned}$$

Teile durch  $\|u - u_h\|_{H^1}$ .

$$\|u - u_h\|_{H^1} \leq \frac{C}{\alpha} \|u - v_h\|_{H^1} \quad \forall v_h \in V_h. \quad \square$$

### 3.6 Quasi-Uniformität

Engl.: „shape regularity“.

- Für ein Gitter  $\mathcal{T}_h$  sei  $h$  der Durchmesser des größten Elements in  $\mathcal{T}_h$ .
- Betrachte die Familie  $\mathcal{T}_{h_1}, \mathcal{T}_{h_2}, \dots$  von Gittern mit  $h_k \rightarrow 0$ .
- Seien  $\rho$  der Inkreisradius des größten Elements und  $r$  der Umkreisradius.

**Definition 3.28.** Eine Familie von Gittern  $\{\mathcal{T}_h\}$  heißt *quasi-uniform*, wenn es eine Zahl  $\kappa$  gibt, so dass

$$\forall \mathcal{T}_h \in \{\mathcal{T}_h\} \forall T \in \mathcal{T}_h: \frac{r_t}{\rho_t} < \kappa.$$

### 3.7 Approximationsfehlerabschätzungen in der $H^1$ -Norm

**Satz 3.29.** Sei  $t \geq 2$  und  $\mathcal{T}_h$  eine Familie von quasi-uniformen Triangulierungen. Dann gilt für  $Iu$ , die Interpolation von  $u$  durch stückweise Polynome vom Grad  $t - 1$

$$\|u - Iu\|_{H^m} \leq ch^{t-m} |u|_{H^t}.$$

für alle  $u \in H^t$  und  $0 \leq m \leq t$ .

Für den Beweis benötigen wir folgendes Lemma.

**Lemma 3.30** (Bramble-Hilbert). Sei  $T \subset \mathbb{R}^2$  ein konvexes Polygon und  $t \geq 2$ .  $I: H^t \rightarrow \Pi_{t-1}$  sei die Interpolation durch Lagrange-Polynome an „sinnvoll“ gewählten Stützstellen. Dann gilt

$$\|u - Iu\|_t \leq c|u|_t \quad \forall u \in H^t(T).$$

*Beweis des Satzes.* Wir teilen das Integrationsgebiet  $\Omega$  in die einzelnen Dreiecke der Triangulierung auf und summieren die Integrale über jedes Dreieck. Dann gilt  $\|u - Iu\|_m = \sum_{i=1}^N \|u - Iu\|_{H^m(T_i)}$ . Eine direkte Anwendung des Bramble-Hilbert-Lemmas funktioniert nicht, da  $c$  unspezifiziert von der Größe von  $T$  abhängt.

Ansatz: Transformiere jedes Dreieck  $T$  auf das Referenzdreieck  $T_{\text{ref}}$  und wende dort das Bramble-Hilbert-Lemma an.  $\square$

**Lemma 3.31.** Seien  $T_1, T_2$  affin äquivalent, d.h. es gebe eine bijektive affine Abbildung  $F: T_1 \rightarrow T_2$ ,  $F\hat{x} = x + B\hat{x}$ . Für  $v \in H^m(T_2)$  ist dann  $\hat{v} = v \circ F \in H^m(T_1)$  und es gibt eine Konstante  $C = C(T_1, m)$ , so dass

$$|\hat{v}|_{H^m(T_1)} \leq C \|B\|^m \cdot |\det B|^{-\frac{1}{2}} |v|_{H^m(T_2)}.$$

Damit:

### 3.7 Approximationsfehlerabschätzungen in der $H^1$ -Norm

- Für  $v \in H^m(T)$  ist dann  $\hat{v} = v \circ F \in H^m(T_{\text{ref}})$  und es gilt

$$|\hat{v}|_{H^m(T_{\text{ref}})} \leq C \|B\|^m (\det B)^{-\frac{1}{2}} |v|_{H^m(T)}.$$

Nun betrachten wir wieder  $|u - Iu|_{H^m}$ . Es gilt mit dem vorherigen Lemma und dem Bramble-Hilbert-Lemma

$$\begin{aligned} |u - Iu|_{H^m(T)} &\leq C \left\| B^{-1} \right\|^m (\det B)^{\frac{1}{2}} \|\hat{u} - I\hat{u}\|_{H^m(T_{\text{ref}})} \\ &\leq C \left\| B^{-1} \right\|^m (\det B)^{\frac{1}{2}} |\hat{u}|_{H^t(T_{\text{ref}})} \\ &\leq C \left\| B^{-1} \right\|^m (\det B)^{\frac{1}{2}} \|B\|^t (\det B)^{-\frac{1}{2}} |u|_{H^t(T)} \\ &= C \left( \left\| B^{-1} \right\| \cdot \|B\| \right)^m \|B\|^{t-m} |u|_{H^t(T)}. \end{aligned}$$

- Verbindung zur Geometrie von  $T$ :  $\|B\| \leq r_T \hat{\rho}^{-1}$ , wobei  $\hat{\rho}$  den Inkreisradius im Referenzelement und  $r_T$  den Umkreisradius im globalen Element  $T$  bezeichnen.

*Beweis.* Sei  $x \in \mathbb{R}^2$  mit  $\|x\| \leq 2\hat{\rho}$ . Daraus folgt  $\|Bx\| \leq 2r_T$ . Daraus folgt

$$\begin{aligned} \|B\| &= \sup_{\|x\|=1} \|Bx\| = \sup_{\|x\|=2\hat{\rho}} \left\| Bx (2\hat{\rho})^{-1} \right\| \\ &= \sup_{\|x\|=2\hat{\rho}} \frac{\|Bx\|}{2\hat{\rho}} \leq \frac{2r_T}{2\hat{\rho}} = \frac{r_T}{\hat{\rho}} \end{aligned}$$

Ebenso:  $\|B\|^{-1} \leq \hat{\rho} \rho_T^{-1}$ , wobei nun  $\hat{\rho}$  den Umkreisradius des Referenzelementes und  $\rho_T$  den Inkreisradius des globalen Elements  $T$ . Nun gilt die Abschätzung (durch Einsetzen)

$$|u - Iu| \leq C \left( \frac{r_T}{\hat{\rho}} \cdot \frac{\hat{\rho}}{\rho_T} \right)^m \left( \frac{r_T}{\hat{\rho}} \right)^{t-m} \cdot |u|_{H^t(T)}$$

mit den Konstanten

$$\hat{\rho} = \frac{1}{\sqrt{2}}, \quad \hat{\rho} = \frac{1}{2 + \sqrt{2}} \approx 0.29.$$

Außerdem  $\|B\| \leq 4h$ . Damit wird die obige Abschätzung für  $|u - Iu|$  zu

$$|u - Iu|_{H^m(T)} \leq c \left( \frac{r_T}{\rho_T} \right) h^{t-m} |u|_{H^t(T)}.$$

Der Ausdruck  $r_T \rho_T^{-1}$  bleibt beschränkt für  $h \rightarrow 0$ , da das Gitter als quasi-uniform angenommen wurde.  $\square$

Uns interessiert  $m = 1, t = 2$

$$\|u - Iu\|_{H^1} \leq ch |u|_{H^2}$$

und  $m = 0, t = 2$ :

$$\|u - Iu\|_{L^2} \leq ch^2 |u|_{H^2}.$$



### 3.8 Diskretisierungsfehler in der $L^2$ -Norm

Das Céa-Lemma sagt nichts über den  $L^2$ -Fehler aus. Stattdessen wenden wir den *Aubin-Nitsche-Trick* an. Als Voraussetzung müssen wir  $H^2$ -Regularität des Randwertproblems annehmen, d.h.  $u \in H^2(\Omega)$  und  $\|u\|_{H^2} \leq c \cdot \|\ell\| = \|f\|_{L^2}$ .

**Lemma 3.32** (Aubin-Nitsche-Trick). *Sei  $V_h \subset H^1$  ein FE-Raum und  $u_h \in V_h$  Lösung des Variationsproblems in  $V_h$ . Für jedes  $g \in L^2$  sei  $\phi_g \in H^1$  die eindeutige Lösung der adjungierten Gleichung*

$$a(w, \phi_g) = \langle g, w \rangle_{L^2} \quad \forall w \in H^1. \quad (\text{I})$$

Dann gilt

$$\|u - u_h\|_{L^2} \leq c \|u - u_h\|_{H^1} \cdot \sup_{g \in L^2(\Omega)} \left[ \frac{1}{\|g\|_{L^2}} \inf_{v \in V_h} \|\phi_g - v\|_{H^1} \right].$$

Was hilft das? Die  $H^1$ -Fehleruntersuchung hatte gezeigt:

$$\|u - u_h\|_{H^1} \leq ch \|u\|_{H^2} \leq ch \|f\|_{L^2}.$$

Diese Abschätzung gilt auch für (I):

$$\begin{aligned} \|\phi_g - \phi_{g,h}\|_{H^1} &\leq ch \|\phi_g\|_{H^2} \leq ch \|g\|_{L^2} \\ \inf_{v \in V_h} \|\phi_g - v\|_{H^1} &\leq \|\phi_g - \phi_{g,h}\|_{H^1} \leq ch \|g\|_{L^2}. \end{aligned}$$

Also ist der Term in  $[\dots] \leq ch$ .

*Beweis des Lemmas.* Teste die adjungierte Gleichung mit  $w = u - u_h \in H^1$ :

$$a(u - u_h, \phi_g) = \langle g, u - u_h \rangle_{L^2}.$$

Sei  $v \in V_h$  beliebig. Mit Galerkin-Orthogonalität:  $a(u - u_h, v) = 0$ . Deshalb:

$$\begin{aligned} \langle g, u - u_h \rangle_{L^2} &= a(u - u_h, \phi_g - v) \\ &\leq C \|u - u_h\|_{H^1} \|\phi_g - v\|_{H^1}. \end{aligned}$$

$v$  war beliebig, deshalb:

$$\langle g, u - u_h \rangle_{L^2} \leq C \|u - u_h\|_{H^1} \cdot \inf_{v \in V_h} \|\phi_g - v\|_{H^1}.$$

Mit der Identität

$$\forall w \in L^2: \|w\|_{L^2} = \sup_{g \in L^2} \frac{\langle w, g \rangle_{L^2}}{\|g\|_{L^2}}$$

und der Cauchy-Schwarz-Ungleichung  $\langle w, g \rangle \leq \|w\| \cdot \|g\|$  folgt

$$\|u - u_h\|_{L^2} = \sup_{g \in L^2} \frac{\langle u - u_h, g \rangle}{\|g\|_{L^2}} \leq C \|u - u_h\|_{H^1} \cdot \sup_{g \in L^2} \left[ \frac{1}{\|g\|_{L^2}} \cdot \inf_{v \in V_h} \|\phi_g - v\|_{H^1} \right]. \quad \square$$

### 3.9 Crouzeix-Raviart-Element (Poisson-Gl.)

$CR = \{v \in L^2(\Omega) \mid v|_T \text{ ist affin } \forall T \in \mathcal{T}_h, v \text{ stetig an Kantenmittelpunkten}\}$ . Wir wollen den Diskretisierungsfehler abschätzen. Dabei stellt sich das Problem, dass das Crouzeix-Raviart-Element nicht konform ist, d.h.  $CR \not\subset H^1$ . Das Céa-Lemma kann folglich nicht benutzt werden. Stattdessen verwenden wir eine Erweiterung des Céa-Lemmas, das 2. Lemma von Strang.

- Problem 1:  $a(u, v) = \ell(v)$  für alle  $v \in H^1$  mit  $a: H^1 \times H^1 \rightarrow \mathbb{R}$ . Ist  $a$  für  $V_h \in CR$  überhaupt definiert? Ersetze  $a(\cdot, \cdot)$  durch eine „sinnvolle“ Erweiterung  $a_h(\cdot, \cdot): (V + V_h) \times (V + V_h) \rightarrow \mathbb{R}$  ( $V + V_h \subset H^1 + CR$  komponentenweise). Für die Poisson-Gleichung  $-\Delta u = f$  ist  $a(v, w) = \int_{\Omega} \nabla v \cdot \nabla w$ . Erweiterung:

$$a_h(v, w) = \sum_{T \in \mathcal{T}_h} \int_T \nabla v|_T \cdot \nabla w|_T.$$

- Problem 2: Die  $H^1$ -Norm ist nicht für alle  $v_h \in V_h = CR \not\subset H^1$  definiert. Stattdessen verwenden wir *gitterabhängige Normen*, z.B.:

$$\|v_h\|_h = \sqrt{\sum_{T \in \mathcal{T}_h} \|v\|_{H^1(T)}^2}$$

Diese Norm ist nur für das konkrete Problem sinnvoll.

**Lemma 3.33** (2. Lemma von Strang). *Sei  $a_h$  elliptisch bzgl. der Norm  $\|\cdot\|_h$ , also  $a_h(v, v) \geq \alpha \|v\|_h^2$  für alle  $v \in V_h$  und  $|a(u, v)| \leq C \|v\|_h \|w\|_h$  für alle  $v \in V + V_h, w \in V_h$  mit von  $h$  unabhängigen Konstanten. Dann gilt*

$$\|u - u_h\|_h \leq c \left[ \inf_{v_h \in V_h} \|u - v_h\|_h + \sup_{w_h \in V_h} \frac{|a_h(u, w_h) - \ell_h(w_h)|}{\|w_h\|_h} \right].$$

Der erste Summand heißt Approximationsfehler; der zweite Summand heißt Konsistenzfehler.

**Bemerkung 3.34.** Bei konformen Elementen ist der zweite Summand stets = 0 und die Aussage des obigen Lemmas wird zur Aussage des Céa-Lemmas.

*Beweis.* Sei  $v_h \in V_h$  beliebig.

$$\begin{aligned} \alpha \|u_h - v_h\|_h^2 &\leq a_h(u_h - v_h, u_h - v_h) \\ &= a_h(u_h, u_h - v_h) - a_h(v_h, u_h - v_h) \\ &= \ell_h(u_h - v_h) - a_h(v_h, u_h - v_h) \\ &= a_h(u - v_h, u_h - v_h) + \ell_h(u_h - v_h) - a_h(u, u_h - v_h) \\ &\leq C \|u - v_h\|_h \|u_h - v_h\|_h + [\dots]. \end{aligned}$$

### 3 Finite Elemente

Wir teilen durch  $\|u_h - v_h\|_h$ :

$$\begin{aligned} \|u_h - v_h\|_h &\leq \frac{1}{2} \left[ C \|u - v_h\|_h + \frac{\ell_h(u_h - v_h) - a_h(u, u_h - v_h)}{\|u_h - v_h\|_h} \right] \\ &\leq \frac{1}{2} \left[ C \|u - v_h\|_h + \sup_{w_h \in V_h} \frac{|\ell_h(w_h) - a_h(u, w_h)|}{\|w_h\|_h} \right] \end{aligned}$$

Anwendung der Dreiecksungleichung:

$$\begin{aligned} \|u - u_h\|_h &\leq \|u - v_h\|_h + \|u_h - v_h\|_h \\ &\leq \left(1 + \frac{C}{\alpha}\right) \|u - v_h\|_h + \sup_{w_h \in V_h} \frac{|\ell_h(w_h) - a_h(u, w_h)|}{\alpha \|w_h\|_h}. \quad \square \end{aligned}$$

**Beispiel 3.35.** Die Poisson-Gleichung  $-\Delta u = f$  auf  $\Omega$  mit  $u = 0$  auf  $\partial\Omega$ . Schwache Form:  $a(v, w) = \ell(v)$  für alle  $v \in V = H_0^1(\Omega)$  mit

$$a(v, w) = \int_{\Omega} \nabla v \nabla w \, dx, \quad \ell(v) = \int_{\Omega} f v \, dx.$$

Crouzeix-Raviart-Elemente ( $\dim \Omega = 2$ , alle  $T \in \mathcal{T}_h$  sind Dreiecke).

$$\begin{aligned} CR &= \left\{ v \in L^2(\Omega) : v|_T \text{ affin } \forall T \in \mathcal{T}_h, v \text{ ist stetig an den Dreiecksseiten} \right\} \\ CR_0 &= \{ v \in CR : v = 0 \text{ in den Mittelpunkten der Kanten auf } \partial\Omega \}. \end{aligned}$$

Erweiterungen von  $a(\cdot, \cdot)$  und  $\ell(\cdot)$ :

$$a_h(v, w) := \sum_{T \in \mathcal{T}_h} \int_T \nabla v \nabla w \, dx, \quad \ell_h(v) := \sum_{T \in \mathcal{T}_h} \int_T f v \, dx.$$

Norm  $\|v\|_h^2 = \sum_T \|v\|_{1,T}^2$ . Wir wollen nun den Diskretisierungsfehler  $\|u - u_h\|_h$  abschätzen. Nach dem 2. Lemma von Strang ist

$$\text{Diskretisierungsfehler} \leq \text{Approximationsfehler} + \text{Konsistenzfehler}$$

1. Approximationsfehler: Da  $V_{h,0}^{(1)} \subset CR_0$  gilt<sup>2</sup>

$$\inf_{v_h \in CR_0} \|u - v_h\|_h \leq \inf_{v_h \in V_{h,0}^{(1)}} \|u - v_h\|_h \leq ch|u|_2.$$

2. Konsistenzfehler:

$$\begin{aligned} L_u(w_h) &= a_h(u, w_h) - \ell_h(w_h) = \sum_{T \in \mathcal{T}_h} \left[ \int_T \nabla u \nabla w_h \, dx - \int_T f w_h \, dx \right] \\ &= \sum_{T \in \mathcal{T}_h} \left[ \int_{\partial T} \partial_n u w_h \, dS - \int_T \Delta u w_h \, dx - \int_T f w_h \, dx \right] \\ &= \sum_{T \in \mathcal{T}_h} \int_{\partial T} \partial_n u w_h \, dS \end{aligned}$$

---

<sup>2</sup> $V_{h,0}^{(1)}$  = Lagrange-Raum erster Ordnung mit Nullrandbedingungen

### 3.9 Crouzeix-Raviart-Element (Poisson-Gl.)

da  $-\Delta u = f$  auf jedem  $T \in \mathcal{T}_h$ . Für jede Kante  $e$  des Gitters sei  $\overline{w_h(e)}$  der Mittelwert von  $w_h$  auf  $e$ .

- Auf Randkanten ist  $\overline{w_h(e)} = 0$  für alle  $w_h \in CR_0$ .
- Innere Kanten  $e$  kommen immer genau zwei mal vor. Jeweils mit dem Integranden  $\partial_n u w_h$ .
- Dabei bleibt  $\partial_n u$  betragsmäßig gleich, wechselt aber das Vorzeichen, also

$$L_u(w_h) = \sum_{T \in \mathcal{T}_h} \sum_{e \in T} \int_e \partial_n u (w_h - \overline{w_h(e)}).$$

Zu  $v \in H^2(\Omega)$  sei  $I_h v \in CR_0 \cap C(\Omega)$  die stetige, stückweise affine Funktion, welche  $v$  in den Eckpunkten der Dreiecke interpoliert. Nach Definition des Mittelwertes  $\overline{w_h(e)}$  ist

$$\int_e (w_h - \overline{w_h(e)}) = 0.$$

Deshalb darf von  $\partial_n u$  eine konstante Funktion abgezogen werden, z.B.  $\partial_n I_h u$  (konstant auf jedem  $e$ !). Dann gilt

$$L_u(w_h) = \sum_{T \in \mathcal{T}_h} \sum_{e \in \partial T} \int \partial_n (u - I_h u) (w_h - \overline{w_h(e)}) dS.$$

Cauchy-Schwarz-Ungleichung:

$$|L_u(w_h)| \leq \sum_{T \in \mathcal{T}_h} \sum_{e \in \partial T} \left[ \int_e |\nabla(u - I_h u)|^2 dS \cdot \int_e |w_h - \overline{w_h(e)}|^2 dS \right]^{\frac{1}{2}}.$$

a) Erster Faktor:

- Sei  $v \in H^2(T_{\text{ref}})$ .
- Dann folgt mit dem Bramble-Hilbert-Lemma:

$$\begin{aligned} \int_{\partial T_{\text{ref}}} |\nabla(v - I_h v)|^2 dS &\leq C \|\nabla(v - I_h v)\|_{1, T_{\text{ref}}}^2 \\ &\leq C \|v - I_h v\|_{2, T_{\text{ref}}}^2 \\ &\leq C |v|_{2, T_{\text{ref}}}^2 \end{aligned}$$

Transformationssätze (siehe Lagrange-Elemente):

$$\int_{\partial T} |\nabla(v - I_h v)|^2 dS \leq \tilde{C} h |h|_{2, T_{\text{ref}}}^2$$

Ebenso:

$$\int_e |w_h - \overline{w_h(e)}|^2 dS \leq ch |w_h|_{1, T}^2 \quad \forall w_h \in CR_0.$$

### 3 Finite Elemente

Einsetzen und erneute Anwendung der Cauchy-Schwarz-Ungleichung (in  $\mathbb{R}^n$ ):

$$\begin{aligned}
 |L_u(w_h)| &\leq \sum_{T \in \mathcal{T}_h} \left[ ch|u|_{1,T}^2 \cdot \tilde{c}h|w_h|_{1,T}^2 \right]^{\frac{1}{2}} \\
 &\leq ch \sum_{T \in \mathcal{T}_h} |u|_{2,T}|w_h|_{1,T} \\
 &\leq ch \left[ \sum_{T \in \mathcal{T}_h} |u|_{2,T}^2 \cdot \sum_{T \in \mathcal{T}_h} |w_h|_{1,T}^2 \right]^{\frac{1}{2}} \\
 &= ch|u|_{2,\Omega} \cdot |w_h|_h \\
 &\leq ch|u|_{2,\Omega} \|w_h\|_h
 \end{aligned}$$

Einsetzen in Strang:

$$\sup_{w_h \in CR_0} \frac{|L_u(w-h)|}{\|w_h\|_h} \leq \sup_{w_h \in CR_0} \frac{ch|u|_{2,\Omega} \|w_h\|_h}{\|w_h\|_h} = ch|u|_2$$

Es folgt  $\|u - u_h\|_h \leq ch|u|_2$ .

### 3.10 Die Stokes-Gleichung

Wir betrachten jetzt das *Stokes-Problem*. Stationäre Strömung einer zähen, inkompressiblen Flüssigkeit (z.B. Honig). Hauptsächlich wirken in einer Flüssigkeit oder einem Gas zwei Kräfte:

- Massenträgheit und
- Reibung der einzelnen Teilchen aneinander im Material.

Bei Gasen ist die Reibung der Teilchen sehr klein. Bei *zähen* Flüssigkeiten ist die Massenträgheit sehr klein und die Reibung sehr groß. Weiterhin sind viele Gase *kompressibel*, d.h. mit einem gewissen Kraftaufwand kann man ein gegebenes Gasvolumen verringern. Flüssigkeiten können in guter Näherung als *inkompressibel* angenommen werden. Wir suchen nun:

- Geschwindigkeitsfeld  $u: \Omega \rightarrow \mathbb{R}^d$  und
- Druckfeld  $p: \Omega \rightarrow \mathbb{R}$ ,

sodass

$$\Delta u + \nabla p = -f \text{ in } \Omega \tag{3.1}$$

$$\nabla \cdot u = 0 \text{ in } \Omega \tag{3.2}$$

$$u = 0 \text{ auf } \partial\Omega.$$

Die erste Gleichung (3.1) beschreibt eine *Impulserhaltung*, die zweite Gleichung (3.2) beschreibt eine *linearisierte Massenerhaltung*.

- Der Druck  $p$  wird nur bis auf eine additive Konstante bestimmt.
- Typischerweise: Fordere Normierung  $\int_{\Omega} p = 0$ .

### 3.10.1 Schwache Formulierung

1. Sei  $q \in L_0^2(\Omega) := \{q \in L^2: \int_{\Omega} q = 0\}$  eine Testfunktion. Teste mit  $q$  die zweite Gleichung (3.2). Aus  $\operatorname{div} u = 0$  wird

$$b(u, q) := \int_{\Omega} \operatorname{div} u \cdot q = 0 \quad \forall q \in L_0^2.$$

2. Sei  $v \in H_0^1(\Omega)^d := H_0^1(\Omega, \mathbb{R}^d) = \mathbf{H}_0^1$  eine Testfunktion für Gleichung (3.1).

$$\int_{\Omega} \Delta u \cdot v + \int_{\Omega} \nabla p \cdot v = - \int_{\Omega} f \cdot v.$$

Wir wenden Integralsätze an:

$$\begin{aligned} - \int_{\Omega} \nabla u \nabla v + \int_{\partial\Omega} \frac{\partial u}{\partial n} \cdot v - \int_{\Omega} p \cdot \operatorname{div} v + \int_{\partial\Omega} v p n &= - \int_{\Omega} f \cdot v \\ - \int_{\Omega} \nabla u \nabla v - \int_{\Omega} p \cdot \operatorname{div} v &= \int_{\Omega} f \cdot v \end{aligned}$$

Definiere  $a(v, w) = \int_{\Omega} \nabla v \nabla w$ . Neue Problemstellung: Finde  $u \in H_0^1(\Omega)^d$  und  $p \in L_0^2(\Omega)$ , sodass

$$\begin{aligned} a(u, v) + b(v, p) &= \langle f, v \rangle_{L^2} & \forall v \in H_0^1(\Omega)^d \\ b(u, q) &= 0 & \forall q \in L_0^2. \end{aligned}$$

### 3.10.2 Exkurs: Sattelpunktprobleme

Wir nehmen als Ausgangspunkt einen euklidischen Raum, etwa  $\mathbb{R}^n$ . Minimiere ein Funktional  $J: \mathbb{R}^n \rightarrow \mathbb{R}$  unter der Nebenbedingung  $g(x) = 0$  für ein  $g: \mathbb{R}^n \rightarrow \mathbb{R}$ .

**Beispiel 3.36.**  $J(x) = x_1^2 + x_2^2$  und  $g(x) = x_1 + x_2 - 2$ . Lösung  $x = (1, 1)$  durch gekonntes Auf-das-Bild-Starren.

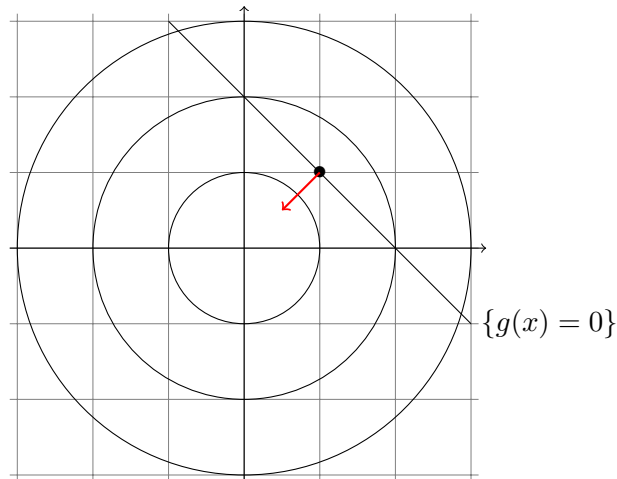


Abbildung 3.1: Skizze zur Minimierung von  $J$ , gekennzeichnet durch die kreisförmigen Höhenlinien des Paraboloids, unter der Nebenbedingung  $g(x) = 0$ .

### Lagrange-Formalismus in eukl. Räumen

Im Minimierer  $x^*$  steht der Gradient von  $J$  senkrecht auf der Nullmenge von  $g$ , d.h.  $\nabla J(x^*)$  und  $\nabla g(x^*)$  sind kollinear:

$$\nabla J = \lambda \nabla g$$

Die Zahl  $\lambda$  heißt *Lagrange-Multiplikator*.

**Lemma 3.37.** *Die Minimierer sind stationäre Punkte des Lagrange-Funktional*

$$\mathcal{L}: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}, \quad \mathcal{L}(x, \lambda) := J(x) + \lambda g(x).$$

*Beweis.* Sei  $x^*, \lambda^*$ , sodass  $\nabla \mathcal{L}(x^*, \lambda^*) = 0$ . Das heißt:

$$0 = \frac{\partial \mathcal{L}}{\partial x} = \nabla J + \lambda \nabla g.$$

$$0 = \frac{\partial \mathcal{L}}{\partial \lambda} = g(x^*). \quad \square$$

Achtung: Diese stationären Punkte von  $\mathcal{L}$  sind keine Extrema, sondern Sattelpunkte.

**Beispiel 3.38** (Fortsetzung).  $\mathcal{L}(x, \lambda) = x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 2)$ .

- Ableitung von  $\mathcal{L}$ :

$$\partial_{x_1} \mathcal{L} = 2x_1 + \lambda, \quad \partial_{x_2} \mathcal{L} = 2x_2 + \lambda, \quad \partial_{\lambda} \mathcal{L} = x_1 + x_2 - 2.$$

- Nullsetzen:

$$\begin{pmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}$$

- Matrix ist regulär, es existiert also eine eindeutige Lösung:

$$(x_1, x_2, \lambda) = (1, 1, -2)$$

- Zweite Ableitung von  $\mathcal{L}$  ist indefinit. Der stationäre Punkt ist ein *Sattelpunkt*.

### Lagrange-Formalismus in Hilberträumen

- Seien  $X$  und  $M$  zwei Hilbert-Räume.
- Sei  $a: X \times X \rightarrow \mathbb{R}$  bilinear, symmetrisch und  $\forall v \in X: a(v, v) \geq 0$ .
- Sei  $b: X \times M \rightarrow \mathbb{R}$  bilinear.
- Seien  $f \in X'$  und  $g' \in M'$ .

Gesucht wird jetzt ein Minimierer von

$$J(v) := \frac{1}{2}a(v, v) - \langle f, v \rangle$$

unter der Nebenbedingung in  $M'$

$$b(v, \cdot) - \langle g, \cdot \rangle = 0,$$

d.h.  $\forall q \in M: b(v, q) - \langle g, q \rangle = 0$ . Lagrange-Funktional hierzu:

$$\mathcal{L}(v, \mu) := \frac{1}{2}a(v, v) - \langle f, v \rangle + \underbrace{\mu (b(v, \cdot) - \langle g, \cdot \rangle)}_{=(b(v, \mu) - \langle g, \mu \rangle)}$$

für  $v \in X$  und  $\mu \in M$ . Erste Ableitung:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial v} &= a(v, \cdot) + b(\cdot, \mu) - \langle f, \cdot \rangle \\ \frac{\partial \mathcal{L}}{\partial \mu} &= b(v, \cdot) - \langle g, \cdot \rangle. \end{aligned}$$

Im Stokes-Problem:

$$\begin{aligned} a(v, w) &= \int \nabla v \nabla w \\ b(v, q) &= \int \operatorname{div} v \cdot q \\ J(v) &= \frac{1}{2} \int |\nabla v|^2 dx - \langle f, v \rangle. \end{aligned}$$

Äquivalente Formulierung des Stokes-Problems: Finde ein divergenzfreies Vektorfeld, das  $J$  minimiert. Der gesuchte Druck  $p \in L_0^2$  in der ursprünglichen Fassung des Problems ist nun der Lagrange-Multiplikator  $\mu$ . Abstrakt ist der Lagrange-Multiplikator die Kraft, welche für die Einhaltung der Nebenbedingung sorgt. Wir beschäftigen uns nun mit der Frage, unter welchen Bedingungen das Stokes-Problem in der schwachen Formulierung wohlgestellt ist. Sei  $X$  ein Banach-Raum und  $V \subset X$  der *Kern* der Nebenbedingung:

$$V := \{v \in X \mid \forall q \in M: b(v, q) = 0\}.$$



### 3 Finite Elemente

**Satz 3.39** (Brezzis Splitting Satz). *Das Sattelpunktproblem hat genau dann eine eindeutige Lösung, wenn folgende Bedingungen erfüllt sind:*

1.  $a(\cdot, \cdot)$  ist  $V$ -elliptisch, d.h.

$$\exists \alpha > 0 \forall v \in V: a(v, v) \geq \alpha \|v\|^2,$$

2. Die Bilinearform  $b(\cdot, \cdot)$  erfüllt die inf-sup-Bedingung

$$\exists \beta > 0: \inf_{\mu \in M} \sup_{v \in X} \frac{b(v, \mu)}{\|v\| \cdot \|\mu\|} > \beta.$$

Die Lösung hängt stetig von  $f$  ab.

*Beweis.* Wird weggelassen. □

#### 3.10.3 Gemischte FE-Methode

Wähle endlich-dimensionale Teilräume  $X_h \subset X$  und  $M_h \subset M$ . Diskretes Problem: Finde  $u_h \in X_h, p_h \in M_h$ , sodass

$$\begin{aligned} a(u_h, v_h) + b(v_h, p_h) &= \langle f, v_h \rangle & \forall v_h \in X_h \\ b(u_h, q_h) &= 0 & \forall q_h \in M_h \end{aligned}$$

Ist dieses Problem wohlgestellt? Bei elliptischen Problemen mit konformer Approximation einfach zu beantworten mit dem Lemma von Lax-Milgram. Hier ist es um einiges schwerer! Tatsächlich stellt der Nachweis der Wohlgestelltheit des diskreten Problems die Hauptarbeit in FE-Methoden dieser Art dar. Wir verwenden Brezzis Splitting Satz.

- Forderung 1):  $a(\cdot, \cdot)$  soll  $V_h$ -elliptisch sein, mit  $V_h = \{v \in X_h \mid \forall q_h \in M_h: b(v, q_h) = 0\}$ . Im Allgemeinen ist  $V_h \not\subset V$ .  $V_h$  kann sogar größer sein als  $V$  (muss separat gezeigt werden).  $a(\cdot, \cdot)$   $V_h$ -elliptisch folgt *nicht* automatisch aus  $V$ -Elliptizität.
- Forderung 2):  $\exists \beta > 0$ :

$$\inf_{\mu \in M} \sup_{v \in X} \frac{b(v, \mu)}{\|v\| \cdot \|\mu\|} > \beta$$

folgt ebenfalls *nicht* automatisch aus inf-sup für  $X$  und  $M$ .

Die Forderungen 1) und 2) wirken in gewisser Weise entgegen. Diese zwei Bedingungen heißen:

- Babuška-Brezzi-Bedingung,
- Brezzi-Bedingung,
- Ladyženskaya-Babuška-Brezzi-Bedingung (LBB-Bedingung).

**Korollar 3.40.** *Wenn  $X_h$  und  $M_h$  die LBB-Bedingung erfüllen, dann ist das diskrete Problem wohlgestellt.*

### 3.10.4 Diskretisierungsfehler

**Satz 3.41** (Céa-Lemma für nicht-elliptische Probleme). *Die Bedingungen des Splitting Satzes von Brezzi seien erfüllt. Die Räume  $X_h \subset X$ ,  $M_h \subset M$  erfüllen die LBB-Bedingung. Dann ist*

$$\|u - u_h\|_X + \|\eta - \eta_h\|_M \leq c \left( \inf_{v_h \in V_h} \|u - v_h\|_X + \inf_{\mu_h \in M_h} \|\eta - \mu_h\|_M \right).$$

*Beweis.*  $(u, \eta)$  löst das Variationsproblem.  $(u_h, \eta_h)$  löst das FE-Problem. Galerkin-Orthogonalität:

$$\begin{aligned} a(u - u_h, v_h) + b(v_h, \eta - \eta_h) &= 0 & \forall v_h \in X_h \\ b(u - u_h, \mu_h) &= 0 & \forall \mu_h \in M_h \end{aligned}$$

bzw.

$$\begin{aligned} a(u, v_h) + b(v_h, \eta) &= a(u_h, v_h) + b(v_h, \eta_h) \\ b(u, \mu_h) &= b(u_h, \mu_h). \end{aligned}$$

Sei  $(w_h, \lambda_h) \in X_h \times M_h$  beliebig. Dann:

$$\begin{aligned} a(u_h - w_h, v_h) + b(v_h, \eta_h - \lambda_h) &= a(u - w_h, v_h) + b(v_h, \eta - \lambda_h) \\ b(u_h - w_h, \mu_h) &= b(u - w_h, \mu). \end{aligned}$$

Definiere  $\ell \in (X \times M)'$  durch

$$\ell(v, \mu) := a(u - w_h, v) + b(v, \eta - \lambda_h) + b(u - w_h, \mu).$$

$a(\cdot, \cdot)$  und  $b(\cdot, \cdot)$  sind stetig. Deshalb

$$\|\ell\| \leq C (\|u - w_h\| + \|\eta - \lambda_h\|).$$

Aus der Wohlgestellttheit folgt

$$\|u_h - w_h\| + \|\eta_h - \lambda_h\| \leq \frac{1}{\alpha} \|\ell\| \leq \frac{C}{\alpha} (\|u - w_h\| + \|\eta - \lambda_h\|).$$

Dreiecksungleichung:

$$\begin{aligned} \|u - u_h\| + \|\eta - \eta_h\| &\leq \|u - w_h\| + \|w_h - u_h\| + \|\eta - \lambda_h\| + \|\lambda_h - \eta_h\| \\ &\leq \|u - w_h\| + \|\eta - \lambda_h\| + \frac{C}{\alpha} (\|u - w_h\| + \|\eta - \lambda_h\|) \\ &\leq \left(1 + \frac{C}{\alpha}\right) (\|u - w_h\| + \|\eta - \lambda_h\|) \end{aligned}$$

für alle  $w_h \in X_h$  und  $\lambda_h \in M_h$ . Daraus folgt die Behauptung.  $\square$

Nachteil des Satzes: Wir wissen nicht, wie klein der Geschwindigkeitsfehler  $\|u - u_h\|$  bzw. der Druckfehler  $\|\eta - \eta_h\|$  einzeln ist, d.h. wie gut  $X$  durch  $X_h$  bzw.  $M$  durch  $M_h$  approximiert wird.

### 3 Finite Elemente

**Satz 3.42.** *Es gelten die gleichen Voraussetzungen wie im vorigen Satz. Zusätzlich sei  $V_h \subset V$ , d.h. für jedes  $v_h \in V_h$  gelte*

$$\forall \mu_h \in M_h: b(v_h, \mu_h) = 0 \implies \forall \mu \in M: b(v_h, \mu) = 0.$$

Dann gilt

$$\|u - u_h\| \leq C \cdot \inf_{v_h \in X_h} \|u - v_h\|.$$

*Beweis.* Seien  $v_h \in V_h$  und  $w_h \in V_h$ .

$$\begin{aligned} a(u_h - v_h, w_h) &= a(u_h, w_h) - a(u, w_h) + a(u - v_h, w_h) \\ &= -b(w_h, \eta_h) + \langle f, w_h \rangle + b(w_h, \eta) - \langle f, w_h \rangle + a(u - v_h, w_h) \\ &= b(w_h, \eta - \eta_h) + a(u - v_h, w_h). \end{aligned}$$

Es gilt  $b(w_h, \eta - \eta_h) = 0$ , denn

- $w_h \in V_h$ , d.h.  $b(w_h, \mu_h) = 0$  für alle  $\mu_h \in M_h$ .
- Nach Annahme  $V_h \subset V \implies b(w_h, \mu) = 0$  für alle  $\mu \in M$ , also  $b(w_h, \eta - \eta_h) = 0$ , da  $\eta - \eta_h \in M$ .

Wähle  $w_h := u_h - v_h \in V_h \subset V$ . Dann folgt

$$\begin{aligned} \alpha \|u_h - v_h\|^2 &\leq a(u_h - v_h, u_h - v_h) \\ &= a(u - v_h, u_h - v_h) \\ &\leq C \|u - v_h\| \cdot \|u_h - v_h\| \\ \implies \|u_h - v_h\| &\leq \frac{C}{\alpha} \|u - v_h\|. \end{aligned}$$

Dreiecksungleichung:

$$\|u - u_h\| \leq \|u - v_h\| + \|v_h - u_h\| \leq \left(1 + \frac{C}{\alpha}\right) \|u - v_h\|. \quad \square$$

Soweit zur Diskretisierungsfehleranalyse. Interessant wird jetzt die Konstruktion von passenden FE-Räumen. Wie sehen diese Räume  $X_h, M_h$  aus?

**Beispiel 3.43.** Motivation: Die im Stokes-Problem auftretenden Bilinearformen sind

$$a(u, v) = \int \nabla u \nabla v, \quad b(u, q) = \int \operatorname{div} u \cdot q.$$

Plausible Wahl:

$$\begin{aligned} X_h &:= P_1 = \{v \in C \text{ stückweise affin}\} \subset H^1, \\ M_h &:= P_0 = \{q \text{ stückweise konstant}\} \subset L^2. \end{aligned}$$

Aber: Erfüllt nicht die inf-sup-Bedingung:

$$\exists q \in M_h \setminus \{0\} \forall v \in X_h: \int \operatorname{div} v \cdot q = 0.$$

*Beweisskizze.* Sei  $\Omega = (0, 1)^2$  und sei auf  $\Omega$  ein uniformes Rechteckgitter mit  $N \in \mathbb{N}$  und  $h = N^{-1}$  gegeben. Sei ferner  $a_{ij}$  der Punkt mit Koordinate  $(ih, jh)$  und  $T_{ij}$  die rechts-obere Gitterzelle bezüglich  $a_{ij}$ .

- Für  $q \in M_h$  sei  $q_{i+\frac{1}{2}, j+\frac{1}{2}}$  der Wert von  $q$  auf  $T_{ij}$ .
- Für  $\mathbf{v} \in X_h$  (Geschwindigkeitsfeld (Vektor)) seien  $(u_{ij}, v_{ij})$  die Werte der zwei Komponenten in  $a_{ij}$ .
- $q$  ist zellenweise konstant, daher gilt (mit dem Satz von Gauß)

$$\begin{aligned} \int_{T_{ij}} \operatorname{div} \mathbf{v} \cdot q &= q_{i+\frac{1}{2}, j+\frac{1}{2}} \int_{T_{ij}} \operatorname{div} \mathbf{v} = q_{i+\frac{1}{2}, j+\frac{1}{2}} \int_{\partial T_{ij}} \langle \mathbf{v}, \mathbf{n} \rangle dS \\ &= \frac{1}{2} h q_{i+\frac{1}{2}, j+\frac{1}{2}} (u_{i+1, j} - u_{i, j} + u_{i+1, j+1} - u_{i, j+1} + v_{i+1, j+1} - v_{i+1, j} + v_{i, j+1} - v_{i+1, j}) \end{aligned}$$

- Addiere über alle Elemente:

$$\int_{\Omega} \operatorname{div} \mathbf{v} q = [\dots] = -h^2 \sum_{0 < i, j < N} (u_{i, j} (\partial_1 q)_{i, j} + v_{i, j} (\partial_2 q)_{i, j})$$

mit den Differenzenoperatoren

$$\begin{aligned} (\partial_1 q)_{i, j} &= \frac{1}{2h} (q_{i+\frac{1}{2}, j+\frac{1}{2}} + q_{i+\frac{1}{2}, j-\frac{1}{2}} - q_{i-\frac{1}{2}, j+\frac{1}{2}} - q_{i-\frac{1}{2}, j-\frac{1}{2}}) \\ (\partial_2 q)_{i, j} &= \frac{1}{2h} (q_{i+\frac{1}{2}, j+\frac{1}{2}} + q_{i-\frac{1}{2}, j+\frac{1}{2}} - q_{i+\frac{1}{2}, j-\frac{1}{2}} - q_{i-\frac{1}{2}, j-\frac{1}{2}}) \end{aligned}$$

Dabei haben wir eine partielle Integratino verwendet:

$$\int \operatorname{div} \mathbf{v} \cdot q = - \int_{\Omega} \langle \mathbf{v}, \nabla q \rangle = - \int_{\Omega} u \cdot \frac{\partial q}{\partial x} + v \cdot \frac{\partial q}{\partial y}$$

- Die  $\partial_1 q, \partial_2 q$  sind Null, wenn

$$q_{i+\frac{1}{2}, j+\frac{1}{2}} = q_{i-\frac{1}{2}, j-\frac{1}{2}}, \quad q_{i-\frac{1}{2}, j+\frac{1}{2}} = q_{i+\frac{1}{2}, j-\frac{1}{2}}.$$

- Dies gilt, wenn alternativ
  1.  $q$  konstant ist (weniger interessant) oder
  2.  $q$  schachbrettartig die Werte  $+1$  und  $-1$  annimmt.
- Solch ein  $q$  erfüllt also  $\int \operatorname{div} \mathbf{v} \cdot q = 0$  für alle  $\mathbf{v} \in X_h$ . Daraus folgt

$$\inf_{\mu \in M_h} \sup_{v \in X_h} \frac{\int \operatorname{div} v \cdot \mu}{\|v\| \cdot \|\mu\|} = 0$$

Man sieht in der Praxis teilweise Schachbrettmuster. Dies führte auf die Bezeichnung *Schachbrettinstabilität*. □

### 3 Finite Elemente

**Beispiel 3.44** (Das Taylor-Hood-Element). Wir verwenden jetzt

$$X_h := P_2 = \{v \in C \cap H_D^1 : v|_T \in \Pi_2\} \subset (H_D^1)^d,$$

$$M_h := P_1 \cap L_0^2 = \{q \in C \text{ st\"uckweise affin}\} \cap L_0^2.$$

Das Taylor-Hood-Element erf\"ullt die inf-sup-Bedingung

$$\inf_q \sup_v \frac{\int \operatorname{div} v \cdot q}{|v| \|q\|_{\mathcal{T}}} \geq \beta > 0$$

mit der gitterabh\"angigen Norm

$$\|\mu\|_{\mathcal{T}} := \sqrt{\sum_{T \in \mathcal{T}} h_T^2 \|\nabla \mu\|_{0,T}^2}$$

mit  $h_T$  dem Durchmesser von  $T$ .

- Im Nenner steht  $|v|_1$  statt  $\|v\|_1$ , aber diese sind auf  $H_D^1$  \\"aquivalent (Friedrich-Ungleichung).

*Beweisskizze.*

- Die gitterabh\"angige Norm  $\|\cdot\|_{\mathcal{T}}$  ist tats\"achlich eine Norm.
- Sei  $q \in P_1 \cap L_0^2$  beliebig. Das Gitter bestehe nur aus Dreiecken.
- Sei  $\mathcal{E}$  die Menge der inneren Kanten.
- F\"ur  $E \in \mathcal{E}$  sei  $t_E$  der Einheitstangentialvektor. (Orientierung vernachl\"assigt).
- Definiere  $u_{\mathcal{T}} \in P_2^2 \cap H_D^1$  durch

$$u_{\mathcal{T}} = \sum_{E \in \mathcal{E}} \psi_E h^2 \frac{\partial q}{\partial t_E} t_E$$

wobei  $\psi_E$  die  $P_2$ -Basisfunktion zu  $E$  ist.

- Rechnen (wenn das Gitter quasi-uniform ist):

$$\int_{\Omega} q \operatorname{div} u_{\mathcal{T}} \geq c \|q\|_{\mathcal{T}}^2$$

- $|u_{\mathcal{T}}|_1^2 \leq c \|q\|_{\mathcal{T}}^2$ . Damit

$$\sup_{u \in X_h} \frac{\int q \operatorname{div} u}{|u|_1} \geq \frac{\int q \operatorname{div} u_{\mathcal{T}}}{|u_{\mathcal{T}}|_1} \geq c \frac{\|q\|_{\mathcal{T}}^2}{|u_{\mathcal{T}}|_1} \geq c \frac{\|q\|_{\mathcal{T}} |u_{\mathcal{T}}|_1}{|u_{\mathcal{T}}|_1} = c \|q\|_{\mathcal{T}} \quad \square$$

Das Taylor-Hood-Element funktioniert.

# 4 Konstruktion von adaptiven Gittern

## 4.1 Residuenfehlerschätzer

### 4.1.1 Effizienz des Residuenschätzers

- Der Residuenschätzer beschränkt den Fehler auch von unten.
- Und das sogar *lokal*, d.h. auf jedem Element separat.

Ab jetzt sei  $V_h$  der Lagrange-Raum zweiter Ordnung.

- Sei  $f_h$  die  $L^2$ -Projektion von  $f$  in  $V_h$ .
- Sei  $\omega_T$  die Menge aller Dreiecke, die mit  $T$  eine gemeinsame Kante haben.
- Sei  $\omega_e$  die Menge aller Dreiecke, die  $e$  als Kante haben.

**Satz 4.1** (Braess [1], Satz III.8.1). *Sei  $\mathcal{T}$  eine quasiuniforme Triangulierung mit Regularitätsparameter  $\kappa$ . Dann gibt es eine Konstante  $c = c(\Omega, \kappa)$  so dass*

$$\eta_{T,R} \leq c \left[ \|u - u_h\|_{1,\omega_T}^2 + \sum_{T' \subset \omega_T} h_T^2 \|f - f_h\|_{0,T'}^2 \right]^{\frac{1}{2}}.$$

- Der Ausdruck  $\|f - f_h\|_{0,T}$  heißt *Datenoszillation*. Er misst, wie gut sich  $f$  mit dem gegebenen FE-Raum bzw. Gitter darstellen lässt.

Für den Beweis brauchen wir eine besondere Funktion („Abschneidefunktion“).

(Für den kompletten Beweis brauchen wir sogar zwei solche Funktionen: Neben der für Elemente auch noch eine zweite für Kanten. Diese kommt aber nur in dem Teil des Beweises vor, den wir der Einfachheit halber weglassen.)

- Die Elementblasenfunktion  $\psi_T = 27\lambda_1\lambda_2\lambda_3$  ( $\lambda_i$  die baryzentrischen Koordinaten auf  $T$ )
- kubisches Polynom
- $\psi_T = 0$  auf  $\partial T$  und außerhalb von  $T$
- $\inf \psi_T = 0, \sup \psi_T = 1$

#### 4 Konstruktion von adaptiven Gittern

**Hilfssatz 4.1.** Sei  $\mathcal{T}$  quasiuniform mit Parameter  $\kappa$ . Dann gilt

- a)  $\|\psi_T v\|_{0,T} \leq \|v\|_{0,T} \quad \forall v \in L^2(T)$
- b)  $\|\psi_T^{\frac{1}{2}} p\|_{0,T} \geq c \|p\|_{0,T} \quad \forall p \in \Pi_2$
- c)  $\|\nabla(\psi_T p)\|_{0,T} \leq c h_T^{-1} \|\psi_T p\|_{0,T} \quad \forall p \in \Pi_2.$

*Beweis von Satz 4.1.* Die Dreiecksungleichung liefert

$$\|\Delta u_h + f\|_{0,T} \leq \|\Delta u_h + f_h\|_{0,T} + \|f - f_h\|_{0,T}.$$

Wir führen deshalb das reduzierte Elementresiduum ein:

$$R_{T,\text{red}} := (\Delta u_h + f_h)|_T \in \Pi_2(T)$$

Außerdem definieren wir

$$w_T := \psi_T \cdot R_{T,\text{red}}.$$

Rechnen:

$$\begin{aligned} c^{-1} \|R_{T,\text{red}}\|_{0,T}^2 &\leq \|\psi_T^{\frac{1}{2}} R_{T,\text{red}}\|_{0,T}^2 && \text{(Hilfssatz 4.1 b)} \\ &= (\psi_T^{\frac{1}{2}} R_{T,\text{red}}, \psi_T^{\frac{1}{2}} R_{T,\text{red}})_{0,T} \\ &= (R_{T,\text{red}}, \psi_T R_{T,\text{red}})_{0,T} \\ &= (R_{T,\text{red}}, w_T)_{0,T} && \text{(Def. von } w_T) \\ &= (\Delta u_h + f_h - f + f, w_T)_{0,T} \\ &= (R_T, w_T)_{0,T} + (f_h - f, w_T)_{0,T} \\ &= (R_T, w_T)_{0,T} + \underbrace{\sum_{e \in \Gamma_h} (R_e, w_T)_{0,e}}_{= 0, \text{ da } w_T = 0 \text{ auf } \Gamma_h} + (f_h - f, w_T)_{0,T} \\ &= l(w_T) + (f_h - f, w_T)_{0,T} \\ &\quad \text{(Mit der Linearform } l(\cdot) \text{ aus dem Beweis der Zuverlässigkeit)} \\ &= (\nabla(u - u_h), \nabla w_T)_{0,T} + (f_h - f, w_T)_{0,T} \\ &\quad \text{(Nach Def. von } l(\cdot) \text{ und } \text{supp } w_T = T) \\ &\leq |u - u_h|_{1,T} \cdot |w_T|_{1,T} + \|f - f_h\|_{0,T} \cdot \|w_T\|_{0,T} \\ &\leq \|u - u_h\|_{1,T} \cdot |w_T|_{1,T} + \|f - f_h\|_{0,T} \cdot \|w_T\|_{0,T} \end{aligned}$$

Wir wollen durch  $\|R_{T,\text{red}}\|_{0,T}$  kürzen. In der Tat gilt:

•

$$\|w_T\|_{0,T} = \|\psi_T R_{T,\text{red}}\|_{0,T} \leq \|R_{T,\text{red}}\|_{0,T} \quad \text{(Hilfssatz 4.1 a)}$$

•

$$\begin{aligned}
 |w_T|_{1,T} &= \|\nabla(\psi_T R_{T,\text{red}})\|_{0,T} \\
 &\leq ch_T^{-1} \|\psi_T R_{T,\text{red}}\|_{0,T} && \text{(Hilfssatz 4.1 c)} \\
 &\leq ch_T^{-1} \|R_{T,\text{red}}\|_{0,T} && \text{(Hilfssatz 4.1 a)}
 \end{aligned}$$

Damit

$$c^{-1} \|R_{T,\text{red}}\|_{0,T}^2 \leq \|u - u_h\|_{1,T} \cdot ch_T^{-1} \|R_{T,\text{red}}\|_{0,T} + \|f - f_h\|_{0,T} \cdot \|R_{T,\text{red}}\|_{0,T}$$

bzw.

$$\|R_{T,\text{red}}\|_{0,T} \leq c \left[ h_T^{-1} \|u - u_h\|_{1,T} + \|f - f_h\|_{0,T} \right].$$

Mit

$$\|R_T\|_{0,T} = \|\Delta u_h + f\|_{0,T} \leq \|\Delta u_h + f_h\|_{0,T} + \|f - f_h\|_{0,T}$$

schließt man

$$h_T \|R_T\|_{0,T} - h_T \|f - f_h\|_{0,T} \leq h_T \|R_{T,\text{red}}\|_{0,T} \leq c \left[ \|u - u_h\|_{1,T} + h_T \|f - f_h\|_{0,T} \right],$$

bzw.

$$h_T \|R_T\|_{0,T} \leq c \left[ \|u - u_h\|_{1,T} + h_T \|f - f_h\|_{0,T} \right]$$

mit erneut geänderter Konstante.

Ähnlich zeigt man (Details bei Braess):

$$h_e^{\frac{1}{2}} \|R_e\|_{0,e} \leq c |u - u_h|_{1,\omega_e} + \sum_{T' \subset \omega_e} h_{T'} \|f - f_h\|_{0,T'}$$

Zusammen erhält man

$$\begin{aligned}
 \eta_{T,R}^2 &:= h_T^2 \|R_T\|_{0,T}^2 + \frac{1}{2} \sum_{e \in \partial T} h_e \|R_e\|_{0,e}^2 \\
 &\leq c \left[ \left[ \|u - u_h\|_{1,T} + h_T \|f - f_h\|_{0,T} \right]^2 + \frac{1}{2} \sum_{e \in \partial T} \left[ \|u - u_h\|_{1,\omega_e} + \sum_{T' \in \omega_e} h_{T'} \|f - f_h\|_{0,T'} \right]^2 \right] \\
 &\leq c \left[ \|u - u_h\|_{1,T}^2 + h_T^2 \|f - f_h\|_{0,T}^2 + \frac{1}{2} \sum_{e \in \partial T} \left[ \|u - u_h\|_{1,\omega_e}^2 + \sum_{T' \in \omega_e} h_{T'}^2 \|f - f_h\|_{0,T'}^2 \right] \right] \\
 &\quad \text{(wegen } (a+b)^2 \leq 2(a^2 + b^2)\text{)} \\
 &\leq c \left[ \|u - u_h\|_{1,T}^2 + h_T \|f - f_h\|_{0,T}^2 + \sum_{T' \in \omega_T} \left[ \|u - u_h\|_{1,T'}^2 + h_{T'}^2 \|f - f_h\|_{0,T'}^2 \right] \right] \\
 &\quad \text{(da } \omega_T = \bigcup_{e \in \partial T} \omega_e\text{; Den Faktor } \frac{1}{2}\text{ kann man einfach durch 1 ersetzen.)} \\
 &\leq c \left[ \|u - u_h\|_{1,\omega_T}^2 + \sum_{T' \in \omega_T} h_{T'}^2 \|f - f_h\|_{0,T'}^2 \right] \quad \square
 \end{aligned}$$





## 5 Finite-Volumen-Verfahren

- Sei  $\Omega$  ein Gebiet in  $\mathbb{R}^d$  und sei  $[0, t_{\text{end}}]$  ein Zeitintervall.
- Gesucht wird Zeitevolution eines Skalarfeldes  $c: \Omega \times [0, t_{\text{end}}] \rightarrow \mathbb{R}$  unter dem Einfluss eines bekannten Geschwindigkeitsfeldes  $v: \Omega \rightarrow \mathbb{R}^d$ .
- Einfachster Fall: lineare Transportgleichung:

$$\frac{\partial c}{\partial t} + \operatorname{div}(cv) = 0.$$

mit Anfangsbedingungen  $c(x, 0) = c_0(x)$ .

Der Rand  $\partial\Omega$  zerfällt in zwei Teile:

$$\begin{aligned} \Gamma_{\text{in}} &:= \{x \in \partial\Omega: \langle v, n \rangle < 0\} \text{ Einflussrand,} \\ \Gamma_{\text{out}} &:= \partial\Omega \setminus \Gamma_{\text{in}} \text{ Ausflussrand} \end{aligned}$$

Randbedingungen nur auf  $\Gamma_{\text{in}}$ : Dirichlet ( $c$  ist gegeben) oder Neumann ( $c \cdot v$  ist gegeben). Zerlege  $\Omega$  in konvexe, abgeschlossene Polytope  $T$ , sodass

1.  $\bar{\Omega} = \bigcup_{T \in \mathcal{T}} T$ .
2.  $\operatorname{int} T_i \cap \operatorname{int} T_j = \emptyset$  für alle  $T_i, T_j$ .

Nicht-konforme Gitter sind erlaubt. Die  $T$  heißen *Kontrollvolumen*. Wir wenden den Satz von Gauß auf jedem  $T_i$  an:

$$\frac{\partial}{\partial t} \int_{T_i} c \, dx + \int_{\partial T_i} (cv) \cdot n \, dS = 0$$

Zeitgitter:  $0 = t_0 < t_1 < t_2 < \dots < t_{\text{end}}$ . Explizites Euler-Verfahren:

$$\int_{T_i} c(x, t_{k+1}) \, dx - \int_{T_i} c(x, t_k) \, dx + \Delta t_k \int_{\partial T_i} (cv)n \, dS = 0$$

für alle  $T_i$ . Approximiere  $c$  durch stückweise konstante Funktion  $C$ , das Integralmittel von  $c$  über  $T_i$ :

$$C_i^k := \frac{1}{|T_i|} \int_{T_i} c(x, t_k) \, dx.$$

Einsetzen:

$$C_i^{k+1}|T_i| - C_i^k|T_i| + \Delta t_k \int_{\partial T_i} (cv)n \, dS = 0.$$

## 5 Finite-Volumen-Verfahren

- Sei  $\gamma_{ij}$  der Schnitt von  $T_i$  mit  $T_j$ .
- Sei  $n_{ij}$  die Einheitsnormale auf  $\gamma_{ij}$ .
- Sei  $v_{ij}$  der Wert von  $v$  im Mittelpunkt von  $\gamma_{ij}$ .

Approximiere den Fluss  $\int_{\gamma_{ij}} (cv)n \, dS$  durch

- $C_i^k v_{ij} n_{ij}$ , falls  $V_{ij} n_{ij} \geq 0$  (Masse strömt aus  $T_i$  heraus).
- $C_j^k v_{ij} n_{ij}$ , falls  $v_{ij} n_{ij} < 0$  (Masse strömt nach  $T_i$  hinein).
- Randterme ...

Formal: diskrete Flussfunktion:

$$\Phi(C_i^k, C_j^k, v_{ij} n_{ij}) := C_i^k \max(0, v_{ij} n_{ij}) - C_j^k \max(0, -v_{ij} n_{ij}).$$

Damit ist

$$C_i^{k+1} |T_i| - C_i^k |T_i| + \Delta t_k \sum_{\gamma_{ij}} |\gamma_{ij}| \Phi(C_i^k, C_j^k, v_{ij} n_{ij}).$$

Auflösen nach  $C_i^{k+1}$ :

$$C_i^{k+1} = C_i^k \left( 1 - \Delta t_k \sum_{\gamma_{ij}} \frac{|\gamma_{ij}|}{|T_i|} \max(0, v_{ij} n_{ij}) \right) + \Delta t_k \sum_{\gamma_{ij}} C_j^k \frac{|\gamma_{ij}|}{|T_i|} \max(0, -v_{ij} n_{ij}).$$

Das Verfahren ist explizit, wir müssen also keine Gleichungssysteme zu lösen. Verfahren nur dann stabil, wenn  $\Delta t_k$  klein genug ist.

# Literatur

- [1] D. Braess. *Finite Elemente*. Springer Verlag, 2013.