Numerik partieller Differentialgleichungen

für Ingenieure

Michael Hinze

Vorlesungsskript zur Vorlesung **Numerische Mathematik II für Ingenieure**,
gehalten im WS 2000/2001 an der TU Berlin, Stand 19.2.01

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Modellierung

Nach Pinnau [18] geschieht die Modellierung eines realen Problems in Form einer Kaskade, bei der jede Stufe mehrere Schritte umfasst. Jede Stufe der Kaskade steht dabei für die Komplexität des gegenwärtigen Modells, wobei die Kaskade von einfachen zu komplexen Modellen führen und jede Stufe wiederum zum besseren Verständnis des realen Problems beitragen sollte. In jeder Stufe der Kaskade sind die folgenden Schritte auszuführen.

1. Verstehe das Problem
2. Diskutiere mögliche mathematische Methoden
3. Formuliere ein mathematisches Modell
4. Löse das mathematische Modell entweder analytisch oder numerisch
5. Interpretiere die Lösung
6. Validiere die Lösung durch Vergleich mit dem Ausgangsproblem
7. Verfeinere eventuell das mathematische Modell $\hookrightarrow$ nächste Stufe der Kaskade und gehe nach 2.
8. Präsentiere die Ergebnisse

Im Folgenden werden Facetten dieses Vorgehens am Beispiel der Modellierung von Stoff- oder Temperaturtransport erläutert. Es stellt sich heraus, daß die mathematischen Modelle i.A. nicht analytisch gelöst werden können und deshalb numerische Methoden zur Lösung verwendet werden sollten. Numerische Methoden zur näherungsweisen Lösung von Transportphänomenen stellen daher den Schwerpunkt dieses Skriptums dar.

1.1 Diffusion und Transport

Im Folgenden wird die Modellbildung für

- Ausbreitung eines Gases in der Atmosphäre,
- Ausbreitung einer Substanz im Lösungsmittel und/oder
- Ausbreitung von Temperatur

durchgeführt. Grundlage für das mathematische Modell sind hier Bilanzen. Der Ausgangspunkt ist dabei das Prinzip der Massenerhaltung. In einem Volumenelement gilt:
Die zeitliche Änderung der Masse einer Substanz ergibt sich aus

i) der in Quellen erzeugten Masse,

ii) der in Senken vernichteten Masse,

iii) der durch die Oberfläche des Volumenelements eintretenden Masse,

iv) der durch die Oberfläche des Volumenelements austretenden Masse.

Bezeichne

$v(t,x)$	$[\frac{m}{s}]$	die Geschwindigkeit des Stoffes am Ort x zur Zeit t,
$\rho(t,x)$	$[\frac{kg}{m^3}]$	dessen Dichte .

Dann heißt

$$w(t,x) \;:=\; v(t,x)\;\rho(t,x)\left[\frac{kg}{m^2 s}\;\right], w = \begin{bmatrix} w^1 \\ w^2 \\ w^3 \end{bmatrix}$$

Teilchenstromdichte.
Zur Beschreibung von iii) und iv) betrachte ein Volumenelement um x zur Zeit t. Das Volumen des

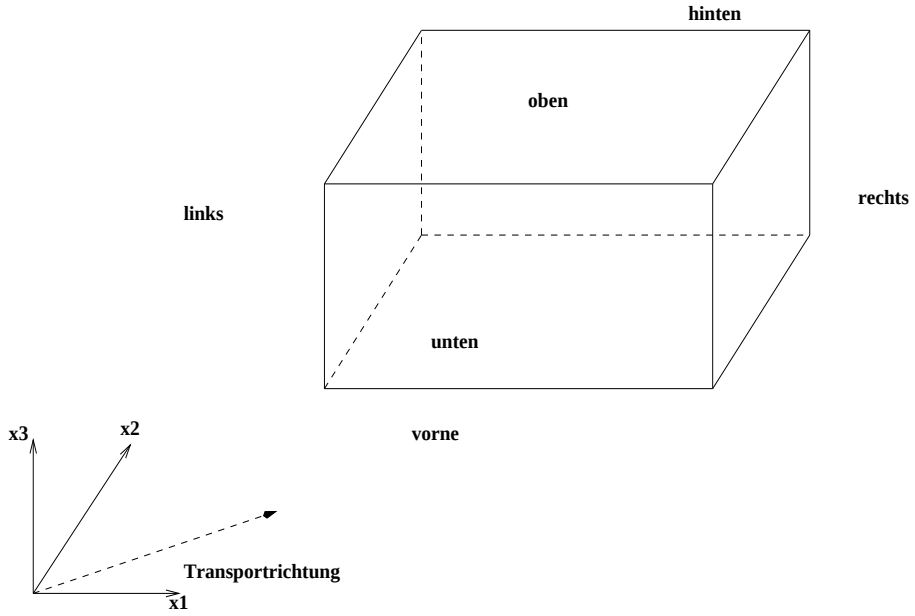


Abbildung 1: Volumenelement, F_1^1 links, F_1^2 rechts, F_2^1 vorne, F_2^2 hinten, F_3^1 unten, F_3^2 oben

Elements ist gegeben durch

$$\Delta V = \Delta x_1 \Delta x_2 \Delta x_3;$$

während des Zeitraums Δt tritt durch

F_1^1	die Masse	$w_1(t, x_1 - \frac{1}{2}\Delta x_1, x_2, x_3)\Delta t \lvert F_1^1\rvert$
F_2^1	die Masse	$w_2(t, x_1, x_2 - \frac{1}{2}\Delta x_2, x_3)\Delta t \lvert F_2^1\rvert$
F_3^1	die Masse	$w_3(t, x_1, x_2, x_3 - \frac{1}{2}\Delta x_3)\Delta t \lvert F_3^1\rvert$

in das Volumenelement ein, analog durch die gegenüberliegenden Flächen Masse aus. Für die Massenänderung im Volumenelement gilt demnach

$$\begin{aligned} \Delta m^o &= \Delta m_1 + \Delta m_2 + \Delta m_3, \\ \Delta m_i &= [w_i(t,\cdot, x_i - \tfrac{1}{2}\Delta x_i, \cdot) - w_i(t,\cdot, x_i + \tfrac{1}{2}\Delta x_i, \cdot)]\Delta t \left\{|F_i^1| + |F_i^2|\right\}/2, \end{aligned}$$

in erster Näherung also

$$\begin{aligned} \Delta m_i &\doteq -\partial_{x_i} w_i(t,x)\Delta t \Delta V, \\ \Delta m^o &\doteq -\mathrm{div}_x w(t,x)\Delta t \Delta V. \end{aligned}$$

Zur Modellierung enthalte das Volumenelement jetzt zusätzlich Quellen und/oder Senken mit Quelldichten

$$q_k[\frac{kg}{m^3 s}], \quad k = 1, \ldots, K.$$

Die im Ort gemittelte Quelldichte ist dann

$$q := \frac{1}{K} \sum_{j=1}^{K} q_j.$$

Während des Zeitraums Δt wird demnach am Ort x die Masse

$$\int_t^{t+\Delta t} q(s,x)ds \doteq q(t,x)\Delta t$$

erzeugt, im Volumenelement in erster Näherung also die Masse

$$\Delta m^t = q(t,x)\Delta t \Delta V.$$

Zusammengefaßt ergibt die Massenbilanz

$$[-\mathrm{div}_x w(t,x) + q(t,x)]\Delta t \Delta V = \Delta m^o + \Delta m^q =: \Delta m. \tag{1}$$

Andererseits ist

$$\text{Masse} = \text{Dichte} \times \text{Volumen} \; (\; \text{Energie} = \text{Temperatur} \times \text{Volumen}),$$

oder

$$\begin{array}{rcl} \Delta m & = & \frac{1}{\Delta t}\,[\rho(t+\Delta t,x) - \delta(t,x)]\rho t \Delta V \\ & \doteq & \rho_t(t,x)\Delta t \Delta V. \end{array} \tag{2}$$

Insgesamt ergibt sich aus (1) und (2)

$$\rho_t(t,x) = -\mathrm{div}_x w(t,x) + q(t,x), \tag{3}$$

und damit ein Zusammenhang zwischen zeitlicher Dichteänderung und Teilchenstromdichte. Es fehlt noch ein Zusammenhang zwischen Teilchenstromdichte und Dichte. Dazu zerlege zunächst die Teilchenstromdichte gemäß

$$w(t,x) = w_D(t,x) + w_T(t,x)$$

in einen Diffusionsstromdichteanteil w_D und einen Transportstromdichteanteil w_T. Für Diffusionsströme postuliere

[F1] Diffusionsströme fließen in Richtung des größten Konzentrationsgefälles und sind in ihrer Stärke proportional zur Konzentrationsdifferenz.

Dieses ist das 1. Fick'sche Gesetz. In Formeln

$$w_D(t,x) = -K(t,x)\nabla_x \rho(t,x),$$

so daß mit (3)

$$\rho_t(t,x) = \mathrm{div}_x(K(t,x)\nabla_x \rho(t,x)) - \mathrm{div}_x w_T(t,x) + q(t,x)$$

erhalten wird. Die Matrix $K = K(t,x)$ heißt Diffusionskoeffizient. Mit

$$w_T(t,x) = v_T(t,x)\rho(t,x) \tag{4}$$

ergibt sich

$$\rho_t(t,x) = \mathrm{div}_x(K(t,x)\nabla_x \rho(t,x)) - v_T(t,x)\nabla_x \rho(t,x) - \rho(t,x)\mathrm{div}_x v_T(t,x) + q(t,x). \tag{5}$$

Zur vollständigen Beschreibung fehlen noch die

- Vorgabe der Dichte zu Beginn des Prozesses,

- Beschreibung des Verhaltens der Dichte auf der Berandung des Behältnisses, in welchem der Prozeß stattfindet.

Bezeichnet Ω das Behältnis, so ist (5) zu ergänzen nach

$$(6)\qquad \begin{cases} \rho_t(t,x) &= \mathrm{div}_x(K(t,x)\nabla_x\rho(t,x)) - v_T(t,x)\nabla_x\rho(t,x) - \\ & \quad \rho(t,x)\mathrm{div}_x v_T(t,x) + q(t,x) & \text{in } \Omega^T \\ RW(\rho)(t,x) &= r(t,x) & \text{in } (\partial\Omega)^T \\ \rho(0,x) &= \rho_0(x) & \text{in } \Omega, \end{cases}$$

wobei $\Omega^T := (0,T)\times\Omega$, $\partial\Omega^T$ analog. Im Folgenden werden Effekte vernachlässigt, welche durch Eigenbewegung des Stoffes induziert werden.

1.2 Spezialfälle und Modellierung des Randverhaltens

Wärmeleitungs- oder Diffusionsgleichung
Dabei ist

$$K(t,x) \equiv K \equiv 1, v_T \equiv 0, q \equiv 0.$$

Es ergibt sich

$$(7)\qquad \begin{cases} \rho_t(t,x) &= \Delta_x\rho(t,x) & \text{in } \Omega^T \\ RW(\rho)(t,x) &= r(t,x) & \text{in } \partial\Omega^T \\ \rho(0,x) &= \rho_0(x) & \text{in } \Omega. \end{cases}$$

Temperatur- oder Konzentrationsvorgabe auf dem Rand vorgeschrieben:

$$(8)\qquad RW(\rho)(t,x) = \rho(t,x) = r_1(t,x), \quad (t,x)\in(\partial\Omega)^T.$$

Temperatur- bzw. Konzentrationsverlauf durch die Berandung ist vorgeschrieben:

$$(9)\qquad RW(\rho)(T,x) = \partial_\eta\rho(t,x) = r_2(t,x), \quad (t,x)\in(\partial\Omega)^T.$$

Temperatur- bzw. Konzentrationsverlauf durch die Berandung ist proportional zur vorhandenen Temperatur/Konzentration:

$$(10)\qquad RW(\rho)(t,x) = \partial_\eta\rho(t,x) + \alpha\rho(t,x) = r_3(t,x), \quad (t,x)\in(\partial\Omega)^T.$$

Behältnis verhält sich wie ein schwarzer Strahler gemäß dem Boltzmann'schen Gesetz:

$$(11)\qquad RW(\rho)(t,x) = \partial_\eta\rho(t,x) + \alpha\rho(t,x)^4 = r_4(t,x), \quad (t,x)\in(\partial\Omega)^T.$$

Konvektions - Diffusionsgleichung
Dabei ist der Diffusionskoeffizient gegeben durch

$$K(t,x) \equiv \varepsilon > 0 \; v_T(t,x) = v_T \; (\Rightarrow \mathrm{div}_x v_T = 0) \; , \; q \equiv 0.$$

Es ergibt sich

$$(12)\qquad \begin{cases} \rho_t(t,x) &= \epsilon\Delta_x\rho(t,x) - v_T\nabla_x\rho(t,x) & \text{in } (\Omega)^T \\ RW(\rho)(t,x) &= r(t,x) & \text{auf } (\partial\Omega)^T \\ \rho(0,x) &= \rho_0(x) & \text{in } \Omega \end{cases}$$

mit Randwerten wie in (8) bis (11).
Diese Gleichung trägt dem Umstand Rechnung, daß sich etwa bei einem chemischen Prozeß Diffusion und Transport auf verschiedenen Skalen abspielen ($0 < \epsilon \ll |v_T|$), d.h., die Diffusionsgeschwindigkeit ist wesentlich geringer als die Strömungsgeschwindigkeit des Mediums. Diesem Umstand muß bei der

numerischen Diskretisierung Rechnung getragen werden.

Konvektion - Diffusion mit Advektion
Hier ist wie oben

$$K(t,x) \equiv \epsilon > 0 \ , \ v_T(t,x) = v_T,$$

aber

$$q(t,x) = q(\rho(t,x)) = a(t,x)\rho(t,x) \ , \ a(t,x) \equiv a$$

d.h., die Quelldichte ist proportional zur vorhandenen Konzentration. Es ergibt sich

$$(13) \qquad \begin{cases} \rho_t(t,x) &= \overbrace{\epsilon\Delta_x\rho(t,x)}^{\text{Diffusion}} - \overbrace{v_T\nabla_x\rho(t,x)}^{\text{Konvektion}} + \overbrace{a\rho(t,x)}^{\text{Advektion}} & \text{in } (\Omega)^T \\ RW(\rho)(t,x) &= r(t,x) & \text{auf } (\partial\Omega)^T \\ \rho(0,x) &= \rho_0(x) & \text{in } \Omega. \end{cases}$$

Beim gestörten Gelfand Problem (Zündung in einem Festkörperbrennstoff) ist, siehe auch Kapitel 5.

$$K(t,x) \equiv K \ , \ v_T \equiv 0$$

und

$$q(t,x) = q(\rho(t,x)) = \delta(t,x)e^{\rho(t,x)} \ , \ \delta(t,x) = \delta > 0.$$

d.h., die Quelldichte hängt von der Konzentration ab und die von der Quelle erzeugte Konzentration ist proportional zur nderung der von der Quelle erzeugten Konzentration. In Formeln

$$\frac{\partial}{\partial\rho}q(\rho) \approx q(\rho).$$

Damit ergibt sich

$$(14) \qquad \begin{cases} \rho_t(t,x) &= K\Delta_x\rho(t,x) + \delta e^{\rho(t,x)} & \text{in } (\Omega)^T \\ RW(\rho)(t,x) &= r(t,x) & \text{in } (\partial\Omega)^T \\ \rho(0,x) &= \rho_0(x) & \text{in } \Omega. \end{cases}$$

Hierbei handelt es sich um ein stark nichtlineares Problem, das in Kapitel 5 wieder aufgegriffen wird. Z.B. kann mit diesem Modell ansatzweise der chemische Ablauf erklärt werden, der zur Challenger-Katastrophe führte.

Transportgleichung
Hier ist

$$K(t,x) \equiv 0, \quad \mathrm{div}_x v_T(t,x) = 0.$$

Damit ergibt sich

$$(15) \qquad \begin{cases} \rho_t(t,x) &= -v_T\nabla_x\rho(t,x) + q(t,x) & \text{in } (\Omega)^T \\ RW(\rho)(t,x) &= r(t,x) & \text{auf } (\partial\Omega)^T \\ \rho(0,x) &= \rho_0(x) & \text{in } \Omega. \end{cases}$$

Im Allgemeinen hat es keinen Sinn, bei dem Transportproblem (15) Randwerte auf den ganzen Rand des Gebietes vorzuschreiben, denn betrachte

$$\rho_t(t,x) + v\rho_x(t,x) = 0, \quad \rho(0,x) = \rho_0(x).$$

Die eindeutige Lösung dieser Gleichung ist gegeben durch

$$\rho(t,x) = \rho_0(x - vt).$$

Sei jetzt $\Omega := (0,1)$ und

$$\rho(t,0) = 0, \quad \rho(t,1) = 1.$$

Dann ist notwendigerweise

$$\begin{aligned} \rho_0(-vt) &= 0 \quad \forall\, t \geq 0, \\ \rho_0(1-vt) &= 1 \quad \forall\, t \geq 0, \end{aligned}$$

woraus zur Zeit $t = \frac{1}{v}$ der Widerspruch

$$1 = \rho_0(0) = 0$$

folgt. Zudem müßte die Anfangsbedingung zu den Randwerten kompatibel sein.

1.3 Modellierung von Verkehrsfluß

Das Verkehrsaufkommen auf einer einspurigen Landstraße kann mit den Überlegungen des Kapitels 1.1 modelliert werden, und zwar wie folgt.
Bezeichnen

$\rho(t,x)$ Fahrzeugdichte (Anzahl Autos/Kilometer Straße)

$q(t,x)$ Fahrzeugstrom (Anzahl Autos/Stunde)

$[a,b]$ Abschnitt Straße (Strecke von Punkt a bis Punkt b)

t_1, t_2 $\Delta\, t = t_2 - t_1$ Zeitraum,

so gilt

$$\#\text{ Fahrzeuge in } [a,b] \text{ zur Zeit } t = \int_a^b \rho(t,x)dx.$$

Sind nun $\int_a^b \rho(t_1,x)dx$ und $\int_a^b \rho(t_2,x)dx$ verschieden, so müssen entweder Autos bei a den Streckenabschnitt erreichen oder bei b verlassen. Weil die # Fahrzeuge, die bei a die Strecke [a,b] im Zeitraum zwischen t_1 und t_2 erreichen $= \int_{t_1}^{t_2} q(t,a)dt$ ist und die # Fahrzeuge, die bei b die Strecke [a,b] im Zeitraum zwischen t_1 und t_2 verlassen $= \int_{t_1}^{t_2} q(t,b)dt$, kann dieser Sachverhalt in Formeln wie folgt ausgedrückt werden:

$$\underbrace{\int_a^b \rho(t_2,x) - \rho(t_1,x)dx}_{\substack{\text{Differenz \# Autos zu den} \\ \text{verschiedenen Zeitpunkten bei } x}} = \underbrace{\int_{t_1}^{t_2} q(t,a) - q(t,b)dt}_{\substack{\text{Autos rein bei } a \text{ minus} \\ \text{Autos raus bei } \ b}}$$

Diese Relation liefert

$$\underbrace{\frac{1}{b-a}\int_a^b \frac{\rho(t_2,x)-\rho(t_1,x)}{t_2-t_1}\,dx}_{\longrightarrow\ \rho_t(t_1,a)\ (t_2\to t_1,\, b\to a)} = \underbrace{\frac{1}{t_2-t_1}\int_{t_1}^{t_2} \frac{q(t,a)-q(t,b)}{b-a}\,dt}_{\longrightarrow\ -q_x(t_1,a)\ (t_2\to t_1,\, b\to a)}.$$

Die gewünschte Erhaltungsgleichung ist in diesem Fall

$$\rho_t(t,x) = -q_x(t,x). \tag{16}$$

An dieser Stelle muß wieder ein phänomenologischer Ansatz bemüht werden, um q mit ρ in Beziehung zu setzen. Als Ansatz dazu diene

$$q(t,x) = h(\rho(t,x))$$

und die folgende Beobachtung.

"Der Fahrzeugstrom q erfährt bei zunehmender Fahrzeugdichte ρ bei der Dichte ρ_m eine Sättigung h_0 und nimmt für größere Dichten ab."

Als einfachen Ansatz für h wähle

$$h(\rho) = -\frac{1}{2}(\rho - \rho_m)^2 + h_0.$$

Damit ergibt sich in (16) unter Verwendung von AW'en

$$\begin{cases} \rho_t(t,x) - (\rho(t,x) - \rho_m)\rho_x(t,x) &= 0 & \text{in } (\Omega)^T \\ \rho(0,x) &= \rho_0(x) & \text{in } \Omega^T, \end{cases}$$

so daß sich schliesslich mit der Transformation $\rho = -(\rho - \rho_m)$ die Burgers Gleichung

$$\begin{cases} \rho_t(t,x) + \rho(t,x)\rho_x(t,x) &= 0 & \text{in } (\Omega)^T \\ \rho(0,x) &= \rho_m - \rho_0(x) & \text{in } \Omega, \end{cases} \tag{17}$$

ergibt.
Allgemeine Modelle für die mathematische Beschreibung von Verkehrsfluß führen auf kinetische Gleichungen der Form

$$f_t(t,x,v) + v(t,x)f_x(t,x,v) = C(f)(t,x,v), \tag{18}$$

wobei f die Anzahl der Fahrzeuge bei x zur Zeit t mit Geschwindigkeit v bezeichnet. Ist wieder $\rho(t,x)$ wieder die Fahrzeugdichte, so gilt

$$\rho(t,x) = \int_0^{v_{max}} f(t,x,v)dv.$$

C bezeichnet dabei den Interaktionsparameter, von welchem in der Modellierung angenommen wird, daß er sich aus einem Anteil zusammensetzt, welcher Bremsvorgänge beschreibt, und aus einem Anteil, der Beschleunigungsvorgänge modelliert, i. e.

$$C(f) = B(f) + A(f). \tag{19}$$

Näheres hierzu ist zu finden in [21], download unter
http://www.mathematik.tu-darmstadt.de/~klar.
Weitere Beispiele für die Modellierung von Transportproblemen und allgemeinen hyperbolischer Erhaltungsgleichungen werden von Kröner in [15] angegeben.

2 Numerische Behandlung von skalaren Erhaltungsgleichungen

In diesem Kapitel werden Schemata für Transportgleichungen in Erhaltungsform angegeben und hinsichtlich ihrer Eigenschaften untersucht. Ein Transportproblem liegt in Erhaltungsform vor, falls es in der Form

$$\begin{cases} \rho_t(t,x) + h(\rho(t,x))_x &= 0 & \text{in} & \quad (0,\infty)\times\mathbb{R} \\ \rho(0,x) &= \rho_0(x) & \text{in } \mathbb{R} \end{cases} \tag{20}$$

geschrieben werden kann.

2.1 Motivation und Herleitung konservativer Verfahren

Liegt eine Differentialgleichung vor, welche einen physikalischen Prozeß beschreibt, so sollte ein Verfahren zur numerischen Simulation des Prozesses idealerweise alle dem beschriebenen System inherenten physikalischen Eigenschaften reproduzieren. Ein solches Verfahren heißt dann konservativ.
Zur Herleitung konservativer numerischer Schemata für (20) wird die Erhaltungsgleichung in integraler Form geschrieben. Dazu sei

$$V := [t_1, t_2] \times [a, b]$$

ein Testvolumen. Die Erhaltungsgleichung in integraler Form ergibt sich durch die Integration der Erhaltungsgleichung in (1.20) über V, i.e.:

$$\int_{t_1}^{t_2}\int_{a}^{b} \rho_t(t,x)dxdt + \int_{t_1}^{t_2}\int_{a}^{b} h(\rho(t,x))_x dxdt = 0,$$

(vergl. Kapitel 1, Modellierung) was wiederum äquivalent zu

$$\int_{a}^{b} \rho(t_2,x) - \rho(t_1,x)dx \; + \; \int_{t_1}^{t_2} h(\rho(t,b)) \; - \; h(\rho(t,a))dt = 0 \tag{21}$$

ist. Um diese Identität für die Konstruktion von numerischen Verfahren zu nutzen, werden die auftretenden Integranden numerisch approximiert. Dazu sei

$$t^n = n\Delta t, \quad x_i = i\Delta x, \quad n \in \mathbb{N}_0, \quad i \in \mathbb{Z}$$

wodurch auf $(0,\infty)\times\mathbb{R}$ ein Gitter mit Zeitschrittweite Δt und Ortsauflösung Δx definiert wird. Für

$$V = \left[t^n, t^{n+1}\right] \; \times \; \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right],$$

wobei

$$x_{i-\frac{1}{2}} := (i - \frac{1}{2})\Delta x, \quad x_{i+\frac{1}{2}} := (i + \frac{1}{2}\Delta x),$$

ergibt sich in (21)

$$\underbrace{\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \rho(t^{n+1},x) - \rho(t^n,x)dx}_{I_1} \; + \; \underbrace{\int_{t^n}^{t^{n+1}} h(\rho(t,x_{i+\frac{1}{2}})) - h(\rho(t,x_{i-\frac{1}{2}}))dt}_{I_2} \; = 0. \tag{22}$$

Aus Kapitel 1 ist bekannt, daß I_1 die Massenänderung in $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ während des Zeitraums Δt beschreibt, $-I_2$ die Masse, welche während Δt über $x_{i-\frac{1}{2}}$ und $x_{i+\frac{1}{2}}$ ein- bzw. austritt. Mit anderen Worten: (22) garantiert, daß die Masse erhalten bleibt. Um (22) numerisch zugänglich zu machen, bezeichnen ρ_i^n eine Approximation von $\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \rho(t^n,x)dx$ und g_i^n eine Approximation von $\frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} h\left(\rho(t,x_i)\right)dt$.

Definition 2.1. (Numerischer Fluß)
Sei $g \in \mathcal{C}(\mathbb{R}^2)$ eine Funktion, mit welcher $g_{i+\frac{1}{2}}^n$ in der Form $g\left(\rho_i^n, \rho_{i+1}^n\right)$ geschrieben werden kann. Dann heißt die Funktion g Numerischer Fluß. Ein Numerischer Fluß heißt konsistent mit der Erhaltungsgleichung (20): $\Longleftrightarrow$

$$g(u,u) = h(u) \quad \text{für alle } u \in \mathbb{R}.$$

Numerische Flüsse modellieren den Transport über die Ränder numerischer Zellen.
Mit den obigen Setzungen wird unter Berücksichtigung von (22) das Schema

$$\rho_i^{n+1} - \rho_i^n \; = \; -\frac{\Delta t}{\Delta x}(g_{i+\frac{1}{2}}^n - g_{i-\frac{1}{2}}^n). \tag{23}$$

zur numerischen Approximation einer Lösung von (20) erhalten.
Die Definition der Konsistenz in Definition 2.1 motiviert sich wie folgt: Bezeichnet $\rho(t,x) \equiv \rho_0$ die Lösung von (20) zu konstanten Anfangswerten (Nachweis, daß mit konstanten Anfangswerten die Lösung auch tatsächlich konstant in Ort und Zeit ist!), so gilt

$$g_{i+\frac{1}{2}}^0 = g(\rho_i^0, \rho_{i+1}^0) = g(\rho_0, \rho_0).$$

Andererseits ist im Fall

$$g(\rho_i^0, \rho_{i+1}^0) = \frac{1}{\Delta t} \int_0^{\Delta t} h(\rho(t, x_{i+\frac{1}{2}}))dt$$

sicher auch

$$g_{i+\frac{1}{2}}^0 = \frac{1}{\Delta t} \int_0^{\Delta t} h(\rho_0) = h(\rho_0),$$

so daß obige Definition von Konsistenz sinnvoll ist.
Da es sich bei der Definition von ρ_i^n und g_i^n um Approximationen handelt, kann Definition 2.1 natürlich auch auf Vorschriften der Gestalt

$$\rho_i^{n+1} - \rho_i^n = -\frac{\Delta t}{\Delta x} \left\{ \Theta(g_{i+\frac{1}{2}}^{n+1} - g_{i-\frac{1}{2}}^{n+1}) + (1-\Theta)(g_{i+\frac{1}{2}}^n - g_{i-\frac{1}{2}}^n) \right\} \tag{24}$$

verallgemeinert werden, wobei $0 \leq \Theta \leq 1$. Für $\Theta = 0$ heißt die Vorschrift explizit, für $0 < \Theta < 1$ semi-implizit und für $\Theta = 1$ implizit.
Daß das Schema (24) tatsächlich konservativ ist, folgt aus

$$\sum_{i \in \mathbb{Z}} \left(g_{i+\frac{1}{2}}^n - g_{i-\frac{1}{2}}^n \right) = 0 \qquad \text{(falls Summe existent)},$$

was wiederum

$$\sum_{i \epsilon \mathbb{Z}} \rho_i^{n+1} = \sum_{i \epsilon \mathbb{Z}} \rho_i^n \quad \text{für alle } n \geq 0$$

impliziert, Massenerhaltung der diskreten Größe also.
Der Ausdruck

$$\delta h_i^n := \frac{1}{\Delta x} \left(g_{i+\frac{1}{2}}^n - g_{i-\frac{1}{2}}^n \right) \tag{25}$$

kann als Approximation von

$$h(\rho(t^n, x_i))_x,$$

angesehen werden.
Beispiele von numerischen Flüssen im Fall $h(\rho) = \rho$ sind etwa

$$\begin{array}{llllllll}
1. & g(u,v) & = & u & \Rightarrow & \delta h_i^n & = & \frac{1}{\Delta x}\left(\rho_i^n - \rho_{i-1}^n\right) \\
2. & g(u,v) & = & v & \Rightarrow & \delta h_i^n & = & \frac{1}{\Delta x}\left(\rho_{i+1}^n - \rho_i^n\right) \\
3. & g(u,v) & = & \frac{1}{2}(u+v) & \Rightarrow & \delta h_i^n & = & \frac{1}{2\Delta x}\left(\rho_{i+1}^n - \rho_{i-1}^n\right).
\end{array}$$

Hilfsatz 2.2. (Konsistenz)
Sei $\rho \in \mathcal{C}^2(\mathbb{R} \times \mathbb{R})$ klassische Lösung von (20) und $g \in \mathcal{C}^2(\mathbb{R}^2)$ konsistenter Numerischer Fluß. Dann ist das Verfahren (23) konsistent von erster Ordnung in Ort und Zeit, d.h., für den Abschneidefehler

$$r(t^n, x_i, \rho, \Delta t, \Delta x) := \frac{\rho_i^{n+1} - \rho_i^n}{\Delta t} + \frac{g(\rho_i^n, \rho_{i+1}^n) - g(\rho_{i-1}^n, \rho_i^n)}{\Delta x} \tag{26}$$

gilt

$$r(t^n, x_i, \rho, \Delta t, \Delta x) = \mathcal{O}(\Delta t + \Delta x) \ (\Delta t, \Delta x \to 0).$$

Dabei ist $\rho_i^n := \rho(t^n, x_i)$.

Beweis. Es wird gezeigt, daß

$$\frac{\rho_i^{n+1} - \rho_i^n}{\Delta t} + \frac{g_{i+\frac{1}{2}}^n - g_{i-\frac{1}{2}}^n}{\Delta x} = \rho_t(t^n, x_i) + h(\rho_t(t^n, x_i))_x + \mathcal{O}(|\Delta t| + |\Delta x|).$$

Sicher gilt

$$\rho_i^{n+1} = \rho_i^n + \rho_t(t^n, x_i)\Delta t + \mathcal{O}(|\Delta t|^2),$$

womit die Konsistenz der Zeitableitung gesichert ist. Als nächstes schreibe

$$\frac{1}{2}\left\{h(\rho_{i+1}^n) - h(\rho_{i-1}^n)\right\} = \partial_x h(\rho_i^n)\Delta x + \mathcal{O}\left(|\Delta x|^2\right)$$

und drücke den Term auf der linken Seite durch die Funktion g aus (was ja wegen der Konsistenz des Numerischen Flußes geht). Es ist

$$\begin{aligned} h(\rho_{i+1}^n) - h(\rho_{i-1}^n) = g(\rho_{i+1}^n, \rho_{i+1}^n) - g(\rho_{i-1}^n, \rho_{i-1}^n) = \\ g(\rho_{i+1}^n, \rho_{i+1}^n) - g(\rho_{i-1}^n, \rho_{i+1}^n) + g(\rho_{i-1}^n, \rho_{i+1}^n) - g(\rho_{i-1}^n, \rho_{i-1}^n) = \\ \underbrace{\partial_1 g(\rho_i^n, \rho_{i+1}^n)}_{\partial_1 g(\rho_i^n, \rho_i^n) + \mathcal{O}(|\rho_{i+1}^n - \rho_i^n|)} (\rho_{i+1}^n - \rho_{i-1}^n) + \underbrace{\partial_2 g(\rho_{i-1}^n, \rho_i^n)}_{\partial_2 g(\rho_i^n, \rho_i^n) + \mathcal{O}(|\rho_i^n - \rho_{i-1}^n|)} (\rho_{i+1}^n - \rho_{i-1}^n) + \mathcal{O}(|\rho_{i+1}^n - \rho_{i-1}^n|^2) = \\ \partial_1 g(\rho_i^n, \rho_i^n)(\rho_{i+1}^n - \rho_{i-1}^n) + \partial_2 g(\rho_i^n, \rho_i^n)(\rho_{i+1}^n - \rho_{i-1}^n) + \mathcal{O}(|\Delta x|^2). \end{aligned}$$

Weiter gilt

$$g\left(\rho_i^n, \rho_{i+1}^n\right) = g\left(\rho_i^n, \rho_i^n\right) + \partial_2 g\left(\rho_i^n, \rho_i^n\right)\left(\rho_{i+1}^n - \rho_i^n\right) + \mathcal{O}\left(|\rho_{i+1}^n - \rho_i^n|^2\right)$$

und

$$g\left(\rho_{i-1}^n, \rho_i^n\right) = g\left(\rho_i^n, \rho_i^n\right) - \partial_1 g\left(\rho_i^n, \rho_i^n\right)\left(\rho_i^n - \rho_{i-1}^n\right) + \mathcal{O}\left(|\rho_i^n - \rho_{i-1}^n|^2\right),$$

so daß schliesslich

$$\begin{aligned} g_{i+\frac{1}{2}}^n - g_{i-\frac{1}{2}}^n = g(\rho_i^n, \rho_{i+1}^n) - g(\rho_{i-1}^n, \rho_i^n) = \\ \frac{1}{2}\partial_1 g(\rho_i^n, \rho_i^n)(\rho_{i+1}^n - \rho_{i-1}^n) + \frac{1}{2}\partial_2 g(\rho_i^n, \rho_i^n)(\rho_{i+1}^n - \rho_{i-1}^n) - \\ \frac{1}{2}\left\{\partial_1 g(\rho_i^n, \rho_i^n) - \partial_2 g(\rho_i^n, \rho_i^n)\right\}(\rho_{i+1}^n - 2\rho_i^n + \rho_{i-1}^n) + \mathcal{O}(|\Delta x|^2) = \\ \frac{1}{2}\left\{h(\rho_{i+1}^n) - h(\rho_{i-1}^n)\right\} + \mathcal{O}(|\Delta x|^2) = \partial_x h(\rho_i^n)\Delta x + \mathcal{O}(|\Delta x|^2). \end{aligned}$$

Damit ergibt sich in (26)

$$r(t^n, x_i, \rho, \Delta t, \Delta x) = \rho_t(t^n, x_i) + h(\rho(t^n, x_i))_x + \mathcal{O}(|\Delta t| + |\Delta x|),$$

und somit die Behauptung. □

2.2 Lösungsbegriffe und numerische Konsequenzen

Wie Aufgabe 221 am Beispiel der **Burgersgleichung** (17) verdeutlicht, darf i. A. selbst bei beliebig glatten Anfangswerten nicht mit einer im klassischen Sinn in $(0, \infty) \times \mathbb{R}$ differenzierbaren Lösung von (20) gerechnet werden. Lokal in der Zeit ist das anders, denn es gilt

Satz 2.3. (Lokale Existenz von glatten Lösungen)
Sei $h \in \mathcal{C}^2(\mathbb{R}), \rho_0 \in \mathcal{C}^1(\mathbb{R})$ und ρ_0 zudem gleichmäßig beschränkt auf $\mathbb{R}$. Dann gibt es ein Zeitintervall $[0, T]$ dergestalt, daß (20) eine Lösung $\rho \in \mathcal{C}^1([0, T] \times \mathbb{R})$ besitzt.

Beweis. [15, Lemma 2.1.2] □

Soll von globalen Lösungen skalarer Erhaltungsgleichungen in der Zeit gesprochen werden können, muß aufgrund der vorangestellten Bemerkung deren Lösungsbegriff modifiziert werden.

Definition 2.4. (Schwache Lösungen)
Sei $\rho_0 \in L^\infty(\mathbb{R})$ (d.h. Lebesgue-meßbar und beschränkt). Dann heißt ρ schwache Lösung von (20) genau dann, wenn $\rho \in L^\infty(\mathbb{R}^+ \times \mathbb{R})$ gilt und

$$\int_{\mathbb{R}^+}\int_{\mathbb{R}} \rho\varphi_t + h(\rho)\varphi_x dx dt + \int_{\mathbb{R}} \varphi(0, x)\rho_0(x) dx = 0 \qquad \text{für alle } \varphi \in \mathcal{C}_0^\infty([0, \infty] \times \mathbb{R}) \tag{27}$$

erfüllt ist.

Dabei bezeichnet $\mathcal{C}_0^\infty([0,\infty]\times\mathbb{R})$ die Menge aller auf $[0,\infty]\times\mathbb{R}$ ∞-oft differenzierbaren Funktionen mit kompakten Trägern in $[0,\infty]\times\mathbb{R}$.
Daß eine klassische Lösung von (20)(das meint eine beschränkte differenzierbare Lösung) auch eine schwache Lösung ist, folgt unmittelbar mit partieller Integration. Schwache Lösungen von (20) dürfen allerdings im Gegensatz zu klassischen Lösungen auch Unstetigkeiten besitzen, welche allerdings nicht beliebig ausfallen(ausarten) dürfen. Zulässige Unstetigkeiten werden durch die sogenannte Rankine-Hugoniot Bedingung charakterisiert, die wie folgt formuliert werden kann.

Hilfsatz 2.5. (Rankine-Hugoniot Bedingung)
Sei $\rho\in L^1_{loc}(\mathbb{R}^+\times\mathbb{R})$ und $\gamma: t\mapsto(t,\gamma(t))$ eine hinreichend glatte Kurve, welche $\mathbb{R}^+\times\mathbb{R}$ in zwei Teile M_L und M_R separiert. Ferner gelte

$$\rho_L := \rho_{|M_L}\ \in\mathcal{C}^1(\bar{M_l}),\quad \rho_R := \rho_{|M_R}\ \in\mathcal{C}^1(\bar{M_r})$$

und ρ_L und ρ_R seien lokal klassische Lösungen in M_L bzw. M_R. Dann erfüllt ρ die Gleichung (27) für alle $\varphi\in\mathcal{C}_0^\infty(\mathbb{R}^+\times\mathbb{R})$ dann und nur dann, wenn

$$\{\rho_L(t,\gamma(t))-\rho_R(t,\gamma(t))\}\,\dot\gamma(t) = h(\rho_L(t,\gamma(t)))-h(\rho_R(t,\gamma(t)))\quad\forall t>0 \tag{28}$$

gültig ist. Liegt auf der Kurve γ eine Unstetigkeit vor, so heißt $\dot\gamma(t)$ deren **Ausbreitungsgeschwindigkeit**.

Beweis. Sei ν die äußere Normale an M_L (dann ist $-\nu$ die äussere Normale an M_R!). Dann gilt für $\varphi\in\mathcal{C}(\mathbb{R}^+\times\mathbb{R})$ (beachte, daß dann $\varphi(0,x)=0$ gilt!)

$$0=\int\limits_{\mathbb{R}^+}\int\limits_{\mathbb{R}}\rho(t,x)\varphi_t(t,x)+h(\rho(t,x))\varphi_x(t,x)\,dxdt =: \int\limits_{M_L}\ldots+\int\limits_{M_R}\ldots$$

Nun ist mit Hilfe des Gauß'schen Satzes angewendet im Raum-Zeit Zylinder

$$\int\limits_{M_L}\rho_L\varphi_t+h(\rho_L)\varphi_x\,dxdt=\int\limits_{M_L}[\rho_L,h(\rho_L)]^t\nabla_{(t,x)}\varphi\,dxdt=$$
$$-\int\limits_{M_L}\mathrm{div}_{(t,x)}[\rho_L,h(\rho_L)]^t\varphi dxdt+\int\limits_{\partial M_L}\nu[\rho_L,h(\rho_L)]\varphi\,d\mathcal{O}_{M_L}=\int\limits_{\partial M_L}\nu[\rho_L,h(\rho_L)]\varphi\,d\mathcal{O}_{M_L}.$$

Analog (beachte, daß ν durch $-\nu$ zu ersetzen ist!)

$$\int\limits_{M_R}\rho_R\varphi_t+h(\rho_R)\varphi_x\,dxdt=-\int\limits_{\partial M_R}\nu[\rho_R,h(\rho_R)]\varphi\,d\mathcal{O}_{M_R}.$$

Wegen $\nu=C[1,-\gamma']^t$ für ein $C=C(t)>0$ folgt die Behauptung mit dem Variationslemma von Du-Bois Reymond. □

Schwache Lösungen von (20) müssen durch ihre Anfangswerte nicht eindeutig festgelegt sein, wie das folgende Beispiel zeigt.

Beispiel Sei $h(v):=\frac{1}{2}v^2$. Dann entspricht (20) der Burgersgleichung. Als Anfangswerte wähle

$$\rho_0(x)=\begin{cases}0 & x\le 0,\\ 1 & x>0\end{cases}$$

und definiere die Funktionen

$$u_1(t,x):=\begin{cases}0 & x<\frac{t}{2},\\ 1 & x\ge\frac{t}{2}\end{cases}$$

und

$$u_2(t,x):=\begin{cases}0 & x<0,\\ \frac{x}{t} & 0\le x\le t,\\ 1 & t<x.\end{cases}$$

Damit ist für u_1 $\gamma(t) = \frac{t}{2}$, $\gamma'(t) = \frac{1}{2}$, u_1 erfüllt also Bedingung (28), die Anfangsbedingung und ist deshalb eine schwache Lösung. Die Funktion u_2 ist eine stückweise klassische Lösung, welche für $t > 0$ keine Sprünge besitzt. Auf den Knicklinien

$$\{(t,x); x = 0\} \text{ und } \{(t,x); x = t\} \ (t > 0)$$

ist u_2 stetig und Bedingung (28) damit trivialerweise erfüllt. Beide Funktionen sind demnach schwache Lösungen. □

Frage: Welche Lösung ist bei Mehrdeutigkeit auszuzeichnen?

Es soll die physikalische Lösung ausgezeichnet werden. Dazu führe Reibung in den durch (20) definierten Transportprozeß ein. Dann hat dieser zu einer gegebenen Anfangsbedingung unter realisitschen Bedingungen an die Anfangswerte und an die Funktion h eine eindeutige Lösung.
Was meint 'Reibung einführen'? Betrachte dazu zu $\epsilon > 0$ das Anfangswertproblem

$$\left\{ \begin{array}{rcll} \rho_t(t,x) - \epsilon\rho_{xx}(t,x) + h(\rho(t,x))_x & = & 0 & \text{in } \mathbb{R}^+ \times \mathbb{R} \\ \rho(0,x) & = & \rho_0(x) & \text{in } \mathbb{R}. \end{array} \right. \tag{29}$$

Sei ρ^ϵ die eindeutige Lösung von (29), wobei davon ausgegangen wird, daß die Daten h und ρ_0 Existenz und Eindeutigkeit einer Lösung von (29) gewährleisten. Existiert jetzt

$$\lim_{\epsilon \to 0} \rho^\epsilon(t,x) =: \rho^*(t,x) \text{ fast überall in } \mathbb{R}^+ \times \mathbb{R} \tag{30}$$

und ist ρ^* schwache Lösung von (20), so heißt ρ^* **Viskositätslösung**, und genau diese sollen berechnet werden.
Viskositätslösungen erfüllen die sogenannte **Entropie-Bedingung**. Zu deren Erläuterung sei S eine glatte Kurve in $\mathbb{R}^+ \times \mathbb{R}$, entlang welcher eine schwache Lösung ρ von (20) eine Unstetigkeit besitze. Dann erfüllt ρ die Entropie-Bedingung in $(t,x) \in S$ genau dann, wenn in jedem Unstetigkeitspunkt (t,x) mit

$$\rho_L := \lim_{\epsilon \to 0} \rho(t, x - \epsilon), \quad \rho_R := \lim_{\epsilon \to 0} \rho(t, x + \epsilon)$$

und

$$s := \frac{h(\rho_L) - h(\rho_R)}{\rho_L - \rho_R}$$

die Ungleichungskette

$$h'(\rho_L) \ > s \ > \ h'(\rho_R) \tag{31}$$

erfüllt ist.
Eine Unstetigkeit, welche (28) und (31) erfüllt, heißt **Schock**.

Satz 2.6. (Eindeutigkeit von Entropielösungen)
Sei $h \in \mathcal{C}^2(\mathbb{R})$ und $h'' > 0$ auf $\mathbb{R}$. Erfüllen zwei schwache Lösungen (31), so sind sie gleich.

Beweis. [15, Theorem 2.1.] □

2.3 Geeignete und ungeeignete Verfahren

Zunächst wird anhand 3er Beispiele gezeigt, daß, soll ein numerisches Verfahren die Viskositätslösung reproduzieren, zentrale Differenzenquotienten zur Approximation des Terms $h(\rho(t,x))_x$ ungeeignet sind.

Beispiel
1.) Sei $h(\rho) := \frac{1}{2}\rho^2$ und

$$\rho_0(x) := \left\{ \begin{array}{ll} 1 & x \geq 0 \\ -1 & x < 0. \end{array} \right.$$

Der numerische Fluß sei definiert durch

$$g(u,v) := \frac{1}{2}\left\{h(u)+h(v)\right\},$$

womit sich mittels (23) das numerische Verfahren

$$\rho_i^{n+1} = \rho_i^n - \frac{\Delta t}{2\Delta x}\left\{h(\rho_{i+1}^n) - h(\rho_{i-1}^n)\right\} = \rho_i^n - \frac{\Delta t}{2\Delta x}(\rho_{i+1}^n - \rho_{i-1}^n)\frac{\rho_{i+1}^n + \rho_{i-1}^n}{2}$$

ergibt. Damit ist aber

$$\rho_i^n = \begin{cases} 1 & x_i \geq 0 \\ -1 & x_i < 0, \end{cases} \quad \forall n \geq 0$$

d.h., die Anfangsbedingung wird reproduziert. Diese erfüllt wegen

$$h'(\rho_L) = -1 \;<\; 0 \;<\; h'(\rho_R) = 1$$

gerade nicht die Entropie-Bedingung (31).
2.) Sei $g(u,v) = av$, $h(u) = au$, $a > 0$, so daß die lineare Gleichung

$$\begin{cases} \rho_t(t,x) + a\rho_x(t,x) &= 0 & \text{in } \mathbb{R}^+ \times \mathbb{R} \\ \rho(0,x) &= \rho_0(x) & \text{in } \mathbb{R} \end{cases}$$

vorliegt. Die exakte Lösung ist durch

$$\rho(t,x) = \rho_0(x - at)$$

gegeben. Mit oben definiertem numerischen Fluß ist (23) äquivalent zu

$$\rho_i^{n+1} = \rho_i^n - a\frac{\Delta t}{\Delta x}(\rho_{i+1}^n - \rho_i^n),$$

woraus mit

$$E\rho_i^n := \rho_{i+1}^n$$

nach kurzer Rechnung

$$\begin{aligned} \rho_i^n &= \left(1 + a\frac{\Delta t}{\Delta x} - a\frac{\Delta t}{\Delta x}E\right)^n \rho_i^0 \\ &= \sum_{m=0}^{n} \binom{n}{m} \left(1 + a\frac{\Delta t}{\Delta x}\right)^m \left(-a\frac{\Delta t}{\Delta x}E\right)^{n-m} \rho_i^0 \\ &= \sum_{m=0}^{n} \binom{n}{m} \left(1 + a\frac{\Delta t}{\Delta x}\right)^m \left(-a\frac{\Delta t}{\Delta x}\right)^{n-m} \rho_{i+(n-m)}^0 \end{aligned}$$

erhalten wird. D.h., daß zur Berechnung von $\rho(t^n, x_i)$ nur Werte von ρ_0 rechts von x_i Verwendung finden. Ist dann z.B.

$$\rho_0(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0, \end{cases}$$

so gilt

$$1 = \rho(t^n, 0) = \rho_0(-at^n) \overset{!}{\approx} \rho_0^n = \sum_{m=0}^{n} \binom{n}{m} \left(1 + a\frac{\Delta t}{\Delta x}\right)^m \left(-a\frac{\Delta t}{\Delta x}\right)^{n-m} \rho_{0+(n-m)}^0 \rho_0(n\Delta x)) = 0.$$

Dieses Verfahren scheint daher auch ungeeignet zu sein, weil es die exakte Lösung nicht reproduziert. Das liegt an der Verwendung vorwärtsgenommener Differenzen für die Diskretisierung der Ortsableitung. Für $a < 0$ und rückwärtige Differenzen ergäbe sich das gleiche Verhalten.

3.) wie 2.) mit $g(u,v) = au$. Dann gilt mit

$$\mathrm{L}\rho_i^n := \rho_{i-1}^n$$

$$\rho_i^n = \sum_{m-0}^{n} \binom{n}{m} \left(1 - a\frac{\Delta t}{\Delta x}\right)^m \left(a\frac{\Delta t}{\Delta x}\right)^{n-m} \rho_{i-(n-m)}^0,$$

und das sieht vernünftig aus, weil jetzt die richtigen Werte zur Berechnung der Lösung in der neuen Zeitschicht herangezogen werden. Die Verfahrensvorschrift ist dabei durch

$$\rho_i^{n+1} = \rho_i^n - a\frac{\Delta t}{\Delta x}(\rho_i^n - \rho_{i-1}^n)$$

gegeben. Dies' ist gleichzeitig eine Approximation zweiter Ordnung an die Differentialgleichung

$$\rho_t(t,x) + a\rho_x(t,x) - \Delta x \frac{1}{2} a \left(1 - \frac{\Delta t}{\Delta x} a\right) \rho_{xx}(t,x) = 0. \tag{32}$$

Ist noch

$$\frac{\Delta t}{\Delta x}|a| \ < \ 1,$$

so stimmt diese Differentialgleichung mit der aus (29) für $\epsilon := \Delta x \frac{1}{2} a \left(1 - \frac{\Delta t}{\Delta x} a\right)$ überein. Obiges Schema führt demnach künstlich Diffusion in den numerischen Prozess ein und berechnet daher eine numerische Approximation an die 'physikalische' Lösung. Der Vorzeichenbedingung werden wir bei Stabilitätsuntersuchungen wieder begegnen, und sie heißt **Courant-Friedrichs-Lewy Bedingung**, kurz **CFL-Bedingung**. □

Die beiden Verfahren, die im Folgenden vorgestellt werden, liefern eine numerische Approximation an die Viskositäts- bzw. Entropielösung von (20), indem sie künstlich Diffusion in das numerische System einführen.

Verfahren 2.7. (Lax-Friedrichs Schema)
Der (konsistente) numerische Fluß

$$g(u,v) := \frac{1}{2}\left\{h(u) + h(v)\right\} + \frac{\Delta x}{2\Delta t}(u - v) \tag{33}$$

definiert das **Lax-Friedrichs Schema**. Als Vorschrift:

$$\rho_i^{n+1} = \rho_i^n - \frac{\Delta t}{2\Delta x}\left\{h(\rho_{i+1}^n) - h(\rho_{i-1}^n)\right\} + \frac{1}{2}\left(\rho_{i+1} - 2\rho_i^n + \rho_{i-1}^n\right). \tag{34}$$

Die Idee bei diesem Verfahren ist es, rückwärtige Differenzen als zentrale Differenzen + Diffusion zu schreiben , d.h.

$$\alpha_i - \alpha_{i-1} = \frac{1}{2}(\alpha_{i+1} - \alpha_{i-1}) - \frac{1}{2}(\alpha_{i+1} - 2\alpha_i + \alpha_{i-1})$$

und diesen Sachverhalt auf die Diskretisierung des Transportterms $h(\rho(t,x))_x$ in (20) zu übertragen. Wie im vorangegangenen Beispiel gezeigt, würde die alleinige Wahl von zentralen Differenzen i.A. nicht ausreichen, um die Viskositätslösung zu approximieren.
Aus dem vorangegangenen Beispiel motiviert sich auch das nächste Verfahren. Es verwendet Upwinding und ist nach **Enquist-Osher** benannt.

Verfahren 2.8. (Enquist-Osher Schema, 1981)
Der konsistente numerische Fluß

$$g(u,v) := h^+(u) + h^-(v) \tag{35}$$

definiert das **Enquist-Osher Schema** [7]. Dabei ist

$$h^+(u) := h(0) \ + \int_0^u \max\left\{h'(s), 0\right\} \, ds, \ \text{ und } h^-(u) := \int_0^u \min\{h'(s), 0\} \, ds.$$

Unter Verwendung von (25) ergibt sich die Vorschrift

$$\rho_i^{n+1} = \rho_i^n - \frac{\Delta t}{\Delta x}\left\{h^+(\rho_i^n) - h^+(\rho_{i-1}^n) + h^-(\rho_{i+1}^n) - h^-(\rho_i^n).\right\} \tag{36}$$

Wie in Aufgabe 6.1 nachzulesen ist, stimmt dieses Verfahren für $h(u) = au$ mit dem Upwind-Schema

$$\rho_i^{n+1} = \rho_i^n - a\frac{\Delta t}{\Delta x}\begin{cases} \rho_i^n - \rho_{i-1}^n & , \quad a > 0 \\ \rho_{i+1}^n - \rho_i^n & , \quad a < 0 \end{cases} \tag{37}$$

überein. Die Idee in Vorschrift (36) ist es

$$h(\rho)_x \approx \frac{1}{\Delta x}\begin{cases} h(\rho_i^n) - h(\rho_{i-1}^n), & h' > 0, \text{ bzw.} \\ h(\rho_{i+1}^n) - h(\rho_i^n), & h' < 0 \end{cases}$$

zu gewährleisten. Weiterführendes findet sich in Kröner [15].

3 Numerische Behandlung elliptischer Differentialgleichungen

Lösungen des Problems (6) heißen stationäre Lösungen, falls sie unabhängig von der Zeit sind. In diesem Kapitel sollen gerade für diese Lösungen numerische Berechnungsverfahren bereitgestellt werden. Wie sich später herausstellen wird, werden diese Methoden auch nützlich bei der numerischen Behandlung zeitdiskretisierter zeitabhängiger Probleme sein.

3.1 Exkurs für eindimensionale Probleme

Betrachtet wird zunächst die Randwertaufgabe

$$\begin{cases} -(ku')' + au &= q \quad \text{in } \Omega = (a,b) \\ u(a) &= \alpha \\ u(b) &= \beta. \end{cases} \tag{38}$$

Ohne Einschränkung kann hier $\alpha = \beta = 0$ angenommen werden, denn die Funktion

$$\tilde{u} = u - l, \quad l(x) = \alpha\frac{b-x}{b-a} + \beta\frac{x-a}{b-a}$$

erfüllt (38) mit $\alpha = \beta = 0$ und $\tilde{q} = q - al + k'l'$.
Zunächst soll (38) mit der Methode der finiten Differenzen approximiert werden. Dazu sei

$$a = x_0 < x_1 \ < \ \ldots \ < \ x_n < x_{n+1} = b \tag{39}$$

ein Gitter über Ω. Die Randwertaufgabe (38) liegt in Divergenzform vor und ist in einem noch zu spezifizierenden Sinn selbstadjungiert. Eine geeignete numerische Diskretisierung von (38) sollte demnach eine Systemmatrix erzeugen, die auch symmetrisch ist. Mit den Hilfsgitterpunkten $x_{i+\frac{1}{2}} := x_i + \frac{1}{2}(x_i + x_{i+1})$, $x_{i-\frac{1}{2}} := x_i - \frac{1}{2}(x_i + x_{i-1})$ wird das wie folgt gewährleistet.

$$\begin{cases} (k(x)u'(x))_i' &\doteq \frac{1}{\Delta x}\left(k_{i+\frac{1}{2}}u'_{i+\frac{1}{2}} - k_{i-\frac{1}{2}}u'_{i-\frac{1}{2}}\right) \\ u'_{i+\frac{1}{2}} &\doteq \frac{1}{\Delta x}(u_{i+1} - u_i) \\ u'_{i-\frac{1}{2}} &\doteq \frac{1}{\Delta x}(u_i - u_{i-1}). \end{cases} \quad \text{Differenzen bzgl. } x \tag{40}$$

Mit (40) ergibt sich

$$-(k(x)u'(x))_i' \quad \doteq \quad \frac{1}{\Delta x^2}\left\{-k_{i-\frac{1}{2}}u_{i-1} + \left(k_{i-\frac{1}{2}} + k_{i+\frac{1}{2}}\right)u_i - k_{i+\frac{1}{2}}u_{i+1}\right\}.$$

Wird noch

$$q_i = q(x_i) \text{ und } a_i = a(x_i)$$

gesetzt, ergibt sich für die Bestimmung von u das lineare Gleichungssystem

$$Au = r \tag{41}$$

mit der Koeffizientenmatrix

$$A = \begin{bmatrix} k_{\frac{1}{2}} + k_{\frac{3}{2}} + a_1\Delta x^2 & -k_{\frac{3}{2}} & & \\ -k_{\frac{3}{2}} & k_{\frac{3}{2}} + k_{\frac{5}{2}} + a_2\Delta x^2 & -k_{\frac{5}{2}} & 0 \\ & \ddots & \ddots & \ddots \\ 0 & \ddots & \ddots & -k_{n-\frac{1}{2}} \\ & & -k_{n-\frac{1}{2}} & k_{n-\frac{1}{2}} + k_{n+\frac{1}{2}} + a_n\Delta x^2 \end{bmatrix} \tag{42}$$

und der rechten Seite

$$r = \Delta x^2 \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_{n-1} \\ q_n \end{bmatrix} + \begin{bmatrix} k_{\frac{1}{2}}\alpha \\ 0 \\ \vdots \\ 0 \\ k_{n+\frac{1}{2}}\beta \end{bmatrix}. \tag{43}$$

Satz 3.1. (Regularität von A)
Sei $k(x) > 0$ in (a, b), $a(x) \geq 0$ in (a, b). Dann ist die Matrix A regulär.

Beweis. Wegen $k(x) > 0$ und $a(x) \geq 0$ ist A irreduzibel diagonaldominant. Irreduzibel, weil Tridiagonalmatrix mit nichtverschwindenden Nicht-Diagonalelementen und strikt diagonaldominant in der ersten Zeile. Nach [9, Satz 2.3.3] ist A regulär. Siehe dazu auch die Veranstaltung Numerische Mathematik I. □

Eine Matrix heißt irreduzibel diagonaldominant, falls sie irreduzibel und diagonaldominant ist, wobei in einer Zeile echte Diagonaldominanz vorliegen muß. Eine Matrix A heißt irreduzibel, falls der ihr zugeordnete gerichtete Graph stark zusammenhängend ist.

Bemerkung 3.2. Ist $k(x) \geq 0, a(x) > 0$, so ist die Matrix A aus (42) offensichtlich strikt diagonaldominant und somit invertierbar.

Eine wichtige Rolle bei der qualitativen Untersuchung von Lösungen des Gleichungssystems (41) spielt die Kondition der Matrix A. Diese ist durch

$$\mathrm{cond}(A) := \|A\| \, \|A^{-1}\|, \tag{44}$$

definiert, wobei $\|\cdot\|$ die einer Vektornorm $|\cdot|$ zugeordnete Matrixnorm bezeichne, i.e.

$$\|A\| := \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{|Ax|}{|x|}.$$

Im Folgenden bezeichnen $|\cdot|$ immer die Euklidische Vektornorm oder Beitragsstriche, sofern nichts anderes gesagt wird. Bezeichnen ΔA und Δr Störungen der Daten in (41), so läßt sich der durch die Störungen in der Lösung verursachte relative Fehler abschätzen;

$$\frac{|\Delta u|}{|u|} \leq \frac{\mathrm{cond}(A)}{1 - \mathrm{cond}(A)\frac{\|\Delta A\|}{\|A\|}} \left\{ \frac{\|\Delta A\|}{\|A\|} + \frac{|\Delta r|}{|r|} \right\}. \tag{45}$$

Dabei ist das gestörte Gleichungssystem durch

$$(A + \Delta A)(u + \Delta u) \quad = \quad r + \Delta r$$

gegeben und $\|A^{-1}\| \, \|\Delta A\| < 1$ vorausgesetzt. Ist etwa $\Delta A = 0$, so wird der relative Fehler in der rechten Seite von (41) im schlimmsten Fall mit $\mathrm{cond}(A)$ verstärkt.

Für reguläre symmetrische Matrizen A ist die Kondition (bzgl. der Euklidischen Norm) durch

$$\text{cond}(A) = \frac{\max\{|\lambda(A)|,\ \lambda \text{ Eigenwert von A}\}}{\min\{|\lambda(A)|,\ \lambda \text{ Eigenwert von A}\}}. \tag{46}$$

Wird nun der in (41) enthaltene Diskretisierungsfehler von der Größenordnung $\mathcal{O}(\Delta x^2)$ sein, ergibt sich für den globalen Fehler Δu

$$\mathrm{A}\Delta u = \Delta x^2 \Delta r,$$

ergo,

$$|\Delta u| \leq \|A^{-1}\|\Delta x^2|\Delta r| \leq \frac{1}{|\lambda_{\min}(A)|}\Delta x^2|\Delta r|, \tag{47}$$

wobei $\lambda_{\min}(A) := \min\{|\lambda(A)|\ ,\ \lambda \text{ Eigenwert von A}\}$. Sind jetzt $k(x) \equiv k$, $a(x) \equiv a$ konstant, so können die Eigenwerte von A direkt abgegeben werden.

Satz 3.3. (Eigenwerte einer Tridiagonalmatrix)
Sei

$$B := \begin{bmatrix} \beta & \gamma & 0 \\ \alpha & \ddots & \gamma \\ 0 & \alpha & \beta \end{bmatrix} \in \mathbb{R}^{nxn}$$

mit $\alpha \cdot \gamma > 0$. Dann besitzt B die Eigenwerte

$$\lambda_i = \beta + 2\sqrt{\alpha\gamma}\text{sign}(\alpha)\cos\frac{i\pi}{n+1},\ 1 \leq i \leq n \tag{48}$$

und die Eigenvektoren v^i mit den Komponenten

$$v^i_j = (\frac{\alpha}{\gamma})^{\frac{j-1}{2}}\sin\frac{ij\pi}{n+1},\ 1 \leq i \leq n, 1 \leq j \leq n. \tag{49}$$

Beweis. [13, S.104] □

Mit diesem Satz ergibt sich für die Eigenwerte von A für $b - a = 1$ mit

$$\alpha = \gamma = -k, \quad \beta = 2k + \Delta x^2 a$$

die Darstellung

$$\lambda_i(A) = 2k + \Delta x^2 a - 2k\cos\frac{i\pi}{n+1},\ 1 \leq i \leq n, \tag{50}$$

in erster Näherung demnach für den kleinsten Eigenwert $\lambda_{\min}(A) = \lambda_1(A)$

$$\lambda_1(A) \doteq a\Delta x^2 + k\pi^2\Delta x^2.$$

Der absolute Fehler in (47) kann somit durch

$$|\Delta u| \leq \frac{1}{a + k\pi^2}|\Delta r|. \tag{51}$$

abgeschätzt werden. Für kleine Fehler Δr bleibt demnach auch der absolute Fehler Δu klein. Für die Kondition von A gilt wegen $\lambda_{\max}(A) = \lambda_n(A) \doteq 4k$

$$\text{cond(A)} \doteq \frac{4k}{k\pi^2\Delta x^2 + a\Delta x^2}.$$

Demnach lieferte (45) die Abschätzung

$$\frac{|\Delta u|}{|u|} \leq \frac{4k}{(k\pi^2 + a)\Delta x^2}\frac{|\Delta r|}{|r|} \quad (\Delta x \to 0). \tag{52}$$

Andererseits gilt mit (51) aber auch

$$\frac{|\Delta u|}{|u|} \leq \frac{1}{(a+k\pi^2)} \frac{|r|}{|u|} \frac{|\Delta r|}{|r|}, \tag{53}$$

und diese Abschätzung ist nur dann pessimistisch, wenn der Quotient $|r|/|u|$ groß ist, also etwa in der Terminology aus Kapitel 1, Quellen/Senken den Stofftransport oder den Wärmestrom durch Diffusion in einer Größenordnung beeinflussen, die kleiner als ihre eigene Quellstärke ist. Das ist allerdings unrealisitsch. Offenbar gilt

Merkregel 3.4. Das Gleichungssystem (41) ist bezüglich des absoluten Fehlers gut konditioniert, bezüglich des relativen Fehlers nur in pathologischen Fällen schlecht.

In (38) sind natürlich auch andere Randbedingungen möglich. Exemplarisch werde die Bedingung

$$u'(b) \quad = \quad 0 \tag{54}$$

untersucht und numerisch approximiert. Letzteres mittels der Differenzenapproximation

$$u'(x_{n+1}) \doteq \frac{u_{n+1} - u_n}{\Delta x}. \tag{55}$$

Diese Approximation von (54) ist von erster Ordnung. Für die Matrix A in (42) ergibt sich eine Änderung in der letzten Zeile, welche jetzt folgende Gestalt hat

$$[0, \ldots 0, -k_{n-\frac{1}{2}}, k_{n-\frac{1}{2}} + a_n \Delta x^2].$$

Darüberhinaus ist der letzte Eintrag von r in (43) gleich q_n.
Es sei bemerkt, daß (55) auch dann noch von erster Ordnung ist, falls (54) durch

$$u'(b) = \alpha \qquad \alpha \neq 0$$

ersetzt wird. Eine Analyse des Störungsverhaltens bzgl. der Randbedingungen kann jetzt wie oben durchgeführt werden. Als Resultat wird erhalten, daß kleine Störungen kleine Abweichungen in der Lösung verursachen. Auch ergibt sich, daß es keinen Sinn macht, die Randbedingungen genauer zu approximieren als den Differentialoperator. Dazu die

Merkregel 3.5. Ableitungen in den Randbedingungen sollten von der gleichen Ordnung approximiert werden wie die Differentialgleichung.

Um eine Approximation von $u'(b) = 0$ von zweiter Ordnung zu erhalten, wird der Hilfsgitterpunkt $x_{n+2} := x_{n+1} + \Delta x$ eingeführt und die Ableitung am Rand gemäß

$$0 = u'(x_{n+1}) \doteq \frac{1}{2\Delta x} \left[u_{n+2} - u_n\right] \tag{56}$$

approximiert. Die Differentialgleichung in (38) wird dann auch im Gitterpunkt x_{n+1} gemäß

$$-(k(x)u'(x))'_{n+1} \doteq \frac{1}{\Delta x^2} \left\{-k_{n+\frac{1}{2}} u_n + \left(k_{n+\frac{1}{2}} + k_{n+\frac{3}{2}}\right) u_{n+1} - k_{n+\frac{3}{2}} u_{n+2}\right\}$$

approximiert, wobei u_{n+2} mittels (56) substituiert wird. Für $\alpha \neq 0$ wird entsprechend vorgegangen. von erster Ordnung.
Zum Abschluß dieser Sektion noch etwas zur stationären **Konvektions-Diffusionsgleichung**

$$\left\{ \begin{array}{rcll} -\epsilon u''(x) + v u'(x) + a u(x) & = & 0 & \text{in } (0,1) \\ u(0) & = & \alpha & \\ u(1) & = & \beta, & \end{array} \right. \tag{57}$$

wobei $v \neq 0, a \geq 0$ vorausgesetzt wird und darüberhinaus $0 < \epsilon << |v|$ gelten soll.

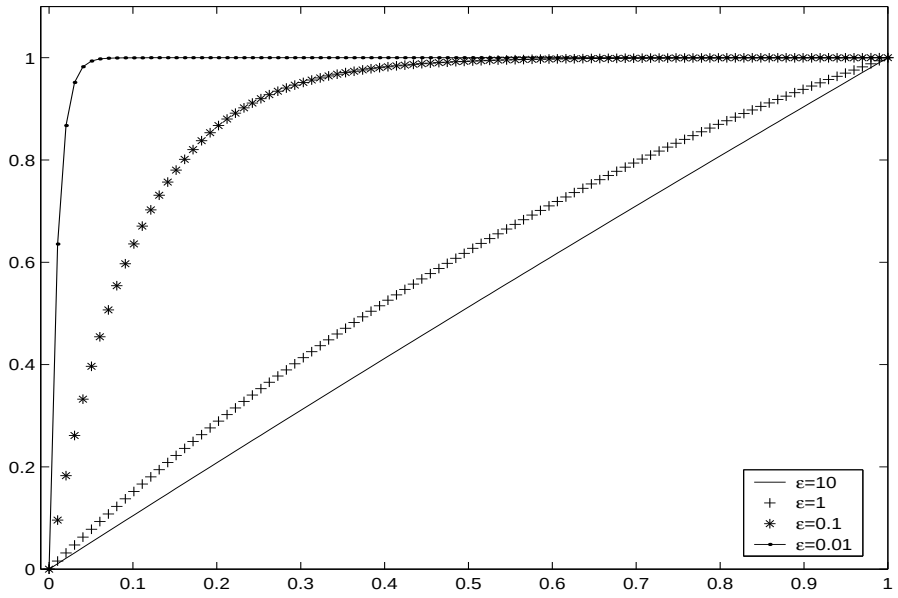


Abbildung 2: Funktion u^ϵ für verschiedene Werte von ϵ

Ist etwa $a \equiv 0, \alpha = 0, \beta = 1, v = -1$, so ist die exakte Lösung von (57) durch

$$u^\epsilon(x) = (1 - e^{-\frac{x}{\epsilon}})/(1 - e^{-\frac{1}{\epsilon}}). \tag{58}$$

gegeben, siehe Fig. 2. Werden zur numerischen Approximation des konvektiven Terms zentrale Differenzen verwendet, ergibt sich mit

$$\begin{aligned} \mathrm{D}^+ u_i &:= (u_{i+1} - u_i)/\Delta x \\ \mathrm{D}^- u_i &:= (u_i - u_{i-1})/\Delta x \\ \mathrm{D}^0 u_i &:= (u_{i+1} - u_{i-1})/(2\Delta x) \end{aligned}$$

und

$$-\epsilon\ \mathrm{D}^+\mathrm{D}^- u_i^\epsilon - \mathrm{D}^0 u_i^\epsilon = 0\ , \quad u_0 = 0\ ,\ u_{n+1} = 1,$$

daß

$$u_i^\epsilon = \frac{1 - d^i}{1 - d^{n+1}}, \quad d := \frac{2\epsilon - \Delta x}{2\epsilon + \Delta x}.$$

Es ergibt sich für das Lösungsverhalten

i.) Oszillation, falls $\Delta x > 2\epsilon$,

ii.) nicht die gewünschte Konvergenz, falls etwa $\epsilon = \Delta x$.

Im letzten Fall gilt

$$\lim_{\Delta x \to 0} u^\epsilon(x_1) = \lim_{n \to \infty} u^\epsilon(\frac{1}{n+1}) = 1 - \frac{1}{e},$$

aber

$$\lim_{\Delta x \to 0} u_1^\epsilon = \frac{2}{3}.$$

Ursache dafür ist das extreme Verhalten der Lösung für $\epsilon \to 0$, denn offensichtlich gilt punktweise

$$\lim_{\Delta x \to 0} u^\epsilon(x) = \begin{cases} 1, & x \in (0,1) \\ 0, & x = 0. \end{cases}$$

Bei $x = 0$ ergibt sich eine sogenannte Grenzschicht, siehe Fig 2. Hinsichtlich der Ausbildung von Grenzschichten gilt für Aufgaben des Typs (57) die

Merkregel 3.6. (Grenzschichten)
Ist $v(x) \geq v_0 > 0$, so bildet sich bei $x = 1$ eine Grenzschicht aus,
Ist $v(x) \leq v_0 < 0$, so bildet sich bei $x = 0$ eine Grenzschicht aus.

Schon bei der numerischen Behandlung von Transportgleichungen wurde festgestellt, daß zentrale Differenzen nicht ohne Zusatzmaßnahmen Verwendung finden sollten. So auch hier nicht.
Die oben aufgezeigten Eigenschaften schlagen sich auch in den Eigenschaften der Systemmatrix A zu (57) nieder. Die i-te Zeile von A hat die Gestalt

$$0, \ldots 0, \; -\frac{\epsilon}{\Delta x^2} - \frac{v_i}{2\Delta x}, \; \frac{2\epsilon}{\Delta x^2} + a_i, \; -\frac{\epsilon}{\Delta x^2} + \frac{v_i}{2\Delta x}, \; 0 \ldots 0,$$

so daß Diagonaldominanz nur für

$$\Delta x |v_i| \; \leq \; 2\epsilon \text{für alle } i$$

gewährleistet werden kann. Diese Forderung ist sehr restrikitv. Werden hingegen für die Diskretisierung von u' die Differenzen

$$\begin{cases} \mathrm{D}^+ u_i & , \quad v < 0 \\ \mathrm{D}^- u_i & , \quad v > 0 \end{cases} \tag{59}$$

benutzt, so ergibt sich etwa für $v > 0$ als i-te Zeile der Matrix A

$$0, \ldots 0, \; -\frac{\epsilon}{\Delta x^2} - \frac{v_i}{\Delta x}, \; \frac{2\epsilon}{\Delta x^2} + \frac{v_i}{\Delta x} + a_i, \; -\frac{\epsilon}{\Delta x^2}, \; 0, \ldots 0,$$

und damit Diagonaldominanz ohne Einschränkung an Δx. Auch haben Diagonal- und Nebendiagonalelemente verschiedene Vorzeichen mit negativen Nebendiagonalelementen. Dieses, zusammen mit der irreduziblen Diagonaldominanz, liefert ein hinreichendes Kriterium für die numerische Stabilität des Diskretisierungsverfahrens, weil die resultierende Systemmatrix dann eine M-Matrix ist.

Definition 3.7. (L- und M-Matrizen)
Sei $A = (a_{ij})_{i,j=1}^n, \in \mathbb{R}^{n \times n}$. Dann heißt A

i) L_0-Matrix, falls $a_{ij} \leq 0$ für $i \neq j$ gilt,

ii) L-Matrix, falls A L_0-Matrix und $a_{ii} > 0$ gilt

iii) M-Matrix, falls A L_0-Matrix, A^{-1}existiert und $A^{-1} \geq 0$

gilt.

Den Zusammenhang zwischen L- und M-Matrizen stellt der folgende Satz her.

Satz 3.8. (L- und M-Matrizen)
Sei A L-Matrix und

i) strikt diagonaldominant oder

ii) irreduzibel diagonaldominant,

dann ist A eine M-Matrix.

Beweis. [17] □

Im Falle der Upwind Diskretisierung (59) wird demnach eine M-Matrix erhalten, weil A irreduzibel diagonaldominant und L-Matrix ist, und zwar unabhängig von der Wahl von Δx relativ zu ϵ. Aus obigen Beobachtungen leitet sich die folgende Merkregel ab.

Merkregel 3.9. (Numerik bei Transportdominanz)
Die Diskretisierung des Transportterms in einer Konvektions-Diffusionsgleichung vom Typ (57) sollte nach den Anforderungen an die Diskretisierung der reduzierten Gleichung geschehen. Dabei ist die zu (57) gehörende reduzierte Gleichung durch

$$vu'(x) + au(x) = 0 \quad \text{in} \quad (a, b) \tag{60}$$

gegeben.

Nun kurz zur Konvergenz von Finite Differenzen Approximationen. Für solche Approxiamtionen kann die exakte Lösung u mit der numerischen Lösung nur auf dem Gitter verglichen werden. Folgende Definition macht demnach Sinn.

Definition 3.10. (Konvergenz von FD-Approximationen)
Ein Differenzenverfahren zur numerischen Lösung von (38) oder (57) heißt konvergent (von der Ordnung k) genau dann, wenn

$$\max_i |(R_{\Delta x}u)(x_i) - u_i| \to 0 \ (\leq c\Delta x^k) \tag{61}$$

für $\Delta x \to 0$ gilt.

In dieser Definition bezeichnet $R_{\Delta x}$ die Restriktion auf das Gitter. $R_{\Delta x}u$ ist demnach ein $n+2$-Vektor. Hinreichend für die Konvergenz der FD-Approximation ist die Konsistenz des FD-Schemas, zusammen mit der Stabilität des Verfahrens.
Es bezeichnen

$$\mathrm{L}u := -\epsilon u'' \ + \ vu' \ + \ au \tag{62}$$

den zu (56) gehörenden Differentialoperator und $A_{\Delta x} = A$ die durch eine FD-Approximation entstandene Systemmatrix, welche als Approximation des Differentialoperators L in (62) aufgefaßt wird.

Definition 3.11. (Konsistenz und Stabilität in der Maximumnorm)
Das FD-Verfahren heißt konsistent (von der Ordnung k), falls

$$|A_{\Delta x}(R_{\Delta x}u) - R_{\Delta x}Lu|_\infty \to 0 \ (\leq c\Delta x^k) \tag{63}$$

für $\Delta x \to 0$ gilt. Es heißt stabil (bezüglich der Matrixnorm), falls aus $A_{\Delta x}w = f$ die Abschätzung

$$|w|_\infty \leq c|f|_\infty \tag{64}$$

mit c unabhängig von Δx folgt.

Offensichtlich sind die oben beschriebenen Differenzenschemata konsistent und stabil, wobei Stabiltät aus der M-Matrix Eigenschaft folgt. Denn es gilt der

Satz 3.12. (M-Kriterium)
Sei A eine L_0-Matrix. Dann ist A eine M-Matrix genau dann, wenn es einen Vektor e gibt mit $Ae > 0$. Es gilt zudem die Abschätzung

$$\|A^{-1}\| \leq \frac{|e|}{\min_k (Ae)_k}.$$

Beweis. [10, Satz 1.5] □

Aus Stabilität und Konsistenz folgt die Konvergenz der Differenzenschemata.

3.2 Poissongleichung in 2 Raumdimensionen

In (6) sei jetzt $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet, u sei eine zeitunabhängige Funktion, $k \equiv diag(1,1)$, $v_T \equiv 0$. Dann ist (6) äquivalent zu

$$\begin{cases} -\Delta u(x) &= q(x) & \text{in } \Omega \\ u(x) &= r(x) & \text{auf } \partial\Omega \end{cases} \tag{65}$$

Diese Gleichung heißt **Poisson'sches Randwertproblem** mit **Dirichlet Randbedingungen**. Zunächst sei

$$\Omega := (0,1) \times (0,1).$$

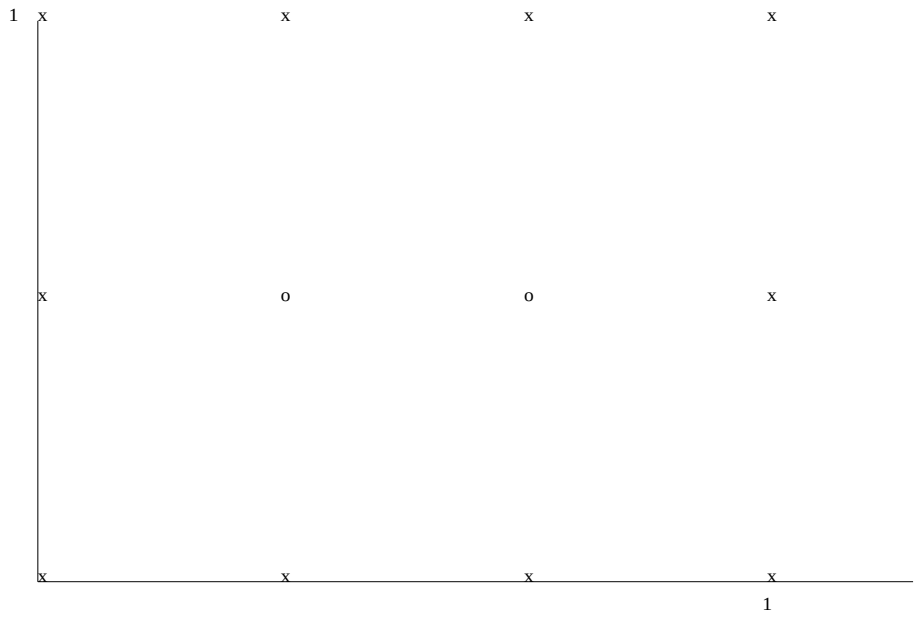


Abbildung 3: Gitter $\bar{\Omega}_{hk}$, o=Ω_{hk}, x=$\bar{\Omega}_{hk} \setminus \Omega_{hk}$

3.2.1 Finite Differenzen Approximation der Poisson Gleichung

Über $\bar{\Omega}$ wird zu $m, n \in \mathbb{N}$ mit $h := \frac{1}{n+1}$, $k := \frac{1}{m+1}$ gemäß

$$\bar{\Omega}_{hk} = \left\{ (x_1^i, x_2^j);\ x_1^i = ih, x_2^j = jk,\ i = 0, \ldots, n+1,\ j = 0, \ldots, m+1 \right\} \tag{66}$$

ein Gitter definiert. Zu beachten sind die unterschiedlichen Skalen von x_1 und x_2 in Abb. 3.2.1.
Gleichung (65) besitze eine eindeutig bestimmte Lösung. Mit

$$u_{ij} \approx u(x_1^i, x_2^j)$$

wird zunächst der 5-Punkte-Stern zur numerischen Approximation von (65) verwendet. Dazu werden die zweiten Ableitungen in die jeweiligen Koordinatenrichtungen mittels zweiter zentraler Finiter Differenzen approximiert. Als diskretes Analogon zu (65) ergibt sich

$$\begin{cases} \dfrac{-u_{i+1j} + 2u_{ij} - u_{i-1j}}{h^2} + \dfrac{-u_{ij+1} + 2u_{ij} - u_{ij-1}}{k^2} = q_{ij} & i = 1, \ldots, n\ ;\ j = 1, \ldots, m \\ u_{ij} = r_{ij} & \begin{array}{l} i = 0, n+1;\ j = 1, \ldots, m \\ j = 0, m+1;\ i = 1, \ldots, n. \end{array} \end{cases} \tag{67}$$

Sei ab jetzt $h = k$. Dann kann die erste Zeile in (67) in der Differenzensternformulierung prägnanter geschrieben werden als

$$\frac{1}{h^2} \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} u_h = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} q. \tag{68}$$

Satz 3.13. Konsistenz des 5-Punkte Sterns
Sei $u \in C^4(\bar{\Omega})$. Dann ist die Differenzenapproximation (68) der Poissongleichung (65) konsistent von zweiter Ordnung.

Beweis. Aufgabe 6.8. □

Bemerkung 3.14. In Satz 3.13 reicht auch $u \in C^{3,1}(\bar{\Omega})$.

Bei der numerischen Diskretisierung von (65) kann auch versucht werden, durch Hinzunahme weiterer Gitterpunkte die Konsistenzordnung zu erhöhen. Werden dabei nur benachbarte Gitterpunkte berücksichtigt, ist diese Strategie nicht von Erfolg gekrönt.

Satz 3.15. (Konsistenz kompakter 9-Punkte-Sterne)
Sei $u \in C^4(\bar{\Omega})$. Eine konsistente 9-Punkte-Diskretisierung von (64) ist von der Form

$$\frac{1}{h^2}\begin{bmatrix} \delta & r & \delta \\ r & \mu & r \\ \delta & r & \delta \end{bmatrix} u_h = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} q \tag{69}$$

mit $\mu + 4r + 4\delta = 0, r + 2\delta = -1$ und besitzt die Ordnung 2, welche nicht verbessert werden kann.

Bei 9-Punkte Schemata kann folglich die Konsistenzordnung nur durch verfeinerte Diskretisierung der rechten Seite q erzielt werden. Dazu das

Beispiel Sei in (68) $r = -\frac{2}{3}, \delta = -\frac{1}{6} \Rightarrow \mu = \frac{10}{3}$. Wird (69) durch

$$\frac{1}{h^2}\begin{bmatrix} -\frac{1}{6} & -\frac{2}{3} & -\frac{1}{6} \\ -\frac{2}{3} & \frac{10}{3} & -\frac{2}{3} \\ -\frac{1}{6} & -\frac{2}{3} & -\frac{1}{6} \end{bmatrix} u_h = \frac{1}{12}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 8 & 1 \\ 0 & 1 & 0 \end{bmatrix} q \tag{70}$$

ersetzt und ist $u \in C^6(\bar{\Omega})$, so ist (70) konsistent von der Ordnung 4. □

Als nächstes wird die Systemmatrix des 5-Punkte-Sterns hergeleitet. Dazu werden die Unbekannten u_{ij} zeilenweise in der Form

$$u_{11}\ u_{21}\ \dots\ u_{n1}\ u_{12}\ u_{22}\ \dots\ u_{n2}\ \dots u_{1n}\ u_{2n}\ \dots\ u_{nn}. \tag{71}$$

angeordnet. Dann ist mit

$$\mathrm{I} := \mathrm{diag}(-1, \dots, -1) \in \mathbb{R}^{n\times n},$$

und

$$\mathrm{T} := \begin{bmatrix} 4 & -1 & & \\ -1 & \ddots & \ddots & 0 \\ \ddots & \ddots & \ddots & -1 \\ 0 & \ddots & -1 & 4 \end{bmatrix}$$

(67) äquivalent zu dem block-tridiagonalen Gleichungssystem

$$\frac{1}{h^2}\begin{bmatrix} \mathrm{T} & \mathrm{I} & & & \\ \mathrm{I} & \ddots & \ddots & 0 & \\ & \ddots & \ddots & \ddots & \\ & 0 & \ddots & \mathrm{T} & \mathrm{I} \\ & & & \mathrm{I} & \mathrm{T} \end{bmatrix} u = \mathrm{R} \tag{72}$$

mit

$$\mathrm{R} = \Big[r_{11} + \frac{1}{h^2}(u_{01} + u_{10}), r_{21} + \frac{1}{h^2}u_{20}, \dots, r_{n1} + \frac{1}{h^2}(u_{n+11} + u_{n0}), r_{12} + \frac{1}{h^2}u_{02}, \dots,$$
$$r_{n2} + \frac{1}{h^2}u_{n+12}, \dots, r_{1n} + \frac{1}{h^2}(u_{0n} + u_{1n+1}), r_{2n} + \frac{1}{h^2}\ u_{2n+1}, \dots, r_{nn} + \frac{1}{h^2}(u_{n+1n} + u_{nn+1})\Big].$$

Dabei beachte man, daß die Struktur der Matrix von der Numerierung der Punkte abhängt.

3.2.2 Neumann Randbedingungen

In (65) werden jetzt die Randbedingungen geändert und **Neumann Randbedingungen** betrachtet:

$$\left\{\begin{array}{rcll} -\Delta u(x) & = & q(x) & \text{in } \Omega \\ \partial_\eta u(x) & = & r(x) & \text{auf } \partial\Omega. \end{array}\right. \tag{73}$$

Hier bezeichnet η die äußere Normale an Ω. An q und r muß die Bedingung

$$\int_\Omega q(x)dx \; - \int_{\partial\Omega} r(x)ds = 0 \tag{74}$$

gestellt werden, denn mittels partieller Integration folgt

$$\int_\Omega q(x) = -\int_\Omega \Delta u(x) = \int_{\partial\Omega} \partial_\eta u(x) = \int_{\partial\Omega} r(x)ds.$$

Die Lösung u von (73) kann nicht eindeutig bestimmt sein, denn mit u ist auch $u+$ Konstante eine Lösung von (73).
Zur Diskretisierung von (73) wird der Laplaceoperator wie in (67) mittels des Differenzensterns (68) diskretisiert. Die Approximation der Randbedingung in (73) geht dann wie folgt:

$$\left.\begin{array}{rcl} u_{0j} - u_{1j} & = & hr_{0j} \\ u_{n+1j} - u_{nj} & = & hr_{n+1j} \end{array}\right\} j = 1, \ldots, n,$$

$$\left.\begin{array}{rcl} u_{j0} - u_{j1} & = & hr_{j0} \\ u_{jn+1} - u_{jn} & = & hr_{jn+1} \end{array}\right\} j = 1, \ldots, n. \tag{75}$$

Damit ergeben sich aus (67) und (76) zusammen $n^2 + 4n$ Gleichungen für ebensoviele Unbekannte. Das System ist allerdings nicht eindeutig lösbar, da das homogene System nicht eindeutig lösbar ist. Damit ist es auch nicht für jede rechte Seite lösbar. Es ist zu erwarten, daß für die diskretisierten Daten q_{ij} und r_{ij} eine zu (74) analoge Beziehung zu gelten hat, wenn eindeutige Lösbarkeit gewährleistet werden soll. Zur Herleitung dieser Beziehung wird (74) mit Hilfe einfacher Quadratur näherungsweise hingeschrieben:

$$\begin{array}{rcll} \int_\Omega q(x)dx & \approx & h^2 & \sum_{i,j=1}^n q_{ij} \\ \int_\Omega -\Delta u(x)dx & \approx & & \sum_{i,j=1}^n u_{i-1j} + u_{i+1j} + u_{ij+1} + u_{ij-1} - 4u_{ij}. \end{array}$$

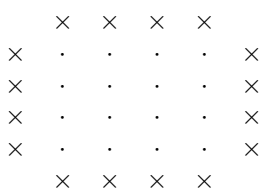
Bei der Summation heben sich Summanden, die zu inneren Gitterpunkten korrespondieren, gegenseitig auf. Umordnen in der letzten Summe ergibt schliesslich

$$\begin{array}{rcl} \int_\Omega -\Delta u(x) & \approx & \sum_{i=1}^n \overbrace{u_{i1} - u_{i0}}^{-hr_{i0}} + \overbrace{u_{in} - u_{in+1}}^{-hr_{in+1}} \\ & + & \sum_{j=1}^n \underbrace{u_{1j} - u_{0j}}_{-hr_{0j}} + \underbrace{u_{nj} - u_{n+1j}}_{-hr_{n+1,j}} \\ & = & -h \sum_{(x_i,y_j)\in\partial\Omega} r_{ij} \\ & \approx & -\int_{\partial\Omega} r(x)ds. \end{array}$$

Als diskretes Analogon zu (74) wird demnach

$$h^2 \sum_{i,j=1}^n q_{ij} + h \sum_{(x_i,y_j)\in\partial\Omega} r_{ij} = 0 \tag{76}$$

erhalten. Diese Identität wird nicht immer erfüllbar sein, für kleine Werte von h wird die linke Seite allerdings gegen 0 konvergieren.
Um auch die Randbedingung von zweiter Ordnung zu approximieren, werden Hilfspunkte eingeführt. Sei also Ω wieder das Einheitsquadrat. Dann wird das Gitter $\bar{\Omega}_{hk}$ gemäß der nachfolgenden Skizze ergänzt.



× Hilfspunkte

· Gitterpunkte

Bei der Verwendung von zentralen Differenzen zur Approximation der Randableitung wird die Herleitung von (76) damit allerdings ein wenig komplizierter.

Bemerkung 3.16. Werden die Daten q und r in (74) gemäß

$$\begin{aligned} q_{ij} &:= \frac{1}{h^2} \int_{x_i^2-h}^{x_i^2+h} \int_{x_i^1-h}^{x_i^1+h} q(x)dx, && i,j = 1,\dots,n \\ r_{il} &:= \frac{1}{h} \int_{x_i^1-\frac{h}{2}}^{x_i^1+\frac{h}{2}} r(x^1, x_l^2)dx^1, && l = 0, n+1 \\ r_{lj} &:= \frac{1}{h} \int_{x_j^2-\frac{h}{2}}^{x_j^2+\frac{h}{2}} r(x_l^1, x^2)dx^2, && l = 0, n+1 \end{aligned}$$

approximiert, so folgt (76) aus (74).

Der Nachweis der Konsistenz für die in (67) und (69) angegebenen Differenzenapproximation von Lösungen zu (65) ist eine leichte Übung. Die genannten Differenzenapproximationen sind auch stabil, d.h., die Systemmatrix A_h erfüllt

$$\|A_h^{-1}\|_\infty \leq C \text{ unabhängig von } h.$$

Diese Eigenschaft kann mit Hilfe diskreter Maximumprinzipien nachgewiesen werden, siehe etwa [10, S. 20 ff]. Hier ist das Konvergenzresultat:

Satz 3.17. (Konvergenz von Finite-Differenzen-Verfahren)
Die FD-Diskretisierung (68) von (64) sei konsistent von zweiter Ordnung. Dann gilt mit der üblichen Notation

$$\max_{0\leq i,j\leq n+1} |u_{ij} - u(x_i^1, x_j^2)| \leq Ch^2. \tag{77}$$

Die Konstante C hängt von den vierten Ableitungen von u und von dem Gebiet Ω ab.

3.2.3 Relaxationsverfahren

Charakteristisch für die bei Finite-Differenzen Approximationen auftretenden Matrizen ist deren **Dünnbesetztheit**. In der Regel sind die Matrizen symmetrisch und haben Bandstruktur. Die Bandbreite hängt dabei von der Numerierung der Gitterpunkte ab und muß nicht klein sein. Bei direkten Lösungsverfahren, etwa der Gauß-Elimination, kann demnach ein Fill-in im Profil der Matrix entstehen. Gerade bei hoher Anzahl von Gitterpunkten sind direkte Verfahren daher ungeeignet. Besser geeignet sind iterative Verfahren, auch, weil sie es erlauben, die numerische Lösung des Gleichungssystems nur bis auf eine vorgelegte Toleranz genau bereitzustellen.
Im Folgenden soll eine wichtige Klasse solcher Verfahren vorgestellt und deren Eigenschaften diskutiert werden. Sei dazu das Gleichungssystem

$$Ax = b, \quad A = L + D + R$$

vorgelegt. Dann gilt

$$\begin{aligned} (L+D+R)x &= b && \Longrightarrow \\ Dx &= b-(L+R)x && \Longrightarrow \\ x &= D^{-1}b - D^{-1}(L+R)x, \text{ falls } D^{-1} \text{ invertierbar.} \end{aligned}$$

Ausgehend von einem Startwert x^0 wird jetzt die (Fixpunkt-)Iteration

$$x^{i+1} = -D^{-1}(L+R)x^i + D^{-1}b \quad i \in \mathbb{N} \tag{78}$$

betrachtet. Dieses Verfahren heißt **Gesamtschritt- oder Jacobiverfahren**.
Wird gemäß

$$\begin{aligned} (L+D+R)x &= b && \Longrightarrow \\ x &= (L+D)^{-1}b - (L+D)^{-1}Rx, \text{ falls } (L+D)^{-1} \text{ existiert} \end{aligned}$$

umgeformt, so wird das **Einzelschritt- oder Gauß-Seidel Verfahren**

$$x^{i+1} = (L+D)^{-1}b - (L+D)^{-1}Rx^i \tag{79}$$

erhalten. Wie sich zeigen wird, hängt die Konvergenzgeschwindigkeit iterativer Verfahren vom Spektralradius der Iterationsmatrix ab. Durch **Relaxation** kann man versuchen, diesen zu verringern. Dazu schreibe (79) als

$$x^{i+1} = -D^{-1}(Lx^{i+1} + Rx^i - b) \tag{80}$$

und füge eine Korrektur mit Hilfe der alten Iterierten ein;

$$x^{i+1} = (1-w)x^i - wD^{-1}(Lx^{i+1} + Rx^i - b) \qquad \Longrightarrow$$

$$x^{i+1} = (D+wL)^{-1}[(1-w)D - wR]x^i + w(D+wL)^{-1}b. \tag{81}$$

Jetzt der Konvergenzsatz für iterative Verfahren.

Satz 3.18. (Konvergenz iterativer Verfahren)
Ein iteratives Verfahren der Form

$$x^{i+1} = Cx^i + b, \quad x^0 \quad \text{gegeben}$$

konvergiert genau dann für jeden Startwert x^0, wenn der **Spektralradius**

$$\rho(C) := \max\{|\lambda|; \lambda \text{ Eigenwert von } \} < 1$$

erfüllt.

Beweis. [4] □

Für strikt diagonaldominante oder irreduzibel diagonaldominante Matrizen sind die Iterationen (78) und (79) konvergent, denn es gilt der

Satz 3.19. (Hinreichende Bedingungen für Konvergenz von Gesamt- und Einzelschrittverfahren)
Ist A strikt diagonaldominant oder irreduzibel diagonaldominant, so konvergieren Gesamt- und Einzelschrittverfahren.

Beweis. Sei $A = (a_{ij})$ strikt diagonaldominant, d.h.,

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \text{ für alle } i.$$

Dann gilt für

$$G := -D^{-1}(L+R),$$

$$\|G\|_\infty = \max_i \sum_j |g_{ij}| < 1.$$

Wegen

$$u(G) \leq \|G\|_\infty$$

folgt für das Gesamtschrittverfahren die Behauptung.
Sei A irreduzibel diagonaldominant. Dann gilt

$$\|G\|_\infty \leq 1,$$

woraus nur $u(G) \leq 1$ geschlossen werden kann. Zeige mittels Beweis durch Widerspruch $u(G) < 1$. Sei dazu λ Eigenwert von G mit zugehörigem Eigenvektor $x \neq 0$, d.h.,

$$(L + R)x = -\lambda D x. \tag{82}$$

Setze

$$M := \max_j |x_j| \ (= |x|_\infty).$$

Dann kann nicht

$$|x_j| = M$$

für alle j gelten, weil sonst für den Index i_0, für welchen in der Matrix strikte Diagonaldominanz gültig ist,

$$|\lambda| \ |a_{i_0 i_0}| M \leq M \sum_{j \neq i_0} |a_{i_0 j}|$$

richtig wäre, ein Widerspruch. Sei jetzt x_{j_0} mit

$$|x_{j_0}| < M$$

und j_k ein beliebiger Index. Da A irreduzibel, gibt es Indizes $j_1, \ldots, j_{k-1}$ dergestalt, daß

$$a_{j_{i+1} j_i} \neq 0, \quad i = 0, \ldots, k-1$$

gilt. Für die Komponente j_1 in (82) gilt

$$\lambda a_{j_1 j_1} x_{j_1} = \sum_{l \neq j_1} a_{j_1 l} x_l.$$

Weil $a_{j_1 j_0} \neq 0$ und $|x_{j_0}| < M$, ergibt sich hier

$$|\lambda| \ |x_{j_1}| \ < \ M,$$

also entweder $|\lambda| < 1$ oder $|x_{j_1}| < M$. Ist Letzteres richtig, so kann induktiv

$$|x_{j_l}| < M$$

geschlossen werden, was, da j_l beliebig gewählt wurde, einen Widerspruch zu $|x|_\infty = M$ darstellt. Also kann $|\lambda| = 1$ nicht richtig sein, woraus die Behauptung folgt.
Der Beweis zum Einzelschrittverfahren ist etwa in [4] zu finden. □

Beispiel Entstammt die Matrix A etwa aus einer konsistenten 9-Punkte Diskretisierung wie in (69), ist sie wegen

$$\mu + 4\delta + 4r = 0$$

diagonaldominant. Irreduzibel ist sie auch. Strikte Diagonaldominanz liegt in den Zeilen vor, für deren Komponenten Dirichlet-Randbedingungen vorgeschrieben werden. □

Jetzt noch zu den Relaxationsverfahren (81) und dazu zunächst zwei Resultate, deren Beweis in [4] gefunden werden kann.

Satz 3.20. (Relaxationsparameter für das relaxierte Einzelschrittverfahren) (Kahan) [4, (8.3.5)Satz]
Soll die Iteration (81) konvergieren, muß schon $w \in (0,2)$ gelten.

Satz 3.21. (Konvergenz für positiv definite Matrizen) (Ostrowski, Reich) [4, (8.3.7)Satz]
Ist A positiv definit, so konvergiert das Relaxationsverfahren (80) für jedes $w \in (0,2)$.

Bemerkung 3.22. Die aus der 5-Punkte-Diskretisierung des Laplaceoperators mit zeilenweiser Numerierung resultierende Matrix ist positiv definit (bei Dirichlet-Randwerten).

Bei der FD-Diskretisierung von elliptischen Randwertaufgaben stößt man häufig auf sogenannte konsistent geordnete Matrizen. Für diese Matrizenklasse können Konvergenzresultate für das Relaxationsverfahren bereitgestellt werden.

Definition 3.23. (Konsistent geordnete Matrix)
Eine Matrix $A = L + D + R$ heißt konsistent geordnet : $\Longleftrightarrow$ Die Eigenwerte von

$$B(\alpha) := \alpha D^{-1} L + \alpha^{-1} D^{-1} R \tag{83}$$

sind unabhängig von α.

Beispiel Sei

$$A = \begin{bmatrix} D_1 & A_{12} & 0 \\ A_{21} & \ddots & A_{n-1,n} \\ 0 & A_{nn-1} & D_n \end{bmatrix}, \quad D_i \text{ diagonal, regulär.} \tag{84}$$

Dann gilt mit

$$S_\alpha := \begin{bmatrix} I_1 & & \\ & \alpha I_2 & 0 \\ & \ddots & \\ 0 & & \alpha^{n-1} I_n \end{bmatrix}$$

offensichtlich

$$B(\alpha) = S_\alpha B(1)\ S_\alpha^{-1}, \tag{85}$$

so daß $B(\alpha)$ und $B(1)$ für alle $\alpha \neq 0$ dieselben Eigenwerte besitzen. □

Matrizen der Form (84) werden bei der 5-Punkte Diskretisierung des Laplaceoperators auf dem Einheitsquadrat etwa bei diagonaler und schachbrettartiger Numerierung erhalten, nicht allerdings bei zeilen- oder spaltenweiser Numerierung in derselben Richtung.
Die Matrix aus (72), welche aus der Diskretisierung des Dirichletproblems (65) unter Verwendung von (68) bei zeilenweiser Numerierung in derselben Richtung resultiert, ist aber auch konsistent geordnet. Das kann wie folgt eingesehen werden. Ist x Eigenvektor von $B(\alpha)$ zum Eigenwert λ,

$$B(\alpha)x = \lambda x,$$

ergibt sich für die (ij)-te Komponente

$$\frac{\alpha}{4}\left[-x_{i-1j} - x_{ij-1}\right] + \frac{1}{4\alpha}\left[-x_{i+1,j} - x_{ij+1}\right] = \lambda \quad x_{ij}.$$

Mit der Transformation

$$\tilde{x}_{ij} := \alpha^{-(i+j)} x_{ij}$$

kann diese Identität umgeschrieben werden zu

$$\frac{1}{4}[-\tilde{x}_{ij-1} - \tilde{x}_{i-1j}] + \frac{1}{4}[-\tilde{x}_{i+1j} - \tilde{x}_{ij+1}] = \lambda \quad \tilde{x}_{ij},$$

also

$$B(1)\tilde{x} = \lambda\tilde{x}.$$

Zeilenweise Numerierung bei wechselnden Richtungen führt in diesem Anwendungsfall nicht auf eine konsistent geordnete Matrix, wie das Beispiel

$$\begin{array}{ccccc} 20 & 19 & 18 & 17 & 16 \\ 11 & 12 & 13 & 14 & 15 \\ 10 & 9 & 8 & 7 & 6 \\ 1 & 2 & 3 & 4 & 5. \end{array}$$

zeigt.

Bemerkung 3.24. Im Gegensatz zum Gesamtschrittverfahren (78) ist das Konvergenzverhalten des relaxierten Einzelschrittverfahrens (79) abhängig von der Numerierung.

Für konsistent geordnete Matrizen können die Eigenwerte der Iterationsmatrix des relaxierten Einzelschrittverfahrens

$$E(\omega) := (D + \omega L)^{-1}[(1-\omega)D - \omega R] \tag{86}$$

aus denen der Gesamtschrittmatrix

$$G := -D^{-1}(L + R) \tag{87}$$

berechnet werden, denn es gilt der

Satz 3.25. (Eigenwerte konsistent geordneter Matrizen)
Sei A konsistent geordnet und sei $\omega \neq 0$. Dann ist $\lambda \neq 0$ Eigenwert von $E(\omega)$ genau dann, wenn ein Eigenwert μ der Matrix G existiert mit

$$\frac{(\lambda + \omega - 1)^2}{\lambda} = \omega^2\mu^2. \tag{88}$$

Beweis. Es gilt

$$\frac{1}{\omega\lambda^{\frac{1}{2}}}(I - \omega D^{-1}L)(\lambda I - E(\omega)) = \{(\lambda+\omega-1)I - \omega(\lambda D^{-1}L + D^{-1}R)\}\frac{1}{\omega\lambda^{\frac{1}{2}}} = \frac{(\lambda+\omega-1)I}{\omega\lambda^{\frac{1}{2}}} - \frac{\omega\lambda^{\frac{1}{2}}}{\omega\lambda^{\frac{1}{2}}}(\lambda^{\frac{1}{2}}D^{-1}L + \lambda^{-\frac{1}{2}}D^{-1}R).$$

Damit folgt unter Verwendung von $\det(AB) = \det(A)\,\det(B)$

$$\det(\lambda I - E(\omega)) \;=\; 0 \iff \det\left[\frac{\lambda+\omega-1}{\omega\lambda^{\frac{1}{2}}}I - (\lambda^{\frac{1}{2}}D^{-1}L + \lambda^{\frac{-1}{2}}D^{-1}R)\right] = 0.$$

Da A konsistent geordnet ist, ist das wiederum äquivalent zu

$$\det\left[\frac{\alpha+\omega-1}{\omega\lambda^{\frac{1}{2}}}I - (D^{-1}(L+R)\right] = 0,$$

d.h.,

$$\frac{\lambda+\omega-1}{\omega\lambda^{\frac{1}{2}}} \text{ und } -\frac{\lambda+\omega-1}{\omega\lambda^{\frac{1}{2}}}$$

sind Eigenwerte von $-G$ (A konsistent geordnet!). Damit ergibt sich die Behauptung. □

Es ist leicht einzusehen, daß für konsistent geordnete Matrizen mit μ auch $-\mu$ Eigenwert von G ist. Sind dann die Eigenwerte μ_j der Matrix G aus (87) bekannt, so ergeben sich die Eigenwerte λ_j der Matrix $E(\omega)$ durch auflösen in (88) zu

$$\lambda_j = \left[\frac{\omega}{2}\mu_j \pm \sqrt{\frac{1}{4}\omega^2\mu_j^2 + 1 - \omega}\right]^2. \tag{89}$$

Der betragsmäßig größte Eigenwert ergibt sich für $\mu_j = u(G)$, ergo (vorausgesetzt, daß μ_j reell)

$$u(E(\omega)) = \left[\frac{\omega}{2}u(G) + \sqrt{\frac{1}{4}\omega^2 u(G)^2 + 1 - \omega}\right]^2. \tag{90}$$

Daraus ergibt sich unmittelbar die

Folgerung 3.26. (Gesamtschrittverfahren = $\frac{1}{2}$ Einzelschrittverfahren)
Es gilt

$$u(E(1)) = u^2(G). \tag{91}$$

Beweis. Folgt unmittelbar aus (89) mit $\omega = 1$. □

Mit anderen Worten: Das Einzelschrittverfahren konvergiert für konsistent geordnete Matrizen etwa doppelt so schnell wie das Gesamtschrittverfahren aus (78). Zudem kann aus (90) ein optimales $w^* \in (1,2)$ abgeleitet werden;

$$\omega^* = \frac{2}{1+\sqrt{1-u(G)^2}}, \quad u(E(\omega^*)) = \omega^* - 1 = \left(\frac{u(G)}{1+\sqrt{1-u(G)^2}}\right)^2, \tag{92}$$

denn für $0 < \omega \leq \omega^*$ ist $u(E(\omega))$ streng monoton fallend. Allgemeiner gilt

$$u(E(\omega)) = \begin{cases} \omega - 1 & \text{für } \omega^* \leq \omega < 2 \\ 1 - \omega + (1/2)\omega^2 u(G)^2 + \omega u(G)\sqrt{1-\omega+(1/4)\omega^2 u(G)^2} & \text{für } 0 < \omega \leq \omega^*. \end{cases}$$

3.3 Die Finite-Element Methode am Beispiel des Poisson Problems

3.3.1 Variationsform und abstraktes Galerkin Verfahren

Sei zunächst $\Omega \subset \mathbb{R}^2$ polygonal berandetes, beschränktes Gebiet. Wieder soll zu vorgelegter rechter Seite q mit Randwerten $r = 0$ die Aufgabe (65) numerisch gelöst werden. Dazu sei u eine Lösung von (65) und $v \in C_0^1(\Omega)$. Multiplikation der ersten Gleichung in (65) mit v, Integration über Ω und anschließende partielle Integration führen auf das **Variationsproblem**
Finde $u \in C^1(\bar{\Omega})$ mit $u(x) = r(x)$ auf $\partial\Omega$ und

$$\int_\Omega \nabla u(x) \nabla v(x) dx = \int_\Omega q(x)v(x)dx \quad \forall v \in C_0^1(\Omega). \tag{93}$$

Definition 3.27. (Variationsformulierung)
(93) ist die zu (65) gehörige **Variationsformulierung**.

Hier bezeichnet $C_0^1(\Omega)$ die Menge aller einmal stetig in Ω differenzierbaren Funktionen, deren Funktionswerte höchstens auf einer kompakt in Ω enthaltenen Teilmenge nicht verschwinden. Nun könnte man denken, daß umgekehrt eine Funktion u, welche (93) erfüllt, auch eine Lösung von (65) darstellt. Für hinreichend glatte Funktionen q ist das auch richtig. Gleichung (93) macht allerdings für alle Funktionen q Sinn, für welche der Integralausdruck auf der rechten Seite Sinn macht.
Die Idee des **Galerkin Verfahrens** zur numerischen Approximation von (65) ist jetzt, Lösungen von (93) in einem geeigneten endlichdimensionalen Raum W_h zu suchen und auch die Menge der **Testfunktionen**, $C_0^1(\Omega)$ in (93), durch einen endlich-dimensionalen Testraum V_h zu ersetzen. Der Index h steht hierbei für die Diskretisierungsfeinheit etwa einer Triangulierung Ω_h von Ω, dazu aber später mehr. Formal ergibt sich:
Finde $u_h \in W_h$ derart, daß

$$\int_{\Omega_h} \nabla u_h(x) \nabla v_h(x) dx = \int_{\Omega_h} q(x)v_h(x)dx \quad \forall v_h \epsilon V_h \tag{94}$$

gültig ist.
Im Folgenden gelte $\Omega_h = \Omega$, falls nichts anderes erwähnt wird. Die Wahl der Testfunktionen ist hier nicht unkritisch. Die Merkregel lautet

Merkregel 3.28. Testfunktionen müssen überall dort verschwinden (= 0 sein), wo die Werte der gesuchten Funktionen durch Dirichlet Randbedingungen festgelegt sind.

Zur weiteren Verdeutlichung dieses Sachverhaltes betrachte das Minimierungsproblem.

Finde u^* mit $u^* = 0$ auf $\partial\Omega$ und

$$\min_u \int_\Omega \frac{1}{2}|\nabla u(x)|^2 - q(x)u(x)dx = J(u^*) \quad \forall u, u_{|_{\partial\Omega}} = 0, \tag{95}$$

welches als Lösung eine Dichteverteilung u^* charakterisiert, deren kinetische Energie minimal ist. Wie kann u^* charakterisiert werden? Zunächst ist klar, daß

$$J(u^*) \leq J(u) \quad \forall u \neq u^*$$

mit $u_{|_{\partial\Omega}} = 0$ zu gelten hat. Ergo auch

$$J(u^*) \leq J(u^* + \epsilon v) \quad \forall v, v_{|_{\partial\Omega}} = 0, \epsilon \neq 0. \tag{96}$$

Damit folgt aber auch

$$0 = \frac{d}{d\epsilon} J(u^* + \epsilon v)_{|\epsilon=0}$$

und das wiederum ist äquivalent zu

$$\int_\Omega \nabla u^*(x)\nabla v(x)dx = \int_\Omega q(x)v(x)dx \quad \forall v, v_{|\partial\Omega} = 0.$$

Dabei sei bemerkt, daß (96) auch richtig bleibt, falls $u^*_{|\partial\Omega} \neq 0$ festgelegt ist.
Nebenbei liefert (96) noch eine variationelle Formulierung der Aufgabenstellung (65).
Jetzt zur Finite-Element Methode.

3.3.2 Finite Elemente in einer Raumdimension

Zur Motivation sei $n = 1$ und $\Omega = (0, 1)$. Definiere als Testraum

$$V_h := \left\{ v \in C^0(\bar{\Omega}); v_{|_{I_j}} \in \mathcal{P}_1 \text{ für } j = 0, \ldots, n \right\}, I_j := [x_j, x_{j+1}),$$

die Menge aller auf $\bar{\Omega}$ stetigen Funktionen, welche eingeschränkt auf jedes Teilintervall I_j einer Zerlegung $\Omega_h := \cup_{j=0}^n I_j$ von Ω mit Gitterweite $h = \frac{1}{n+1}$ ein Polynom ersten Grades darstellen. Eine Basis von V_h ist durch

$$b_i(x) := \begin{cases} 1, & x = x_i \\ 0, & x \in \Omega \setminus [x_{i-1}, x_i] \cup [x_i, x_{i+1}] \end{cases} \quad i = 1, \ldots, n$$

und der Forderung $b_i \in V_h$ festgelegt. In (94) wird jetzt $W_h = V_h$ gewählt und der Ansatz

$$u_h(x) := \sum_{i=1}^n u_i b_i(x)$$

für u_h in (94) eingesetzt. Es ergibt sich ein lineares Gleichungssystem

$$A_h u = q,$$

wobei

$$A_h = (a_{ij})_{i,j=1}^n, \quad a_{ij} = \int_\Omega b_i'(x)b_j'(x)dx,$$

und

$$q \;=\; (q_1, \ldots, q_n)^t, \quad q_i \;:=\; \int_\Omega q(x) b_i(x) dx.$$

Die Matrix A_h ist in diesem Fall positiv definit (Nachweis!), die Einträge a_{ij} der Matrix A_h können in diesem Fall exakt, die Einträge q_i der rechten Seite etwa mittels einer Quadraturformel (vergleiche (112)) numerisch berechnet werden.
Analog soll jetzt für das zweidimensionale Problem vorgegangen werden. Dazu müssen die Räume V_h und W_h in (94) in geeigneter Weise definiert werden. Als erstes wird dazu das Gebiet Ω in Teilgebiete zerlegt.

3.3.3 Triangulierungen

Es gelte

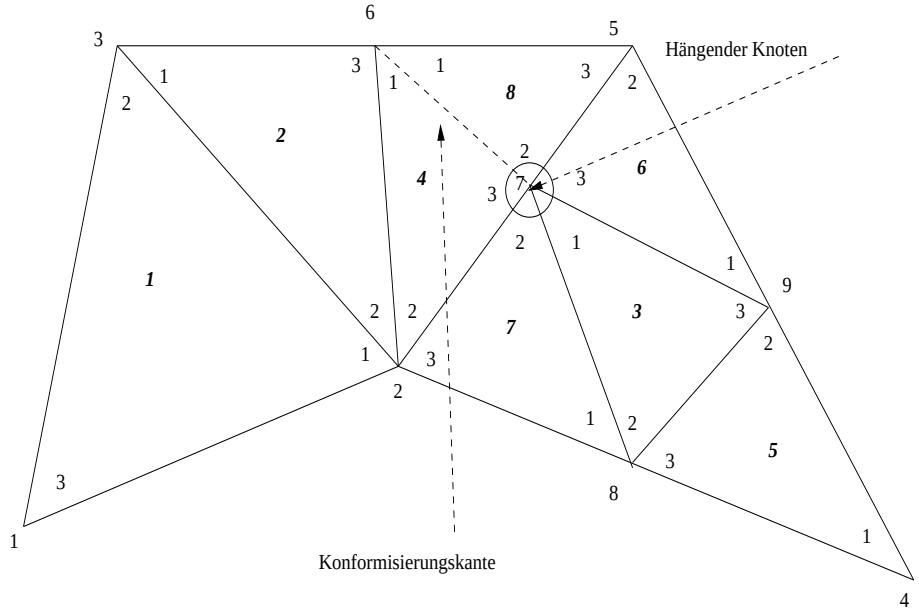


Abbildung 4: Triangulierung mit hängendem Knoten, Konformisierungskante, lokaler und globaler Numerierung

$$\bar{\Omega} = \bigcup_{j=1}^{m} \bar{T}_j \;,\; T_i \cap T_j = \emptyset \;,\; i \neq j, \tag{97}$$

d.h., Ω sei in disjunkte Teilgebiete zerlegt, vergleiche Fig. 4.

Definition 3.29. (Zulässige Zerlegungen)
Eine Zerlegung von Ω der Form (97) heißt zulässig: $\Longleftrightarrow$ Zwei beliebige Teilgebiete T_i, T_j erfüllen genau eine der Bedingungen

$$\left\{ \begin{array}{rlll} \text{i)} & T_i & = & T_j \\ \text{ii)} & \bar{T}_i & \cap & \bar{T}_j \quad \text{Kante} \\ \text{iii)} & \bar{T}_i & \cap & \bar{T}_j \quad \text{Ecke} \\ \text{iv)} & \bar{T}_i & \cap & \bar{T}_j \quad \text{leer.} \end{array} \right. \tag{98}$$

Im Folgenden bestehen Zerlegungen ausschließlich aus Dreiecken und heissen **Triangulierungen**. Die Anzahl der Dreiecke einer Triangulierung wird nachfolgend mit nt bezeichnet, die Anzahl der Ecken mit nv.

Als nächstes sollen Methoden angegeben werden, wie aus einer gegebenen zulässigen Triangulierung

$$Z_h = \{T_1, \ldots, T_{nt}\} \tag{99}$$

mit Gitterweite

$$h := \max_{1 \le i \le nt} \operatorname{diam}(T) \tag{100}$$

eine zulässige Triangulierung mit geringerer Gitterweite erzeugt werden kann.

Algorithmus 3.30. (Kongruentes Verfeinern)

1. $Z^0_{h_0} = \{T_1, \ldots, T_{nt_0}\}$, $l = 0$, $k_{max} \ge 0$ gegeben (Makrotriangulierung).
2. Teile jedes $T \in Z_{h_l}$ in 4 kongruente Dreiecke.
3. $Z^0_{h_l}$ = neue Menge von Dreiecken.
4. $l = l + 1$
5. $l = k_{max}$ stop, sonst gehe nach 2.

Der Verwaltungsaufwand einer Triangulierung in zwei Raumdimensionen ist wesentlich höher als der für eindimensionale Gitter. Ist eine Makrotriangulierung $Z^0_{h_0}$ mit nt Dreiecken und nv Knoten gegeben, so werden zu deren Verwaltung die Felder

$x(2,i)$	$i = 1, \ldots, nv$	Knotenkoordinaten
$node(j,i)$	$j = 1,2,3,\ i = 1, \ldots, nt$	Lokale zu globaler Numerierung
$edge(j,i)$	$j = 1,2,3,\ i = 1, \ldots, nt$	Lage der Dreiecke zueinander
$ivert(i)$	$i = 1, \ldots, nv$	Lage globaler Knoten in der Triangulierung

eingeführt. Der Zähler j entspricht dabei der lokalen Numerierung eines Dreiecks. Die Felder seien wie folgt belegt, vergleiche Fig. 4:

$$\left\{ \begin{array}{l} node(j,i) = \quad k, \text{falls, der } j\text{-te Knoten des } i\text{-ten Dreiecks die globale Nummer } k \text{ hat} \\ \\ edge(j,i) = \left\{ \begin{array}{ll} k, & \text{falls der } j\text{-te Knoten des } i\text{-ten Dreiecks gegenüber} \\ & \text{von Dreieck } k \text{ liegt} \\ 0, & \text{falls der } j\text{-te lokale Knoten des } i\text{-ten Dreiecks gegenüber} \\ & \text{dem Dirichletrand liegt} \\ -1, & \text{falls der } j\text{-te lokale Knoten des } i\text{-ten Dreiecks gegenüber} \\ & \text{dem Neumannrand liegt} \end{array} \right. \\ \\ ivert(i) = \left\{ \begin{array}{ll} 1, & \text{falls } i\text{-ter globaler Knoten in } \Omega \text{ liegt} \\ 0, & \text{falls } i\text{-ter globaler Knoten auf dem Dirichletrand liegt} \\ -1, & \text{falls } i\text{-ter globaler Knoten auf dem Neumannrand liegt} \end{array} \right. \end{array} \right. \tag{101}$$

Bemerkung 3.31. Dirichleträndер sind immer abgeschlossen, Neumannränder immer offen. Treten innerhalb des Dirichletrandes Unterteilungen auf, so ist immer der Teil des Randes mit homogenen Randwerten abgeschlossen.

Ein weiterer Verfeinerungsalgorithmus basiert auf Bisektion von Dreiecken. Dabei wird zunächst jedem Dreieck T_i einer vorgelegten Triangulierung $Z_h = \{T_{1,\ldots,}T_{nt}\}$ eine Verfeinerungskante zugeordnet, welche bei Bisektion gemäß Fig. (5) vererbt wird. In der Formulierung des Algorithmus' bezeichnet nnt den aktuellen Zähler für die Dreiecke während des Verfeinerungsprozesses. Das Feld $list(1:nt)$ enthält einen Zeiger auf alle Dreiecke, die verfeinert werden sollen.

$$list(i) = \left\{ \begin{array}{lll} 1 & , & i\text{-tes Dreieck wird verfeinert} \\ 0 & , & i\text{-tes Dreieck wird nicht verfeinert} \end{array} \right. , \quad i = 1, \ldots, nt \tag{102}$$

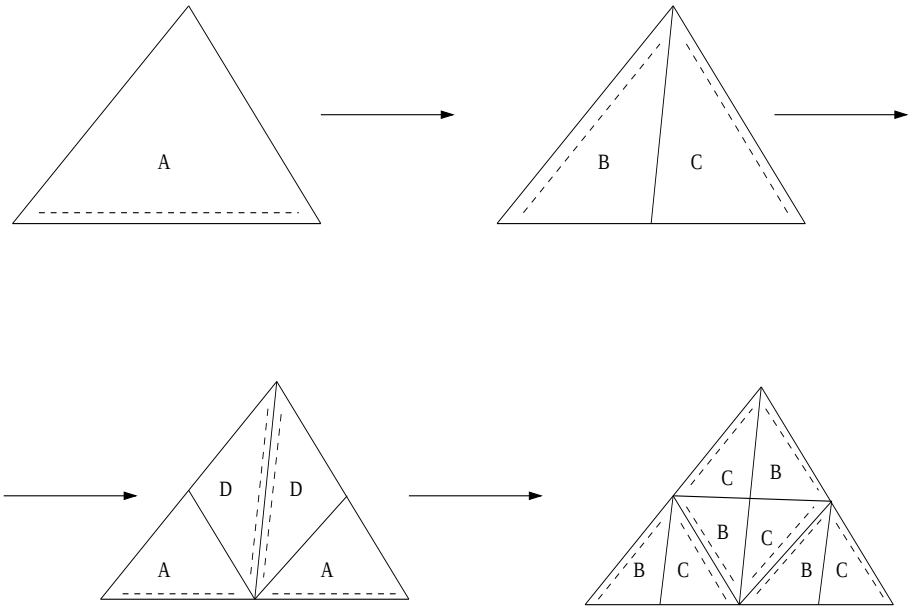


Abbildung 5: Bisektion, Vererbung der Verfeinerungskanten und ähnliche Dreiecke

Algorithmus 3.32. (Verfeinerung mittels Bisektion)

1. $Z_h^0 = \{T_{1,\ldots,}T_{nt}\}$, k gewünschte Anzahl von Verfeinerungen, $l = 0$, *list* enthält die Verfeinerungskanten.
2. α) Halbiere jedes Dreieck in Z_h^l nach seiner Verfeinerungskante hin, siehe Abb. 5 und datiere *list* auf.
   β) $H = \{T; T$ besitzt hängenden Knoten$\}$â
   γ) Ist $H \neq \emptyset$, verfeinere H
3. Bezeichne mit Z_h^{l+1} die neue Menge von Dreiecken
4. $l = l + 1$
5. Ist $l < k$, gehe nach 2., sonst stop.

Hängende Knoten sind in Abb. 4 erläutert.

Hilfsatz 3.33. Punkt 2. von Algorithmus 3.32 garantiert nach der zweiten Iteration eine reguläre Triangulierung Z_h^{l+1}.

Beweis. Folgt unmittelbar aus der Zuweisung von Verfeinerungskanten gemäß Abb. 5 □

Die Algorithmen 3.30 und 3.32 eignen sich auch zur lokalen Verfeinerung von gegebenen Triangulierungen. Zur Verwaltung von lokaler Verfeinerung wird das Feld $list(1 : nt)$ (102) verwendet. Es enthält einen Zeiger auf die Dreiecke der Triangulierung Z_h, welche verfeinert werden sollen.

Algorithmus 3.34. (Kongruent mit lokaler Verfeinerung)

1. $Z_h^0 = \{T_{1,\ldots,}T_{nt}\}, k$ Anzahl der Verfeinerungen, $list(1 : nt)$ Zeiger auf zu verfeinernde Dreiecke, $l = 0$
2. Teile alle Dreiecke in *list* in 4 kongruente Dreiecke und beseitige hängende Knoten durch konformen Abschluß, siehe Abb. 4. Wird dabei ein Dreieck halbiert, das bereits aus einem konformen Abschluß hervorgegangen ist, so löse die Situation rückwirkend auf und verfeinere auch dieses Dreieck kongruent.

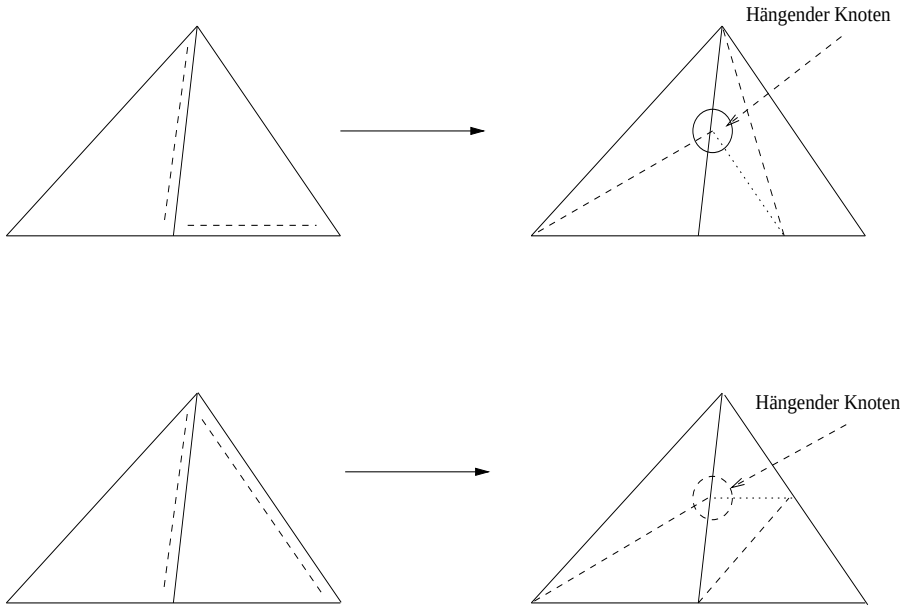


Abbildung 6: Mögliche Teilungen bei Bisektion

3. Z_h^l = neue Menge von Dreiecken.
4. Datiere $list$ auf, $l = l + 1$
5. gehe zu 2., falls $l < k$

Algorithmus 3.35. (Bisektion mit lokaler Verfeinerung)

1. $Z_h^0 = \{T_{1,\ldots,}T_{nt}\}, k$ Anzahl der Verfeinerungen, $list(1 : nt)$ Zeiger auf zu verfeinernde Dreiecke, $l = 0$
2. Halbiere jedes Dreieck in $list$ und verfeinere danach alle Dreiecke mit hängenden Knoten nach ihrer Verfeinerungskante. Beseitige hängende Knoten.
3. Z_h^l = neu Menge von Dreiecken.
4. Datiere $list$ auf, $l = l + 1$
5. gehe zu 2., falls $l < k$

Die in Algorithmus 3.35 auftretenden kritischen Situationen sind in Abb. 6) angedeutet. Dabei ist der konforme Abschluß durch gestrichelte Linien angedeutet. Die Aufdatierung der Numerierung während des Verfeinerungsprozesses ist in Abb. 4 für Bisektion dargestellt.
Zu einer gegebenen Triangulierung Z_h seien

$$\text{(103)} \qquad \begin{cases} h_T & := & \operatorname{diam}(T); T \in Z_h \\ \rho_T & := & \sup\{\operatorname{diam}(K); \bar{K} \subset T \in Z_H\} \\ \sigma_T & := & h_T/\rho_T. \end{cases}$$

Definition 3.36. (stabile Triangulierungen)
Eine Folge $\left\{Z_{h_k}^k\right\}_{k\in\mathbf{N}}$ von Triangulierungen heißt stabil: $\Longleftrightarrow$

$$\exists c, C > 0 \; \forall_{k\in\mathbb{N}} \; : \; c \le \sigma_T^k \le C \; \forall \, T \in Z_{h_k}^k .$$

Setze

$$\left\{\begin{array}{lcl} h & := & \sup\limits_{T\in Z_h} h_T \\ \rho & := & \inf\limits_{T\in Z_h} \rho_T \\ \sigma & := & h/\rho. \end{array}\right. \tag{104}$$

Es gilt der

Satz 3.37. (Bisektion und kongruentes Teilen sind stabil)
Sei Ω polygonal berandet und Z_h^0 eine zulässige Makrotriangulierung. Dann erzeugen die Algorithmen 3.30, 3.32, 3.34 und 3.35 Folgen stabiler Triangulierungen.

Beweis. Für die Algorithmen 3.30, 3.34 sei

$$h^0 := \sup_{T\in Z_h^0} h_T,\ \rho^0 := \inf_{T\in Z_h^0} \rho_T \text{ und } \sigma^0 := h^0/\rho^0.$$

Dann gilt sicher

$$c \le \sigma^0 \le C$$

und da alle erzeugten Dreiecke kongruent sind, gilt das auch für σ^k, $k \in \mathbb{N}$.
Bei den Algorithmen 3.32, 3.35 ist die Situation etwas komplizierter. Sofort kann allerdings eingesehen werden, daß aus einem Dreiecke während des Verfeinerungsprozesses höchstens vier neue Dreieckstypen entstehen können, siehe Abb. 5. Daher gilt

$$2^{-k}\min\{h_A,h_B,h_C,h_D\} \le h_T^k \le 2^{-k}\max\{h_A,h_B,h_C,h_D\}$$

und

$$2^{-k}\min\{\rho_a,\rho_b,\rho_c,\rho_d\} \le \rho_T^k \le 2^{-k}\max\{\rho_A,\rho_B,\rho_C,\rho_D\}$$

und somit

$$\underbrace{\frac{\min\{h_A,h_B,h_C,h_D\}}{\max\{\rho_A,\rho_B,\rho_C,\rho_D\}}}_{=:c} \le \sigma^k \le \underbrace{\frac{\max\{h_A,h_B,h_C,h_D\}}{\min\{\rho_A,\rho_B,\rho_C,\rho_D\}}}_{=:C}$$

□

Das Aufdatieren der Felder *node* und *edge* soll an einem Beispiel erläutert werden, vergleiche Abb. 4.
Es werde das Dreieck mit der globalen Nummer i behandelt. Dann gilt:

$$edge(j, nnt+j) = i, \quad node(j, nnt+j) = node(j,i), \quad j = 1,2,3$$

Prüfe, ob ein Dreieck gegenüber liegt und schon behandelt wurde oder ob es sich um ein Randstück handelt:

$$edge(j,i) < i \text{ (gegenüberliegendes Dreieck noch nicht behandelt) oder } edge(j,i) \le 0$$

Im Randfall oder falls das gegenüberliegenden Dreieck noch nicht behandelt wurde, erzeuge neuen Knoten:

$$x(l, nnv+j) = \frac{1}{2}\left(x(l, node(j\oplus 1, i)) + x(l, node(j\oplus 2, i))\right).$$

Dabei ist $\oplus = +mod3$. Datiere *node* und *edge* auf:

$$\left.\begin{array}{l} node(j,i) \\ node(j\oplus 1, nnt+j\oplus 2) \\ itnoch(j\oplus 2, nnt+j\oplus 1) \end{array}\right\} = nnv+j,$$

$$\left.\begin{array}{l} edge(j, nnt+j\oplus 1) \\ edge(j, nnt+j\oplus 2) \end{array}\right\} = edge(j,i) \text{ Randfall}$$

$edge(j,i) > i$, d.h. gegenüberliegendes Dreieck bereits behandelt.

$$\left.\begin{array}{l} node(j,i) \\ node(j\oplus 1, nnt+j\oplus 2) \\ node(j\oplus 2, nnt+j\oplus 1) \end{array}\right\} = \text{ bereits erzeugter Knoten,}$$

setze $ie := edge(j,i)$. Ist dann $edge(l,ie) = i$, so setze $ieg = edge(l,ie)$ und

$$edge(j, nnt+j\oplus 1) = nt + 3*(ieg-1) + l\oplus 2, \quad edge(j, nnt+j\oplus 2) = nt+3*(ieg-1)+l\oplus 1.$$

Zuletzt datiere *edge* der alten Dreiecke auf:

$$edge(j,i) = nnt + j, \quad j = 1,2,3.$$

3.3.4 Finite Element Räume

Als nächstes werden zu einer gegebenen Triangulierung Finite Element Räume konstruiert. Diese Räume sollen aus Funktionen bestehen, die, eingeschränkt auf jedes Dreieck der Triangulierung, Polynome in zwei Veränderlichen darstellen und global stetig sind. Als Basis sollen Funktionen mit lokalen Trägern konstruiert werden, ähnlich den Hütchenfunktionen in einer Raumdimension.
Betrachte zunächst die Situation auf einem Dreieck $T(P_0, P_1, P_2)$ mit Eckpunkten $P_0(x_1^0, x_2^0)$, $P_1(x_1^1, x_2^1)$, $P_2(x_1^2, x_2^2)$. Dieses Dreieck kann bijektiv auf das Einheitsdreieck $\hat{T}$ mit den Ecken $(0,0),(1,0),(0,1)$ abgebildet werden, siehe Abb.7. Dabei ist $F_T : \hat{T} \to T$ gegeben durch

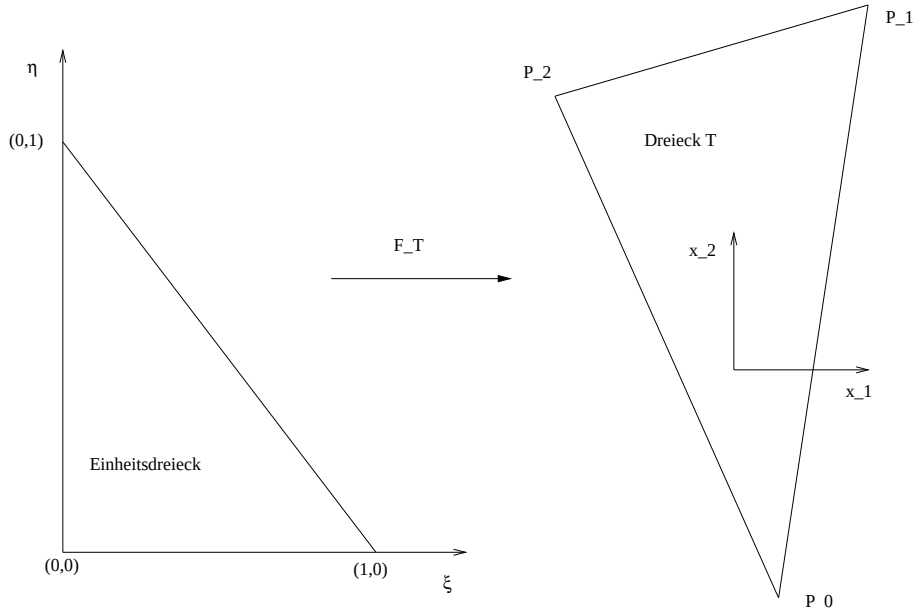


Abbildung 7: Abbildung auf das Einheitsdreieck

$$F_T(\xi,\eta) := \begin{bmatrix} x_1^0 \\ x_2^0 \end{bmatrix} + \xi \begin{bmatrix} x_1^1 - x_1^0 \\ x_2^1 - x_2^0 \end{bmatrix} + \eta \begin{bmatrix} x_1^2 - x_1^0 \\ x_2^2 - x_2^0 \end{bmatrix} \tag{105}$$

mit

$$|DF_T(\xi,\eta)| = \det \begin{bmatrix} x_1^1 - x_1^0 & x_1^2 - x_1^0 \\ x_2^1 - x_2^0 & x_2^2 - x_2^0 \end{bmatrix} = 2|T| \neq 0. \tag{106}$$

Ferner gilt natürlich $F_T(0,0) = P_0$, $F_T(1,0) = P_1$, $F_T(0,1) = P_2$. Ist jetzt $\hat{t}$ ein Polynom über $\hat{T}$, so ist wegen der Affinität von F_T auch $t := \hat{t} \circ F_T^{-1}$ ein Polynom vom gleichen Höchstgrad über T. Die Differentialoperatoren rechnen sich beim Koordinatenwechsel wie folgt um

$$\begin{bmatrix} \partial x_1 \\ \partial x_2 \end{bmatrix} = \begin{bmatrix} \xi_{x_1} & \eta_{x_1} \\ \xi_{x_2} & \eta_{x_2} \end{bmatrix} \begin{bmatrix} \partial \xi \\ \partial \eta \end{bmatrix}, \tag{107}$$

wobei

$$\begin{array}{rclcrcl} \xi_{x_1} &=& \frac{1}{2}(x_2^2 - x_2^0)/|T| &,& \eta_{x_1} &=& -\frac{1}{2}(x_2^1 - x_2^0)/|T|, \\ \xi_{x_2} &=& -\frac{1}{2}(x_1^2 - x_1^0)/|T| &,& \eta_{x_2} &=& \frac{1}{2}(x_1^1 - x_1^0)/|T|. \end{array}$$

Nun zu den Freiheitsgraden von Polynomen über Dreiecken. In den nachfolgenden Beispielen bezeichnen $P_i, 1 = 1, \ldots, 6$ die Ecken $(i = 1, 2, 3)$ bzw. die Seitenmittelpunkte $(i = 4, 5, 6)$ des jeweiligen Dreiecks

Beispiel Lineare Elemente (Courant 1950)

$$t_i(x_1, x_2) = a + bx_1 + cx_2, \quad t_i(P_j) \stackrel{!}{=} \delta_{ij}.$$

Es ergeben sich 3 Freiheitsgrade. Im Einheitsdreieck $\hat{T}$ ergeben sich die **baryzentrischen Koordinaten**

$$\begin{array}{rclcrcl} \lambda_1 &=& \hat{t}_1(\xi, \eta) &=& 1 - \xi - \eta \\ \lambda_2 &=& \hat{t}_2(\xi, \eta) &=& \xi \\ \lambda_3 &=& \hat{t}_3(\xi, \eta) &=& \eta \end{array} \tag{108}$$

□

Beispiel Quadratischer Ansatz

$$t_i(x_1, x_2) = a + bx_1 + cx_2 + dx_1x_2 + ex_1^2 + fx_2^2, \quad t_i(P_j) \stackrel{!}{=} \delta_{ij}.$$

Dabei gilt mit (106)

$$\begin{array}{rclcrcl} \hat{t}_1(\xi, \eta) &=& (1 - \xi - \eta)(1 - 2\xi - 2\eta) &=& \lambda_1(2\lambda_1 - 1) \\ \hat{t}_2(\xi, \eta) &=& \xi(2\xi - 1) &=& \lambda_2(2\lambda_2 - 1) \\ \hat{t}_3(\xi, \eta) &=& \eta(2\eta - 1) &=& \lambda_3(2\lambda_3 - 1) \\ \hat{t}_4(\xi, \eta) &=& 4\xi(1 - \xi\eta) &=& 4\lambda_1\lambda_2 \\ \hat{t}_5(\xi, \eta) &=& 4\xi\eta &=& 4\lambda_2\lambda_3 \\ \hat{t}_6(\xi, \eta) &=& 4\eta(1 - \xi - \eta) &=& 4\lambda_1\lambda_3 \end{array}$$

□

Der Grad der Ansatzfunktion kann hier beliebig hochgetrieben werden. Auch können Elemente durch Vorgaben von Ableitungswerten definiert werden (**Hermite-Interpolation**). Näheres dazu in [5] und in [10].
Nach diesen Vorbereitungen können jetzt die Räume W_h und V_h für die diskrete Variationsformulierung (94) bereitgestellt werden. Dazu sei Z_h eine zulässige Triangulierung von Ω. Setze für $l \geq 1$

$$\left\{ \begin{array}{rcl} W_h &:=& \left\{ v \in C^0(\bar{\Omega}) \;;\; v_{|_T} \in \mathcal{P}_l(T) \;;\; T \in Z_h \right\}, \\ V_h &:=& \left\{ v \in W_h \;;\; v_{|_{\partial\Omega}} = 0 \right\}. \end{array} \right. \tag{109}$$

Funktionen in W_h brauchen nicht mehr stetig differenzierbar sein. Damit (94) trotzdem Sinn macht, muß der Differenzierbarkeitsbegriff abgeschwächt werden.

Definition 3.38. (Schwache Differenzierbarkeit)
Eine Funktion $f \in L^p(\Omega)$ heißt schwach partiell differenzierbar bzgl. der Koordinatenrichtung x_i, falls es eine Funktion $g \in L^p(\Omega)$ gibt mit

$$\int_\Omega f \partial_{x_i} v \, dx = -\int_\Omega gv \, dx \quad \forall v \in C_0^\infty(\Omega).$$

Die Funktion g heißt i-te schwache partielle Ableitung von f. Existieren die schwachen partiellen Ableitungen in alle Koordinatenrichtungen, so ist $f \in H^{1,p}(\Omega)$ und das ist eine Definition. $H^{l,p}(\Omega)(l > 1)$ ist analog definiert.

Es gilt der

Satz 3.39. (Stückweise differenzierbar und stetig $\Longleftrightarrow$ schwach differenzierbar)
Sei $l \geq 1$ und Ω beschränkt. Eine stückweise beliebig oft differenzierbare Funktion $v : \bar{\Omega} \to \mathbb{R}$ gehört genau dann zu $H^{l,p}(\Omega)$, wenn $v \in C^{l-1}(\bar{\Omega})$ gilt.

Beweis. [3, Satz 5.2] □

Wird die diskrete Lösung u_h zu (65) in W_h gesucht, so besitzt sie nf **Freiheitsgrade**, welche sich zusammenstellen aus nf_I Freiheitsgraden im Innern und $nf - nf_I$ Freiheitsgraden auf dem Rand von Ω. Die Dirichlet Randbedingungen aus (65) werden direkt in den Ansatz für u_h integriert. Dieser liest sich dann als

$$u_h(x) = \sum_{i=1}^{nf} u_i b_i(x) = \sum_{j=nf_I+1}^{nf} r(P_j) b_j(x) + \sum_{j=1}^{nf_I} u_j b_j(x), \tag{110}$$

und es wird direkt ersichtlich, daß $u_h \in C^0(\bar{\Omega})$ gilt. Ferner liefert $u_{h_{|_T}} \in \mathcal{P}_l$, daß $u_h \in H^{1,p}(\Omega)(1 \leq p \leq \infty)$ gilt. Dabei sind die Funktionen b_i durch

$$b_i \in W_h, \qquad b_i(P_j) = \delta_{ij} \quad (1 \leq i, j \leq nf) \tag{111}$$

festgelegt, wobei P_j die Punkte im Koordinatensystem bezeichnen, die mit den Freiheitsgraden korrespondieren. Es gilt $nf = nv$ nur bei linearen Elementen. In (94) ergibt sich demnach das Gleichungssystem

$$A_h u = q_h$$

mit

$$A_h = (a_{ij})_{i,j=1}^{nf_I}, \quad a_{ij} = \int_{\Omega_h} \nabla b_i(x) \nabla b_j(x)\, dx \quad \text{und} \quad q_h = (q_i)_{i=1}^{nf_I},$$

$$q_i = \int_{\Omega_h} q(x) b_i(x)\, dx - \sum_{j=nf_I+1}^{nf} r(P_j) \int_{\Omega} \nabla b_j \nabla b_i\, dx.$$

Definition 3.40. (Masse- und Steifigkeitsmatrix)
Die Matrix

$$A_h := (a_{ij})_{i,j=1}^{nf}, \quad a_{ij} := \int_{\Omega_h} \nabla b_i(x) \nabla b_j(x) dx$$

heißt **Steifigkeitsmatrix**, die Matrix

$$M_h := (m_{ij})_{i,j=1}^{nf}, \quad m_{ij} := \int_{\Omega_h} b_i(x) b_j(x) dx$$

heißt **Massematrix**.

Mit den Ansatzfunktionen b_i aus (111) gilt

$$m_{ij} = 0, \text{ falls } P_i \text{ und } P_j \text{ nicht zum selben Dreieck gehören,}$$

analog auch

$$a_{ij} = 0, \text{ falls } P_i \text{ und } P_j \text{ nicht zum selben Dreieck gehören.}$$

Als nächstes wird gezeigt, wie sich die Einträge der Matrizen A_h und M_h bei gegebener Triangulierung berechnen lassen. Bezeichne dazu

$$s_i := \text{supp } b_i, \quad i = 1, \ldots, nf.$$

Dann gilt

$$m_{ij} = \int_{s_i \cap s_j} b_i(x) b_j(x)\, dx = \sum_{T \subset s_i \cap s_j} \int_T b_i(x) b_j(x)\, dx$$

und mit (106)

$$\int_T b_i(x)b_j(x)\,dx = \int_{\hat{T}} \hat{t}_{k_i}(\xi,\eta)\hat{t}_{k_j}(\xi,\eta)\,d(\xi,\eta)\;2|T|.$$

Die Berechnung von a_{ij} ist ein wenig aufwendiger;

$$a_{ij} = \int_{s_i\cap s_j} \nabla b_i(x)\nabla b_j(x)\,dx = \sum_{T\subset s_i\cap s_j} \int_T \nabla b_i \nabla b_j\,dx,$$

und

$$\int_T \nabla b_i(x)\nabla b_j(x)\,dx =$$

$$2|T| \int_{\hat{T}} (\hat{t}_{k_i\xi}\xi_{x_1} + \hat{t}_{k_i\eta}\eta_{x_1})(\hat{t}_{k_j\xi}\xi_{x_1} + \hat{t}_{k_j\eta}\eta_{x_1}) + (\hat{t}_{k_i\xi}\xi_{x_2} + \hat{t}_{k_i\eta}\eta_{x_2})(\hat{t}_{k_j\xi}\xi_{x_2} + \hat{t}_{k_j\eta}\eta_{x_2})d(\xi,\eta)$$

mit ξ_{x_i}, η_{x_i} aus (107). Dabei bezeichnet $\hat{t}_{k_j}$ die Ansatzfunktion über dem Einheitsdreieck $\hat{T}$, die zu der lokalen Knotennummer k_j des die Funktion b_j definierenden Freiheitsgrads P_j in T korrespondiert. Sind also die Ansatzfunktionen Polynome vom Grade l, verbleibt die Berechnung von Integralen der Form

$$\int_{\hat{T}} \xi^i\eta^j\,d\xi d\eta \quad 0 \le i,j \le l.$$

Dazu dient der

Hilfsatz 3.41. (Exakte Integration)
Sei $\hat{Q}$ das Einheitsquadrat, $\hat{T}$ das Einheitsdreieck. Dann gilt

$$\int_{\hat{T}} \xi^i\eta^j d\xi d\eta = \frac{i!j!}{(i+j+2)!} \text{ und } \int_{\hat{Q}} \xi^i\eta^j d\xi d\eta = \frac{1}{(i+1)(j+1)}.$$

Beweis. Mittels Induktion ergibt sich

$$I_{ji} = \int_{\hat{T}} \xi^i\eta^j\,d\xi d\eta \frac{i(i-1)\ldots 2\cdot 1}{(j+1)(j+2)\ldots(j+i)} I_{j+i,0} \text{ und}$$

$$I_{j+i,0} = \int_0^1 \int_0^{1-\xi} \xi^{i+j} d\xi d\eta = \frac{1}{(i+j+1)(i+j+2)}.$$

Damit ergibt sich die Behauptung für das Einheitsdreieck. Die Formel für den Einheitsquader ist evident. □

In der Regel kann die rechte Seite in (94) nicht exakt ausgewertet werden, weil $q_{|_T}, T \in Z_h$ nicht unbedingt ein Polynom zu sein braucht. Um die rechte Seite dennoch genau genug berechnen zu können, werden **Quadraturformeln** verwendet:

$$\text{(112)} \qquad \begin{cases} \int_T \chi(x)dx & \cong & \chi(P_S)|T|, \text{ exakt in } \mathcal{P}_1(T), P_S \text{ Schwerpunkt von } T \\ \int_T \chi(x)dx & \cong & \frac{|T|}{3} \sum\limits_{0\le i<j\le 2} \chi(P_{ij}), \text{ exakt in } \mathcal{P}_2(T), P_{ij} \text{ Kantenmittelpunkte von } T \\ \int_T \chi(x)dx & \cong & \frac{|T|}{60}\left\{3\sum\limits_{i=0}^{2}\chi(P_i) + 8\sum\limits_{0\le i<j\le 2}\chi(P_{ij}) + 27\chi(P_S)\right\} \text{ exakt in } \mathcal{P}_3(T). \end{cases}$$

Jetzt wird noch angegeben, wie die Matrizen M_h und A_h programmtechnisch aufgebaut werden können. Ausserdem wird die Realisierung einer Matrix-Vektor Multiplikation kodiert. Dazu sei

$$\begin{array}{lcl} \text{bandb} & = & \max\{\sharp \text{ Nachbarn eines Freiheitsgrades } P_j\} \\ \text{memo } (j,i) & = & k, \text{ falls } j\text{-ter Nachbar des } i\text{-ten globalen Freitsgrades globale Nummer } k \text{ hat} \end{array}$$

und

$$em(i,j) = \int_T b_i(x) b_j(x)\, dx \quad ea(i,j) = \int_T \nabla b_i(x) \nabla b_j(x)\, dx \quad i,j \text{ lokale Nummerierung}$$

Als Nächstes ein FORTRAN Pseudocode zur Generierung der Matrizen.

CODE 3.42. (Kompilierung von Masse- und Stetigkeitsmatrix)

```
                Do i = 1, nt
                  call elem(i, ea)
                  Do m = 1, nloc
                    mi = node(m, i)
                    Do l = 1, nloc
                      li = node(l, i)
                      Do j = 1, bandb
                        im = memo(j, li)
                        if(im.eq.0.or.im.eq.mi)then
                          a(li, j) = a(li, j) + ea(l, m)
                          memo(j, li) = mi
                          goto100
                        endif
j                     continue
100                   continue
l                   continue
m                 continue
i               continue.
```

Dabei belegt das Unterprogramm *elem* das Feld *ea* (bzw. *em*) für das angegebene Dreieck.

Die Beiträge der einzelnen Integrale können aber mit der Formel (105) elementweise berechnet werden. Im Falle linearer Elemente gilt daher mit Hilfsatz 3.41

$$\begin{aligned} ea = 2|T| \Bigg\{ & \tfrac{1}{2}\left(\xi_{x_1}^2 + \xi_{x_2}^2\right) \begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \tfrac{1}{2}\left(\xi_{x_1}\eta_{x_1} + \xi_{x_2}\eta_{x_2}\right) \begin{bmatrix} 2 & -1 & -1 \\ 1 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \\ & + \tfrac{1}{2}\left(\eta_{x_1}^2 + \eta_{x_2}^2\right) \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix} \Bigg\}. \end{aligned} \tag{113}$$

Die Matrizen M_h und A_h aus Definition 3.31 haben die folgenden Eigenschaften.

Hilfsatz 3.43. (Eigenschaften von Masse- und Steifigkeitsmatrix)
Sei Z_h eine zulässige Triangulierung. Dann ist die Massematrix M_h symmetrisch und positiv definit, die Steifigkeitsmatrix ist symmetrisch und positiv semidefinit.

Beweis. Aufgabe 6.26. □

Es sei bemerkt, daß

$$dim \text{ ker } A_h = 1.$$

und der Kern von A_h durch den Vektor $(1,\ldots,1)^t \in \mathbb{R}^{nf}$ aufgespannt wird. Die Matrix A_h ist die Systemmatrix der Finite Element Approximation für das Poisson Problem (65) mit Neumann Randbedingungen. Dabei wird die Bedingung $u = r$ auf $\partial\Omega$ in (65) ersetzt durch $\partial_\eta u = 0$ auf $\partial\Omega$.

Die Matrix A_h ist die Systemmatrix für das Variationsproblem (95), falls über dem Raum W_h (siehe (109)) minimiert wird, d.h., wird das Variationsproblem

$$\min_{u\in W_h} \int_\Omega \frac{1}{2}|\nabla u(x)|^2 - q(x)u(x)dx \tag{114}$$

betrachtet, so ergibt sich mit dem Ansatz

$$u(x) = \sum_{i=1}^{nf} u_i b_i(x)$$

das Minimierungsproblem

$$\min_{(u_1,\dots u_{nf})\in \mathbb{R}^{nf}} \frac{1}{2} u^t A_h u - q^t u$$

für den Koeffizientenvektor $(u_{1,\dots,}u_{nf}) = u$ und als notwendige Bedingung für die Löbarkeit des Gleichungssystems

$$A_h u = q,$$

wegen $(1,\dots,1)^t A_h = 0$ und $\sum\limits_{i=1}^{nf} b_i(x) = 1$ auch, daß

$$0 = (1,\dots,1)^t A_h u \;=\; (1,\dots,1)^t q \;=\; \int_\Omega q(x)dx$$

gelten muß, vergleiche (73), (74).
Jetzt noch etwas zum Umgang mit Randbedingungen bei der numerischen Implementierung. Als Beispielproblem dafür diene

$$\left\{\begin{array}{rcll} -\Delta u + v_T^t \nabla u + cu &=& q & \text{in } \Omega \\ u &=& r_D & \text{auf } \Gamma_D \\ \partial_\eta u + \alpha u &=& r_N & \text{auf } \Gamma_N \end{array}\right. \tag{115}$$

mit $\Gamma_D \cup \Gamma_N = \partial\Omega, \Gamma_D \cap \Gamma_N = \emptyset$, vergleiche (10). Zur Herleitung der variationellen Formulierung sei in Anlehnung an Merkregel 3.28 der Raum der Testfunktion durch

$$V = \left\{ v \in H^{1,2}(\Omega); v_{|\Gamma_D} = 0 \text{ im Spursinne} \right\}$$

gegeben, wobei der Raum $H^{1,2}(\Omega)$ nach Definition 3.38 eingeführt wurde. Das zu (115) korrespondierende Variationsproblem liest sich dann als
Finde $u \in H^{1,2}(\Omega)$ mit $u = r_D$ auf Γ_D und

$$\int_\Omega \nabla u \nabla v + v_T^t \nabla u v + cuvdx + \int_{\Gamma_N} \alpha uvds = \int_\Omega qvdx + \int_{\Gamma_N} r_N vds \quad \forall v \in V. \tag{116}$$

Zur Diskretisierung dieses Problems wird jetzt für u der Ansatz

$$u_h(x) = \sum_{i=1}^{nf_I+nf_N} u_i b_i(x) + \sum_{i=nf_I+nf_N+1}^{nf} r_{D_i} b_i(x)$$

gemacht und als Testraum

$$V_h = \left\{ v \in W_h; v_{|\Gamma_D} = 0 \right\}$$

gewählt. Aus (116) ergibt sich damit das lineare Gleichungssystem

$$(117)\quad \underbrace{(\tilde{A}_h + S_h + \tilde{M}_h + R_h)u}_{D_h} = q + r_N - \left[\sum_{i=nf_I+nf_N+1}^{nf} r_{D_i} \int_\Omega \nabla b_i \nabla b_j + v_T^t \nabla b_i b_j + c b_i b_j dx\right]_{j=1}^{nf_I+nf_N} - \left[\sum_{i=nf_I+nf_N+1}^{nf} r_{N_i} \int_{\Gamma_N} \alpha b_i b_j ds\right]_{j=1}^{nf_I+nf_N}$$

wobei mit $nf_I = \sharp$ innere Freiheitsgrade, $nf_N = \sharp$ Neumannfreiheitsgrade gilt:

$$\begin{aligned}
\tilde{A}_h &= (\tilde{a}_{ij})_{i,j=1}^{nf_I+nf_N}, & \tilde{a}_{ij} &= \int_\Omega \nabla b_i \nabla b_j dx \\
S_h &= (s_{ij})_{i,j=1}^{nf_i+nf_N}, & s_{ij} &= \int_\Omega v_T^t \nabla b_i b_j dx \\
\tilde{M}_h &= (\tilde{m}_{ij})_{i,j=1}^{nf_I+nf_N}, & \tilde{m}_{ij} &= \int_\Omega c b_i b_j dx \\
R_h &= (r_{ij})_{i,j=1}^{nf_I+nf_N}, & r_{ij} &= \int_{\Gamma_N} \alpha b_i b_j ds.
\end{aligned}$$

Bemerkung 3.44. Gilt $\Gamma_D \neq \emptyset$, so ist die Steifigkeitsmatrix $\tilde{A}_h$ symmetrisch und positiv definit, denn

$$z^t A_h z = \int_\Omega |\nabla u|^2 dx, \quad u = \sum_{i=1}^{nf_I+nf_N} z_i b_i(x),$$

also $\tilde{A}_h$ positiv semidefinit. Ist mit $z \neq 0$ jetzt $\tilde{A}_h z = 0$, so folgt $u \equiv const$ und $u_{|\Gamma_D} = 0$, also $const \equiv 0$, also schon $z = 0$.

Die Systemmatrix D_h in (117) ist nicht symmetrisch, falls $v_T \neq 0$ gilt. Diese Matrix ist allerdings dünn besetzt. Ferner enthält sie nur für die Indexpaare (i, j) Nicht-Nullelemente, deren entsprechende Freiheitsgrade P_i und P_j Nachbarn sind. Als nächstes werden zwei Möglichkeiten angegeben, eine Matrix-Vektor-Multiplikation mit D_h effizient zu realisieren. Dazu sei

$$n_0 = \#\text{ Nicht-Nullelemente von } D_h, \quad n = \#\text{ Zeilen von } D_h$$

und H ein Vektor der Länge n_0, welcher die Nicht-Nullelemente von D_h speichern soll. Ferner seien

$$CN(i) = j, \text{ falls das i-te Element von } H \text{ in der j-ten Spalte von } D_h \text{ steht}$$

und

$$(118)\quad RS(k) = \begin{cases} l & , \text{ falls das l-te Element von } H \text{ das erste} \\ & \quad \text{Nicht-Nullelement der k-ten Zeile von } D_h \text{ ist} \\ n_0 + 1 & , \quad k = n + 1. \end{cases}$$

Ist die Matrix D_h symmetrisch, reicht es, H stufenweise mit dem linken unteren Dreieck von D_h (einschließlich Diagonale) zu belegen. Es wird jetzt die Matrix-Vektor Multiplikation

$$z = D_h d$$

auf zwei verschiedene Arten kodiert.

CODE 3.45. (Matrix-Vektor Sparse mit memo und CN,RS)

i.)

```
      Do k = 1, n
            z(k) = 0
            Do j = RS(k), RS(k+1) - 1
            Z(k) = Z(k) + H(j) * d(CN(j))
j     continue
k   continue
```

ii.)
```
      Do k = 1, n
        z(k) = 0
        Do j = 1, bandb
        Z(k) = Z(k) + a(k, j) * d(memo(j, k))
j       continue
k   continue
```

Dabei ist *bandb* vor FORTRAN 3.33 erklärt, ebenso *memo*. Die Matrix D_h ist bei ii.) analog zu Code 3.42 auf das Feld $a(n, bandb)$ gelegt.

3.4 Numerische lineare Algebra

Gilt in (115) $v_T \equiv 0$, so ist die Matrix D_h in (117) symmetrisch und, falls entweder $\Gamma_D \neq \emptyset$ oder $c > 0$, auch positiv definit. Dieser Fall tritt sehr häufig auf. Zunächst wird daher das Gleichungssystem

$$Ax = b \tag{119}$$

mit symmetrischer, positiv definiter Matrix A betrachtet. Der Standardalgorithmus zur Lösung von (119) ist das Verfahren der konjugierten Gradienten, kurz **CG-Verfahren** (Conjugate Gradient Method).

Algorithmus 3.46. (Basis CG-Verfahren)
Gegeben seien $b, x^0 \in \mathbb{R}^n$ und eine positiv definite Matrix $A \in \mathbb{R}^{n\times n}$. Berechnet wird die Lösung x von (119).

Initialisierung

$$\begin{aligned} r^0 &= b - Ax^0 \\ p^0 &= r^0 \\ i &= 0 \end{aligned}$$

do while $i \leq n$ und $r^i \neq 0$

$$\begin{aligned} \alpha_i &= (r^i, p^i)/(p^i, Ap^i) \\ x^{i+1} &= x^i + \alpha_i p^i \\ r^{i+1} &= r^i - \alpha_i Ap^i \\ \beta_i &= (r^{i+1}, Ap^i)/(p^i, Ap^i) \\ p^{i+1} &= r^{i+1} - \beta_i p^i \\ i &= i + 1 \end{aligned}$$

endwhile

Der numerische Aufwand in jedem Iterationsschritt besteht in einer Matrix-Vektor-Multiplikation und der Berechnung von 3 Skalarprodukten. Ein Skalarprodukt wird eingespart, falls

$$\begin{cases} \alpha_i &= |r^i|^2/(p^i, Ap^i) \\ \beta_i &= |r^{i+1}|^2/|r^i|^2 \\ p^{i+1} &= r^{i+1} + \beta_i p^i \end{cases} \tag{120}$$

im obigen Algorithmus gesetzt wird. Aufgrund von Orthogonalitätseigenschaften der Iterierten ergibt sich mit (120) ein zum Algorithmus 3.46 äquivalenter Algorithmus. Die Erfahrung zeigt allerdings, daß der numerische Mehraufwand stabilisierend wirken kann, d.h., daß die sparsame Variante häufiger versagt als die spendablere. Der Algorithmus 3.46 hat folgende Eigenschaften.

Satz 3.47. (CG-Verfahren ist ein direktes Verfahren)
Sei $A \in \mathbb{R}^{n\times n}$ symmetrisch und positiv definit. Dann liefert das CG-Verfahren in Algorithmus 3.35 nach höchstens n Schritten die exakte Lösung x^* des Gleichungssystems (119), wobei die Wahl des Startwertes x^0 beliebig ist.

Beweis. [10, Satz 5.2] □

Das CG-Verfahren ist demnach ein direktes Verfahren oder kann als solches aufgefaßt werden. Seine rekursive Formulierung läßt jedoch zu, die Lösung von (119) bis zu einer vorgelegten Genauigkeit zu berechnen. Die Anzahl der dazu benötigten Iterationen kann abgeschätzt werden.

Satz 3.48. (Konvergenzgeschwindigkeit des CG-Verfahrens)
Sei $A \in \mathbb{R}^{n\times n}$ positiv definit und $x^0 \in \mathbb{R}^n$ beliebiger Startwert. Die Iterierten x^k des CG-Verfahrens erfüllen

$$(A(x^k - x^*), x^k - x^*) \;\leq\; 2\left\{\frac{\sqrt{\lambda_{max}(A)} - \sqrt{\lambda_{min}(A)}}{\sqrt{\lambda_{max}(A)} + \sqrt{\lambda_{min}(A)}}\right\}^k (A(x^0 - x^*), x^0 - x^*) \tag{121}$$

Beweis. [10, Satz 5.3] □

Es bezeichne

$$|x|_A := \sqrt{x^t A x}$$

die zur Matrix A assoziierte Matrixnorm und

$$\kappa(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \tag{122}$$

die Kondition der symmetrischen, positiv definiten Matrix A. Aus (120) wird eine Abschätzung erhalten dafür, wieviele Iterationen des CG-Verfahrens höchstens durchzuführen sind, um den Ausgangsfehler auf die Größenordnung ϵ zu reduzieren. Hinreichend dafür ist, daß k

$$\left\{\frac{\sqrt{\lambda_{max}(A)} + \sqrt{\lambda_{min}(A)}}{\sqrt{\lambda_{max}(A)} - \sqrt{\lambda_{min}(A)}}\right\}^k \geq \frac{2}{\epsilon}$$

erfüllt. Mit Hilfe von (122) kann das auch umgeschrieben werden zu

$$\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^k \;\geq\; \frac{2}{\epsilon},$$

also

$$k \;\geq\; \ln\frac{2}{\epsilon} \;/\; \ln\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}.$$

Wird noch

$$\ln\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1} \geq \frac{2}{\sqrt{\kappa}}, \quad \kappa > 1,\ \kappa = \kappa(A)$$

berücksichtigt, so ist für

$$k \;\geq\; \frac{1}{2}\sqrt{\kappa(A)}\ln\left(\frac{2}{\epsilon}\right) + 1 \tag{123}$$

sichergestellt, daß

$$|x^k - x^*|_A \;\leq\; \epsilon|x^1 - x^*|_A$$

gilt.

Die maßgebliche Größe für die Konvergenzgeschwindigkeit ist hier die Kondition der Matrix A. Ihre Güte hängt maßgeblich von der Breite des Spektrums von A ab, d.h., je weiter $\lambda_{max}(A)$ und $\lambda_{min}(A)$ auseinanderliegen, desto pessimistischer sind die Erwartungen hinsichtlich der Fehlerreduktion. Die Konvergenzgeschwindigkeit kann mittels Vorkonditionierung verbessert werden. Dazu bemerke, daß die eindeutige Lösung x^* von (119) auch das Minimum der Funktion

$$f(x) := \frac{1}{2}x^t Ax - b^t x$$

darstellt. Sei jetzt C eine nichtsinguläre Matrix. Dann ist mit

$$\tilde{A} := C^{-1}AC^{-t}, \quad \tilde{b} := C^{-1}b$$

das eindeutige Minimum y^* der Funktion

$$\tilde{f}(y) := \frac{1}{2}y^t \tilde{A}y - \tilde{b}^t y$$

durch

$$y^* = \tilde{A}^{-1}\tilde{b} = C^t x^*$$

gegeben, wobei x^* die Lösung von (119) darstellt. Ferner gilt wegen der positiven Definitheit von $\tilde{A}$ mit (121)

$$|y^k - y^*|_{\tilde{A}} \leq 2\left\{\frac{\sqrt{\kappa(\tilde{A})}-1}{\sqrt{\kappa(\tilde{A})}+1}\right\}^k |y^0 - y^*|_{\tilde{A}} \tag{124}$$

und zu gegebenen $\epsilon > 0$ gilt für die in (122) hergeleitete Zahl $k = k(\epsilon)$

$$k \geq \frac{1}{2}\sqrt{\kappa(\tilde{A})}ln\left(\frac{2}{\epsilon}\right) + 1.$$

Ist $\kappa(\tilde{A}) < \kappa(A)$, so wird wohl y^* schneller gut approximiert als x^*. Weil mit

$$x^* = C^{-t}y^*, \quad x^k = C^{-t}y^k$$

$$|y^k - y^*|_{\tilde{A}} = |x^k - x^*|_A$$

folgt, motiviert sich aus diesen Überlegungen unter Beachtung von

$$C^{-t}\tilde{A}C^t = C^{-t}C^{-1}A =: W^{-1}A$$

(d.h., WA und $\tilde{A}$ haben dieselben Eigenwerte) der

Algorithmus 3.49. (Vorkonditioniertes CG-Verfahren)
Gegeben seien $b, x^0 \in \mathbb{R}^n$, eine positiv definite Matrix $A \in \mathbb{R}^{n\times n}$ und ein geeigneter **Vorkonditionierer** $W \in \mathbb{R}^{n\times n}$, i.e. W positiv definit. Berechnet wird die Lösung x^* von (119).

Initialisierung

$$r^0 = b - Ax^0$$
$$p^0 = W^{-1}r^0$$
$$\beta_0 = (p^0, r^0)$$
$$i = 0$$

do while $i \leq n$ und $r^i \neq 0$

$$\alpha_i = \beta_i/(p^i, Ap^i)$$
$$x^{i+1} = x^i + \alpha_i p^i$$

$$
\begin{aligned}
r^{i+1} &= r^i - \alpha_i A p^i \\
q^{i+1} &= W^{-1} r^{i+1} \\
\beta_{i+1} &= (q^{i+1}, r^{i+1}) \\
p^{i+1} &= q^{i+1} + \tfrac{\beta_{i+1}}{\beta_i} p^i \\
i &= i + 1
\end{aligned}
$$

end while

Die Kunst besteht jetzt darin, die Matrix W so zu wählen, daß

i) $\kappa(\tilde{A})$ nahe bei 1 bzw. $K(\tilde{A}) << K(A)$ und/oder
ii) $q = W^{-1} r$ leicht auflösbar

gewährleistet werden kann. Diese beiden Eigenschaften widersprechen sich natürlich im Allgemeinen und gesuchtt ist hier der Königsweg.
Abschließend werden noch einige Vorkonditionierer angegeben und ihre Eigenschaften vorgestellt.

Beispiel 1. Diagonale Skalierung. Dabei wird

$$
W := \operatorname{diag}(A) \tag{125}
$$

gewählt.
2. SOR-Vorkonditionierung. Dabei wird ausgehend von der Zerlegung $A = D + L + R$

$$
W := \frac{1}{2-\omega} \left(\frac{1}{\omega} D + L\right) \left(\frac{1}{\omega} D\right)^{-1} \left(\frac{1}{\omega} D + R\right) \tag{126}
$$

gewählt. Im symmetrischen Fall heißt das SSOR Vorkonditionierung.
3. Vorkonditionierung mittels unvollständiger LR-Zerlegung. Dabei wird $A = M + R$ mit $M = \tilde{L}\tilde{R}$ und $R = A - \tilde{L}\tilde{R}$ angesetzt und

$$
W := \tilde{L}\tilde{R} \quad (= \tilde{L} D \tilde{L}^t \text{ für symmetrische Matrizen}) \tag{127}
$$

gewählt, vergleiche [16]. Zur Berechnung der Faktoren $\tilde{L}$ und $\tilde{R}$ wird die L-R-Zerlegung etwa nur auf den Nicht-Nullelementen von A durchgeführt. □

Es gilt

Satz 3.50. (Konditionsverbesserung)
Für die SSOR-Vorkonditionierung gilt

$$
\min_{0<w<2} \kappa(\tilde{A})(\omega) \leq \frac{1}{2} + \sqrt{\frac{1}{2}\kappa(A)}, \tag{128}
$$

für die Vorkonditionierung mittels unvollständiger LR-Zerlegung

$$
\kappa(\tilde{A}) \leq C\sqrt{\kappa(A)}, \tag{129}
$$

falls die LR-Zerlegung nur auf den Nicht-Nullelementen der Ausgangsmatrix A durchgeführt wird.

Beweis. [16, (1.73 c)](1.73 c) für (128), (129) in [3]. □

Beispiel Wähle

$$
A = \frac{1}{h^2} \begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & 0 \\ & \ddots & & -1 \\ 0 & & -1 & 2 \end{bmatrix}, \quad h = \frac{1}{N+1}.
$$

Dann ist A positiv definit und Satz 39 liefert für die Eigenwerte

$$\begin{aligned} \lambda_{max}(A) &= 2 - 2\cos\tfrac{\pi N}{N+1} && \leq 4 \\ \lambda_{min}(A) &= 2 - 2\cos\tfrac{\pi}{N+1} &= \tfrac{1}{2}\tfrac{\pi^2}{(N+1)^2} + \mathcal{O}\left(\tfrac{1}{N^4}\right), \end{aligned}$$

also mit (121)

$$\kappa(A) \approx \frac{8}{\pi^2} h^{-2}.$$

Benötigt demnach das CG-Verfahren k Iterationen zur Reduktion des Ausgangsfehlers auf ϵ Teile seiner Ausgangsgröße, so kommen die mit SSOR und unvollständiger LR-Zerlegung vorkonditionierten Varianten mit $\mathcal{O}(\sqrt{\kappa})$ Iterationen aus. □

Bemerkung 3.51. (Mehrgittermethoden)
Mit der Hilfe von Mehrgittermethoden lassen sich Vorkonditionierer entwickeln, mit deren Hilfe die Anzahl der zur Lösung notwendigen Iterationen im CG-Verfahren unabhängig von der Dimension des Gleichungssystems gehalten werden kann. Eine Ausführliche Diskussion zu Mehrgitterverfahren gibt Hackbusch in [12].

Sei jetzt die Matrix A in (119) regulär, aber nicht mehr notwendig symmetrisch. Das ist etwa der Fall, wenn in (115) $v_T \neq 0$ gilt. Das iterative Verfahren der Wahl ist hier **GMRES** (Generalized Minimal Residual). Die Idee des Verfahrens ist die Folgende.

1. Berechne eine Orthonormalbasis des **Krylov-Raumes**

   (130) $$K(A, r^0, i) := \text{span}\ \left\{r^0, Ar^0, \ldots, A^{i-1}r^0\right\}.$$

2. Transformiere die Matrix A auf obere Hessenbergform $\tilde{A}$.

3. Berechne mit Hilfe von $\tilde{A}$ eine das Residuum bzgl. der Euklidischen Norm minimierende Iterierte $x^{i+1} \in x^0 + K(A, r^0, i)$.

Algorithmus 3.52. (Vorkonditioniertes GMRES)
Gegeben seien $b, x^0 \in \mathbb{R}^{n\times n}$, eine reguläre Matrix $A \in \mathbb{R}^{n\times n}$ und ein Vorkonditionierer $W \in \mathbb{R}^{n\times n}$. Berechnet wird die Lösung x^* von $Ax^* = b$. Die natürliche Zahl $m \geq 1$ sei ein Restart Index.

$r^0 = b - Ax^0$
restart $v^1 = r^0/\|r^0\|_2$
$i = 1$
do while $1 \leq i \leq m$ and $r^{i-1} \neq 0$
  for $j = 1, i$
    $z^j = W^{-1}v^j$
    $w = Az^j$
    for $l = 1, j$
      $h_{lj} = w^t v^l$
      $w = w - h_{l_j} v^l$
    endfor $_l$
    $h_{j+1,j} = \|w\|_2$
    $v^{j+1} = w/h_{j+1,j}$
  endfor$_j$
  $H^i = (h_{\alpha\beta})_{\alpha=1,i+1\,\beta=1,i} \in \mathbb{R}^{i+1,i}$

$V^i = [v^1, \ldots, v^i] \in \mathbb{R}^{n,i}$

$e_1^{(i+1)} = (\|r^0\|_2, 0, \ldots, 0)^t \in \mathbb{R}^{i+1}$

$\beta^i = \arg\min_{\beta \in \mathbb{R}^i} \|e_1^{(i+1)} - H^i \beta\|_2$

$x^i = x^0 + W^{-1} V^i \beta^i$

$r^i = b - Ax^i$

$i = i + 1$

end while

If $r^m \neq 0$

$x^0 = x^m$

$r^0 = r^m$

goto restart

endif

Bemerkung 3.53. Der restart Index m sollte groß genug gewählt werden, damit das GMRES-Verfahren seine guten Konvergenzeigenschaften entwickeln kann. Diese kommen zum Tragen, wenn das Residuum in eine *gute* Richtung zeigt.

Herleitung des Algorithmus: Die i-te Iterierte soll die Gestalt

$$x^i = x^0 + \sum_{j=0}^{i-1} \alpha_j A^j r^0 \tag{131}$$

haben, d.h. im Raum $x^0 + K(A, r^0, i)$ liegen und soll

$$\|r^i\|_2 = \|b - Ax^i\|_2 = \min_{\alpha \in \mathbb{R}^i} \|r^0 - \sum_{j=0}^{i-1} \alpha_j A^{j+1} r^0\|_2 \tag{132}$$

erfüllen. Um dieses Minimierungsproblem zu lösen, wird die Struktur von r^i ausgenutzt. Dazu wird zunächst mit Hilfe des **Gram-Schmidt Orthonormalisierungsverfahrens** eine Orthonormalbasis von $K(A, r^0, i)$ berechnet. Algorithmisch

$$\begin{aligned} &v^1 = r^0 / \|r^0\|_2 \\ &\text{for } j = 1, i-1 \\ &\quad w^j = Av^j \\ &\quad \hat{v}^{j+1} = w^j - \sum_{l=1}^{j} (v^l)^t w^j v^l \\ &\quad v^{j+1} = \hat{v}^{j+1} / \|\hat{v}^{j+1}\|_2 \\ &\text{endfor}_j \end{aligned} \tag{133}$$

Hilfsatz 3.54. (GMRES iteriert nur endlich oft)
Das Gram-Schmidt-Verfahren (133) bricht ab, falls für ein $1 \leq k \leq i-1$ der Vektor $A^k r^0 \in K(A, r^0, k)$ erfüllt. Dann ist x^k aber die gesuchte Lösung, d.h. $x^k = x^*$.

Beweis. Sei $1 \leq k \leq i-1$ mit $A^k r^0 \in K(A, r^0, k)$ und k der kleinste Index dieser Art. Dann gilt

$$A^k r^0 = \sum_{j=0}^{k-1} \gamma_j A^j r^0, \quad \gamma_j \text{ geeignet},$$

wobei $\gamma_0 \neq 0$, denn sonst wäre

$$A^k r^0 = \sum_{j=1}^{k-1} \gamma_j A^j r^0 = A \sum_{j=1}^{k-1} \gamma_j A^{j-1} r^0,$$

woraus wegen der Regularität von A

$$A^{k-1}r^0 = \sum_{j=1}^{k-1} \gamma_j A^{j-1} r^0 = \sum_{j=0}^{k-2} \gamma_{j+1} A^j r^0$$

folgt, $A^{k-1}r^0$ also ein Element aus $K(A, r^0, k-1)$ sein müßte, ein Widerspruch zur Minimalität von k. Für r^k ergibt sich

$$\begin{aligned} r^k &= b - A(x^0 + \sum_{j=0}^{k-1} \alpha_j A^j r^0) \\ &= r^0 - \sum_{j=0}^{k-1} \alpha_j A^{j+1} r^0 \\ &= r^0 - \sum_{j=0}^{k-2} \alpha_j A^{j+1} r^0 - \alpha_{k-1} A^k r^0 \\ &= r^0 - \sum_{j=0}^{k-2} \alpha_j A^{j+1} r^0 - \alpha_{k-1} \sum_{j=0}^{k-1} \gamma_j A^j r^0 \\ &= (1 - \alpha_{k-1}\gamma_0) r^0 - \sum_{j=1}^{k-1} (\alpha_{j-1} + \alpha_{k-1}\gamma_j) A^j r^0. \end{aligned}$$

Wähle jetzt

$$\alpha_{k-1} = 1/\gamma_0, \ \alpha_{j-1} = -\alpha_{k-1}\gamma_j = -\gamma_j/\gamma_0, \quad j = 1, \ldots, k-2$$

so daß $r^k = 0$ folgt. Offensichtlich ist der so definierte Vektor α der in (131) gesuchte, weil dieser $\|r^k\|_2 = 0$ erfüllt. □

Folgerung 3.55. (GMRES ist nach spätestens n Iterationen fertig)
Ist $A \in \mathbb{R}^{n\times n}$ regulär, so berechnet GMRES nach höchstens n Schritten die exakte Lösung x^* von (118).

Beweis. Es existieren maximal n linear unabhängige Vektoren im $\mathbb{R}^n$, also bricht Gram-Schmidt nach höchstens n Schritten ab. □

Bemerkung 3.56. Da in jedem GMRES-Schritt bzgl. aller bisherigen Suchrichtungen minimiert wird, ist die Folge $\{\|r^i\|_2\}_{i\in\mathbb{N}}$ monoton fallend.

Jetzt weiter mit der Herleitung. Angenommen, der Prozeß (132) breche nicht innerhalb der ersten i-Schritte ab und erzeuge ein ONB $\{v^1, \ldots, v^i\}$ von $K(A, r^0, i)$. Dann gilt mit geeignetem β_j

$$x^i = x^0 + \sum_{j=0}^{i-1} \alpha_j A^j r^0 = x^0 + \sum_{j=1}^{i} \beta_j v^j$$

und

$$r^i = r^0 - \sum_{j=1}^{i} \beta_j A v^j = r^0 - \sum_{j=1}^{i} \beta_j w^j, \ w^j = A v^j.$$

Aus (133) folgt

$$w^j = \|\hat{v}^{j+1}\|_2 v^{j+1} + \sum_{l=1}^{j} (v^l)^t w^j v^l.$$

Definiere jetzt die Matrix $H^i \in \mathbb{R}^{i+1,i}$ gemäß

$$\begin{aligned} h_{lj} &= (v^l)^t w^j \quad , \ l \le j \le i, \\ h_{j+1,j} &= \|\hat{v}^{j+1}\|_2 \quad , \ 1 \le j \le i, \\ h_{lj} &= 0 \quad , \ 1 \le j+1 < l \le i+1 \end{aligned}$$

und setze $V^{i+1} = [v^1, \ldots, v^{i+1}] \in \mathbb{R}^{n\times(i+1)}$. Dann läßt sich w^j mit $\beta = (\beta_1, \ldots, \beta_i)^t$ schreiben als

$$w^j = h_{j+1,j}\ v^{j+1} + \sum_{l=1}^{j} h_{lj} v^l = \sum_{l=1}^{j+1} h_{lj} v^l,$$

ergo

$$\begin{aligned} r^i &= r^0 - \sum_{j=1}^{i} \beta_j \sum_{l=1}^{j+1} h_{lj} v^l \\ &= r^0 - \sum_{l=1}^{j+1} v^l \sum_{j=1}^{i} h_{lj} \beta_j \\ &= r^0 - V^{i+1} H^i \beta. \end{aligned}$$

Mit $v^1 = r^0/\|r^0\|_2$ und $e_1^{(i+1)} = (\|r^0\|_2, 0, \ldots, 0)^t \in \mathbb{R}^{i+1}$ gilt

$$r^0 = V^{i+1} e_1^{(i+1)},$$

und da die Spalten von V^{i+1} orthonormal sind, folgt

$$\|r^i\|_2 = \|r^0 - V^{i+1} H^i \beta\|_2 = \|V^{i+1}(e_1^{(i+1)} - H^i \beta)\|_2 = \|e_1^{(i+1)} - H^i \beta\|_2. \tag{134}$$

Damit ist das Minimierungsproblem (132) durch das Minimierungsproblem (134) mit oberer Hessenberg-Matrix H^i ersetzt worden. Die Herleitung des Algorithmus 3.52 ist damit abgeschlossen.

Verbleibt die Lösung des Least-Squares-Problems (134). Seien dazu $H \in \mathbb{R}^{n+k,n}$ obere Hessenberg-Matrix und $a \in \mathbb{R}^{n+k}$ gegeben. Betrachtet wird jetzt das Least-Squares Problem

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^n} \|a - H\beta\|_2,$$

dessen Lösung mit Hilfe einer QR-Zerlegung realisiert werden soll. Sei dazu Q eine orthogonale Matrix, d.h. Q erfülle

$$Q^t Q = Id_{n+k}.$$

Dann gilt wegen der Orthogonalität von Q und der Definition der Euklidischen Norm

$$\|Q(a - H\beta)\|_2 = \|a - H\beta\|_2.$$

Wähle jetzt die Matrix Q so, daß

$$QH = \begin{bmatrix} R \\ 0 \end{bmatrix}, \quad R \in \mathbb{R}^{n \times n},\ 0 \in \mathbb{R}^{k \times n}$$

mit oberer Dreiecksmatrix R. Schreibe noch

$$Qa = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \in \mathbb{R}^{n+k},\ a_1 \in \mathbb{R}^n,\ a_2 \in \mathbb{R}^k.$$

Dann gilt mit der Lösung β^* von

$$R\beta^* = a_1 \tag{135}$$

auf jeden Fall

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^n} \|a - H\beta\|_2$$

und

$$\min_{\beta \in \mathbb{R}^n} \|a - H\beta\|_2 = \|a_2\|_2.$$

Darüber hinaus ist das Gleichungssystem (135) leicht auflösbar.

Es verbleibt somit die Konstruktion einer Matrix Q, welche H auf die gewünschte Gestalt transformiert. Dazu wird Q aus Drehungen komponiert. Es bezeichne

$$D(p,q,\phi) := \begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & \cos\phi & & \dots & & -\sin\phi & & \leftarrow p \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & \sin\phi & & \dots & & \cos\phi & & \leftarrow q \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \\ & & & & & & & & & & 1 \end{bmatrix} \tag{136}$$

eine Drehung um den Winkel ϕ in der Ebene, welche durch die p-te und q-te Koordinatenrichtung aufgespannt wird. Durch Multiplikation von Links mit $D(p,q,\phi)$ soll H so modifiziert werden, daß für

$$H' = D(p,q,\phi)H$$

mit

$$H = (h_{ij})_{i=1,n+k;j=1,n}$$

die Beziehungen

$$0 = h'_{qp} = -h_{pp}\sin\phi + h_{qp}\cos\phi$$

gelten. Mit $c := \cos\phi$ und $s := \sin\phi$ ergibt sich

$$c = \frac{sh_{pp}}{h_{qp}} = \sqrt{1-c^2}\frac{h_{pp}}{h_{qp}},$$

ergo

$$c = \frac{|\alpha|}{\sqrt{1+\alpha^2}}, \quad s = \frac{|\alpha|}{\alpha\sqrt{1+\alpha^2}}, \text{ wobei } \alpha := \frac{h_{pp}}{h_{qp}}.$$

Damit ist aber der Winkel ϕ in $D(p,q,\phi)$ festgelegt.
Diese Methode soll jetzt in (134) angewendet werden, um $H' \in \mathbb{R}^{i+1,i}$ auf obere Dreicksform zu transformieren. Der nötige numerische Aufwand ergibt sich aus

Hilfsatz 3.57. (QR-Zerlegung einer Hessenbergmatrix)
Sei $H \in \mathbb{R}^{n\times n}$ obere Hessenbergmatrix. Eine Zerlegung

$$H = QR$$

mit orthogonalem Q und oberer Dreiecksmatrix R ist mit den $n-1$ Rotationsmatrizen $D(1,2;\phi_1)$, $D(2,3;\phi_2)$, …, $D(n-1,n;\phi_{n-1})$ durchführbar.

Beweis. Übung. □

Als direkte Konsequenz ergibt sich für den numerischen Aufwand in (134):

i.) H^i auf obere Dreiecksform R^i : i-mal Wurzel ziehen und $4i$ Multiplikationen

ii.) Rückwärts einsetzen mit R^i.

Bemerkung 3.58. Der numerische Aufwand von GMRES steigt mit wachsender Anzahl von Iterationen. Schnelle Konvergenz wird allerdings erst bei relativ großen Werten von i beobachtet, ein Restart sollte demnach in Algorithmus 3.52 nicht mit zu kleinem m erfolgen.

Eine ganz andere Beobachtung hat Jan Brandts 1998 gemacht. Bezeichnen $CL(n)$ n Schritte eines klassischen Verfahrens (etwa des Gauß-Seidel oder des SOR-Verfahrens) und GMRES (k) k Schritte GMRES, so stellt sich häufig folgendes Phänomen heraus.

Beobachtung 3.59. (Pre-Processing für GMRES)
Bei der numerischen Lösung von Gleichungssystemen mit nichtsymmetrischen Matrizen unter Verwendung von GMRES beobachtete Jan Brandts 1998 das Verhalten

$$\text{GMRES}(k-n) \circ CL(n) r^0 \approx \text{GMRES}(k) r^0.$$

Vom Standpunkt des numerischen Aufwands aus ist $\text{GMRES}(k-n) \circ CL(n) r^0$ wesentlich besser als $\text{GMRES}(k) r^0$. Gerade bei Matrizen, welche aus der Diskretisierung von Konvektions-Diffusionsproblemen resultieren, scheint diese Beobachtung zuzutreffen.

Bemerkung 3.60. Für reguläre indefinite Matrizen, wie sie häufig bei Sattelpunktproblemen auftauchen, gibt es Varianten des CG-Verfahrens, welche nicht soviel Speicherplatz wie GMRES benötigen. Genannt seien hier

- **BiCG (Biconjugate Gradients)**
- **BiCG$_3$(Biconjugate Gradients mit 3 Matrixmultiplikationen)**
- **BiCGSTAB(Biconjugate Gradients Stabilisiert)**
- **CGS(Conjugate Gradient Squared)**
- **CSBCG(Composite Step Biconjugate Gradients)**
- **LAL(Look ahead Lanzcos)**
- **QMR(Quasi Minimal Residual)**
- **TFQMR(Transpose Free QMR)**
- **SymmLQ(Symmetrische LQ-Zerlegung)**

Eine vergleichende Besprechung dieser Verfahren findet etwa in [20, 11] statt.

3.5 Kondition, Fehlerabschätzungen und Fehlerschätzer

In diesem Kapitel wird kurz auf die Kondition der aus der Finite Elemente Diskretisierung von (115) entstehenden Systemmatrizen eingegangen. Danach werden Fehlerabschätzungen für Finite-Elemente Approximationen angegeben. Diese wiederum werden dazu benutzt, einfache Fehlerindikatoren und Fehlerschätzverfahren zur Gittersteuerung zu motivieren.

3.5.1 Kondition

Im Folgenden sei Z_h eine zulässige Triangulierung eines polygonal berandeten Gebietes $\Omega \subset \mathbb{R}^2$ mit Gitterweite h. Über der Triangulierung seien global stetige, stückweise polynomiale Ansatzfunktionen erklärt dergestalt, daß sich auf jedem Dreieck $T_l \in Z_h$ genau n_l Freiheitsgrade ergeben. Es gilt

Satz 3.61. (Kondition von Masse- und Steifigkeitsmatrix)
Bezeichne M_h die Masse- und A_h die Steifigkeitsmatrix der Diskretisierung von (115) mit o.g. Finiten Elementen. Dann gilt

$$\kappa(M_h) = O(1), \text{ und } \kappa(A_h) = O(h^{-2}). \tag{137}$$

Ist die zugrundeliegende Triangulierung Element einer Folge stabiler Triangulierungen (siehe Definition 3.36), so können die Konstanten in den Landau'schen Symbolen unabhängig von der Gitterweite h gewählt werden.

Beweis. Die Matrizen A und M sind symmetrisch und positiv definit. Zunächst wird die Kondition von M abgeschätzt. Es gilt

$$\kappa(M) = \frac{\lambda_{max}(M)}{\lambda_{min}(M)},$$

wobei

$$\lambda_{max}(M) = \max_{|\alpha|=1} \alpha^t M\alpha \text{ und } \lambda_{min}(M) = \min_{|\alpha|=1} \alpha^t M\alpha.$$

Das Skalarprodunkt $\alpha^t M\alpha$ kann mit Hilfe der Element-Massen-Matrizen $M^{(l)}(1 \le l \le nt)$, vgl. Code 3.42, in der Form

$$\alpha^t M\alpha = \sum_{l=1}^{nt} \alpha^{(l)t} M^{(l)} \alpha^{(l)},$$

geschrieben werden, wobei für jedes Element T_l der Vektor $\alpha^{(l)} \in \mathbb{R}^{n_l}$ aus den Komponenten von α besteht, deren Freiheitsgrade zu T_l gehören. Damit ergibt sich sofort

$$\alpha^t M\alpha \ge \sum_{l=1}^{nt} \lambda_{\min}^{(l)} |\alpha^l|^2 \ge p_1 \min_{1\le l\le nt} \lambda_{\min}^{(l)} |\alpha|^2,$$

mit $p_1 := \min_{1\le i\le nf} \{\#T; P_i \text{ Freiheitsgrad in } T\}$ und $\lambda_{\min}^{(l)} = \min_{|\alpha^{(l)}|=1} \alpha^{(l)t} M^{(l)} \alpha^{(l)}$. Analog

$$\alpha^t M\alpha \le p_2 \max_{1\le l\le nt} \lambda_{max}^{(l)} |\alpha|^2,$$

mit $p_2 := \max_{1\le i\le nf} \{\#T; P_i \text{ Freiheitsgrad in } T\}$ und $\lambda_{\max}^{(l)} = \max_{|\alpha^{(l)}|=1} \alpha^{(l)t} M^{(l)} \alpha^{(l)}$. Damit ergibt sich aber für $\lambda_{min}(M)$ und $\lambda_{max}(M)$

$$p_1 \min_{1\le l\le nt} \lambda_{min}^{(l)} \le \lambda_{min}(M) \le \lambda_{max}(M) \le p_2 \max_{1\le l\le nt} \lambda_{max}^{(l)}. \tag{138}$$

Bezeichnet $\hat{T}$ das Einheitsdreieck, so gilt mit (110)

$$m_{n_i n_j}^{(l)} = \int_{T_l} b_i b_j = 2|T_l| \int_{\hat{T}} \tilde{b}_{n_i} \tilde{b}_{n_j} = 2|T_l| \tilde{m}_{n_i n_j},$$

ergo

$$M^{(l)} = 2|T_l| \tilde{M}.$$

Hier bezeichnen n_i die zu den globalen Knotennummern i gehörigen lokalen Knotennummern. Die Matrix $\tilde{M}$ habe die Eigenwerte

$$\tilde{\lambda}_1 \le \tilde{\lambda}_2 \le \ldots \le \tilde{\lambda}_{n_l}.$$

Dann folgt in (138) wegen $\lambda_j^{(l)} = 2T_l|\tilde{\lambda}_j|$ auch

$$2p_1\tilde{\lambda}_1 \min_{1\le l\le nt} |T_l| \le \lambda_{min}(M) \le \lambda_{max}(M) \le 2p_2\tilde{\lambda}_{n_l} \max_{1\le l\le nt} |T_l|,$$

also wegen $\tilde{\lambda}_1 > 0$ die Behauptung für die Kondition der Massematrix M.
Wir zeigen nur $\kappa(A) \le ch^{-2}$. Dazu wird die sogenannte Poincaré-Ungleichung benötigt.

Hilfsatz 3.62. (Poincaré-Ungleichung)
Sei $\Omega \subset \mathbb{R}^n$ beschränktes Gebiet und $y \in H^1(\Omega), y_{|\Gamma_1} = 0, \emptyset \ne \Gamma_1 \subset \partial\Omega$. Dann gibt es eine positive Konstante $c_p = c_p(\Omega)$ mit

$$\left(\int_\Omega |y|^2 dx\right)^{\frac{1}{2}} \le c_p \left(\int_\Omega |\nabla y|^2 dx\right)^{\frac{1}{2}}. \tag{139}$$

Diese Aussage gilt auch für Funktionen $y \in H^1(\Omega)$ mit $\int_\Omega y(x)dx = 0$.

Beweis. [10, Lemma 3.2] □

Weiter im Beweis von Satz 3.61. Mit

$$y_h(x) = \sum_{i=1}^{nf} \alpha_i b_i(x)$$

gilt wegen (139)

$$\alpha^t A \alpha = \int_\Omega |\nabla y_h(x)|^2 dx \geq \frac{1}{c_p^2} \int_\Omega |y_h(x)|^2 dx = \frac{1}{c_p^2} \alpha^t M \alpha,$$

also auch wegen der Uniformität der Triangulierungen

$$\min_{|\alpha|=1} \alpha^t A \alpha =: \mu_{min}(A) \geq \frac{1}{c_p^2} \lambda_{min}(M) \geq \frac{2}{c_p^2} \tilde{\lambda}_1 \min_{1 \leq l \leq nt} |T_l| \geq c_1 h^2.$$

Verbleibt die Abschätzung für

$$\mu_{max}(A) = \max_{|\alpha|=1} \alpha^t A \alpha.$$

Analog zur Behandlung von M bezeichne

$$\mu_1^{(l)} := \min_\beta \frac{\beta^t A^{(l)} \beta}{|\beta|^2}, \quad \mu_{nl}^{(l)} := \max_\beta \frac{\beta^t A^{(l)} \beta}{|\beta|^2}.$$

Dann gilt

$$\mu_{max}(A) \leq p_2 \max_{1 \leq l \leq nt} \mu_{nl}^{(l)} \leq p_2 \|A^{(l)}\|_\infty \leq p_2 n_l \max_{1 \leq r,s \leq n_l} |a_{rs}^{(l)}|. \tag{140}$$

Ferner gilt

$$|a_{rs}^{(l)}| = |\int_{T_l} \nabla b_{k_r} \nabla b_{k_s}| \leq C |T_l| \underbrace{max \left\{ |\xi_{x_1}|, |\xi_{x_2}|, |\eta_{x_1}|, |\eta_{x_2}| \right\}^2}_{= \text{ geom}},$$

vgl. (105), wobei k_r die zum lokalen Freiheitsgrad r gehörige globale Freiheitsgradnummer bezeichnet. Aus der Definition von geom folgt direkt

$$\text{geom } = \mathcal{O}(h^{-2})$$

und daher wegen $|T_l| = \mathcal{O}(h^2)$ sofort

$$\mu_{max}(A) = \mathcal{O}(n_l).$$

Insgesamt demnach

$$\frac{\mu_{max}(A)}{\mu_{min}(A)} \leq c h^{-2}.$$

Eine genauere Analyse zeigt, daß in (140) auch entsprechend nach unten abgeschätzt werden, so daß sich schließlich die Behauptung ergibt, siehe etwa [16, S.236 ff]. □

Bemerkung 3.63. Für Triangulierungen in zwei Raumdimensionen gilt

$$\det F_T \text{ geom } \leq \frac{4}{\sin \alpha},$$

wobei α den kleinsten in der Triangulierung auftretenden Winkel bezeichnet, siehe ebenfalls [16].

Hilfsatz 3.64. (Masse- und Steifigkeitsmatrizen stabiler Triangulierungen)
Sei $\left\{ Z_{h_k}^k \right\}_{k \in \mathbb{N}}$ eine Folge stabiler Triangulierungen (im Sinne von Definition 3.36). Dann gilt für die Masse- und Steifigkeitsmatrizen M_k bzw. A_k, $k \in \mathbb{N}$

$$\begin{cases} \kappa(M_k) &= \mathcal{O}(1) \\ \kappa(A_k) &= \mathcal{O}(h_k^{-2}) \end{cases} \quad \forall\, k \in \mathbb{N}. \tag{141}$$

Beweis. Folgt unmittelbar aus dem Beweis zu Satz 3.61, da dort alle auftretenden Konstanten in Termen der lokalen Freiheitsgrade und der Quotienten σ_k, siehe (103), ausgedrückt werden können. □

3.5.2 Fehlerabschätzungen

Zur Illustration wird wieder das Poissonproblem (65) betrachtet. Es lautet: Finde u mit

$$\left\{ \begin{array}{rcll} -\Delta u(x) & = & q(x) & \text{in} \quad \Omega \subset \mathbb{R}^2 \\ u(x) & = & r(x) & \text{auf} \quad \partial\Omega. \end{array} \right. \tag{142}$$

Im Folgenden sei o.E. $r(x) \equiv 0$, so daß ein homogenes Dirichletproblem vorliegt. Die Diskretisierung von (142) mit finiten Elementen lautet (vgl. 94)

$$\text{Finde } u_h \in W_h \text{ mit } u_h(x) = 0 \text{ auf } \partial\Omega \text{ und}$$

$$\int_{\Omega_h} \nabla u_h(x) \nabla v_h(x) dx = \int_{\Omega_h} q(x) v_h(x) dx \; \forall \, v_h \in V_h, \tag{143}$$

wobei die Räume W_h und V_h in (109) eingeführt wurden. Bemerke, daß hier $V_h = W_h$ gewählt werden kann. Zur Erinnerung

$$V_h = \left\{ v_h \in C^0(\bar{\Omega}_h) \, ; v_{h_{|_T}} \in \mathcal{P}_l(T) \; \forall \, T \in Z_h \right\},$$

wobei Z_h eine Triangulierung von Ω bezeichnet und über die Vereinigung ihrer Dreiecke Ω_h definiert. Aus (143) und der zu (142) assoziierten variationellen Formulierung ergibt sich für die Differenz $u - u_h$ im Falle $\Omega = \Omega_h$

$$\int_{\Omega} \nabla (u - u_h)(x) \nabla v_h(x) dx = 0 \; \forall \, v_h \in V_h, \tag{144}$$

d.h., daß diese Differenz orthogonal zu V_h bzgl. des durch die sogenannte Energie(halb)norm

$$|v|_1^2 := \int_{\Omega} |\nabla v(x)|^2 dx$$

induzierten Skalarproduktes ist. Das Ziel besteht nun darin, den Fehler $u - u_h$ in geeigneten Normen mit bekannten Größen in Termen der Gitterweite h abzuschätzen. Typische Normen wären hier

$$\begin{array}{rcll} ||v||_\infty & := & \operatorname{ess\,sup}_{x \in \Omega} |v(x)|, & \\ ||v||_1 & := & \left(||v||_0^2 + |v|_1^2 \right)^{\frac{1}{2}} & , \quad \text{wobei} \\ ||v||_0 & := & \left(\int_{\Omega} |v(x)|^2 dx \right)^{\frac{1}{2}} & , \quad |v|_1 := (\int_{\Omega} |\nabla v(x)|^2 dx)^{\frac{1}{2}}. \end{array}$$

Es sei vorweggenommen, daß Fehlerabschätzungen bzgl. $|| \cdot ||_\infty$ schwer herzuleiten sind. Daher werden hier zunächst die Normen $|| \cdot ||_0$ und $|| \cdot ||_1$ betrachtet. Mit Hilfe der Poincaré-Ungleichung (139) ergibt sich zunächst

$$||u - u_h||_0^2 \leq c_p^2 |u - u_h|_1^2, \tag{145}$$

so daß zur Abschätzung von $||u - u_h||_1$ die Abschätzung der Halbnorm $|u - u_h|_1$ verbleibt. Grundlegend dafür ist

Satz 3.65. (Céa's Lemma)
Bezeichne u die eindeutig bestimmte variationelle Lösung von (142), u_h die diskrete Lösung von (143). Dann gilt

$$|u - u_h|_1 \leq \inf_{v_h \in W_h} |u - v_h|_1. \tag{146}$$

Beweis. Es gilt

$$\begin{aligned} |u-u_h|_1^2 &= \int_\Omega \nabla(u-u_h)(x)\nabla(u-u_h)(x)dx \\ &= \int_\Omega \nabla(u-u_h)(x)\nabla(u-v_h)(x)dx + \int_\Omega \nabla(u-u_h)(x)\nabla \underbrace{(v_h-u_h)}_{\in V_h}(x)dx \\ &\overset{(3.105)}{=} \int_\Omega \nabla(u-u_h)(x)\nabla(u-v_h)(x)dx \\ &\leq |u-u_h|_1|u-v_h|_1 \ , \ v_h \in W_h \text{ beliebig.} \end{aligned}$$

Damit

$$|u-u_h|_1 \leq |u-v_h|_1 \ , \ v_h \in W_h \text{ beliebig,}$$

also die die Behauptung. □

Sei jetzt u stetig auf $\bar{\Omega}$. Dann kann die Interpolierende

$$(I_h u)(x) := \sum_{i=1}^{nf} u(P_i)b_i(x) \tag{147}$$

sinnvoll definiert werden und es gilt natürlich $I_h u \in W_h$. Damit folgt direkt

$$\inf_{v_h \in W_h} \|u-v_h\|_1 \leq \sqrt{1+c_p^2}\,|u-I_h u|_1,$$

also auch

$$|u-u_h|_1 \leq \sqrt{1+c_p^2}\,|u-I_h u|_1, \text{ unabhängig von } h \text{ und } u. \tag{148}$$

Es verbleibt demnach, den Interpolationsfehler $u-I_h u$ abzuschätzen. Bevor hier ein allgemeines Resultat vorgestellt wird, soll das Vorgehen anhand eindimensionaler Probleme erläutert werden. Dazu sei

$$a = x_0 < x_1 < \ldots < x_n < x_{n+1} = b$$

eine Zerlegung des Intervalls $[a,b]$ und $I_h u$ sei wie in (147) definiert. Dann gilt sicher (für stückweise lineare Interpolation) in (x_i, x_{i+1})

$$\begin{aligned} u'(x) - (I_h u)'(x) &= u'(x) - \tfrac{u(x_{i+1})-u(x_i)}{x_{i+1}-x_i} \\ &= u'(x) - u'(\xi) \qquad , \quad \xi \in (x_i, x_{i+1}) \\ &= \int_\xi^x u''(t)dt, \end{aligned}$$

so daß

$$\begin{aligned} |u'(x) - (I_h u)'(x)| &\leq \left(\int_\xi^x 1^2 dt\right)^{\frac{1}{2}} \left(\int_\xi^x u''(t)^2 dt\right)^{\frac{1}{2}} \\ &\leq |x_{i+1}-x_i|^{\frac{1}{2}} \ \|u''\|_{L^2((x_i,x_{i+1}))} \ , \ x \in [x_i, x_{i+1}]. \end{aligned}$$

Damit folgt sofort

$$\|u' - (I_h u)'\|_0 \leq h\|u''\|_0, \tag{149}$$

vgl. Übungsaufgaben 6.40, 6.41. Für den Interpolationsfehler in der L^2-Norm ergibt sich mit einem ähnlichen Argument

$$\|u - I_h u\|_0 \leq h^2\|u''\|_0, \tag{150}$$

also mehr, als sich aus (149) zusammen mit (145) direkt ergeben würde.
Es gilt der folgende allgemeine Interpolationssatz.

Satz 3.66. (Interpolationsfehler)
Sei $\Omega \subset \mathbb{R}^2$ polygonal berandetes, beschränktes Gebiet und Z_h eine zulässige Triangulierung von Ω mit nf Freiheitsgraden und den dazugehörigen Basisfunktionen $b_1, \ldots, b_{nf}$. Ferner sei $k \in \mathbb{N}$ die kleinste ganze Zahl mit $\mathcal{P}_k(\Omega) \subset V_h = \text{span}\{b_1, \ldots, b_{nf}\}$. Dann gilt für alle $u \in H^{k+1}(\Omega), s = 0, 1$

$$|u - I_h u|_s \leq c[\sin \alpha]^{-s} h^{k+1-s} |u|_{k+1}, \tag{151}$$

wobei α den kleinsten Winkel in Z_h und h die Gitterweite der Triangulierung bezeichnen. Die Konstante c ist positiv und unabhängig von h und α.

Beweis. [16, Theorem 5.6] □

Mit diesem Satz folgt für $s = 1$ aus (146) und (145) sofort

Satz 3.67. (Fehlerabschätzung für das Poissonproblem)
Mit der Notation aus Satz 3.66 gilt für den Fehler $u - u_h$

$$\|u - u_h\|_1 \leq c|\sin \alpha]^{-1} h^k |u|_{k+1}. \tag{152}$$

Insbesondere wird für den L^2-Fehler $\|u - u_h\|_0$ nur die Ordnung 1 erhalten.

Mit dem sogenannten **Nitsche-Trick** (siehe [3]) kann die Fehlerabschätzung für die L^2-Norm verbessert werden.

Satz 3.68. (L^2-Fehler)
Es gelten die Voraussetzungen von Satz 3.66 mit $k = 1$. Dann gilt

$$\|u - u_h\|_0 \leq c\ h^2 |u|_2. \tag{153}$$

Beweis. [3] □

Unter gewissen Voraussetzungen an das Gebiet Ω und die Daten können auch Fehlerabschätzungen in der L^∞-Norm bewiesen werden. So gilt etwa für Interpolation in einer Raumdimension im Intervall $[x_{i-1}, x_i]$ (Interpolation linear)

$$u(x) - (I_h u)(x) = \frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x} \int_{x_{i-1}}^{xi} \int_{\eta}^{\xi} u''(s)\ ds\ d\eta\ d\xi, \tag{154}$$

vgl. Übungsaufgabe 6.40. Damit ergibt sich

$$\sup_{x \in I} |u(x) - (I_h u)(x)| \leq c\ h^{\frac{3}{2}} \|u''\|_0, \tag{155}$$

falls $u \in H^2(I)$ ist, und auch

$$\sup_{x \in I} |u(x) - (I_h u)|(x)| \leq c\ h^2 \|u''\|_\infty, \tag{156}$$

falls $u \in H^{2,\infty}(I)$ ist.
Mehr als die in (155) und (156) auftretenden Ordnungen darf man auch in höheren Raumdimensionen nicht erwarten. Es gilt

Satz 3.69. (L^∞-Fehlerabschätzungen)
Sei $u \in H^2(\Omega)$ Lösung des Modellproblems (142) und u_h aus (144) zugehörige lineare Finite-Element Approximation. Dann gilt

$$||u - u_h||_\infty \leq C\ h\ |u|_2 \tag{157}$$

mit einer von h unabhängigen positiven Konstanten C. Ist $u \in H^{2,\infty}(\Omega)$, so kann

$$||u - u_h||_\infty \leq ch^2\ |\ln h|\ |u|_{2,\infty} \tag{158}$$

nachgewiesen werden.

Beweis. [10, Satz 4.5, Bemerkung 4.] □

Zur Erinnerung hier noch einmal die Definition der verwendeten Halbnormen; $|u|_2 := \|D^2 u\|_0$ und $|u|_{2,\infty} := \|D^2 u\|_\infty$.

3.5.3 Fehlerabschätzer und Fehlerindikatoren

Die Fehlerabschätzungen aus der vorhergehenden Sektion sollen jetzt dazu verwendet werden, die Triangulierungen des diskretisierten Problems adaptiv im Sinne von Gleichverteilung des Fehlers über die Dreiecke der Triangulierung anzupassen. Ein solches Verfahren kann notwendig werden, wenn die zu approximierende Funktion etwa Singularitäten (in ihren Ableitungen) besitzt. Beispiele dafür finden sich in der Übungsaufgabe 6.30.
Hier zunächst das Prinzip.

Prinzip Zu einer vorgelegten Toleranz *tol* soll die Triangulierung adaptiv so eingestellt werden, daß der Gesamtfehler die Toleranz unterschreitet und der Fehlerbeitrag auf jedem Dreieck der finalen Triangulierung in etwa gleich groß ist. Idealerweise sollte die Triangulierung optimal sein in dem Sinne, daß es keine Triangulierung mit weniger Freiheitsgraden gibt, bei deren Verwendung die Toleranz unterschritten wird. □

Als Modellproblem wird in Übungsaufgabe 6.30 die Teilaufgabe a) mit $c = 0$ betrachtet, d.h. gesucht ist eine Funktion u mit

$$\left\{ \begin{array}{rcll} -\Delta u & = & 1 & \text{in} \quad \Omega \\ u & = & g & \text{auf} \quad \Gamma_D \\ \partial_\eta u & = & 0 & \text{auf} \quad \Gamma_N. \end{array} \right. \tag{159}$$

Dabei ist $\Omega := \{r\cos\phi, r\sin\phi); 0 \le r \le 1, 0 < \phi < \alpha\pi\}$ und $\Gamma_N = \emptyset, \Gamma_D = \partial\Omega$. Die Randwerte seien durch die Funktion

$$g(r,\varphi) := r^{\frac{1}{\alpha}} \sin\frac{\varphi}{\alpha} - \frac{r^2}{4}.$$

festgelegt. Dann gilt wegen

$$\Delta = \partial_{rr} + \frac{1}{r}\partial_r + \frac{1}{r^2}\partial_{\varphi\varphi},$$

daß $u(r,\varphi) := g(r,\varphi)$ das Problem (159) löst. Eine leichte Rechnung zeigt, daß mit $\beta := \frac{1}{\alpha}$

$$u \in H^{1,p}(\Omega) \quad \forall_{1\le p<\frac{2}{1-\beta}}, \beta < 1$$

und für $2 > \beta \ge 1$ gilt

$$u \in H^{2,p}(\Omega) \quad \forall_{1\le p<\frac{2}{2-\beta}}.$$

Die Lösung genügt demnach nicht den Regularitätsanforderungen des Satzes 3.67. Vielmehr gilt hier für $\beta \le 1$

$$|u - u_h|_1 \le c\, h^{\beta-\epsilon}\, |u|_{1,\frac{2}{1-\beta}-\epsilon} \text{ für alle } \epsilon > 0, \tag{160}$$

d.h., die **Approximationseigenschaft der numerischen Lösung nimmt mit der Regularität der kontinuierlichen Lösung ab**. Tatsächlich ist der Regularitätsverlust bei diesem Beispiel aber ein lokales Phänomen bei $x = 0$, siehe etwa [14]. Es besteht begründete Hoffnung, daß ein lokal feines Gitter bei $x = 0$ die Approximationseigenschaften der numerischen Lösung verbessert.

Bemerkung 3.70. Im Allgemeinen ist die Lage von Singularitäten nicht bekannt. Es ist daher wünschenswert, diese durch den Verfeinerungsprozeß zu lokalisieren.

Die in den nachfolgend dargestellten Zugängen verwendeten finiten Elemente sind linear.

Fehlerindikator nach Erikson/Johnson, a-priori Zugang: Sei $Z_h = \{T_1, \ldots, T_{nt}\}$ eine Triangulierung und u_h eine auf Z_h zu u berechnete Finite-Element Approximation. Fordere für eine vorgelegte Toleranz *tol*

$$|u - u_h|_1^2 = \sum_{T\in Z_h} |u - u_h|_{1,T}^2 \le C^2 \sum_{T\in Z_h} h_T^2 |u|_{2,T}^2 \stackrel{!}{=} tol^2. \tag{161}$$

Um den Fehler auf jedem Dreieck gleich zu verteilen, fordere, daß

$$C\ h_T|u|_{2,T}$$

auf jedem Dreieck die gleiche Größe hat. Verfeinere also das Dreieck T, falls

$$h_T|u|_{2,T} > \frac{tol}{C\sqrt{nt}} \tag{162}$$

gültig ist. Liegt eine L^∞-Fehlerabschätzung wie in (157) oder (158) vor, fordere einfach, daß der Fehler über jedem Dreieck höchstens so groß wie *tol* ist. Es ergibt sich als Forderung für Verfeinerung

$$h_T|u|_{2,T} > \frac{tol}{C} \tag{163}$$

im Fall von (156) und

$$h_T^2\ |\ln h_T|\ |u|_{2,\infty,T} > \frac{tol}{C} \tag{164}$$

im Fall (158). In die Verfeinerungskriterien (162) bis (164) geht noch die gesuchte kontinuierliche Lösung u ein. Da diese i.d.R. nicht bekannt ist, werden die auftretenden zweiten Ableitungen von u durch Differenzenquotienten, gebildet mit ∇u_h, ersetzt. Um dieses Vorgehen näher zu illustrieren, betrachte die lokale Situation an einem beliebigen Dreieck T der Triangulierung. Es bezeichnen T_1, T_2, T_3 die Nachbardreiecke von T und S_1, S_2, S_3, S die entsprechenden Schwerpunkte, siehe Abb. 8. Dann wird gesetzt

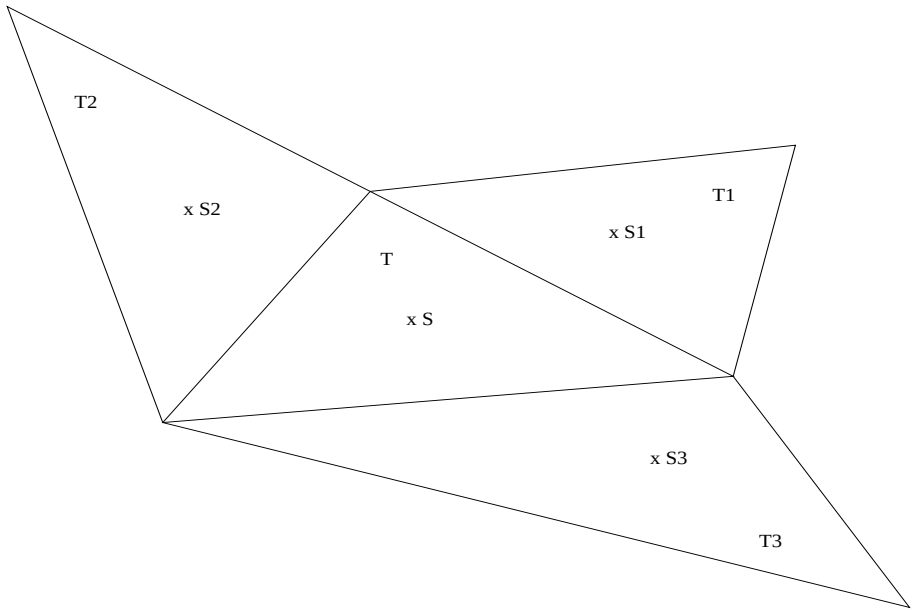


Abbildung 8: Differenzenapproximation von D^2u

$$H_j := |< S - S_1, e_j >| + |< S - S_2, e_j >| + |< S - S_3, e_j >|$$

und (beachte lineare finite Elemente!)

$$|u_{x_i x_j}(S)| \approx \sum_{k=1}^{3} \frac{|< S - S_k, e_j >|}{H_j} \frac{|< (\nabla u_h(S) - \nabla u_h(S_k))e_i >|}{H_j} =: D_{ij}u_h(T), \tag{165}$$

womit sich

$$|u|_{2,\infty,T} \approx \max_{1\le i,j\le 2} D_{ij}u_h(T) =: D^2_\infty(u_h, T) \tag{166}$$

und bei Verwendung der Schwerpunktregel für die Quadratur

$$|u|_{2,T} \approx \left\{ \sum_{i,j=1}^{2} \left(D_{ij} u_h(T)\right)^2 |T| \right\}^{\frac{1}{2}} =: D_2^2(u_h, T) \tag{167}$$

ergibt. In algorithmischer Form liest sich ein Lösungsalgorithmus mit adaptiver Verfeinerung wie folgt.

Algorithmus 3.71. (Adaptive Verfeinerung)
Sei eine Ausgangstriangulierung Z_{h_0} gegeben. Setze $Z_h = Z_{h_0}$ und gebe die Toleranz *tol* und die Konstante $c > 0$ vor.

i.) Berechne die FE-Lösung u_h zu Z_h

ii.) Berechne für alle $T \in Z_h$ entweder

$D_\infty^2(u_h, T)$

oder

$D_2^2(u_h, T)$

Das ist abhängig von der gewünschten Norm.

iii.) Ist (hier wird nur der Fall (164) betrachtet)

$h_T^2 |\ln h_T| D_\infty^2(u_h, T) > \frac{tol}{C}$,

so markiere T zur Verfeinerung.

iv.) Verfeinere alle markierten Dreiecke in Z_h, mache die resultierende Triangulierung regulär, nenne sie $Z_{\frac{h}{2}}$.

v.) Ist $Z_{\frac{h}{2}} = Z_h$, stop. Sonst setze $Z_h := Z_{\frac{h}{2}}$ und gehe zu i.)

Weitere Einzelheiten liefern Erikson und Johnson in [8].
Ein wesentliches Kriterium dafür, ob Algorithmus 3.71 optimale Triangulierungen liefert, beschreibt die asymptotische Exaktheit eines Fehlerindikators oder Fehlerschätzers. So würde etwa der Fehlerindikator, basierend auf (167), **asymptotisch exakt** sein, wenn mit Konstanten $E_h, D_h > 0$

$$D_h \sum_{T \in Z_h} h_T^2 \left[D_2^2(u_h, T)\right]^2 \leq |u - u_h|_1^2 \leq E_h \sum_{T \in Z_h} h_T^2 \left[D_2^2(u_h, T)\right]^2 \tag{168}$$

und

$$D_h / E_h \to 1 \quad (h \to 0)$$

richtig wäre. Gilt nur (168) mit $\frac{D_h}{E_h} \geq \epsilon > 0 \quad \forall_{h>0}$, würde der **Fehlerindikator stabil oder brauchbar** genannt. Es sei bemerkt, daß der auf (167) basierende Indikator in diesem Sinne i.A. nicht brauchbar ist. Im Folgenden wird ein Fehlerschätzer vorgestellt, der auf Bank und Weiser [1], zurückgeht und für parallele Triangulierungen asymptotisch exakt ist ([6]). Parallele Triangulierungen sind solche, die durch 3 Scharen paralleler Geraden erzeugt werden können.

Fehlerschätzer nach Bank/Weiser, a-posteriori Zugang: Wurde beim Fehlerindikator von Erikson/Johnson noch mit Hilfe von a-priori Fehlerabschätzungen argumentiert und das Gitter gesteuert, wird beim nachfolgend skizzierten Zugang eine Abschätzung des lokalen Fehlers in Termen aktuell zu berechnender Größen ausgenutzt. Als Modellproblem diene hier

$$\begin{cases} -\Delta u &= q & \text{in } \Omega \\ u &= 0 & \text{auf } \Gamma_D \\ \partial_\eta u &= g & \text{auf } \Gamma_N \end{cases} \tag{169}$$

Sei u_h Finite Element Lösung zu (169) auf Z_h und

$$e := u - u_h.$$

Dann gilt für e (beachte, daß u_h stückweise linear)

$$\int_\Omega \nabla e \nabla v dx = \int_\Omega q v dx + \int_{\Gamma_N} g v d\Gamma - \sum_{T \in Z_h} \int_{\partial T} \partial_\eta u_h v d\Gamma \tag{170}$$

für alle $v \in H^1_{\Gamma_D}(\Omega) := \left\{ v \in H^1(\Omega); v_{|_{\Gamma_D}} = 0 \right\}$. Dabei ist für $v \in H^1_{\Gamma_D}(\Omega)$

$$\int_\Omega \nabla u_h \nabla v dx = \sum_{T \in Z_h} \int_T \nabla u_h \nabla v dx = \sum_{T \in Z_h} \left\{ \underbrace{\int_T -\Delta u_h v dx}_{=0} + \int_{\partial T} \partial_\eta u_h v d\Gamma \right\}.$$

Bezeichnet E_T die Menge der Kanten des Dreiecks T, so kann (170) mit Hilfe von

$$J_l := \begin{cases} \partial_\eta u_{h_{|_{T_1}}} - \partial_\eta u_{h_{|_{T_2}}} & , \quad \bar{T}_1 \cap \bar{T}_2 = l \\ 2(g - \partial_\eta u_{h_{|_T}}) & , \quad \bar{T} \cap \Gamma_N = l \\ 0 & , \quad \bar{T} \cap \Gamma_D = l \end{cases}$$

umgeschrieben werden als

$$\int_\Omega \nabla e \nabla v dx = \sum_{T \in Z_h} \left\{ \int_T q v dx + \frac{1}{2} \sum_{l \in E_T} \int_l J_l v d\Gamma \right\} \quad \forall v \in H^1_{\Gamma_D}(\Omega). \tag{171}$$

Der Fehlerschätzer wird jetzt wie folgt konstruiert. Setze

$$\mathcal{P}_2^0(T) := \{ v \in \mathcal{P}_2(T) \ , \ v(P) = 0 \ , \ P \text{ Ecke von } T \},$$

löse lokal

$$\int_T \nabla e_T \nabla v dx = \int_T q v dx + \frac{1}{2} \sum_{l \in E_T} \int_l J_l v d\gamma \quad \forall v \in \mathcal{P}_2^0(T) \tag{172}$$

und setze

$$\begin{cases} \eta_T & := & ||\nabla e_T||_{0,T}, \\ \eta & := & \left(\sum_{T \in Z_h} \eta_T^2 \right)^{\frac{1}{2}}. \end{cases} \qquad \textbf{Fehlerschätzer} \tag{173}$$

Verfeinert werden sollen alle Dreiecke, für die

$$\eta_T > \frac{tol}{\sqrt{nt}} \tag{174}$$

gilt. Der Ausdruck η heißt **Fehlerschätzer** (hier für die Energienorm $|\cdot|_1$). Wird in Algorithmus 3.71 die Abfrage unter iii.) durch (174) ersetzt, so kann für **parallele Triangulierungen** (Vereinigung zweier benachbarter Dreiecke ist ein Parallelogramm) gezeigt werden [6], daß

$$\frac{|u_h - u|_1}{\eta} \ \to \ 1 \quad (h \to 0). \tag{175}$$

Die lokale Verfeinerung der Triangulierungen in Algorithmus 3.71 kann mit Algorithmus 3.34 oder Algorithmus 3.35 erfolgen.
Die Berechnung von e_T in (172) verlangt auf jedem Dreieck die Lösung eines (3 x 3) Gleichungssystems mit positiv definiter Koeffizientenmatrix. Der Raum $\mathcal{P}_2^0(\hat{T})$ wird dabei aufgespannt durch die Basisfunktionen

$$\begin{aligned} b_1(\xi,\varphi) &= 4\xi(1-\xi-\varphi) \\ b_2(\xi,\varphi) &= 4\eta(1-\xi-\varphi) \\ b_3(\xi,\varphi) &= 4\xi\varphi, \end{aligned}$$

vgl. den quadratischen Ansatz nach (108).

4 Numerische Behandlung parabolischer Probleme

Im vorliegenden Kapitel geht es um die numerische Behandlung der Gleichung (6). Zu diesem Zweck definiere formal

$$(Lu)(t,x) := -\mathrm{div}(K(t,x)\nabla_x u(t,x)), \tag{176}$$

setze

$$b := -v_T(t,x) \quad \equiv \quad const$$

und fordere, daß die Matrix $K(t,x)$ symmetrisch und gleichmäßig positiv definit in (t,x) ist, d.h. für ein $\alpha > 0$ gilt $\xi^t K(t,x)\xi \geq \alpha|\xi|^2$ für alle (t,x). Damit kann (6) geschrieben werden als

$$\left\{ \begin{array}{rcll} u_t(t,x) + (Lu)(t,x) + b^t\nabla_x u(t,x) + cu(t,x) &=& q(t,x) & \text{in } \Omega^T \\ RW(u)(t,x) &=& r(t,x) & \text{auf } \partial\Omega^T \\ u(0,x) &=& u_0(x) & \text{in } \Omega. \end{array} \right. \tag{177}$$

Dabei bezeichnet Ω^T den zu Ω gehörigen Raum-Zeit-Zylinder, i.e.

$$\Omega^T := (0,T) \times \Omega$$

und $(0,T)$ ist der Zeithorizont, auf dem das Problem (6) betrachtet werden soll. In der Literatur wird häufig auch Q als Bezeichnung für Ω^T verwendet.
Zunächst soll eine geeignete variationelle Formulierung von (177) bereitgestellt werden. Zu diesem Zweck sei X ein beliebiger Banachraum (=vollständiger normierter Raum). Dann besteht der Raum

$$L^2(0,T;X) \quad (\text{kurz } L^2(X))$$

aus allen meßbaren Funktionen $u : (0,T) \to X$, für die

$$\|u\|_{L^2(X)} := \left(\int_0^T \| u(t) \|_X^2 \, dt \right)^{\frac{1}{2}}$$

endlich ist. Damit ist $L^2(X)$ selbst ein Banachraum. Im Folgenden sei X separabel und reflexiv. Ist dann X^* der Dualraum von X (Menge der stetigen linearen Funktionale auf X), so ist $L^2(X^*)$ der Dualraum von $L^2(X)$, i.e.

$$L^2(X^*) = L^2(X)^*.$$

Zur variationellen Formulierung sei jetzt

$$V := H_0^1(\Omega), \quad H := L^2(\Omega).$$

Dann gilt $V^* = H^{-1}(\Omega)$. Definiere

$$W(0,T) := \left\{ v \in L^2(V);\ v_t \in L^2(V^*) \right\}$$

und für $t \in (0,T)$ die **Bilinearform** $a(t) : V \times V \to \mathbb{R}$ gemäß

$$a(t)(w,v) := \int_\Omega K(t,x)\nabla w(x)\nabla v(x)dx.$$

Die variationelle Formulierung von (177) für das Dirichletproblem ($RW(u)(t,x) = u(t,x)$) lautet dann:

$$\text{(178)} \qquad \begin{cases} \text{Finde } u \in W(0,T) \text{ mit } u(t,x) = r(t,x), x \in \partial\Omega, u(0,x) = u_0(x) \text{ und} \\ \frac{d}{dt}\int_\Omega u(t,x)v(x)dx \; + a(t)(u(t,\cdot),v) \; + \int_\Omega b^t\nabla_x u(t,x)v(x)dx \\ + \; c\int_\Omega u(t,x)v(x)dx \; = \int_\Omega q(t,x)v(x)dx \quad \forall\, v \in V, \text{ für fast alle } t \in (0,T). \end{cases}$$

Um Schreiberei zu reduzieren, sei ab jetzt $b = 0, c = 0$.

4.1 Vertikale Linienmethode

Die Aufgabe (178) wird jetzt gemäß (77) im Ort (semi-) diskretisiert. Dazu sei $V_h \subset V$ Finite Element Raum und

$$"W_h = I_h r + V_h"$$

der affine Raum, der aus V_h und einer Finite Element Interpolation $I_h r$ der (nach Ω fortgesetzten) Randwerte gebildet wird. Hier kann auch eine Projektion verwendet werden, siehe (181). Das zu (178) korrespondierende diskrete Problem (für $b = 0, c = 0$) lautet dann

$$\text{(179)} \qquad \begin{cases} \text{Finde } u_h(t) \in W_h \text{ mit } u_h(0) \; = \; u_h^0 \in V_h \; (\in W_h) \\ \text{und} \\ \frac{d}{dt}\int_\Omega u_h(t,x)v_h(x)dx \; + \; a(t)(u_h(t,\cdot),v_h) \; = \int_\Omega q(t,x)v_h(x)dx \; \forall\, v_h \in V_h, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{für fast alle } t \in (0,T). \end{cases}$$

Die Anfangswerte u_h^0 für u_h können dabei aus der Finite-Element-Interpolation

$$\text{(180)} \qquad u_h^0 := I_h u_0,$$

oder aus der L^2-Approximation u_h^0 im Finite-Element-Raum definiert durch

$$\text{(181)} \qquad \int_\Omega u_h^0(x)v_h(x)dx := \int_\Omega u_0(x)v_h(x)dx \qquad \forall\, v_h \in V_h$$

gewonnen werden. Aus (179) ergibt sich ein System gewöhnlicher Differentialgleichungen. Dazu seien $b_{1,\ldots}b_{nf}$ die zu den Freiheitsgraden einer gegebenen Zerlegung Z_h von Ω gehörenden FE-Basisfunktionen. Mit dem Ansatz

$$\text{(182)} \qquad u_h(t,x) = \sum_{i=1}^{ni} u_i(t)b_i(v) + \sum_{i=ni+1}^{nf} r(t,P_i)b_i(x)$$

ergibt sich in (4.4) mit

$$U(t) := (u_1(t), \ldots, u_{ni}(t))^t$$

das System gewöhnlicher Differentialgleichungen

$$\text{(183)} \qquad \begin{cases} M\dot{U} + A(t)U \; = \; F(t) \quad \text{in } (0,T), \\ U(0) \; = \; U_0, \end{cases}$$

wobei

$$M = (m_{ij})_{i,j=1}^{ni}, \qquad m_{ij} := \int_\Omega b_i(x)b_j(x)dx$$
$$A(t) = (a_{ij}(t))_{i,j=1}^{ni}, \qquad a_{ij}(t) := a(t)(b_i,b_j)$$

und

$$F(t) = (f_i)_{i=1}^{ni} \quad , \quad f_i := \int_\Omega q(t,x) b_i(x) dx \; - \sum_{j=ni+1}^{nf} r(t,P_j) a(t)(b_j, b_i) - \sum_{j=ni+1}^{nf} r_t(t,P_j) \int_\Omega b_j(x) b_i(x) dx.$$

Die Anfangswerte sind dabei im Fall (180) durch

$$U_0 = (u_0(P_1), \ldots, u_{ni}(P_{ni}))^t,$$

gegeben und im Fall von (181) durch das lineare Gleichungssystem

$$MU_0 = \left[\int_\Omega u_0(x) b_i(x) dx \right]_{i=1}^{ni}$$

bestimmt.
Für den Fehler der Semi-Diskretisierung gilt

Satz 4.1. (L^2-Fehler der Semi-Diskretisierung)
Sei U Lösung von (183) und $u_h(t,x)$ die zugeordnete Finite Elemente Approximation. Dann gilt

$$\| u(t,\cdot) - u_h(t,\cdot) \|_0 \;\leq\; e^{-\alpha t} \| u_0 - u_h^0 \| + ch^s \left\{ \| u(t) \|_s \; + \int_0^t e^{-\alpha(t-\kappa)} \| u_t \|_s \, d\kappa \right\}, \tag{184}$$

wobei $u(t,x)$ als ausreichend glatt vorausgesetzt ist. Der Exponent s ist hier wieder die größte ganze Zahl mit $\mathcal{P}_s(\Omega) \subset V_h$. Die Konstanten c und α sind positiv, wobei α der durch die Matrix $K(t,x)$ definierten Positivitätskonstanten entspricht.

Beweis. [10, Satz 6.5] □

Zur numerischen Integration von (183) können jetzt alle in Anhang (A) vorgestellten Verfahren verwendet werden. Exemplarisch werde hier das **Zwischenschritt-σ-Schema** betrachtet. Dabei wird die Zeitableitung $\dot{U}$ in (183) durch Vorwärtsdifferenzen mit Zeitschrittweite Δt ersetzt. Sei jetzt

$$t_0 := 0, \quad t_i := i\Delta t, \text{ wobei } n\Delta t = T,$$

ein Zeitgitter und $U^k := U(t_k)$. Das System (183) wird dann wie folgt diskretisiert:

$$M \frac{U^{k+1} - U^k}{\Delta t} + \sigma A(t_{k+1}) U^{k+1} + (1-\sigma) A(t_k) U^k = \underbrace{\sigma F(t_{k+1}) + (1-\sigma) F(t_k)}_{F^{k+\sigma}}, \quad \sigma \in [0,1]. \tag{185}$$

Für $\sigma = 0$ ist das Verfahren explizit (obwohl in jedem Schritt mit der Massematrix M gelöst werden muß) und heißt **Explizites Euler Verfahren**, sonst implizit. Für $\sigma = \frac{1}{2}$ ergibt das **Crank-Nicolson Verfahren**, für $\sigma = 1$ das **Implizite Euler Verfahren**.
Bevor wir uns um die Stabiliät und Konsistenz dieses Verfahrens kümmern, hier bereits eine Fehlerabschätzung (und demnach auch ein Konvergenzresultat) für das voll diskretisierte System. Für den Fehler $u_h^k(x) - u(t_k, x)$ gilt

Satz 4.2. (L^2-Fehler für das Voll-Diskrete Problem)
Es sei für $k = 0, \ldots, n-1$ der Vektor U^k die Lösung von (185), $U^k = (u_1^k, \ldots, u_{ni}^k)$, und $u_n^k(x) := \sum_{i=1}^{ni} u_i^k b_i(x)$. Dann gilt mit der Lösung u von (179) und für $0 < \sigma \leq 1$

$$\|u(t_k \cdot) - u_h^k\| \leq \|u_h^0 - u_0\| + c_1 h^s \left(\|u_0\|_s + \int_0^{t_k} \|u_t\|_s dt \right) + c_2 \Delta t \int_0^{t_k} \|u_{tt}\|_0 dt, \tag{186}$$

wobei s wie in Satz 4.1 definiert ist. Für $\sigma = \frac{1}{2}$ gilt sogar

$$\|u(t_k, \cdot) - u_h^k\| \leq \|u_h^0 - u_0\| + c_1 h^s \left(\|u_0\|_s + \int_0^{t_k} \|u_t\|_s ds \right) + C_3 \Delta t^2 \int_0^{t_k} \|u_{ttt}\|_0 + \|\Delta u_{tt}\|_0 dt. \tag{187}$$

Beweis. [10, Satz 6.7] □

Bevor die Stabilität und Konstistenz untersucht wird, noch eine kurze Diskussion des numerischen Aufwands des Zwischenschritt-σ-Verfahrens. Der numerische Aufwand in (185) besteht in der Lösung der Gleichungssysteme (U^0 gegeben)

$$\{M + \sigma\Delta t A(t_{k+1})\}\, U^{k+1} = \{M - (1-\sigma)\Delta t A(t_{k+1})\}\, U^k + \Delta t F^{k+\sigma}, \quad k = 0, \ldots, n-1. \tag{188}$$

Die Koeffizientenmatrix ist für alle $0 \leq \sigma \leq 1$ positiv definit, da die Bilinearform $a(t)$ als gleichmässig positiv definit vorausgesetzt wurde.
Um die Darstellung übersichtlicher zu gestalten, gelte jetzt

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

so daß es sich in (185) um die in Ort und Zeit diskretisierte **Wärmeleitungsgleichung** handelt. Die Matrix A ist damit unabhängig von t und stimmt mit der Steifigkeitsmatrix (Definition 3.40) überein. (185) geht dann über in

$$(M + \sigma\Delta t A)U^{k+1} \;=\; (M - (1-\sigma)\Delta t A)U^k + \Delta t F^{k+\sigma}, \quad U^0 \text{ gegeben.}$$

Das ist eine Vorschrift der Art

$$U^{k+1} \;=\; QU^k + \Delta t G^k, \quad U^0 \text{ gegeben,} \tag{189}$$

wobei in (185) mit $\tilde{A} := M^{-1}A$

$$Q = (I + \sigma\Delta t\tilde{A})^{-1}(I - (1-\sigma)\Delta t\tilde{A}) \text{ und } G^k = (I + \sigma\Delta t\tilde{A})^{-1}M^{-1}F^{k+\sigma}. \tag{190}$$

gilt. Unser nächstes wird es sein, die Stabilität von (188) zu untersuchen. Zuvor allerdings definieren wir Stabilität und Konsistenz und geben den Satz von **Lax** and, der besagt, daß ein konsistentes und stabiles Verfahren auch konvergiert.

Definition 4.3. (Stabilität)
Die Vorschrift (189) heißt stabil (bzgl. der Norm $\|\cdot\|$), falls mit einer positiven Konstanten K

$$\|U^k\| \leq K \left(\|U^0\| \;+\; \Delta t \sum_{j=0}^{k-1} \|G^j\| \right), \quad k \in \mathbb{N}$$

richtig ist.

Definition 4.4. (Konsistenz)
Das Differenzenschema (189) sei aus der Diskretisierung von (177) mit Ortsschrittweite Δx und Zeitschrittweite Δt hervorgegangen. Das Schema (189) heißt konsistent (bzgl. der Norm $\|\cdot\|$), falls für die Lösung u von (177)

$$u^{k+1} = Qu^k + \Delta t\; G^k + \Delta t \tau^k$$

gilt und der **Abschneidefehler** $\|\tau^k\| \to 0$ $(\Delta x, \Delta t \to 0)$ erfüllt, sowie $\|u_h^0 - u\| \to 0$ für $h \to 0$. Dabei bezeichnet $u^k \in \mathbb{R}^n$ den Vektor mit den Einträgen $u(k\Delta t, P_i), i = 1, \ldots, ni$.
Gilt

$$\|\tau^k\| = O(\Delta x^p) + O(\Delta t^q) \quad (\Delta x, \Delta t \to 0) \text{ und } \|u_h^0 - u\| = O(\Delta x^p),$$

so heißt (189) konsistent von der Ordnung (p, q).

Definition 4.5. (Konvergenz)
Voraussetzungen und Notation wie in Def. 4.3, 4.4. Dann heißt das Schema (189) bzgl. $\|\cdot\|$ konvergent (von der Ordnung (p, q)), falls

$$\|U^k - u^k\| \to 0 \quad (= O(\Delta x^p) + O(\Delta t^q))$$

für $\Delta x, \Delta t \to 0$ gilt.

Wichtig ist der folgende Satz, der besagt, daß **Konsistenz und Stablilität Konvergenz liefern**.

Satz 4.6. (Lax)
Voraussetzungen und Notation wie in Definition 4.5. Ist (189) konsistent (von der Ordnung (p, q)) und stabil, so ist (189) konvergent (von der Ordnung (p, q)).

Allgemeinere Definitionen von Stabilität und Konvergenz finden sich in [19, Chapter 2].
Zunächst macht die Definition von Q in (190) nur Sinn, falls $(I + \sigma\Delta t\tilde{A})^{-1}$ existiert. Ein hinreichendes Kriterium dafür ist etwa gegeben durch

$$\sigma\Delta t\|\tilde{A}\| < 1.$$

Aus Hilfsatz 3.43 wissen wir allerdings, daß A und M positiv definit sind, demnach $(I+\sigma\Delta t\tilde{A})^{-1}$ wegen $\sigma\Delta t \geq 0$ immer existiert. Die Definition von Q in (190) ist daher immer sinnvoll.
Als nächstes soll untersucht werden, unter welchen Bedingungen an σ und an die zugrundeliegende Triangulierung Z_h das Verfahren (188) stabil bzgl. der Euklidischen Norm ist.

Satz 4.7. (Stabilität für das Zwischenschritt-σ-Verfahren, Euklidische Norm)
Es bezeichne ξ den Durchmesser des kleinsten Inkreises in der Triangulierung Z_h, Δt die Zeitschrittweite. Dann ist das Schema (188) zur Diskretisierung von (177) mit $L = -\Delta, b \equiv 0, c = 0$ stabil, falls

$$\frac{1}{2} \leq \sigma \leq 1,$$

und konditional stabil für $0 \leq \sigma < \frac{1}{2}$, s.h. stabil, falls zusätzlich

$$12(1-2\sigma)\Delta t < \xi^2 \text{ geht vielleicht auch mit 9 anstelle von 12} \tag{191}$$

gilt.

Beweis. Zunächst bemerke, daß mit λ Eigenwert von $\tilde{A} := M^{-1}A$ die Zahl

$$s := (1 + \Delta t\sigma\lambda)^{-1}(1 - \Delta t(1-\sigma)\lambda) \tag{192}$$

Eigenwert der Matrix Q in (190) ist, denn Q und $\tilde{A}$ besitzen dieselben Eigenvektoren. Wegen der Symmetrie von Q ist

$$|s| \leq 1$$

ein hinreichendes Kriterium für die Stabilität, was wegen (190) wiederum äquivalent zu

$$\begin{cases} 1 - \Delta t(1-\sigma)\lambda & \leq & 1 + \Delta t\sigma\lambda & \text{(immer erfüllt) und} \\ 1 - \Delta t(1-\sigma)\lambda & \geq & -1 - \Delta t\sigma\lambda, \end{cases}$$

ist, d.h.

$$\Delta t\lambda(1-2\sigma) \overset{!}{\leq} 2. \tag{193}$$

Wegen der Positivität von λ und Δt ist diese Ungleichung immer erfüllt, falls nur

$$\frac{1}{2} \leq \sigma \leq 1$$

richtig ist, womit der erste Teil der Behauptung bewiesen ist.
Sei jetzt $0 \leq \sigma < \frac{1}{2}$. Aus (193) kann eine hinreichende Stabilitätsbedingung gewonnen werden, falls $\lambda_{max}(\tilde{A})$ in (193) durch eine obere Schranke (in Termen von ξ und/oder h) ersetzt wird. Dazu dient der folgende Hilfsatz.

Hilfsatz 4.8. (Inverse Ungleichung)
Es sei V_h der über Z_h definierte Finite Element Raum mit linearen Elementen. Dann gilt für alle $v \in V_h$

$$\|\nabla v\|^2 \leq \frac{18}{\xi^2}\|v\|^2, \tag{194}$$

wobei ξ den Radius des kleinsten Inkreises in Z_h bezeichnet.

Beweis. Auf einem beliebigen Dreieck $T \in Z_h$ gilt

$$\int_T |\nabla v|^2 = \int_T |v_1 \nabla b_1 + v_2 \nabla b_2 + v_3 \nabla b_3|^2 \leq 3|T| \left\{ |v_1 \nabla b_1|^2 + |v_2 \nabla b_2|^2 + |v_3 \nabla b_3|^2 \right\}.$$

Ferner gilt (siehe Abb. 9)

$$|\nabla b_1| = \frac{|b_1(P_1) - b_1(Q_1)|}{|\overline{P_1 Q_1}|} = \frac{1}{|\overline{P_1 Q_1}|} = \frac{|\overline{P_2 P_3}|}{2|T|}.$$

Ergo

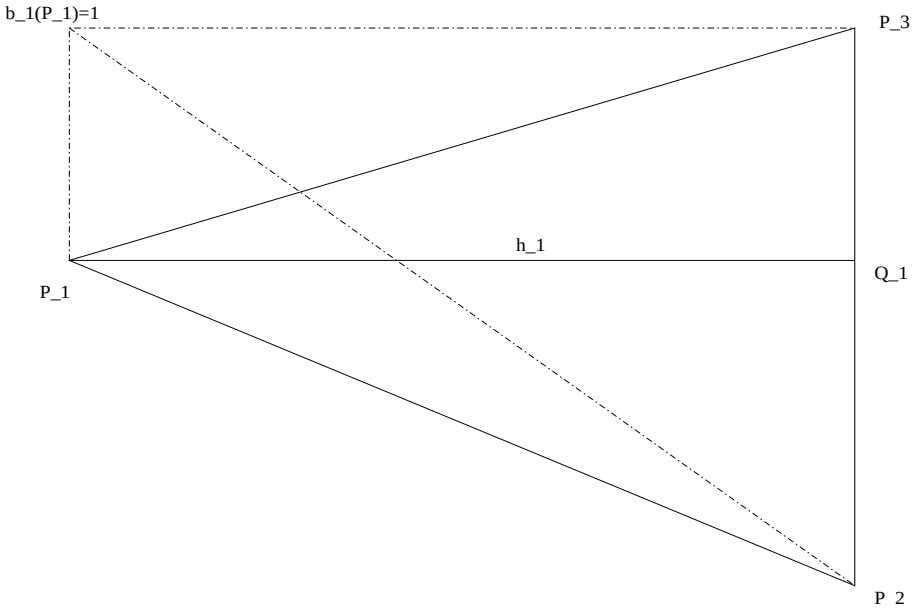


Abbildung 9: Auswertung von ∇b_1

$$\int_T |\nabla v|^2 \leq \frac{3h_T^2}{4|T|} \left\{ v_1^2 + v_2^2 + v_3^2 \right\}.$$

Es ist nun

$$\int_T v^2 = 2|T| \int_{\hat{T}} v_1^2 b_1^2 + v_2^2 b_2^2 + v_3^2 b_3^2 + 2v_1 v_2 b_1 b_2 + 2v_1 v_3 b_1 b_3 + 2v_2 v_3 b_2 b_3$$
$$= \frac{|T|}{6} \left\{ v_1^2 + v_2^2 + v_3^2 + v_1 v_2 + v_1 v_3 + v_2 v_3 \right\}.$$

Diesen Term nach unten abschätzen:

$$\begin{aligned} &\alpha) \qquad v_1, v_2, v_3 \geq 0 : \qquad \int_T |v|^2 \geq \tfrac{|T|}{6} \left\{ v_1^2 + v_2^2 + v_3^2 \right\} \\ &\beta) \quad v_1 < 0, v_2, v_3 \geq 0 : \qquad \int_T |v|^2 \geq \tfrac{|T|}{24} \left\{ v_1^2 + v_2^2 + v_3^2 \right\}, \end{aligned}$$

wobei bei β) die Young'sche Ungleichung

$$\epsilon a^2 + \frac{1}{4\epsilon} b^2 \ \geq ab \quad (\epsilon > 0)$$

mit $\epsilon = \frac{1}{3}$ benutzt wurde. Es ergibt sich

$$\frac{\int\limits_T |\nabla v|^2}{\int\limits_T |v|^2} \quad \leq \quad \frac{18 h_T^2}{|T|^2},$$

also auch

$$\int\limits_\Omega |\nabla v|^2 = \sum_{T \in Z_h} \int\limits_T |\nabla v|^2 \leq \sum_{T \in Z_h} \frac{18}{\xi_T^2} \int\limits_T |v|^2 \leq \frac{18}{\xi^2} \|v\|_0^2.$$

Bei der vorletzten Ungleichung wurde noch benutzt, daß für ein Dreieck T mit Seitenlängen a, b, c

$$\xi_T = \frac{2|T|}{a+b+c} \quad \text{und} \quad 2h_T \leq a+b+c$$

gilt, womit direkt

$$\frac{h_T}{|T|} \quad \leq \quad \frac{1}{\xi_T}$$

folgt. Damit ist der Hilfsatz bewiesen. □

Um den Beweis von Satz 4.7 abzuschließen bemerke, daß für alle Eigenwerte λ von $\tilde{A}$ mit zugehörigem Eigenvektor v

$$Av = \lambda M v$$

gilt. Für

$$\varphi := \sum_{i=1}^{nf} v_i b_i$$

gilt dann

$$\frac{\|\nabla \varphi\|^2}{\|\varphi\|^2} = \lambda \leq \frac{18}{\xi^2}.$$

Damit in (193) ergibt die hinreichende Stabilitätsbedingung

$$\Delta t \lambda (1-2\sigma) \leq \Delta t \frac{18}{\xi^2}(1-2\sigma) \overset{!}{\leq} 2,$$

also

$$9\Delta t(1-2\sigma) \leq \xi^2 (\text{ insbesondere auch } 12\Delta t(1-2\sigma) \leq \xi^2.$$

Das ist die Behauptung und der Satz 4.7 damit bewiesen. □

Jetzt soll noch die Stabilität in der L^∞-Norm untersucht werden. Stabilität bzgl. der L^∞-Norm ist eine strengere Forderung als Stabilität bzgl. der L^2-Norm (=Euklidischen Norm). Haupthilfsmittel zum Nachweis der L^∞-Stabilität von Schema (185) ist die **Inversmonotonie** der Verfahrensmatrizen.

Definition 4.9. (Inversmonotonie)
Eine Matrix $A \in \mathbb{R}^{n \times n}$ heißt inversmonoton: $\Longleftrightarrow$

$$Ax \leq Ay \quad \Rightarrow \quad x \leq y \tag{195}$$

im Sinne der Halbordnung von Vektoren. Äquivalent zu (195) ist

$$A^{-1} \text{ existiert und } A^{-1} \geq 0 \text{ elementweise.} \tag{196}$$

Es gilt

Hilfsatz 4.10. (Inversmonotone L_0-Matrizen sind M-Matrizen)
Sei $A \in \mathbb{R}^{n \times n}$ L_0-Matrix (Definition 3.7), d.h. $a_{ij} \leq 0, i \neq j$. Dann ist A inversmonoton genau dann, wenn es einen Vektor $e > 0$ gibt mit der $Ae > 0$ erfüllt. Zudem gilt dann die Abschätzung

$$\|A^{-1}\| \leq \frac{\|e\|}{\min\limits_k (Ae)_k}. \tag{197}$$

Ist A symmetrische, positiv definite L_0-Matrix, so ist A inversmonoton.

Beweis. Hier nur der Beweis der Aussage für positiv definite L_0-Matrizen, weil der Rest nicht gebraucht wird.
Setze

$$\mu := \max\ a_{ii} > 0 \quad \text{und} \quad S := \mu I - A.$$

Dann gilt wegen $a_{ij} \leq 0, i \neq j$

$$S \geq 0, \text{ und } S \text{ symmetrisch},$$

S hat also nur reelle Eigenwerte. Ferner gilt

$$\lambda_{max}(S) > 0,$$

denn ist x Eigenvektor von S und λ zugehöriger Eigenwert, so folgt

$$0 < |\lambda||x|^2 = |x^t S x| \leq |x^t|S|x| \leq |x|^2 \lambda_{max}(S),$$

wobei hier

$$|x| := (|x_1|, \ldots, |x_n|)^t.$$

Demnach gilt

$$0 < \lambda_{max}(S) = \sigma(S) = \mu - \lambda_A$$

für einen Eigenwert $\lambda_A > 0$ von A. Daraus folgt insbesondere

$$\sigma(S) < \mu.$$

Damit ergibt sich allerdings die Existenz von A^{-1}, denn es gilt

$$A^{-1} = (\mu I - S)^{-1} = \mu^{-1}(I - \mu^{-1}S)^{-1}.$$

Wegen $\sigma(\mu^{-1}S) < 1$ existiert A^{-1} und hat die Darstellung

$$(I - \mu^{-1}S)^{-1} = \sum_{j=0}^{\infty} \mu^{-j-1} S^j \geq 0,$$

also auch $A^{-1} \geq 0$. □

Der Nachweis der L^∞-Stabilität verlangt noch die Voraussetzung, daß die betrachteten Triangulierungen Z_h vom spitzen Typ sind. Eine Triangulierung Z_h heißt vom **spitzen Typ**, falls der größte in Z_h auftretende Winkel echt kleiner als $\frac{\pi}{2}$ ist, vom **schwach spitzen Typ**, falls der größte auftretende Winkel kleiner oder gleich $\frac{\pi}{2}$ ist.

Satz 4.11. (Stabilität von (185) in der L^∞-Norm)
Sei $\{U^j\}_{j\in\mathbb{N}}$ Lösung von (185) und die zugrundeliegende Triangulierung Z_h sei vom spitzen Typ. Ferner bezeichne ξ den kleinsten Inkreis, θ den größten Winkel in Z_h. Dann gilt

$$\|U^{j+1}\|_\infty \ \leq\ \|U^0\|_\infty \ +\ c\Delta t \sum_{k=0}^{j} \|F^{k+\sigma}\|_\infty, \tag{198}$$

falls

$$\begin{cases} \frac{3}{2}\Delta t(1-\sigma) & \leq & \xi^2 \quad \text{und} \\ \Delta t\sigma & \geq & \frac{1}{3\cos\theta}\xi^2 \end{cases}$$

erfüllt ist.

Beweis. Schreibe wieder

$$U^{j+1} = \overbrace{\underbrace{(M + \Delta t\sigma A)^{-1}}_{S_1^{-1}} \underbrace{(M - \Delta t(1-\sigma)A)}_{S^2}}^{S} \quad U^j + \Delta t(M + \Delta t\sigma A)^{-1} F^{j+\sigma}.$$

Beweisidee: Zeige $S\mathbb{1} \leq \mathbb{1}, \mathbb{1} := (1, \ldots, 1)^t$ und $S \geq 0$, denn dann folgt

$$\|SU^j\|_\infty = \|[\sum_l S_{il} U_l^j]_i\|_\infty \leq [\sum_l S_{il} |U_l^j|]_i \leq \|U^j\|_\infty \|\mathbb{1}\|_\infty.$$

Zum Nachweis von $S\mathbb{1} \leq \mathbb{1}$ und $S \geq 0$ genügt gemäß Hilfsatz 4.10 der Nachweis von

$$\underbrace{S_2 \geq 0,\ S_{1ij} \leq 0,\ (i \neq j),\ S_{ii} \geq 0,\ S_1 \text{ positiv definit},}_{\Rightarrow S \geq 0}$$

und

$$S\mathbb{1} \leq \mathbb{1}$$

bzw. hinreichende Bedingungen dafür herzuleiten. Die Herleitung dieser Eigenschaften geschieht in mehreren Schritten.
$S_2 \geq 0$: Es gilt

$$\begin{aligned} m_{ij} &= \sum_{T \subset s_i \cap s_j} 2|T| \int_{\hat{T}} \hat{b}_i \hat{b}_j \geq 0, \quad j \neq i, \\ m_{ii} &> 0 \quad \text{klar, und} \\ a_{ij} &= \sum_{T \subset s_i \cap s_j} |T| \nabla b_i \nabla b_j \leq 0, \quad j \neq i, \end{aligned}$$

denn

$$\nabla b_i \nabla b_j = |\nabla b_i||\nabla b_j| \cos(\pi - \alpha_{ij}) = -|\nabla b_i \|\nabla b_j| \cos \alpha_{ij} \leq 0,$$

siehe Fig. 10, weil $\alpha_{ij} \leq \frac{\pi}{2}$. Daher gilt sicher $S_2 \geq 0$, falls

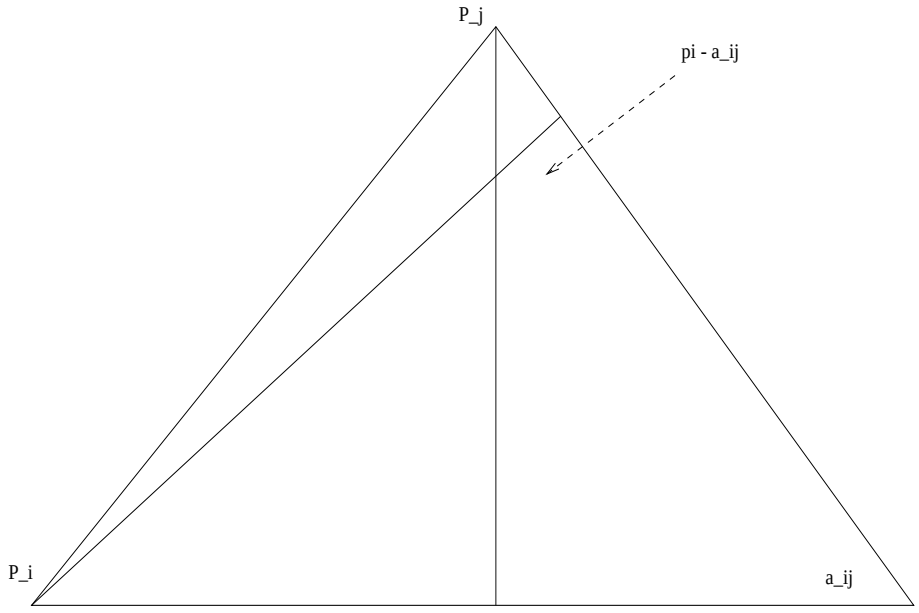


Abbildung 10: Winkel in der Triangulierung

$$m_{ij} - \Delta t(1-\sigma) a_{ij} \geq 0 \quad \forall i, j.$$

Diese Ungleichung ist sicher für alle $i \neq j$ erfüllt, weshalb nur der Nachweis der Bedingung

$$m_{ii} - \Delta t(1-\sigma) a_{ii} \geq 0$$

verbleibt, welche wiederrum äquivalent ist zu

$$\int_\Omega b_i^2 \geq \Delta t(1-\sigma) \int_\Omega |\nabla b_i|^2. \tag{199}$$

Wie im Beweis von Hilfsatz 4.8 wird mit Hilfe von $\frac{h_T}{|T|} \leq \frac{1}{\xi_T}$ abgeschätzt;

$$\int_T |\nabla b_i|^2 \quad \leq \quad \frac{h_T^2}{4|T|} \quad \leq \quad \frac{h_T}{4\xi_T}.$$

Ferner ist

$$\int_T b_i^2 = |T|/6,$$

so daß

$$\frac{\int_T |\nabla b_i|^2}{\int_T b_i^2} \leq \frac{3}{2} \quad \frac{h_T}{|T|\xi_T} \leq \frac{3}{2} \quad \frac{1}{\xi_T^2}.$$

Damit ergibt sich

$$\int_\Omega |\nabla b_i|^2 \leq \frac{3}{2} \frac{1}{\xi^2} \int_\Omega b_i^2.$$

Das in (199) ergibt als hinreichende Bedingung

$$\frac{3}{2}\Delta t(1-\sigma) \leq \xi^2. \tag{200}$$

$S_1^{-1} \geq 0$**:** Zeige $S_{1_{ij}} \leq 0, S_{1_{ii}} \geq 0, S_1$ positiv definit, denn dann folgt mit Hilfsatz 4.10 $S_1^{-1} \geq 0$.
Wegen Für $i = j$ ist das wegen $S_{1_{ii}} = m_{ii} + \Delta t\sigma a_{ii} \geq 0$ erfüllt. Für $i \neq j$ ergibt sich $S_{1_{ij}} \leq 0$ $(i \neq j)$ falls

$$m_{ij} + \Delta t\sigma a_{ij} \leq 0 \ \forall i \neq j,$$

was ausgeschrieben

$$\int_\Omega b_i b_j + \Delta t\sigma \int_\Omega \nabla b_i \nabla b_j \ \leq \ 0 \ \forall i \neq j \tag{201}$$

bedeutet. Wie gehabt, vergleiche Fig. 10,

$$\begin{aligned}
\int_T b_i b_j \quad &= \quad \tfrac{|T|}{12}, \\
\int_T \nabla b_i \nabla b_j \quad &= \quad |T| \nabla b_i \nabla b_j \\
&= \quad |T| \ |\nabla b_i| \ |\nabla b_j| \cos \sphericalangle(\nabla b_i, \nabla b_j) \\
&\overset{\text{Fig. 10}}{=} \quad -|T||\nabla b_i||\nabla b_j| \cos \alpha_{ij} \\
&= \quad -|T| \tfrac{h_i h_j}{4|T|^2} \cos \alpha_{ij} \\
&= \quad -\tfrac{h_i h_j}{4|T|} \cos \alpha_{ij}.
\end{aligned}$$

Das in (201) ergibt die Bedingung

$$|T| \left\{ \frac{1}{12} - \cos \alpha_{ij} \Delta t\sigma \frac{h_i h_j}{4|T|^2} \right\} \leq 0,$$

womit wie folgt abgeschätzt werden kann.

$$\Delta t\sigma \geq \frac{4|T|^2}{12 \cos \alpha_{ij} h_i h_j} \geq \frac{|T|^2}{3h_T^2 \cos \alpha_{ij}} \geq \frac{\xi_T^2}{3 \cos \alpha_{ij}}.$$

Damit ist (θ maximaler Winkel) (201) erfüllt, falls

$$\Delta t\sigma \geq \frac{\xi^2}{3 \cos \theta}.$$

$S_{1ii} > 0$ ist klar, also gilt $S_1^{-1} \geq 0$.
$S\mathbb{1} \leq \mathbb{1}$: Dazu zeige

$$A\,\mathbb{1} \geq 0,$$

denn damit wäre

$$(M + \Delta t\sigma A)\,\mathbb{1} \geq (M - \Delta(1-\sigma)A)\,\mathbb{1},$$

also $S\,\mathbb{1} \leq \mathbb{1}$. Es ist

$$A\mathbb{1} = \left[\sum_{l=1}^{nv} \int_\Omega \nabla b_i \nabla b_l\right]_{i=1}^{nv} = \left[\underbrace{\sum_{l=1}^{nv+nr} \int_\Omega \nabla b_i \nabla b_l}_{=0} - \underbrace{\sum_{l=nv+1}^{nv+nr} \int_\Omega \nabla b_i \nabla b_l}_{\leq 0}\right]_{i=1}^{nv} \geq 0$$

weil $\sum_{l=1}^{nv+nr} b_l = 1$. Damit ist alles gezeigt und Satz 4.11 ist bewiesen. □

Bemerkung 4.12. Die Bedingungen aus Satz 4.11 sind nicht immer vereinbar. Ist etwa $\sigma = \frac{1}{2}$, so lautet die Forderung

$$\frac{1}{2}\Delta t \leq \frac{2}{3}\xi^2 \quad \text{und} \quad \frac{1}{2}\Delta t \geq \frac{1}{3\cos\theta}\xi^2,$$

was $\cos\theta \geq \frac{1}{2}$ impliziert. Sind die Dreiecke nicht gleichseitig, so gibt es eine Winkel $> \frac{\pi}{3}$, also

$$\frac{1}{2} \quad \leq \quad \cos\theta \quad < \quad \cos\frac{\pi}{3} \quad = \quad \frac{1}{2},$$

ein Widerspruch.

Die Stabilitätseigenschaften für das Crank-Nicolson Verfahren sind demnach nicht so gut wie die des impliziten Euler-Verfahrens. Die zweite Bedingung in Satz 4.11 resultiert aus der Forderung (201). Die Einschräankung an die Schrittweite resultiert dabei aus der Positivität der Nebendiagonalelemente der Massematrix.
Der Massematrix in (185) kann man sich entledigen. Dazu sei daran erinnert, daß M aus (179) bei Verwendung des Ansatzes (182) entsteht;

$$\int_\Omega u_{h_t} b_i dx = \sum_{j=1}^{nv} \dot{u}_j(t) \underbrace{\int_\Omega b_i b_j}_{m_{ij}} \quad (i = 1, \ldots, nv).$$

Ersetzt man jetzt in

$$\int_\Omega b_i b_j dx = \sum_{T \subset s_i \cap s_j} \int_T b_i b_j dx$$

die Integration durch die Quadraturformel (P_1, P_2, P_3 Eckpunkte des Dreiecks T)

$$\int_T f(x)dx \approx \frac{|T|}{3}\{f(P_1) + f(P_2) + f(P_3)\} =: Q(f),$$

so folgt

$$\left.\begin{array}{rcl} Q(b_i b_j) & = & 0,\ i \neq j, \\ Q(b_i^2) & = & \frac{|T|}{3} \end{array}\right\} \text{ auf } T.$$

Die Matrix M in (183) bei Verwendung dieser Quadraturformel gegen eine Diagonalmatrix $\tilde{M}$ ausgetauscht, für deren Einträge

$$\tilde{m}_{ii} = \frac{1}{3}|s_i| \tag{202}$$

gilt ($s_i = \operatorname{supp} b_i$).

Bemerkung 4.13. Erinnert man sich jetzt an die Einträge der Massematrix, sieht man sofort, daß

$$\tilde{m}_{ii} = \sum_{j=1}^{nv} m_{ij}$$

richtig ist. Die Diagonalmatrix $\tilde{M}$ entsteht demnach aus der Massematrix M durch **Lumping** (Konzentration) der Nebendiagonalelemente von M auf die Diagonale. Die genaue Bezeichnung dieses Vorgehens ist **Baryzentrisches Lumping**.

Folgende Überlegungen führen auch auf Lumping; in (179) wird als örtliche Approximation für die Zeitableitung u_{h_t} ein Ansatz gemacht, der geringere Regularität besitzt als der Ansatz für u_h. Das rechtfertigt sich daraus, daß die Regularität der Zeitableitung u_t der kontinuierlichen Lösung bzgl. des Ortes in etwa der der zweiten Ortsableitung D^2u der kontinuierlichen Lösung u entspricht. Zur Herleitung des entsprechenden numerischen Schemas machen wir für die diskrete Lösung die zwei verschiedenen Ansätze

$$\tilde{u}_h(t,x) = \sum_{i=1}^{nv} u_i(t)\phi_i(x)$$

und

$$u_h(t,x) = \sum_{i=1}^{nv} u_i(t)b_i(x).$$

Wir verlangen, daß $\tilde{V}_h = \text{span}\{\phi_i, i = 1,\dots,nv\}$ dieselbe Dimension wie $V_h = \text{span}\{b_i, i = 1,\dots,nv\}$ besitzt, die Funktionen ϕ_i lokalen Träger haben, der Durchschnitt der Träger zweier verschiedener Funktionen ϕ_i und ϕ_j leer ist und die jede Funktion ϕ_i stückweise konstant ist.
Zur Approximation der Zeitableitung u_t wird jetzt $\tilde{u}$ verwendet und in der Variationsformulierung (179) der Zeitableitungsterm mit Funktionen aus $\tilde{V}_h$ getestet (eine Art **Petrov-Galerkin Verfahren**). Es ergibt sich (hier wieder am Beispiel der Wärmeleitungsgleichung)

$$\int_\Omega \tilde{u}_{h_t}\tilde{v}_h + \nabla u_h \nabla v_h = \int_\Omega q v_h \quad \forall\, v_h \in V_h, \tilde{v}_h \in \tilde{V}_h,$$

was wegen der Eigenschaften des Ansatzraumes $\tilde{V}_h$ dem System gewöhnlicher Differentialgleichungen

$$D_M\dot{U} + AU = F, \quad U(0) = U_0$$

entspricht. Hier ist D_M eine Diagonalmatrix (die Träger der Φ_i sollen ja disjunkt sein).
Verbleibt noch die die Konstruktion der Funktionen ϕ_i. Dazu wird die zu Z_h **Duale Vernetzung** betrachtet, die entsteht, wenn man die Schwerpunkte benachbarter Dreiecke in Z_h miteinander verbindet, siehe Fig. 11. Die Funktion ϕ_i ist die charakteristische Funktion des Bereichs, der durch das den Punkt P_i umschliessende Polygon der dualen Vernetzung berandet wird. Daraus folgt direkt

$$\int_\Omega \phi_i\phi_j = \delta_{ij}\frac{1}{3}|s_i|,$$

wobei wieder s_i den Träger der Funktion b_i bezeichnet. Das heisst aber, daß die Einträge von D_M mit denen von $\tilde{M}$ übereinstimmen. Der Begriff baryzentrisches Lumping ist somit klar.

Bemerkung 4.14. Ein weitere Möglichkeit des Lumpings ist durch **Circumzentrisches Lumping** gegeben. Dabei werden die Umkreismittelpunkte benachbarter Dreiecke miteinander verbunden und dann weiter wie beim baryzentrischen Lumping vorgegangen. Bei diesem Zugang sollten allerdings die Zerlegung Z_h so beschaffen sein, daß die Umkreismittelpunkte der Dreiecke auch in diesen enthalten sind. Zerlegungen dieser Art heissen vom **Differenzen-Typ**.

Wird jetzt in (185) $\tilde{M}$ anstelle von M verwendet, so ergeben sich die Stabilitätsbedingungen

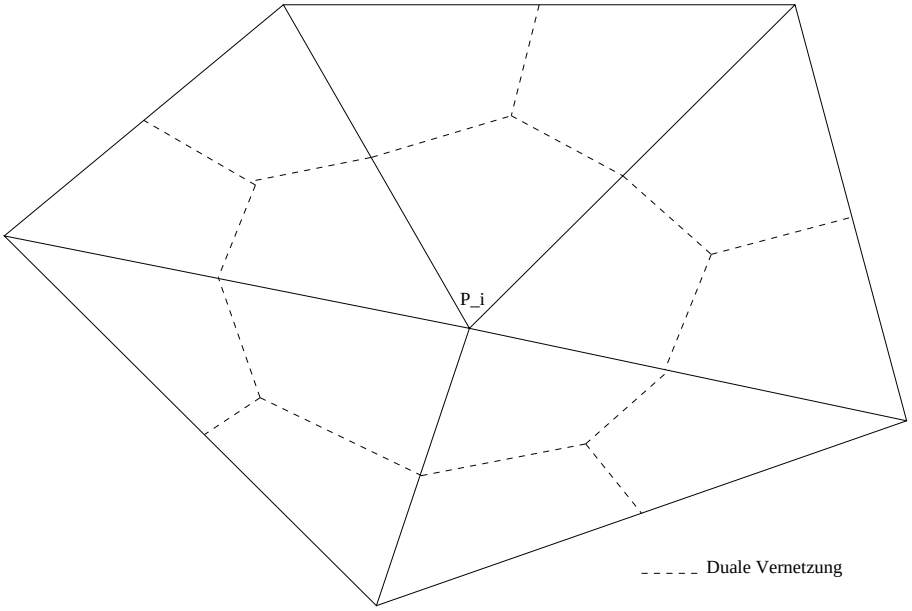


Abbildung 11: Duale Vernetzung

Folgerung 4.15. (Lumping und Stabilität)
Schema (185) mit $\tilde{M}$ anstelle von M ist L^2-stabil, falls

$$\frac{3}{4}\Delta t(1-2\sigma) \leq \xi^2 \quad \text{baryzentrisch, bzw.} \quad \Delta t(1-2\sigma) \leq \xi^2 \text{ circumzentrisch,} \tag{203}$$

erfüllt ist. Ist Z_h schwach spitz, so ist das Schema unter der Bedingung

$$\frac{3}{4}\Delta t(1-\sigma) \leq \xi^2 \quad \text{baryzentrisch, bzw.} \quad \Delta t(1-\sigma) \leq \xi^2 \text{ circumzentrisch} \tag{204}$$

stabil bezüglich der L^∞-Norm.

Beweis. Analog zu dem von Satz 4.11. □

Bemerkung 4.16. (204) ist für das implizite Euler-Verfahren immer erfüllt.

Satz 4.17. (L^2-Fehler bei Lumping)
Die Aussagen der Sätze 4.1 und 4.2 gelten auch unter Verwendung der Lumpingtechnik. Im semi-diskreten Fall kann jedoch das exponentielle Abklingen des Anfangsfehlers für beliebige Triangulierungen nicht mehr gewährleistet werden. Bei entsprechender Regularität der Lösung u gelten die Fehlerabschätzungen

$$\|u_h(t,\cdot) - u(t,\cdot)\| \;\leq\; c\left\{\|u_h^0 - u^0\| + h^2\|u^0\|_2 + \|u(t)\|_2 + \left(\int_0^t \|u_t\|_2^2 ds\right)^{\frac{1}{2}}\right\}, \tag{205}$$

vergleiche (184), und

$$\|u(t_k,\cdot) - u_h^k\| \leq c\left(\|u_h^0 - u^0\| + h^2\left\{\|u^0\|_2 + \|u(t_k)\|_2 + \left(\int_0^{t_k} \|u_t\|_2^2 ds\right)^{\frac{1}{2}}\right\} + \Delta t\left(\int_0^{t_k} \|u_{tt}\|^2\right)^{\frac{1}{2}}\right) \tag{206}$$

vergleiche (185). Eine (186) entsprechende Abschätzung gilt für das Crank-Nicolson Verfahren. Beweise finden sich in [19, Theorem 15.1, Theorem 15.4].

4.2 Horizontale Linienmethode

Das Pendant zur Vertikalen Linienmethode ist die Horizontale Linienmethode. Anders als in (179) wird jetzt (178) zunächst in der Zeit diskretisiert.
Wieder gelte $b \equiv 0, c = 0$ und $K(t,x) = Id_{\mathbb{R}^2}$. Das Problem (178) wird wie üblich in der Zeit diskretisiert, i.e. $0 = t_0 < t_1 < \ldots < t_n = T$, $t_i = i\Delta t$, $\Delta t = \frac{T}{n}$ ist das Zeitgitter. So wird eine Folge von stationären Variationsproblemen erhalten;

$$\text{(207)} \qquad \begin{cases} \text{Gegeben } u_0 = 0, \text{ finde } u_{j+1} \text{ mit } u_{j+1}(x) = r(t_{j+1}, x) \text{ auf } \partial\Omega \text{ und} \\ \int\limits_\Omega \frac{u_{j+1}-u_j}{\Delta t} v dx + \int\limits_\Omega \nabla u_{j+1} \nabla v dx = \int\limits_\Omega q(t_{j+1}, x) v(x) dx \qquad \forall v \in V = H_0^1(\Omega). \end{cases}$$

Mit Hilfe der Funktionen u_j und linearen finiten Elementen wird die sogenannte **Rothe-Funktion** gebildet,

$$\text{(208)} \qquad u^{\Delta t}(t,x) := \sum_{j=1}^{n} u_j(x)\varphi_j(t),$$

wobei

$$\varphi_j(t) := \begin{cases} \frac{t-t_{j-1}}{\Delta t} & , \quad t \in [t_{j-1}, t_j] \\ \frac{t_{j+1}-t}{\Delta t} & , \quad t \in [t_j, t_{j+1}] \\ 0 & , \quad \text{sonst.} \end{cases}$$

Die Funktionen φ_j sind die eindimensionalen linearen Finite-Element Basisfunktionen, hier über dem Zeitintervall definiert. Es soll jetzt der Fehlerzwischen u und $u^{\Delta t}$ in der L^2-Norm abgeschätzt werden. Dazu sei vorausgesetzt, daß q unabhängig von der Zeit ist und zusätzlich $q \in H^2(\Omega) \cap V$ gelte.

Satz 4.18. (L^2-Fehler Semidiskretisierung horizontale Linienmethode)
Mit obigen Voraussetzungen und Notationen gilt

$$\text{(209)} \qquad \|u(t,\cdot) - u^{\Delta t}(t,\cdot)\|_0 \leq c\Delta t \quad \forall t \in [0,T]$$

mit einer von $\|q\|_{H^2 \cap V}$ abhängigen positiven Konstanten c.

Beweis. In (t_{j-1}, t_j) gilt $u - u^{\Delta t} = u_{j-1}(x)\varphi_{j-1}(t) + u_j(x)\varphi_j(t)$. Demnach mit (207) und der Definition von φ_j

$$\text{(210)} \qquad \int\limits_\Omega (u^{\Delta t} - u)_t v dx + \int\limits_\Omega \nabla(u^{\Delta t} - u)\nabla v dx = \Delta t \int\limits_\Omega \frac{u_j - 2u_{j-1} + u_{j-2}}{\Delta t^2} \frac{t_j - t}{\Delta t} v dx \quad \forall\, v \in V, t \in [t_{j-1} t_j].$$

Gelingt es jetzt, den Term

$$\text{(211)} \qquad \frac{u_j - 2u_{j-1} + u_{j-2}}{\Delta t^2} = \frac{1}{\Delta t}\left\{\frac{u_j - u_{j-1}}{\Delta t} - \frac{u_{j-1} - u_{j-2}}{\Delta t}\right\}$$

gleichmässig bzgl. Δt in der L^2-Norm zu beschränken, ist man am Ziel, falls in (210) mit $u^{\Delta t} - u$ getestet wird.
Sei

$$U_i := \frac{u_i - u_{i-1}}{\Delta t}, \quad W_i := \frac{U_i - U_{i-1}}{\Delta t}.$$

Damit gilt wegen $u_0 = 0$ und (207)

$$\frac{1}{\Delta t}\int\limits_\Omega u_1 v + \int\limits_\Omega \nabla u_1 \nabla v = \int\limits_\Omega qv \quad \forall\, v \in V.$$

Damit, und wegen $u_1 \in V$, folgt

$$\|u_1\| \leq \Delta t \|q\|.$$

Subtrahiert man (207) für zwei aufeinanderfolgende Indizies, ergibt sich

$$\frac{1}{\Delta t} \int_\Omega (u_j - u_{j-1})v + \int_\Omega \nabla (u_j - u_{j-1}) \nabla v = \frac{1}{\Delta t} \int_\Omega (u_{j-1} - u_{j-2})v,$$

also durch die Wahl von $v = u_j - u_{j-1}$ als Testfunktion

$$\|u_j - u_{j-1}\| \leq \|u_{j-1} - u_{j-2}\|$$

und induktiv

$$\|u_j - u_{j-1}\| \leq \Delta t \|q\|,$$

bzw.

$$\|U_j\| \leq \|q\| \quad \forall j.$$

Analog gilt für $U_j - U_{j-1}$

$$\frac{1}{\Delta t} \int_\Omega (U_j - U_{j-1})v + \int_\Omega \nabla (U_j - U_{j-1}) \nabla v = \frac{1}{\Delta t} \int_\Omega (U_{j-1} - U_{j-2})v,$$

also auch

$$\|U_j - U_{j-1}\| \leq \|U_{j-1} - U_{j-2}\|.$$

Weiter ist

$$\frac{1}{\Delta t} \int_\Omega (U_2 - U_1)v + \int_\Omega \nabla (U_2 - U_1) \nabla v = \frac{1}{\Delta t} \int_\Omega (U_1 - q)v dx,$$

also

$$\|U_2 - U_1\| \leq \|U_1 - q\|.$$

Wegen

$$\frac{1}{\Delta t} \int_\Omega (U_1 - q)v dx + \int_\Omega \nabla (U_1 - q) \nabla v dx = - \int_\Omega q v dx$$

folgt aber

$$\|U_1 - q\| \leq \Delta t \|q\|,$$

und somit induktiv

$$\|W_j\| \leq \|q\|.$$

In (211) steht

$$\frac{u_j - 2u_{j-1} + u_{j-2}}{\Delta t} = W_j,$$

also folgt mit $v = u^{\Delta t} - u$ in (210)

$$\|u^{\Delta t} - u\| \leq \Delta t \| \frac{u_j - 2u_{j-1} + u_{j-2}}{\Delta t} \| = \Delta t \|W_j\| \leq \Delta t \|q\|. \tag{212}$$

□

Bemerkung 4.19. (207) ist das kontinuierliche Analogon zum impliziten Eulerverfahren (188) mit $\sigma = 1$, kontinuierlich, weil nicht diskretisiert bzgl. der Zeit.

Beispiel Betrachte das Anfangs-Randwertproblem

$$\begin{aligned} u_t - u_{xx} &= \sin x \text{ in } (0,T)\times(0,\pi), \\ u(x,0) &= 0, \\ u(0,t) &= 0, \\ u(\pi,t) &= 0, \end{aligned}$$

dessen eindeutige Lösung durch

$$u(t,x) = (1-e^{-t})\sin x.$$

gegeben ist. Die Rothe-Methode angewendet auf diese Aufgabe liefert für $j \in \mathbb{N}$

$$\frac{u_j(x) - u_{j-1}(x)}{\Delta t} - u_j''(x) = \sin x, \quad u_j(0) = u_j(\pi) = 0.$$

Die eindeutigen Lösungen dieser stationären Randwertaufgaben sind durch

$$u_j(x) = \left\{1 - \frac{1}{(1+\Delta t)^j}\right\}\sin x, \quad j \in \mathbb{N},$$

gegeben. Für die zugehörige Rothe-Funktion

$$u^{\Delta t}(t,x) = u_{j-1}(x) \; + \; \frac{t-t_{j-1}}{\Delta t}\left(u_j(x) - u_{j-1}(x)\right) \qquad \text{auf } [t_{j-1}, t_j]$$

ergibt sich wegen

$$\lim_{\Delta t \to 0} \frac{1}{(1+\Delta t)^{t_j/\Delta t}} = e^{-t_j}$$

direkt

$$\lim_{\Delta t \to 0} u^{\Delta t}(t,x) = u(t,x).$$

□

5 Ein nichtlineares Problem

Die numerische Diskretisierung eines nichtlinearen zeitabhängigen Problems mittels finiter Elemente soll am Beispiel (13) erläutert werden, wobei Transport und Advektion nicht berücksichtigt werden sollen, d.h.,

$$v_T \equiv 0 \text{ und } a = 0.$$

Ferner soll $u(t,x)$ eine Temperatur beschreiben und die Quelldichte der Wärmequellen, $q(t,x)$, von der Temperatur selbst abhängen, und zwar autonom, d.h.,

$$q(t,x) = q(u(t,x)).$$

Darüberhinaus wird verlangt, daß die Wärmequellen qualitativ nach dem phänomenologischen Gesetz

Die von der Wärmequelle bereitgestellte Temperatur $q(u)$ ist proportional zur Änderung der von der Wärmequelle erzeugten Temperatur $\frac{\partial}{\partial u}\, q(u)$

verhalten. Nach **Arrhenius** können solche Quellen gemäß (siehe auch [2])

$$q(u(t,x)) = \delta e^{u(t,x)}, \quad \delta > 0 \text{ geeignet},$$

modelliert werden. Als Modell für die Temperaturentwicklung ergibt sich das Anfangs-Randwertproblem

$$\left\{\begin{aligned} u_t(t,x) - \Delta_x u(t,x) &= \delta e^{u(t,x)} && \text{in } \Omega^T, \\ u(t,x) &= 0 && \text{auf } \partial\Omega^T, \\ u(0,x) &= u_0(x) && \text{in } \Omega. \end{aligned}\right. \tag{213}$$

Problem (213) modelliert etwa den Zündungsvorgang in einem Festkörperbrennstoff, dessen Ausgangstemperaturverteilung vorgelegt ist oder auch die Wärmeausbreitung in einem Kohlehaufen [18].
Wärmediffusion ist in diesem Modell der Mechanismus für Wärmetransport. Nur über sie kann, wenn überhaupt, gewährleistet werden, daß Kühlung am Rand $\partial\Omega$ von Ω Einfluß auf den Zündungsprozeß nehmen kann. Das wiederum hängt von der Relation zwischen Diffusionsgeschwindigkeit und Reaktionsgeschwindigkeit im Festkörperbrennstoff ab. Letztere ist proportional zur Wärmeerzeugung modelliert durch den exponentiellen Term.
Auftreten können demnach folgende Situationen:

i) Zeitskala Diffusion $\approx$ Zeitskala chemische Reaktion.
Hier läuft die Zündung subkritisch ab, ein stabiles Energiegleichgewicht wird erhalten. Kühlung von außen kann den Prozeß unter Kontrolle halten.

ii) Zeitskala Diffusion $>>$ Zeitskala chemische Reaktion.
Hier läuft die Zündung superkritisch ab, d.h., die Wärme kann nicht schnell genug abgeführt werden und der exponentielle Term ist für explosionsartige Wärmeentwicklung verantwortlich.

Als Beispiel für ii) sei die Challenger-Katastrophe genannt. Aber eine undichte Stelle einer Dichtung wurden obenliegende Brennkammern unkontrolliert gezündet und über den Auslasstrichter konnte nicht genug Wärme abgeführt werden. Resultat war die Explosion.
Sei jetzt $\Omega = B(0) \subset \mathbb{R}^n, n \leq 2$. Es wird zunächst das stationäre Problem betrachtet,

$$\left\{ \begin{array}{rcll} -\Delta u & = & \delta e^u & \text{in } \Omega, \qquad \delta > 0, \\ u & = & 0 & \text{auf } \partial\Omega. \end{array} \right. \tag{214}$$

Dieses harmlos aussehende Problem hat wegen seiner Nichtlinearität interessante Eigenschaften. Denn es gibt ein $\delta^* > 0$, so daß

$\alpha.)$ (214) genau 2 Lösungen hat, falls $\delta \in (0, \delta^*)$

$\beta.)$ (214) genau eine Lösung hat, falls $\delta = \delta^*$

$\gamma.)$ (214) keine Lösung hat, falls $\delta > \delta^*$.

Ferner sind die Lösungen für $0 \leq \delta \leq \delta^*$ radial symmetrisch und können für $n = 1, 2$ explizit angegeben werden. Dazu schreibe (214) in Polarkoordinaten und erhalte aus (214) das eindimensionale Randwertproblem

$$\left\{ \begin{array}{rcl} u''(r) + \frac{n-1}{r} u'(r) & = & -\delta e^{u(r)}, \ r \in (0,1), \\ u'(0) & = & 0, \\ u(1) & = & 0. \end{array} \right. \tag{215}$$

Mit

$$\alpha := u(0) \text{ und } \beta := -u'(1)$$

ist für $n = 1$

$$u(r) = \alpha - 2 \ln \cosh \left[\frac{1}{2} r \sqrt{2\delta e^\alpha} \right],$$

wobei α, β und δ durch die Beziehungen

$$\delta = \frac{1}{2} e^{-\alpha} \left\{ \ln \frac{1 + \sqrt{1 - e^{-\alpha}}}{1 - \sqrt{1 - e^{-\alpha}}} \right\}^2$$

und

$$\beta = \sqrt{\beta^2 + 2\delta} \tanh \left\{ \frac{1}{2} \sqrt{\beta^2 + 2\delta} \right\}$$

gekoppelt sind.
Ist $n = 2$, so gilt

$$u(r) = \alpha - 2 \ln \left(1 + \frac{1}{8} \delta e^\alpha r^2 \right),$$

wobei α, β und δ jetzt durch

$$\delta = 8\left(e^{-\frac{1}{2}\alpha} - e^{-\alpha}\right)$$

und

$$\beta^2 - 4\beta + 2\delta = 0$$

gekoppelt sind.
Jetzt zurück zum instationären Problem. Was kann über das Verhalten der Lösungen von (213) ausgesagt werden?
Sei δ^* der oben eingeführte kritische Parameter, d.h. der Parameter, an dem Mehrdeutigkeit der Lösung in Eindeutigkeit übergeht. Dann gilt

i) Ist $\delta \in (\delta^*, \infty)$, so gibt es ein $T \in [\frac{1}{\delta}, \infty)$ derart, daß (213) genau eine Lösung $u \in C^{2,1}(\bar{\Omega} \times [0, T))$ besitzt mit

$$\lim_{t \nearrow T} \max_{x \in \bar{\Omega}} u(t, x) = \infty$$

Blow-up, superkritisches Verhalten.

ii) Ist $\delta \in (0, \delta^*)$ und sei φ die kleinere der beiden Lösungen des stationären Problems (214). Gilt dann für den Anfangswert

$\alpha.$) $\quad u_0(x) \leq \varphi(x) \ \ \forall x \in \bar{\Omega}$,

so hat (213) genau eine Lösung $u \in C^{2,1}(\bar{\Omega} \times [0, \infty))$ und

$$\lim_{t \to \infty} |u(t, \cdot) - \varphi|_\infty = 0.$$

Ist χ die größere der stationären Lösungen von (214), $\chi \neq \varphi$ und

$$u_0(x) < \chi(x) \quad \forall\, x \in \bar{\Omega},$$

so gilt die Aussage von $\alpha.$). Ist

$$u_0(x) > \chi(x) \quad \forall\, x \in \bar{\Omega},$$

so gibt es entweder einen Blow-Up nach endlicher Zeit T mit eindeutiger Lösung $u \in C^{2,1}(\bar{\Omega} \times [0, T))$, oder es gilt $u \in C^{2,1}(\bar{\Omega} \times [0, \infty))$ mit

$$\lim_{t \to \infty} |u(t, x)|_\infty = \infty.$$

Zusammengefaßt

i) Die stationären Lösungen von (213) charakterisieren das dynamische Verhalten von (213) zusammen mit dem Anfangswert u_0.

ii) Ist u_0 "zu groß", ist das Verhalten der Lösung superkritisch.

Zur numerischen Approximation von (213) im Ort werden finite Elemente verwendet. Dazu sei wieder Z_h eine Triangulierung von Ω und für die diskrete Lösung werde der Ansatz

$$u_h(t, x) = \sum_{i=1}^{nf} u_i(t) b_i(x)$$

gemacht. Um die Randbedingungen aus (213) einzuarbeiten, wird, wie gehabt, $u_i(t) = 0$ für alle auf dem Rand liegenden Freiheitsgrade gesetzt. Gehe jetzt wie zur Herleitung von (183) vor. Dann ergibt sich das Differentialgleichungssystem

$$\left\{ \begin{array}{rcl} M\dot{U} + AU + G(U) & = & 0 \\ U(0) & = & U_0, \end{array} \right. \tag{216}$$

wobei wieder M die Masse- und A sie Steifigkeitsmatrix bezeichnet und

$$U = [u_1, \ldots u_{n_i}]^t, \quad n_i = \#\text{ Unbekannte}.$$

Die Funktion G besitzt dabei die Gestalt

$$-G(U)_j = \delta \int_\Omega e^{\sum_{i=1}^{nf} u_i(t) b_i(x)} b_j(x) dx, \quad j = 1, \ldots, n_i.$$

Zur zeitlichen Diskretisierung von (216) wird jetzt wie bei der Herleitung von (185) verfahren. Es ergibt sich mit dem Zeitgitter

$$t_0 := 0, \quad t_i := i\Delta t, \quad n\Delta t = T$$

und der Setzung

$$U^k \approx U(t_k)$$

das Schema

$$\left\{ \begin{array}{rl} M \frac{U^{k+1}-U^k}{\Delta t} + \sigma A U^{k+1} + \sigma G(U^{k+1}) & = -(1-\sigma)G(U^k) - (1-\sigma)AU^k, \\ U^0 & = U(0). \end{array} \right. \tag{217}$$

In jedem Zeitschritt muß demnach ein nichtlineares Gleichungssystem der Form

$$H(U^{k+1}) = 0 \tag{218}$$

gelöst werden, wobei klar ist, wie sich H zusammensetzt. Wird im k-ten Schritt das **Newton-Verfahren** zur Lösung von (218) verwendet, ergibt sich der Algorithmus

Algorithmus 5.1. (Newton-Verfahren)

1. U_0^{k+1} gegeben, $l = 0$
2. Löse $DH(U_l^{k+1})\delta U_l = -H(U_l^{k+1})$
3. setze $U_{l+1}^{k+1} = U_l^{k+1} + \delta U_l$ und $l = l + 1$
4. Ist das Abbruchkriterium erfüllt, so gehe zu 2.

Mögliche Abbruchkriterien sind etwa ($\epsilon > 0$ vorgelegte Fehlertoleranz)

- $||H(U_l^{k+1})||/||H(U_0^{k+1})|| \leq \epsilon$,
- $||H(U_l^{k+1}) - H(U_{l-1}^{k+1})||/||H(U_l^{k+1})|| \leq \epsilon$ oder
- $||\delta U_l||/||\delta U_0|| \leq \epsilon$.

Verbleibt noch die Frage nach der Struktur der Matrix $DH(U)$ in Algorithmus 5.1). Es gilt

$$DH(U) = \frac{1}{\Delta t} M + \sigma A + \sigma DG(U), \tag{219}$$

wobei $DG(U)$ dasselbe Belegungsmuster wie M und A hat, denn es gilt

$$-\frac{\partial}{\partial u_l} G(U)_j = \frac{\partial}{\partial u_l} \left\{ \delta \int_\Omega e^{\sum_{i=1}^{ni} u_i b_i(x)} b_j(x) dx \right\} = \delta \int_\Omega e^{\sum_{i=1}^{ni} u_i b_i(x)} b_l(x) b_j(x) dx.$$

Damit ist $\frac{\partial}{\partial u_l} G(U)_j = 0$, falls $s_l \cap s_j = \emptyset$.
Bei der Auswertung von G sind die Integrale

$$\int_\Omega e^{\sum_{i=1}^{ni} u_i b_i(x)} b_j(x) dx$$

numerisch zu approximieren. Das kann mit den in (112) angegebenen Quadraturformeln geschehen.

Bemerkung 5.2. Zur Diskretisierung bzgl. der Zeit können natürlich alle im Anhang A besprochenen Verfahren verwendet werden. Insbesondere sei in diesem Zusammenhang auf das sogenannte **Fractional-Step-Θ-Schema** hingewiesen, welches für die numerische Integration von Zeitschritt t_k nach t_{k+1} zwei Hilfszeitpunkte $t_k < t_{k+\Theta} < t_{k+1-\Theta} < t_{k+1}$ einfügt.

Algorithmus 5.3. Fractional-Step-Θ-Schema
Auf das System (216) angewendet hat das Verfahren die Gestalt

$$\begin{aligned} t_k \to t_{k+\Theta}: &\quad \left[M + \alpha\Theta\Delta t A^{k+\Theta}\right] U^{k+\Theta} = \left[M - \beta\Theta\Delta t A^{k}\right] U^{k} \\ t_{k+\Theta} \to t_{k+1-\Theta}: &\quad \left[M + \beta\Theta'\Delta t A^{k+1-\Theta}\right] U^{k+1-\Theta} = \left[M - \alpha\Theta'\Delta t A^{k+\Theta}\right] U^{k+\Theta} \\ t_{k+1-\Theta} \to t_{k+1}: &\quad \left[M + \alpha\Theta\Delta t A^{k+1}\right] U^{k+1} = \left[M - \beta\Theta\Delta t A^{k}\right] U^{k+1-\Theta}. \end{aligned}$$

Dabei bezeichnet $A^k := A + G(U^k)$. Für die Parameter gilt $\Theta = 1 - \sqrt{2}/2$, $\Theta' = 1 - 2\Theta$, $\alpha \in (1/2, 1]$ und $\beta = 1 - \alpha$. Für den Spezialfall $\alpha = (1 - 2\Theta)/(1 - \Theta)$ gilt $\alpha\Theta = \beta\Theta'$.

Bemerkung 5.4. Im Gegensatz zum Crank-Nicolson Verfahren ist das Fractional-Step-Θ-Schema für jede oben angegebene Parameterwahl streng A-stabil und weist wesentlich bessere Genauigkeit bei gleicher Schrittlänge auf (1 Schritt Fractional-Step-Θ-Schema ≈ 3 Schritte Crank-Nicolson Verfahren).

6 Übungsaufgaben

6.1 Kapitel

1

Aufgabe 6.1. (Modellierung des Verkehrsflusses auf einer Autobahn)
Es bezeichnen $u = u(t, x)$ die Anzahl der Fahrzeuge pro Kilometer (Fahrzeugdichte) und $q = q(t, x)$ die Anzahl der Fahrzeuge pro Stunde (Fahrzeugfluß) auf einer Autobahn. Dabei ist x die eindimensionale Ortskoordinate und t die Zeit. Ferner seien $[a, b]$ ein endlicher Autobahnabschnitt und $t_1 < t_2$ zwei Zeitpunkte.

1. Wie groß ist die Anzahl der Fahrzeuge in dem Autobahnabschnitt $[a, b]$ zu den Zeitpunkten t_1 und t_2 ?

2. Wie groß ist die Anzahl der Fahrzeuge, die während der Zeitspanne $\delta t = t_2 - t_1$ über a in den Autobahnabschnitt ein- bzw. über b aus dem Autobahnabschnitt ausfahren?

3. Drücken Sie das folgende Erhaltungsgesetz

   Die Anzahl der Fahrzeuge in $[a, b]$ zur Zeit t_2 minus der Anzahl der Fahrzeuge in $[a, b]$ zur Zeit t_1 entspricht der Differenz der in a ein- und in b ausfahrenden Fahrzeuge

   in mathematischer Schreibweise aus (mit Herleitung).

Aufgabe 6.2. (Modellierung und Konsequenzen aus dem Modell)
Mit u und q aus Aufgabe 6.1 sei jetzt q eine Funktion von u, d.h. $q(t, x) = f(u(t, x))$. Die Erhaltungsgleichung aus Aufgabe 6.1 kann damit geschrieben werden als

$$(P_1) \quad \begin{cases} \frac{\partial}{\partial t} u(t, x) &= -\frac{\partial}{\partial x} f(u(t, x)), \qquad x \in [a, b], \quad t \in (t_1, \infty), \\ u(t, x_1) &= v(x), \end{cases}$$

wobei v die Fahrzeugdichte zur Zeit $t = t_1$ bezeichne.

1. Interpretieren Sie qualitativ den Ansatz

   $$f(u) = -(u - u_m)^2 + f_0,$$

   wobei $f_0, u_m > 0$ konstant geeignet gewählt sind.

2. Sei u eine Lösung von (P_1) mit $a = -2, b = 2, f(u) = \frac{1}{2}u^2$ und ∞-oft differenzierbaren Anfangswerten

$$v(x) = \begin{cases} 1, & x \in [-2,-1) \\ 0, & x \in [1,2] \\ g(x) & x \in [-1,1). \end{cases}$$

Zeigen Sie, dass es keine Lösung $u \in C^0([a,b] \times \mathbb{R}^+)$ des Problems (P_1) gibt.
Hinweis: Zeigen Sie zuerst ganz allgemein, dass u entlang der Kurve $(\gamma(t), t)$ konstant ist, wobei $\gamma(t)$ eine Lösung der Anfangswertaufgabe

$$\frac{\partial}{\partial t}\gamma(t) = \frac{\partial}{\partial u} f(u(\gamma(t),t)), \qquad \gamma(t_1) = s \in [a,b]$$

ist.
Danach sollten Sie zeigen, daß diese Kurven $(\gamma(t), t)$ Geraden in der x-t-Ebene sind. Wenn Sie nun diese beiden Aussagen auf das Beispiel anwenden, erhalten Sie die gewünschte Aussage.

3. Wenden Sie das Ergebnis aus b) auf das Problem (P_1) mit f aus a) an und interpretieren Sie das Ergbnis qualitativ.

Aufgabe 6.3. (Einschrittverfahren)
Betrachtet werden Einschrittverfahren zur Integration von Anfangswertproblemen

$$(P_2) \quad \begin{cases} y'(t) &= f(t,y), & t \in (0,\infty), \\ y(0) &= y_0. \end{cases}$$

1. Geben Sie die allgemeine Form eines Einschrittverfahrens mit der Verfahrensfunktion $\Phi(t,y)$ an.

2. Es sei

$$\Phi(t,y) = c_2 f(t,y) + c_3 f(t + c_1 h, y + c_1 h f(t,y)).$$

Leiten Sie Bedingungen an die Koeffizienten c_1, c_2, c_3 her, damit das Einschrittverfahren, das durch dieses $\Phi(t,y)$ definiert wird, die Konsistenzordnung 2 hat, d.h., dass für den lokalen Diskretisierungsfehler gilt:

$$L(t,h) = O(h^2).$$

3. Geben Sie die Wahl der c_i bei einem Ihnen bekannten Einschrittverfahren der Ordnung 2 an.

Aufgabe 6.4. (Mehrschrittverfahren)
Betrachten Sie das Milne-Simpson-Verfahren

$$y_{k+1} = y_{k-1} + \frac{h}{3}\left(f_{k+1} + 4f_k + f_{k-1}\right).$$

Zeigen Sie, dass dieses Verfahren zur Lösung des Anfangswertproblems (P_2) von der Konsistenzordnung 4 ist.

6.2 Kapitel

2

Aufgabe 6.5. Zeigen Sie, dass für $h(u) = au$ das Enquist-Osher-Verfahren zur numerischen Behandlung der Erhaltungsgleichung (20) aus der Vorlesung äquivalent ist zu dem Upwind-Diskretisierungsschema

$$(\text{UW}) \qquad \frac{\rho_i^{n+1} - \rho_i^n}{\Delta t} = \begin{cases} -a\,\frac{\rho_i^n - \rho_{i-1}^n}{\Delta x}, & a > 0 \\ -a\,\frac{\rho_{i+1}^n - \rho_i^n}{\Delta x}, & a < 0 \end{cases}$$

für die Erhaltungsgleichung

$$\rho_t(t,x) + a\,\rho_x(t,x) = 0, \qquad \text{in } \mathbb{R}^+ \times \mathbb{R}.$$

Geben Sie die Konsistenzordnung des Verfahrens (UW) an.

Aufagbe 6.6. Gegeben sei die Transportgleichung des ersten Aufgabenblattes

$$(P_1) \quad \begin{cases} \frac{\partial}{\partial t}u(x,t) &= -\frac{\partial}{\partial x}f(u(x,t)), \qquad x \in [0,6], \quad t \in (0,6], \\ u(x,0) &= v(x). \end{cases}$$

In dieser Differentialgleichung soll die Ortsableitung mit dem rückwärtsgenommenen Differenzenquotienten erster Ordnung zur Schrittweite h approximiert werden. Das führt dann in jedem Ortspunkt x_i auf eine gewöhnliche Differentialgleichung in t. Diese können dann mit dem vorwärtsgenommenen Eulerverfahren numerisch gelöst werden. Diese Vorgehensweise heisst **Linienmethode**.
Ihr Programm soll für $f(u) = \frac{1}{2}u^2$ die Transportgleichung mittels der Linienmethode lösen. Die Anfangs- und Randwerte seien dabei:

1. $$v(x) = \begin{cases} -0.2x + 1.0, & x \leq 5 \\ 0.0, & x > 5 \end{cases}$$
   mit den Randwerten $u(0,t) = 1.0,\ u(6,t) = 0.0$.
2. $$v(x) = \begin{cases} -0.2x + 2.0, & x \leq 5 \\ 1.0, & x > 5 \end{cases}$$
   mit den Randwerten $u(0,t) = 2.0,\ u(6,t) = 1.0$.

Für die Ortsdiskretisierung wählen Sie bitte $h = 0.05$ und als Zeitschrittweite $k = 2.0 \cdot h^2$.
Nun zur Ausgabe, die das Programm liefern soll:
Geben Sie unter Verwendung eines Grafikprogrammes (z.B. *xgraph*) die Anfangsverteilung $u(x,0)$ und die Endverteilung $u(x,6)$ aus. Danach erzeugen Sie bitte unter *matlab* mit dem *movie*-Befehl einen Film, der die zeitliche Entwicklung der Lösung zeigt. Dazu stellen Sie bitte nur jede **zehnte** Zeitschicht dar. Anstelle von *matlab* kann auch der *xanim*-Befehl benutzt werden.

6.3 Kapitel

3

6.3.1 Kapitel

3.1

Aufagbe 6.7. Es sei $u : \mathbb{R} \to \mathbb{R}$ eine hinreichend oft differenzierbare Funktion.

1. Welche Ordnung haben die Approximationen der zweiten Ableitungen
$$\begin{aligned} u''(x) &= \frac{u(x+h) - 2u(x) + u(x-h)}{h^2} + O(h^q) \\ u''(x) &= \frac{2u(x) - 5u(x+h) + 4u(x+2h) - u(x+3h)}{h^2} + O(h^r)? \end{aligned}$$
2. Zeigen Sie, dass
$$u^{(iv)}(x) = \frac{u(x+2h) - 4u(x+h) + 6u(x) - 4u(x-h) + u(x-2h)}{h^4} + O(h^2)$$
   richtig ist.
3. Es sei $u : [0,\infty) \to \mathbb{R}$ hinreichend glatt und $u'(0) =: \alpha$. Geben Sie eine finite Differenzenapproximation von $u''(0)$ in Termen von $u(0), u(h)$ und α an. Welche Ordnung hat Ihre Approximation? Welche Bedingungen müssen an $\alpha, u(h)$ und $u(-h)$ gestellt werden, damit die Approximation von 2ter Ordnung ist?

6.3.2 Kapitel

3.2

Aufgabe 6.8. Es sei $u : \mathbb{R}^2 \to \mathbb{R}$ eine hinreichend oft differenzierbare Funktion. Zeigen Sie, dass die die auf dem Fünf-Punkte-Stern basierende finite Differenzenapproximation

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

des Laplaceoperators Konsistent von 2ter Ordnung ist.

Aufgabe 6.9. Seien $G, G' \subset \mathbb{R}^2$ Gebiete und $f \in C^1(G', G)$ eine konforme Abbildung, d.h. f erfüllt

$$f^1_{\xi_1} = f^2_{\xi_2}, \qquad f^1_{\xi_2} = -f^2_{\xi_1}.$$

1. Zeigen Sie, dass
$$|f_{\xi_1}| = |f_{\xi_2}|, \qquad f_{\xi_1} \cdot f_{\xi_2} = 0$$
gilt.

2. Sei $u \in C^2(G)$ und $\hat{u} := u \circ f \in C^1(G')$. Weiter seien
$$\Delta_\xi = \sum_{i=1}^{2} \frac{\partial^2}{\partial \xi_i^2} \qquad \text{und} \qquad \Delta_x = \sum_{i=1}^{2} \frac{\partial^2}{\partial x_i^2}$$
die Laplace-Operatoren bzgl. der Koordinaten von G' bzw. G. Zeigen Sie, dass für konforme Abbildungen $f \in C^1(G', G)$
$$\Delta_\xi \hat{u} = |f_{\xi_1}|^2 \, (\Delta_x u) \circ f \qquad \forall u \in C^2(G)$$
erfüllt ist.

3. Zeigen Sie, dass der Laplace-Operator in zwei Raumdimensionen *rund* ist, d.h. er ist invariant unter Drehungen des Koordinatensystems.

Aufgabe 6.10. Es bezeichne $\Omega := (0,1)^2$ das offene Einheitsquadrat in $\mathbb{R}^2$ mit dem Rand $\Gamma := \partial\Omega$. Mit Ω_h bzw. Γ_h werden die inneren Gitterpunkte bzw. die Randgitterpunkte des achsenparallelen Gitters zur Gitterweite $h := \frac{1}{n+1}$, $n \in \mathbb{N}$, bezeichnet. Sei jetzt die Gitterfunktion u_h eine **nichtkonstante** Lösung der Differenzengleichung

$$\begin{aligned} -\Delta_h u_h(x) &= f(x), \qquad x \in \Omega_h \\ u_h(x) &= \Phi(x), \qquad x \in \Gamma_h, \end{aligned}$$

wobei Δ_h den in Aufgabe 3, Blatt 2 erklärten Differenzenoperator bezeichne.

1. Zeigen Sie, dass für den Fall $f \equiv 0$ die Extrema $\max_{x \in \bar{\Omega}_h} u_h(x)$ und $\min_{x \in \bar{\Omega}_h} u_h(x)$ auf Γ_h angenommen werden, wobei $\bar{\Omega}_h := \Omega_h \cup \Gamma_h$.

2. Weisen Sie ferner nach, dass zwei Lösungen u_h^1, u_h^2 der Differenzengleichung zu verschiedenen Randwerten Φ^1, Φ^2 die Abschätzung
$$\|u_h^1 - u_h^2\|_\infty \le \max_{x \in \Gamma_h} \|\Phi^1(x) - \Phi^2(x)\|$$
erfüllen.

3. Zeigen Sie noch, dass aus $\Phi^1 \le \Phi^2$ auf Γ_h die Tatsache $u_h^1 \le u_h^2$ in Ω_h folgt.

Aufgabe 6.11. Diskretiesieren Sie das Dirichlet-Problem

$$\begin{aligned} \Delta u &= 0, \qquad x \in \Omega \\ u &= g, \qquad x \in \Gamma, \end{aligned}$$

mit Hilfe des gewöhnlichen *Fünf-Punkte-Differenzensterns* Δ_h, wobei die Gitterpunktmengen $\Omega_h \cup \Gamma_h$ durch die nachfolgenden Skizzen gegeben seien. Wählen Sie dabei die Reihenfolge der Gitterpunkte so, dass die Bandbreite des jeweils entstehenden Gleichungssystems minimal wird und geben Sie die resultierenden Gleichungssysteme an.

Aufagbe 6.12. Es sei $\Omega := (0,1) \times (0,1) \subset \mathbb{R}^2$. Benutzen Sie zur Diskretisierung des Dirichlet-Problems

$$\begin{aligned} \Delta u &= f, \qquad x \in \Omega \\ u &= 0, \qquad x \in \Gamma, \end{aligned}$$

den *Neun-Punkte-Differenzenstern* $\Delta_9 := \frac{1}{3}\Delta_{\mathrm{X}} + \frac{2}{3}\Delta_5$ und leiten Sie für das quadratische Gitter bei zeilenweiser Numerierung das resultierende Gleichungssystem her.

6.3.3 Kapitel

3.3

Aufagbe 6.13. Finite Elemente
Gegeben sei das folgende Problem

$$(P_1) \quad \begin{cases} -u''(x) + qu(x) = f(x), & x \in (a,b), \quad q \geq 0 \\ u(a) = 0, \\ u(b) = 0. \end{cases}$$

1. Beweisen Sie den Satz: Ist $u(x)$ Lösung der Variationsformulierung und $u \in C^2(a,b)$, dann ist $u(x)$ auch Lösung des eigentlichen Problems. Damit sind dann beide Probleme äquivalent.

2. Diskretisieren Sie die Variationsformulierung des Problems (P_1) mittels linearer finiter Elemente. Dabei ist der *Ritz-Galerkin-Ansatz* zu wählen. Geben Sie sowohl die Systemmatrix, alsauch die rechte Seite an.

3. Wählen Sie $q = 0$ und geben Sie die Systemmatrix und die rechte Seite an. Was fällt Ihnen verglichen mit dem resultierenden Gleichungssystem des finiten Differenzen Ansatzes auf?

Aufagbe 6.14. Stabilität im 1D
Es sei $I = (a,b) \subset \mathbb{R}$ ein Intervall, I_h ein äquidistantes Gitter zu Gitterweite $h := \frac{b-a}{n+1}$ und Γ_h der Rand von I_h. Die skalierte euklidische Norm einer Gitterfunktion u_h sei definiert durch

$$\|u_h\|_{2,I_h} := \left(h \sum_{x \in I_h} |u_h(x)|^2 \right)^{\frac{1}{2}}.$$

1. Zeigen Sie, dass für zwei Gitterfunktionen u_h, v_h, die auf Γ_h verschwinden, gilt:

$$\sum_{x \in I_h} \left(D^+ u_h\right)(x) v_h(x) = -\sum_{x \in I_h} u_h(x) \left(D^- v_h\right)(x).$$

2. (Diskrete Poincare-Ungleichung)
Zeigen Sie, dass für Gitterfunktionen u_h, die auf Γ_h verschwinden, die Ungleichung

$$\|u_h\|_{2,I_h} \leq (b-a) \|D^- u_h\|_{2,I_h}$$

Gültigkeit besitzt.

3. Es sei L_h die auf zentralen Differenzen basierende Diskretisierung des Differentialoperators $Lu := -u'' + qu,\ q \geq 0$. Zeigen Sie, dass für jeden Eigenwert λ von L_h die Ungleichung

$$\lambda \geq \frac{1}{(b-a)^2}$$

gilt.

4. Zeigen Sie, dass L_h stabil bzgl. $\|\cdot\|_2$ ist.

Aufagbe 6.15. (Satz von Weinstein)
Beweisen Sie den Satz von Weinstein für den Spezialfall symmetrischer Matrizen:
Sei A eine symmetrische Matrix und $x \neq 0$ ein beliebiger Vektor, so liegt in dem Kreis

$$K := \left\{ \lambda \in \mathbb{R}; |\lambda - R(x)| \leq \left(\frac{\|Ax\|_2^2}{\|x\|_2^2} - R(x)^2 \right)^{\frac{1}{2}} \right\}$$

mindestens ein Eigenwert von A. Dabei definiert $R(x) := \frac{x^T Ax}{\|x\|_2^2}$ den zu x gehörigen Rayleigh-Quotienten von A.

Aufagbe 6.16. Betrachtet wird das Dirichletproblem

$$\begin{cases} -\Delta u(x,y) &= f(x,y), \qquad (x,y) \in \Omega := (0,1) \times (0,1), \\ u(x,y) &= 0, \qquad (x,y) \in \Gamma := \partial\Omega. \end{cases}$$

Die Diskretisierung erfolge mittels des 9-*Punkte-Differenzensterns* auf einem quadratischen Gitter der Schrittweite h nach den folgenden beiden Schemata (vgl. Übung)

$$\frac{1}{6h^2} \begin{bmatrix} -1 & -4 & -1 \\ -4 & 20 & -4 \\ -1 & -4 & -1 \end{bmatrix} u_h = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} f \qquad \text{bzw} \qquad = \frac{1}{12} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 8 & 1 \\ 0 & 1 & 0 \end{bmatrix} f$$

1. Lösen Sie die diskreten Probleme für $n = 10$ und $h = \frac{1}{n+1}$ numerisch mit Hilfe des *SOR-Verfahrens* jeweils für die Relaxationsparameter $\omega = 1.0, \omega = 1.8$ zu den rechten Seiten

   (a) $f(x,y) = sin(\pi x) sin(\pi y)$ und
   (b) $f(x,y) = 2(x + y - x^2 - y^2)$.

   Brechen Sie die Iteration ab, falls der relative Fehler zweier benachbarter Iterierter, gemessen in der Maximumnorm, kleiner als 10^{-6} ausfällt. Dokumentieren Sie die Anzahl der benötigten Iterationen und stellen Sie die numerischen Lösungen graphisch mit Hilfe eines Softwarepaketes Ihrer Wahl dar.

2. Ermitteln Sie experimentell die Konvergenzordnungen der beiden Diskretisierungen.

Aufagbe 6.17. Sei $A \in \mathbb{R}^{n \times n}$ eine Matrix mit einem Spektralradius $\rho(A) < 1$. **Beweisen** Sie die folgenden Aussagen:

1. Ist A symmetrisch, so gilt
$$\lim_{i \to \infty} A^i = 0.$$

2. Es existiert ein $A \in \mathbb{R}^{n \times n}$ mit $\rho(A) \leq 1$, so dass für eine Matrixnorm $\|.\|$ gilt
$$\lim_{i \to \infty} \|A^i\| = \infty.$$

Aufagbe 6.18. Konvergenz iterativer Verfahren
Wir betrachten das Laplace-Problem

$$(P_1) \quad \begin{cases} -\Delta u &= f \qquad \text{in } \Omega := (0,1)^2 \\ u &= 0 \qquad \text{auf } \Gamma := \partial\Omega. \end{cases}$$

Das Problem (P_1) wird unter Verwendung von Δ_5 auf einem lexikographisch geordneten Gitter zur Schrittweite $h = \frac{1}{n+1}$ diskretisiert. Die resultierende Matrix ist dann eine Block-Tridiagonal-Matrix (siehe VL und UE).

1. Weisen Sie nach, dass zur Matrix $J := -D^{-1}(L + R)$ die Eigenwerte durch

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

und die (i,j)-te Komponente der Eigenvektoren durch

$$z_{i,j}^{(k,l)} = \sin\frac{ki\pi}{n+1}\sin\frac{lj\pi}{n+1}, \qquad 1 \le i,j \le n,$$

gegeben sind.

2. Bestimmen Sie die Spektralradien der Iterationsmatrizen des Einzelschrittverfahrens und des Relaxationsverfahrens zum optimalen Parameter ω_b.
Was lässt sich über die Konvergenzgeschwindigkeit der beiden Verfahren aussagen wenn die Schrittweite h gegen Null geht?

3. Bestimmen Sie die Zahl κ (siehe Übung), die angibt wieviele Schritte das Einzelschrittverfahren benötigt, um dieselbe Fehlerreduktion zu schaffen wie ein Schritt mit dem optimalen Relaxationsverfahren.

Aufagbe 6.19. Es bezeichne $\Omega \subset \mathbb{R}^2$ ein beschränktes, hinreichend glatt berandetes Gebiet, Ω_h ein äquidistantes Gitter mit der Schrittweite h auf Ω. Γ_h sei die Menge der Randgitterpunkte und R_h die Restriktion auf Ω_h. Ferner sei L_h die auf dem Fünf-Punkte Stern basierende Diskretisierung des Laplaceoperators $L = -\Delta$ mit der Konsitenzordnung 2.

1. Sei μ ein Eigenwert von L zur Eigenfunktion u. Zeigen Sie, dass es eine positive Konstante c und eine Gitterfunktion z_h gibt, so dass

$$L_h R_h u - \mu R_h u = ch^2 z_h$$

gilt.

2. Sei λ_h Eigenwert von L_h. Zeigen Sie, dass dann mit μ und $x := R_h u$ gilt

$$|\lambda_h - \mu| \le |\lambda_h - R(x)| + ch^2 \frac{|< R_h u, z_h >|}{\|R_h u\|^2},$$

wobei $R(x)$ der Rayleigh-Quotient zum Operator L_h ist.

3. Zeigen Sie, dass zu jedem Eigenwert μ von L ein Eigenwert λ_h von L_h existiert, so dass gilt

$$|\mu - \lambda_h| = O(h^2).$$

Hinweis: Satz von Weinstein (Aufgabe 6.15).

Aufagbe 6.20. Noch einmal Finite Differenzen

1. Geben Sie Funktionen $p(x) > 0$ und $f(x)$ an, so dass die Randwertaufgabe

$$(RWA_1) \quad \begin{cases} u''(x) + \frac{2x}{1+x^2}u'(x) &= \frac{2+6x^2+2x\cos x}{1+x^2} - \sin x, \qquad x \in I := (0,1), \\ u'(0) &= 1, \\ u(1) &= 0 \end{cases}$$

äquivalent in der Form

$$(RWA_2) \quad \begin{cases} -(p(x)u'(x))' &= f(x), \qquad x \in I, \\ u'(0) &= 1, \\ u(1) &= 0 \end{cases}$$

geschrieben werden kann.

2. $\bar{I}_h$ bezeichne das äquidistante Gitter zur Schrittweite $h = \frac{1}{n+1}$ zum Intervall I. Diskretisieren Sie (RWA_2) durch die Vorschrift

$$(p(x)u'(x))'_i = \frac{1}{h}\left(p_{i+\frac{1}{2}}u'_{i+\frac{1}{2}} - p_{i-\frac{1}{2}}u'_{i-\frac{1}{2}}\right),$$

wobei $u'_{i+\frac{1}{2}} = \frac{1}{h}(u_{i+1} - u_i), u'_{i-\frac{1}{2}} = \frac{1}{h}(u_i - u_{i-1})$ und $q_{i+\frac{1}{2}} = q(x_i + \frac{1}{2}h)$ gesetzt sei. Eliminieren Sie u_{-1} durch die Ansätze

$$u'(0) = \frac{1}{2h}(u_1 - u_{-1}) \qquad \text{bzw.} \qquad u'(0) = \frac{1}{h}(u_0 - u_{-1})$$

und stellen Sie für beide Fälle das zugehörige Gleichungssystem auf.

Aufagbe 6.21. Betrachtet wird das Dirichletproblem

$$(P_1) \quad \begin{cases} -(p(x)u'(x))' + q(x)u(x) &= f(x), \qquad x \in I := (a,b), \\ u(a) = \alpha, \qquad u(b) &= \beta. \end{cases}$$

1. Zeigen Sie, daß das Problem (P_1) ohne Einschränkung der Allgemeinheit in der Form

$$(P_2) \quad \begin{cases} -(p(x)u'(x))' + q(x)u(x) &= g(x), \qquad x \in I := (a,b), \\ u(a) = u(b) &= 0. \end{cases}$$

geschrieben werden kann. Wie sieht die rechte Seite $g(x)$ aus?

2. Geben Sie die exakte Lösung $u \in C^1(a,b)$ des Problems (P_2) an, wenn die Werte wie folgt gewählt werden: $a = -1, b = 1, p(x) = q(x) \equiv 1$,

$$g(x) = \begin{cases} -1, & x \in (-1,0] \\ 1, & x \in (0,1) \end{cases}$$

Aufagbe 6.22. Neumannrandbedingungen im 2D
Es sei das Neumannproblem

$$(P_3) \quad \begin{cases} -\Delta u &= f(x), \qquad x \in \Omega := (0,1)^2, \\ \frac{\partial u}{\partial n} &= g(x), \qquad x \in \Gamma := \partial\Omega \end{cases}$$

vorgelegt. Der Laplaceoperator soll mittels Δ_5 approximiert werden. Wie lautet die zu (91) (siehe VL) analoge Bedingung, wenn die Neumannableitungen durch zentrale Differenzen approximiert werden?

Aufagbe 6.23. Finite Elemente und Finite Differenzen
Betrachtet wird das Dirichletproblem

$$(P_2) \quad \begin{cases} -(p(x)u'(x))' + q(x)u(x) &= g(x), \qquad x \in I := (a,b), \\ u(a) = u(b) &= 0. \end{cases}$$

1. Schreiben Sie ein Programm zur Lösung des Problems (P_2) basierend auf der in Aufgabe 1b) angegebenen Approximation der Ableitungsterme. Verwenden Sie dabei ein äquidistantes Gitter

$$a = x_0 < x_1 < ... < x_n < x_{n+1} = b, \quad x_j = x_0 + jh, \quad h = \frac{b-a}{n+1}, n \in \mathbb{N}.$$

2. Zu $n \in \mathbb{N}$ sei das Intervall wie in Aufgabenteil a) diskretisiert. Schreiben Sie ein Programm zur Lösung des Finiten Elemente-Ansatzes (1.Aufgabe, 4.Übung) zur Lösung des Problems (P_2). Als Grundlage dienen wiederum die linearen Finiten Elemente mit dem *Ritz − Galerkin − Ansatz*.

3. Testen Sie Ihre Programme anhand der Daten aus Aufgabe 2b).

4. Vergleichen Sie die Lösungen von a) und b) für $n = 10, 11$ und $n = 30, 31$. Geben Sie alle Lösungen und Fehler graphisch aus. Was fällt Ihnen bei der Betrachtung der Fehler auf? Woran liegt das?

Aufagbe 6.24. Seien $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet, f, g, h, p vorgelegte Funktionen und $c \geq 0$. Betrachtet wird das Problem mit gemischten Randbedingungen

$$(P_1) \quad \begin{cases} -\Delta u + cu &= f, \qquad \text{in } \Omega, \\ u &= g, \qquad \text{auf } \Gamma_1 \\ \frac{\partial u}{\partial n} + pu &= h, \qquad \text{auf } \Gamma_2. \end{cases}$$

Dabei sei $\partial\Omega = \Gamma_1 \cup \Gamma_2, \Gamma_1 \cap \Gamma_2 = \emptyset$.

1. Leiten Sie die Variationsformulierung analog zu (93) der Randwertaufgabe (P_1) her.

2. Geben Sie das zu Aufgabenteil a) korrespondierende Minimierungsproblem an und zeigen Sie, dass die Bedingung
$$\frac{d}{d\epsilon} J(u^* + \epsilon v) = 0, \qquad \forall v \in V$$
an die Richtungsableitung Ihres Funktionals J wieder auf die Variationsformulierung führt.

Aufgabe 6.25. 2D lineare Finite Elemente
Sei $\Omega \subset \mathbb{R}^2$ ein beschränktes, polygonal berandetes Gebiet und sei $\tau = \{T_1, ..., T_{N_T}\}$ eine zulässige Triangulierung von Ω mit Knoten $K = \{P_1, ..., P_{N_P}\}$. Desweiteren bezeichnen $\phi_i, \quad 1 \leq i \leq N_P$ die 2D-linearen Finiten Elemente.

1. Zeigen Sie, dass die Menge $\{\phi_i | 1 \leq i \leq N_P\}$ linear unabhängig in $C^0(\bar{\Omega})$ ist.

2. Leiten Sie durch die Wahl eines geeigneten Ansatzraumes aus der schwachen Formulierung von (P_1) aus Aufgabe 6.24a) ein Gleichungssystem her. Die auftretenden Integrale sollen dabei nicht berechnet werden.

Aufgabe 6.26. Eigenschaften der Diskretisierung
Es seien $\phi_i, \quad 1 \leq i \leq N_P$ wiederum die 2D-linearen Finiten Elemente. Zeigen Sie die folgenden Eigenschaften der Diskretisierung aus Aufgabe 2b):

1. Die Matrix $M = (m_{i,j})_{i,j=1,...,N_P}, m_{i,j} := \int_\Omega \phi_i \phi_j dx$ ist symmetrisch und positiv definit.

2. Die Matrix $A = (a_{i,j})_{i,j=1,...,N_P}, a_{i,j} := \int_\Omega \nabla\phi_i \nabla\phi_j dx$ ist symmetrisch und positiv semidefinit.

3. Die Matrix $A = (a_{i,j})_{i,j=1,...,N_I}, a_{i,j} := \int_\Omega \nabla\phi_i \nabla\phi_j dx$ ist symmetrisch und positiv definit. Dabei bezeichne N_I die Anzahl der in Ω liegenden Knoten der Triangulierung τ.

Aufgabe 6.27. Triangulierungen

1. Welche der beiden folgenden Triangulierungen sind zulässig (gemäss Definition 3.29)? Bennenen Sie ggf. die Dreiecke und Punkte, die gegen die Bedingungen verstossen.

Abbildung 12: Gebiet aus der 4. Numerischen Aufgabe

2. Verfeinern Sie das rechte Gebiet einmal durch kongruente Verfeinerung und einmal durch Bisektion nach der längsten Kante eines Dreiecks der Makrotriangulierung. Machen Sie bei Ihrer Verfeinerung deutlich, wo hängende Knoten entstehen und wie diese beseitigt werden. Geben Sie die Kanten an, nach denen im nächsten Bisektionsschritt verfeinert werden würde.

3. Geben Sie die Belegung von *itnode* und *itedge* von Dreieck 2 und Dreieck 8 an. Die lokale Numerierung sei dabei vom grössten Winkel aus gesehen entgegen des Uhrzeigersinns. Die Ränder des Gebietes seinen Dirichleträndern.

Aufgabe 6.28. Transformation auf das Einheitsdreieck
$\hat{T}$ bezeichne das Einheitsdreieck mit den Koordinaten (ξ, η), T bezeichne ein beliebiges Dreieck einer Triangulierung mit den Koordinaten (x, y).

1. Leiten Sie die Abbildung (105) her, die die Transformation auf das Einheitsdreieck darstellt.

2. Wie lassen sich die Ableitungen u_{x_1}, u_{x_2} einer Funktion durch die Ableitungen u_ξ, u_η ausdrücken? (d.h.: Leiten Sie die Formel (107) her.)

Aufgabe 6.29. Integrale auf dem Einheitsdreieck bzw -quadrat
$\hat{T}$ bezeichne das Einheitsdreieck, $\hat{Q}$ das Einheitsquadrat.

1. Beweisen Sie den Hilfssatz 3.2 der Vorlesung.

2. Mit Hilfe dieses Satzes lassen sich die Masseelementarmatrix e_m und die Steifigkeitselementarmatrix e_a berechnen:

$$\begin{aligned} e_m(i,j) &= \int_T b_{m_i} b_{m_j} dx \\ e_a(i,j) &= \int_T \nabla b_{m_i} \nabla b_{m_j} dx, \end{aligned}$$

   wobei mit b_{m_i} die 2D linearen finiten Elemente bezeichnet werden und i, j die lokale Numerierung ist. Diese Matrizen sollen Sie berechnen.

3. Zeigen Sie, dass für lineare Funktionen $p(x_1, x_2)$

$$\int_T p \; dx = |T| \; p(x_s^T)$$

   gilt, wobei x_s^T der Schwerpunkt des Dreiecks T ist.

Aufgabe 6.30. Berechnen Sie mit Hilfe von linearen finiten Elementen Approximationen der Lösung des Problems

$$(P) \quad \begin{cases} -\triangle u + cu &= 1, & \text{in } \Omega := \{(r\cos(\phi), r\sin(\phi)) \; 0 < r < 1, 0 < \phi < \alpha\pi\} \\ u &= g & \text{auf } \Gamma_D \\ \frac{\partial u}{\partial n} &= 0 & \text{auf } \Gamma_N \end{cases}$$

jeweils für $\alpha = \frac{1}{2}, \frac{2}{3}, 2$ auf dem Gebiet Ω aus Abbildung 1, wobei

1. $\Gamma_D = \partial\Omega, \Gamma_N = \emptyset, g(r,\phi) = r^{\frac{1}{\alpha}} \sin\frac{\phi}{\alpha} - \frac{r^2}{4}$,

2. $\Gamma_D = \Gamma, \Gamma_N = \Gamma^+ \cup \Gamma^-, g(r,\phi) = r^{\frac{1}{\alpha}} \cos\frac{\phi}{\alpha} - \frac{r^2}{4}$,

3. $\Gamma_D = \Gamma^+ \cup \Gamma, \Gamma_N = \Gamma^-, g(r,\phi) = r^{\frac{1}{2\alpha}} \sin\frac{\phi}{2\alpha} - \frac{r^2}{4}$

gewählt werden. Berechnen Sie die numerischen Lösungen mit Hilfe des CG-Verfahrens aus Algorithmus 6.41 und geben Sie diese in allen Fällen graphisch aus. Als freiwillige Zusatzaufgabe kann das cg-Verfahren mit dem SOR-Verfahren vorkonditioniert werden.
Untersuchen Sie das Verhalten der Lösung für verschiedene c (z.B. c=0, c=1, c=10, c=100). Die Triangulierung hat so zu geschehen, dass auch die krummen Ränder verfeinert werden

6.3.4 Kapitel

3.4

Aufgabe 6.31. cg-Verfahren
Berechnen Sie mit Hilfe des cg-Verfahrens Algorithmus 6.41 die **exakte** Lösung des Gleichungssystems

$$\begin{bmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Aufgabe 6.32. Eigenschaften des cg-Verfahrens
Betrachten Sie das cg-Verfahren aus Algorithmus 6.41.

1. Zeigen Sie, dass sowohl im Basis cg-Verfahren (Algorithmus 6.41) als auch in der Formulierung (120) stets r^{i+1} orthogonal zu p^i ist.

2. Zeigen Sie, dass beide Formulierungen des cg-Verfahrens äquivalent sind.

Aufgabe 6.33. Vorkonditioniertes cg-Verfahren
Die vorliegende Makrotriangulierung des Einheitsquadrates soll durch kongruentes Verfeinern trianguliert werden.

1. Zeigen Sie, dass das Laplaceproblem

$$\begin{cases} -\triangle u &= f, & \text{in } \Omega := (0,1) \times (0,1) \\ u &= 0 & \text{auf } \Gamma := \partial\Omega \end{cases}$$

diskretisiert mittels linearer Finiter Elemente im wesentlichen auf den 5-Punkte-Stern führt. Worin liegt der Unterschied? Wie kann man ihn interpretieren?

2. Wieviel Schritte (in Abhängigkeit von $\kappa(A)$) des cg-Verfahrens sind nötig, um den Fehler auf $\frac{1}{10}$ des Anfangsfehlers zu reduzieren?

3. Berechnen Sie exemplarisch zu den Schrittweiten $h = \frac{1}{10}$ und $h = \frac{1}{100}$ die Anzahl der Schritte, die nötig sind, um den Anfangsfehler der Diskretisierung von **Aufgabenteil a)** auf $\frac{1}{10}$ zu reduzieren. Vergleichen Sie diese Zahlen mit den nötigen Schritten falls mittels des SSOR-Verfahrens vorkonditioniert wird.

Aufgabe 6.34. Effizientes Speichern dünnbesetzter Matrizen
Geben sei die folgende 9×9-Matrix A:

$$A := \begin{bmatrix} 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 4 & -1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 4 & 0 & 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 4 & -1 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -1 & 4 & -1 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & -1 & 4 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 4 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & -1 & 4 \end{bmatrix}$$

1. Geben Sie die *sparse*-Vektoren H, CN und RS (siehe (118)) an. Nutzen Sie aus, dass die Matrix A symmetrisch ist.

2. Welche Dimension haben die Vektoren H, CN und RS? Was sind die Vorteile dieser Speicherungsmethode?

Aufgabe 6.35. Energienorm
Es sei $A \in \mathbb{R}^{n \times n}$ eine reguläre Matrix. Dann wird durch

$$\|u\|_A := \|Au\|_2, \qquad \forall u \in \mathbb{R}^n$$

die sogenannte **Energienorm** definiert. Weisen Sie die Normeigenschaft nach.

Aufgabe 6.36. Hessenbergmatrix und LR-Zerlegung
Es sei $H \in \mathbb{R}^{n \times n}$ eine obere Hessenbergmatrix.

1. Geben Sie einen Algorithmus für eine $LR-Zerlegung$ dieser Hessenbergmatrix an. L sei dabei eine untere Dreiecksmatrix mit einsen auf der Diagonalen, R eine rechte obere Dreiecksmatrix. Welches Besetzungsmuster haben L und R?

2. Wie kann man diese Zerlegung benutzen, um das lineare Gleichungssystem $Hx = b$ zu lösen? Wie viele Operationen sind notwendig für die Lösung des Gleichungssystems (LR-Zerlegung **und** auflösen)?

Aufgabe 6.37. Berechnen Sie mit Ihrem Algorithmus aus Aufgabe 6.36 zuerst die LR-Zerlegung der Hessenbergmatrix und bestimmen Sie danach die Lösung des Gleichungssystems

$$\begin{bmatrix} 2 & 1 & 2 & 0 \\ 1 & 2 & 1 & 2 \\ 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix} x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Aufgabe 6.38. QR-Zerlegung von Hessenberg-Matrizen
Beweisen Sie den $Hilfsatz$ 3.58 aus der Vorlesung.
Bemerkung: Es muss auch gezeigt werden, dass das Produkt der Rotationsmatrizen eine orthogonale Matrix bildet.

6.3.5 Kapitel

3.5

Aufgabe 6.39. Fehlerdarstellung im 1D
Es sei u hinreichend glatt und $(I_h u)(x)$ linear interpolierend auf $[x_{i-1}, x_i]$. Zeigen Sie die folgende Fehlerdarstellung

$$u(x) - (I_h u)(x) = \frac{1}{x_i - x_{i-1}} \int_{x_{i-1}}^{x} \int_{x_{i-1}}^{x_i} \int_{\eta}^{\xi} u''(s)\, ds\, d\eta\, d\xi.$$

Aufgabe 6.40. Gegeben sei die Fehlerdarstellung aus Aufgabe 6.39. Zeigen Sie die folgenden Abschätzungen:

1. $\|u(x) - (I_h u)(x)\|_{L^2} \leq ch^2 \|u''\|_{L^2}$,
2. $\|\left(u(x) - (I_h u)(x)\right)'\|_{L^2} \leq ch \|u''\|_{L^2}$,

wobei $h := \max |x_i - x_{i-1}|$, $1 \leq i \leq n+1$ gilt.

Aufgabe 6.41. Fehlerabschätzungen im 2D, C'eas Lemma.
Sei $\Omega \subset \mathbb{R}^2$ ein beschränktes, polygonal berandetes Gebiet und $\tau := \{\tau_h\}_{h>0}$ eine Familie zulässiger Triangulierungen von Ω. Ferner erfülle die stückweise lineare, global stetige Interpolation $I_h u$ der Lösung $u \in H^2(\Omega)$ des Problems

$$(P) \begin{cases} -\Delta u &= f & \text{in } \Omega \\ u &= 0 & \text{auf } \partial\Omega \end{cases}$$

mit $f \in L^2(\Omega)$ die Abschätzung

$$\|u - I_h u\|_{0,\Omega} + h\|\nabla(u - I_h u)\|_{0,\Omega} \leq ch^2 \|u\|_{2,\Omega},$$

wobei c eine positive Konstante bezeichnet.
Zeigen Sie, dass die stückweise lineare, global stetige Finite-Element-Approximation u_h zur Lösung u von (P) die Abschätzungen

1. $\|u - u_h\|_{1,\Omega} \leq c_1 h \|u\|_{2,\Omega}$
2. $\|u - u_h\|_{0,\Omega} \leq c_2 h^2 \|u\|_{2,\Omega}$

erfüllt. Zum Nachweis von b) sei vorausgesetzt, dass jede Lösung u von (P) der Abschätzung

$$\|u\|_{2,\Omega} \leq c_3 \|f\|_{0,\Omega}$$

genügt. Dabei bezeichnen die c_i positive Konstanten.

Aufgabe 6.42. Friedrichsche Ungleichung.
Es sei $H_0^1(0,\pi)$ wie folgt definiert:

$$H_0^1(0,\pi) := \left\{v \in H^1(0,\pi); v(0) = v(\pi) = 0\right\}.$$

Bestimmen Sie die optimale Konstante in der Friedrichschen Ungleichung für Funktionen aus $H_0^1(0,\pi)$, d.h., bestimmen Sie das kleinste $c > 0$ derart, daß

$$\|v\|_0 \leq c\|v'\|_0 \qquad \forall v \in H_0^1(0,\pi)$$

gilt.
Hinweis: Betrachten Sie $I(v) := \frac{\|v'\|_0^2}{\|v\|_0^2}$, $\quad v \in H_0^1(0,\pi),\ v \neq 0$ und minimieren Sie $I(v)$.

Aufgabe 6.43. Adaptive Gittererzeugung
Vorgelegt ist das folgende Problem

$$(P) \quad \begin{cases} -\epsilon\triangle u + cb^T \nabla u &= f, & \text{in } \Omega := \{(r\cos(\phi), r\sin(\phi))\ 0 < r < 1, 0 < \phi < \alpha\pi\} \\ u &= g & \text{auf } \Gamma_D \\ \frac{\partial u}{\partial n} &= 0 & \text{auf } \Gamma_N \end{cases}$$

Abbildung 13: Gebiet aus der 5. Numerischen Aufgabe

jeweils für $\alpha = \frac{1}{2}$ und $\frac{2}{3}$ auf dem Gebiet Ω aus Abbildung 1, wobei

$$\Gamma_D = \partial\Omega,\ \Gamma_N = \emptyset,\ g(r,\phi) = r^{\frac{1}{\alpha}} \sin\frac{\phi}{\alpha} - \frac{r^2}{4}$$

gewählt werden. Ihre Aufgabe ist nun wie folgt:

1. Sei $c = 0$. Berechnen Sie eine adaptive Gitterverfeinerung mittels des Fehler-Indikators von Johnson und mittels des Fehlerschätzers von Bautz/Weiser. Als Normen sollen sowohl die $\|.\|_{L^\infty}$- als auch die $\|.\|_{H^1}$-Norm verwendet werden. Die rechte Seite sei $f = \epsilon$, die exakte Lösung ist dann gleich $g(r,\phi)$.
2. Sei $c = 1$ und $b^T := (1,1)$. Berechnen Sie die numerische Lösung zur rechten Seite $f = 1$.
   Wie ist das Verhalten der Lösung, wenn ϵ sehr klein wird? Hier soll als Löser das GMRES-Verfahren aus Algorithmus 83 programmiert werden.

Aufgabe 6.44. Vorgelegt ist das folgende Problem

$$(P_1) \quad \begin{cases} -\triangle u &= 1, & \text{in } \Omega := \{(r\cos(\phi), r\sin(\phi))\ 0 < r < 1, 0 < \phi < \alpha\pi\} \\ u &= g & \text{auf } \Gamma := \partial\Omega \end{cases}$$

mit $g(r,\phi) := r^{\frac{1}{\alpha}} \sin\frac{\phi}{\alpha} - \frac{r^2}{4}$. β sei durch $\beta := \frac{1}{\alpha}$ definiert.

1. Zeigen Sie, dass durch $u(r,\phi) := g(r,\phi)$ die Lösung des Problems (P_1) gegeben ist.
2. Berechnen Sie den Gradienten in den Koordinaten r und ϕ und den Laplace-Operator in den Koordinaten r und ϕ.
3. Weisen Sie nach, dass für $\beta < 1$ gilt:
$$u(r,\phi) \in H^{1,p}(\Omega), \qquad \forall 1 \le p < \frac{2}{1-\beta}.$$
4. Analog lässt sich zeigen, dass für $\beta \ge 1$ gilt:
$$u(r,\phi) \in H^{2,p}(\Omega), \qquad \forall 1 \le p < \frac{2}{2-\beta}.$$
   Das sollen Sie beweisen.
5. Diskutieren Sie die Auswirkung der oben gezeigten Eigenschaften der Lösung $u(r,\phi)$ auf die Fehlerabschätzungen.

6.4 Kapitel

4

Aufgabe 6.45. Wärmeleitungsgleichung, Konsistenz
Betrachten Sie jetzt die Anfangs-Randwertaufgabe

$$(P_2) \quad \begin{cases} u_t - u_{xx} &= f(x,t), & \text{in } Q := (0,1)\times(0,T) \\ u(0,t) &= u_1(t) & t \in (0,T) \\ u(1,t) &= u_2(t) & t \in (0,T) \\ u(x,0) &= u_0(x) & x \in (0,1). \end{cases}$$

Die Diskretisierung von (P_2) erfolge in Ortsrichtung mit der äquidistanten Schrittweite $h = \frac{1}{n+1}$ und in Zeitrichtung mit der äquidistanten Schrittweite $k = \frac{T}{m}$. Desweiteren bezeichne u_i^j den Näherungswert der exakten Lösung in (x_i, t_j) (f_i^j sei analog definiert).
Zur numerischen Lösung des Problems (P_2) sei das folgende Schema definiert:

$$\begin{cases} \frac{1}{k}\left(u_i^{j+1} - u_i^j\right) &= D^+D^-(\sigma u_i^{j+1} + (1-\sigma)u_i^j) + \tilde{f}_i^j, & i = 1, ..., n, \quad j = 1, ..., m \\ u_0^j &= u_1(t_j) & j = 0, ..., m \\ u_{n+1}^j &= u_2(t_j) & j = 0, ..., m \\ u_i^0 &= u_0(x_i) & i = 0, ..., n+1, \end{cases}$$

wobei D^+D^- den zentralen Differenzenquotienten in Ortsrichtung bezeichnet. Zeigen Sie, dass für den Konsistenzfehler

1. $O(h^2 + k)$ bei beliebigem $\sigma, \tilde{f}_i^j := f(x_i, t_j), u \in C^{4,2}(\bar{Q})$, bzw
2. $O(h^2 + k^2)$ bei $\sigma = \frac{1}{2}, \tilde{f}_i^j := f(x_i, t_j + \frac{k}{2}), u \in C^{4,3}(\bar{Q})$ gilt.
3. Was bedeuten die beiden Aussagen für die Wahl von h und k?

Aufgabe 6.46. Linienmethode
Gegenstand dieser Aufgabe ist wieder die inhomogene Wärmeleitungsgleichung (P_2) aus der Aufgabe 2 mit $u_1(t) = u_2(t) = 0$. Allerdings wird jetzt zeitkoninuierlich diskretisiert, d.h. in Ortsrichtung wird wieder der zentrale Differenzenquotient angesetzt, die Zeitrichtung bleibt jedoch kontinuierlich (vgl. Linienmethode).

1. Wie lautet das entstehende (zeitkontinuierliche) lineare Gleichungssystem?
2. Entkoppeln Sie dieses Gleichungssystem, indem Sie alle auftretenden Funktionen nach einer Orthogonalbasis von Eigenvektoren der Systemmatrix entwickeln.
   Hinweis: Die Eigenvektoren und die Eigenwerte müssen nicht berechnet werden. Nutzen Sie lediglich die Eigenschaften der Eigenvektoren aus.
3. Lösen Sie die auf diese Weise erhaltenen gewöhnlichen Differentialgleichungen für die **homogene** Wärmeleitungsgleichung. Geben Sie damit die homogene Lösung von (P_2) entwickelt nach Eigenvektoren an.

Aufgabe 6.47. Dualräume, Dualoperatoren
Seien X, Y zwei beliebige Banachräume und $T \in L(X, Y)$. Für jedes $y^* \in Y^*$ definiert

$$< Tx, y^* >_{Y \times Y^*} = < x, x^* >_{X \times X^*} \qquad \forall x \in X$$

ein eindeutiges $x^* \in X^*$. Die lineare Abbildung $T^* : Y^* \to X^*$ mit $T^*y^* = x^*$ definiert den dualen Operator.

1. Zeigen Sie, dass der Dualoperator T^* der Gleichung $\|T^*\|_{X^* \leftarrow Y^*} = \|T\|_{Y \leftarrow X}$ genügt.
2. Aus welchen Dualräumen stammen die folgenden Funktionale?
   (a) $J_1(u, v) = \|u\|^2_{H^2(\Omega)} + \frac{1}{2}\|v\|^2_{H^2(\Omega)}$
   (b) $J_2(u) = \int_\Omega \nabla u \nabla v d\Omega - \int_{\partial\Omega} \frac{\partial u}{\partial n} v dS \qquad \forall v$
   (c) $J_3(u) = \|u\|_{H^0(\Omega)}$

Aufgabe 6.48. Stabilität
Betrachten Sie noch einmal die Aufgabe 6.45. Diskutieren Sie mittels des Seperationsansatzes die Stabilität des Diskretisierungsschemas. Für welche σ ist das Verfahren stabil? Welche Bedingungen werden an die Schrittweiten h und k gestellt?

Aufgabe 6.49. Gegeben sei die Anfangs-Randwertaufgabe

$$(P) \quad \begin{cases} u_t - Du &= f, \quad \text{in } \Omega \times (0,T) \\ u(0,t) &= 0 \quad t \in (0,T) \\ u(1,t) &= 0 \quad t \in (0,T) \\ u(x,0) &= u_0(x) \quad x \in \Omega, \end{cases}$$

wobei abhängig von der Raumdimension $\Omega = (0,1) \subset \mathbb{R}, D = \frac{\partial^2}{\partial x^2}$ oder $\Omega = (0,1)^2 \subset \mathbb{R}^2, D = \Delta$ gelte. Die Diskretisierung von (P) erfolge in Ortsrichtung mittels der äquidistanten Schrittweite h.

1. Stellen Sie für beide Fälle mittels der Linienmethode das System gewöhnlicher Differentialgleichungen auf. Geben Sie die entstehenden Matrizen an.

2. Die unter a) aufgestellten Systeme gewöhnlicher Differentialgleichungen sind exponentiell stabil. Das sollen Sie zeigen. **Hinweis:** Aufgabe 6.14.

A Ein- und Mehrschrittverfahren für Anfangswertaufgaben

Sei $f : [0,T] \times \mathbb{R}^n \to \mathbb{R}^n$ hinreichend glatt. Betrachtet wird das Anfangswertproblem

$$(A) \begin{cases} y'(t) &= f(t, y(t)), \\ y(0) &= y_0. \end{cases}$$

Zur numerischen Diskretisierung definiere durch

$$t_0 := 0, \quad t_{j+1} := t_j + \Delta t_j, \ j \in \mathbb{N}$$

ein Zeitgitter $I_{\Delta t}$, welches hier der Einfachheit halber zunächst als äquidistant, d.h., $\Delta t_j = \Delta t \ \forall j \in \mathbb{N}$, vorausgesetzt wird.
Ziel ist es nun, eine Gitterfunktion

$$y_{\Delta t} : I_{\Delta t} \to \mathbb{R}^n$$

zu finden, welche die Lösung y von (A) in den Gitterpunkten t_j möglichst gut approximiert, d.h.,

$$y_{\Delta t}(t_j) - y(t_j) \overset{!}{=} \text{“klein“},$$

und für zunehmende Feinheit des Gitters in den Gitterpunkten gegen die Lösung konvergiert, d.h.,

$$y_{\Delta t}(t_j) - y(t_j) \overset{!}{\to} 0 \quad (\Delta t \to 0).$$

Zur Abkürzung schreibe

$$y_j \quad := \quad y_{\Delta t}(t_j).$$

Mit diesen Notationen können Einschrittverfahren definiert werden.

Definition A.1. (Einschrittverfahren)
Ein Einschrittverfahren zur Bestimmung einer Näherungslösung $y_{\Delta t}$ auf einem Gitter $I_{\Delta t}$ hat die Form

$$(220) \qquad \begin{cases} y_0 : &= y(0) \\ t_{j+1} : &= t_j + \Delta t_j \\ y_{j+1} : &= y_j + \Delta t_j \varphi(t_j, y_j, \Delta t_j). \end{cases} \qquad j = 0, 1, \ldots, m-1$$

Dabei heißt

$$\varphi(\cdot, \cdot, \Delta t) \ [0,T] \times \mathbb{R}^n \ \to \mathbb{R}^n$$

Verfahrensfunktion des zur Differentialgleichung $y' = f(t,y)$ bei der Schrittweite Δt gehörenden numerischen Verfahrens.

Beispiel

1. Euler'sches Verfahren $\varphi(t, y, \Delta t) = f(t, y)$
2. Verbessertes Euler'sches Verfahren $\varphi(t, y, \Delta t) = f(t + \frac{\Delta t}{2}, y + \frac{\Delta t}{2} f(t, y))$

Verfahrensfunktionen können auch implizit sein;

3. Implizites Eulerverfahren

$$\left.\begin{array}{l} y_0 = y(0), \\ t_{j+1} = t_j + \Delta t_j \\ y_{j+1} = y_j + \Delta t_j f(t_{j+1}, y_{j+1}), \end{array}\right\} j = 0, 1, \ldots, m-1, \text{ und}$$

4. Trapezverfahren

$$\left.\begin{array}{l} y_0 = y(0) \\ t_{j+1} = t_j + \Delta t_j \\ y_{j+1} = y_j + \frac{\Delta t_j}{2}[f(t_j, y_j) + f(t_{j+1}, y_{j+1})] \end{array}\right\} j = 0, 1, \ldots, m-1.$$

□

Hier ist wie oben

$$T = t_m.$$

Ist z.B.

$$f(t, y) := \lambda y, \quad \lambda \in \mathbb{R},$$

so gilt in 3.

$$\varphi(t, y, \Delta t) = \frac{\lambda}{1 - \Delta t \lambda}\ y,$$

in 4.

$$\varphi(t, y, \Delta t) = \frac{\lambda}{1 - \frac{\Delta t \lambda}{2}}\ y.$$

Wichtige Begriffe sind Konsistenz und Konvergenz. Zu deren Erluterung ist die Einführung des Verfahrensfehlers hilfreich.

Definition A.2. (Verfahrensfehler)
Sei y die Lösung von (A) (welche als eindeutig existent vorausgesetzt wird), und sei φ die Verfahrensfunktion eines Einschrittverfahrens zur Integration von

$$y' = f(t, y), \quad y(0) = y_0.$$

Dann heißt

$$r(t, y(t), \Delta t) := \frac{y(t + \Delta t) - y(t)}{\Delta t} - \varphi(t, y(t), \Delta t)$$

lokaler Verfahrensfehler des Einschrittverfahrens an der Stelle $(t, y(t))$.

Definition A.3. (Konsistenz)
Sei y die Lösung der Anfangswertaufgabe (A) und sei φ die Verfahrensfunktion eines Einschrittverfahrens zur Integration von (A). φ heißt konsistent, genau dann, wenn

$$\sup_{t \in [0,T]} |r(t, y(t), \Delta t)| \quad \to 0 \qquad (\Delta t \to 0), \tag{221}$$

und ist konsistent von der Ordnung $p > 0$ genau dann, wenn

$$\sup_{t \in [0,T]} |r(t, y(t), \Delta t)| \quad \leq \quad K \Delta t^p \qquad (\Delta t \to 0). \tag{222}$$

Definition A.4. (Konvergenz)
Ein numerisches Verfahren, daß zu jedem Gitter $I_{\Delta t}$ eine Gitterfunktion $y_{\Delta t}$ berechnet, heißt konvergent für die AWA (A), falls für

$$\varepsilon_{\Delta t}(t) := \ y(t) - y_{\Delta t}(t), \quad t \in [0, T]$$

die Beziehung

$$||\varepsilon_{\Delta t}||_{\Delta t} := \max_{t \in I_{\Delta t}} |y_{\Delta t}(t) - y(t)| \to 0 \quad (\Delta t \to 0)$$

gilt. Die Konvergenz ist von der Ordnung $p > 0$ genau dann, wenn

$$\|\varepsilon_{\Delta t}\|_{\Delta t} = \mathcal{O}(\Delta t^p) \quad (\Delta t \to 0).$$

Beispiel
1. Euler

$$\begin{aligned} r(t, y, \Delta t) &= \frac{y(t + \Delta t) - y(t)}{\Delta t} - \underbrace{f(t, y(t))}_{y'(t)} \\ &= \frac{\Delta t}{2} y''(t) + \mathcal{O}(\Delta t^2) \\ &= \mathcal{O}(\Delta t) \quad (\Delta t \to 0). \end{aligned}$$

2. Verbesserter Euler

$$r(t, y, \Delta t) = \mathcal{O}(\Delta t^2) \quad (\Delta t \to 0).$$

□

Runge-Kutta-Verfahren sind wie folgt definiert.

Definition A.5. (Runge-Kutta-Verfahren)
Sei $s \in \mathbb{N}$. Ein Runge-Kutta-Verfahren hat die Gestalt

$$\left.\begin{array}{ll} y_0 & = y(0) \\ t_{j+1} & = t_j + \Delta t \\ y_{j+1} & = y_j + \Delta t \varphi(t_j, y_j, \Delta t) \end{array}\right\} \ j = 0, 1, \ldots, m-1,$$

wobei $\Delta t = \Delta t_j$ und für $\kappa = s$ oder $\kappa = i - 1$

$$\varphi(t, y, \Delta t) := \sum_{i=1}^{s} b_i v_i(t, y)$$

mit

$$v_i(t, y) := f\left(t + c_i \Delta t, y + \Delta t \sum_{k=1}^{\kappa} a_{ik} v_k(t, y)\right), \quad i = 1, \ldots, s.$$

Als Schema

c_1	a_1	$\ldots$	a_{1s}
$\vdots$			
c_S	a_{s1}	$\ldots$	a_{ss}
	b_1	$\ldots$	b_s.

Das Verfahren heißt

implizit für $\kappa = s$,

explizit für $\kappa = i - 1$.

Einschrittverfahren benutzen nur Informationen vom vorherigen Zeitschritt. Verfahren, welche auch Informationen von mehreren vergangenen Zeitschritten nutzen, heißen Mehrschrittverfahren. Hier werden nur lineare Mehrschrittverfahren definiert.

Definition A.6. (Lineare Mehrschrittverfahren)
Ein Verfahren der Form

$$y_{j+k} = y_{j+q} + \Delta t \sum_{i=0}^{k} b_i f_{j+i}, \quad t_{j+k} \in I_{\Delta t} = \{t_0, \ldots, t_m\}$$

mit $k \geq 1, 0 \leq q < k, f_{j+i} := f(t_{j+i}, y_{j+i})$, daß zu bekannten Werten $y_0, \ldots, y_{k-1}$, die Werte $y_k, \ldots, y_m$ zu berechnen gestattet, heißt lineares k-Schrittverfahren. Ist $b_k = 0$, so heißt es explizit, sonst implizit. Der lokale Abschneidefehler für Mehrschrittverfahren ist definiert durch

$$r(t, y(t), \Delta t) := \frac{y(t + k\Delta t) - y(t + q\Delta t)}{\Delta t} - \sum_{i=0}^{k} b_i f(t + i\Delta t, y(t + i\Delta t)).$$

Konsistenzordnung ist wie in (221) und (222) definiert.

MSV können über Integrationsformeln hergeleitet werden. Integration von (A) ergibt

$$y(t_{j+k}) = y(t_{j+q}) + \int_{t_{j+q}}^{t_{j+k}} f(t, y(t))dt. \tag{223}$$

Schreibe jetzt

$$\int_{t_{j+q}}^{t_{j+k}} f(t, y(t))dt \approx \Delta t \sum_{i=0}^{k} b_i f(t_{j+i}, y_{j+i}).$$

Um eine solche Näherung zu erhalten, wird etwa der Integrand durch ein Interpolationspolynom in den Stellen

$$(t_{j+i}, f_{j+i}) \begin{cases} i = 0, 1, \ldots, k-1 & \text{(expliziter Fall)} \\ i = 0, 1, \ldots, k & \text{(impliziter Fall)} \end{cases}$$

ersetzt und exakt ausgewertet.

Literatur

[1] Bank/Weiser. Some A Posteriori Error Estimators for Elliptic Partial Differential Equations. *Mathematics of Computation No. 44*, pages 283–30, 1985. 63

[2] Eberly Bebernes. *Mathematical Problems from Combustion Theory*. Springer, New York, 1989. 80

[3] Braess. *Methode der finiten Elemente*. Springer, 1994. 41, 49, 60

[4] Bulirsch/Stoer. *Numerische Mathematik II*. Springer, 1990. 28, 29, 30

[5] Ciarlet. *The finite Element method for elliptic Problems*. North Holland, 1978. 40

[6] Durán/Rodriguez. Asymptotische Exaktheit des Bank-Weiser Fehlerschätzers. *Numerische Mathematik 62*, 1992. 63, 64

[7] Enquist/Osher. One-sided difference approximations for nonlinear conservation laws. *Math. Comput.*, 36:321–351, 1981. 16

[8] L Eriksson, K./Johnson. An adaptive finite element method for linear elliptic problems. *Mathematics of Computation No. 50*, pages 361–383, 1985. 63

[9] Golub/Ortega. *Scientific Computing*. Teubner, 1996. 18

[10] Großmann/Roos. *Numerik partieller Differentialgleichungen*. Teubner, 1994. 23, 27, 40, 47, 56, 60, 67

[11] Sabine Gutsch. Ein Vergleich CG-ähnlicher Verfahren zur Lösung indefiniter Probleme. Master's thesis, Institut für Informatik und Praktische Mathematik, Universität Kiel, 1994. 55

[12] Wolgang Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Teubner, 1993. 50

[13] Hämmerlin/Hoffmann. *Numerische Mathematik*. Springer, 1991. 19

[14] C. Johnson. *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press, 1987. 61

[15] Kroener. *Numerical Schemes for Conservation Laws*. Wiley/Teubner, 1997. 9, 12, 14, 17

[16] V.A. Barker O. Axelsson. *Finite Element Solution of Boundary Value Problems*. Academic Press, 1984. 49, 57, 60

[17] Ortega/Reinhold. *Iterative solution of nonlinear equations in several variables*. Academic Press New York, 1970. 22

[18] René Pinnau. Mathematische Modellierung. Vorlesungskript, Fachbereich Mathematik, TU Darmstadt, 2000. 3, 81

[19] Von Thomeé. *Galerkin Finite Element Methods for Parabolic Problems*. Springer, 1997. 69, 77

[20] Barret; Chan; Demmel; Donato; Dongarra; Eijkhout; Pozo; Romine & van der Vorst. Templates for the solution of linear systems: Building blocks for iterative methods. preprint. 55

[21] A. Klar/R. Wegner. Enskog-like kinetic models for vehicular traffic, 1999. 9