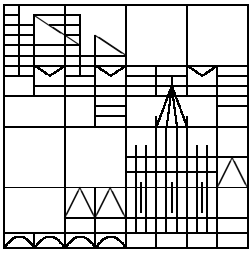
Skript zum Numerik-Teil

der Vorlesung

Theorie und Numerik partieller Differentialgleichungen

Wintersemester 2007/08

Johannes Schropp



Universität Konstanz

Fachbereich Mathematik und Statistik

Stand: 28. November 2008

Inhaltsverzeichnis

8 Finite Differenzenverfahren für elliptische Differentialgleichungen . . . 3

a) Das klassische Differenzenverfahren 3

b) Konsistenz und Stabilität des klassischen Differenzenverfahrens für die Dirichletsche Randwertaufgabe 12

9 Finite Elemente Methoden für elliptische Differentialgleichungen . . . 20

a) Das Ritz'sche Verfahren 20

b) Finite Elemente Methoden 24

c) Theoretische Grundlagen zur Finite Elemente Methode 28

d) Stabilität und Konvergenz der Finite Elemente Methode 35

10 Finite Differenzenverfahren für parabolische Differentialgleichungen . . 42

a) Das Prinzip der Linienmethode 42

b) Konsistenz der Differenzenverfahren 49

c) Stabilität und Konvergenz der Differenzenverfahren 52

11 Hyperbolische Differentialgleichungen 58

a) Differenzenverfahren für die Wellengleichung 58

b) Die Courant-Friedrichs-Levy Bedingung 65

A Anhang 67

a) Iterative Lösung großer Gleichungssysteme 67

8. Finite Differenzenverfahren für elliptische Differentialgleichungen

a) Das klassische Differenzenverfahren

Wir betrachten die Dirichlet'sche Randwertaufgabe

$$\begin{aligned} -\Delta u(x,y) &= g(x,y), \quad (x,y) \in \Omega, \\ u(x,y) &= \gamma(x,y), \quad (x,y) \in \partial\Omega. \end{aligned} \tag{8-1}$$

Hierbei ist $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit gegebenen Funktionen $g : \Omega \longrightarrow \mathbb{R}$, $\gamma : \partial\Omega \longrightarrow \mathbb{R}$. Eine Funktion $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$, welche (8-1) erfüllt, heißt klassische Lösung von (8-1).

8.1 Satz. *Vorgelegt sei (8-1) mit $g \in C^0(\overline{\Omega})$, $\gamma \in C^0(\partial\Omega)$ und $\partial\Omega$ sei stückweise stetig differenzierbar. Dann existiert genau eine klassische Lösung von (8-1).*

Für diese Lösung gilt das Maximum-Minimum Prinzip, d.h.

$$\begin{aligned} g \geq 0 &\implies u \geq \min\{\gamma(x) \,|\, x \in \partial\Omega\}, \\ g \leq 0 &\implies u \leq \max\{\gamma(x) \,|\, x \in \partial\Omega\}. \end{aligned}$$

Dabei setzen wir hier für ein $f : A \longrightarrow \mathbb{R}$:

$$f \geq 0 \iff f(x) \geq 0 \quad \forall x \in A.$$

Seien nun

$$\begin{aligned} L : C^2(\Omega) &\longrightarrow C^0(\Omega), \quad Lu = -\Delta u, \\ R : C^0(\overline{\Omega}) &\longrightarrow C^0(\partial\Omega), \quad Ru = u|_{\partial\Omega}, \end{aligned}$$

so lässt sich (8-1) als

$$Lu = g, \quad Ru = \gamma$$

schreiben.

Sei Ω ein beschränktes Gebiet mit stückweise glattem Rand. Das Paar (L, R) ist invers monoton, d.h. für die nach Satz 8.1 eindeutige Lösung u von $Lu = g$, $Ru = \gamma$ gilt

$$(g, \gamma) \geq 0 \implies u \geq 0.$$

Offensichtlich gilt

$$\gamma, g \geq 0 \implies u \geq \min\{\gamma(x) \,|\, x \in \partial\Omega\} \geq 0.$$

© Johannes Schropp 28. November 2008

Diese Eigenschaft der Inversmonotonie wollen wir auch in den numerischen Verfahren zu (8-1) wiederfinden.

Nun wenden wir uns der numerischen Behandlung des Problems (8-1) zu.

Sei $h > 0$. Wir überziehen zunächst $\mathbb{R}^2$ mit einem äquidistanten Gitter

$$\mathbb{R}^2_h := \{(ih, jh) \,|\, i, j \in \mathbb{Z}\}.$$

Es kann sein, dass $\partial\Omega \cap \mathbb{R}^2_h = \varnothing$ gilt, d.h. dass $\mathbb{R}^2_h$ überhaupt keine Randpunkte von Ω enthält. Um dieser Schwierigkeit erst einmal aus dem Weg zu gehen, nehmen wir an

$$\Omega = (0,1)^2$$

und definieren das Gitter

$$\Omega_h = \{(ih, jh) \,|\, i, j \in \{1, \ldots, M-1\}\}$$

für ein $h = \frac{1}{M} > 0$.

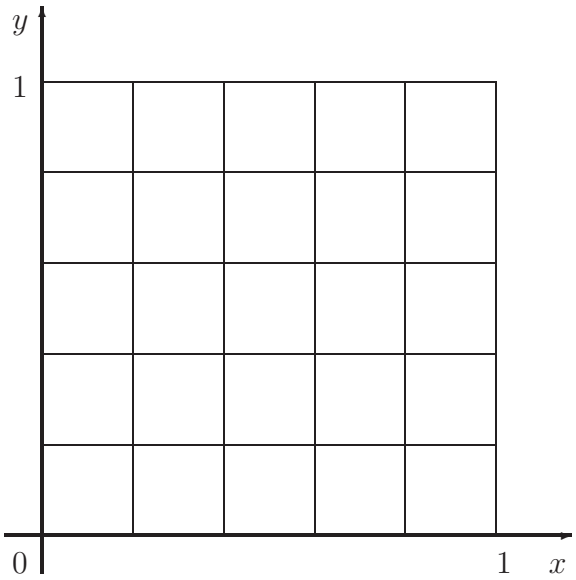


Abbildung 1: Beispiel für ein Gitter über $(0,1)^2$ mit $M = 5$, $h = 0.2$, $(M-1)^2 = 16$

Die Methode der Finiten Differenzen beruht nun darauf, Ableitungen von (8-1) durch Differenzenquotienten zu ersetzen.

Sei nun $v \in C^4([a,b], \mathbb{R})$, so gilt nach der Taylor-Formel

$$v(s \pm h) = v(s) \pm hv'(s) + \frac{h^2}{2}v''(s) \pm \frac{h^3}{6}v^{(3)}(s) + \frac{h^4}{24}v^{(4)}(\eta_\pm)$$

© Johannes Schropp 28. November 2008

für $s, s \pm h \in [a,b]$ mit $\eta_- \in [s-h, s]$, $\eta_+ \in [s, s+h]$.

Dann gilt

$$\begin{aligned}\left|\frac{1}{h^2}(-v(s-h)+2v(s)-v(s+h))+v''(s)\right| &= \\ &= \left|h^{-2}\Big[-v(s)+hv'(s)-\frac{h^2}{2}v''(s)+\frac{h^3}{6}v^{(3)}(s)-\frac{h^4}{24}v^{(4)}(\eta_-)+2v(s)\right. \\ &\quad \left.-v(s)-hv'(s)-\frac{h^2}{2}v''(s)-\frac{h^3}{6}v^{(3)}(s)-\frac{h^4}{24}v^{(4)}(\eta_+)\Big]+v''(s)\right| \\ &= \frac{h^2}{24}\left|v^{(4)}(\eta_-)+v^{(4)}(\eta_+)\right| \le Ch^2 = O(h^2).\end{aligned}$$

Somit ist

$$-v''(s) = \frac{1}{h^2}\big(-v(s-h)+2v(s)-v(s+h)\big)+O(h^2), \tag{8-2}$$

falls $v \in C^4([a,b])$.

Mit der Formel (8-2) für $-v''$ bekommen wir für $(x,y) \in \Omega_h$ folgende Approximationen:

$$\begin{aligned}-u_{xx}(x,y) &\sim \frac{1}{h^2}\big(-u(x-h,y)+2u(x,y)-u(x+h,y)\big), \\ -u_{yy}(x,y) &\sim \frac{1}{h^2}\big(-u(x,y-h)+2u(x,y)-u(x,y+h)\big),\end{aligned}$$

also

$$\begin{aligned}-\Delta u(x,y) \sim \frac{1}{h^2}\big(&-u(x-h,y)-u(x+h,y) \\ &-u(x,y-h)-u(x,y+h)+4u(x,y)\big)\end{aligned}$$

oder mit der Schreibweise $u_{ij} = u(ih, jh)$

$$-(\Delta u)_{ij} \sim \frac{1}{h^2}\big(-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1}+4u_{ij}\big).$$

An den inneren Gitterpunkten

$$\mathring{\Omega}_h = \{(ih,jh) \in \Omega_h \,|\, i,j \in \{2,\dots,M-2\}\}$$

ersetzen wir die Differentialgleichung (8-1) durch

$$h^{-2}\big(-u_{i-1,j}-u_{i+1,j}-u_{i,j-1}-u_{i,j+1}+4u_{i,j}\big) = g_{ij}$$

mit $g_{ij} = g(ih, jh)$.

© Johannes Schropp 28. November 2008

Bei den randnahen Gitterpunkten $\Omega_h^R = \Omega \setminus \overset{\circ}{\Omega}_h$ wählen wir dieselbe Ersetzung, wobei jetzt

$$u_{i\pm1,j} = \gamma_{i\pm1,j} \text{ bzw. } u_{i,j\pm1} = \gamma_{i,j\pm1}$$

einzusetzen ist, falls $((i \pm 1)h, jh)$ bzw. $(ih, (j \pm 1)h) \in \partial\Omega$.

Insgesamt erhalten wir ein lineares Gleichungssystem

$$Au = r \tag{8-3}$$

der Dimension $(M-1)^2$ für die gesuchte Approximation u.

Wir können das Differenzenverfahren auch präziser formulieren.

8.2 Definition. Eine Abbildung

$$\begin{aligned} w &: \Omega_h \longrightarrow \mathbb{R}, \\ w &= (\underbrace{(w(ih, jh))}_{=w_{ij}}, (i,j) \in \{1, \ldots, M-1\}) \end{aligned}$$

nennen wir eine Gitterfunktion und schreiben kurz $w \in \mathbb{R}^{\Omega_h}$ oder genauer $w^h \in \mathbb{R}^{\Omega_h}$.

Somit gilt für A, u, r aus (8-3)

$$A \in \mathbb{R}^{\Omega_h, \Omega_h}, u, r \in \mathbb{R}^{\Omega_h}.$$

Die explizite Darstellung der Matrix A und des Vektors r in (8-3) hängt von der Nummerierung der Gitterpunkte ab.

Beispielsweise erhalten wir für die natürliche, zeilenweise von links unten nach rechts oben laufende Nummerierung folgende Darstellung:

$$\begin{aligned} u &= (u^1, \ldots, u^{M-1}), \quad u^j = (u_{1,j}, u_{2,j}, \ldots, u_{M-1,j}) \in \mathbb{R}^{M-1}, \\ g &= (g^1, \ldots, g^{M-1}), \quad g^j = (g_{1,j}, g_{2,j}, \ldots, g_{M-1,j}) \in \mathbb{R}^{M-1} \end{aligned}$$

sowie

$$B = h^{-2} \begin{pmatrix} 4 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 4 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 4 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 4 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 4 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 4 \end{pmatrix} \in \mathbb{R}^{M-1,M-1},$$

$$C = h^{-2} I \in \mathbb{R}^{M-1,M-1}$$

© Johannes Schropp 28. November 2008

und in Blockschreibweise

$$\underbrace{\begin{pmatrix} B & -C & 0 & \dots & 0 & 0 & 0 \\ -C & B & -C & \dots & 0 & 0 & 0 \\ 0 & -C & B & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & B & -C & 0 \\ 0 & 0 & 0 & \dots & -C & B & -C \\ 0 & 0 & 0 & \dots & 0 & -C & B \end{pmatrix}}_{=:A} \begin{pmatrix} u^1 \\ u^2 \\ u^3 \\ \vdots \\ u^{M-3} \\ u^{M-2} \\ u^{M-1} \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} g^1 \\ g^2 \\ g^3 \\ \vdots \\ g^{M-3} \\ g^{M-2} \\ g^{M-1} \end{pmatrix} + h^{-2} \begin{pmatrix} \widetilde{\gamma}^1 \\ \widetilde{\gamma}^2 \\ \widetilde{\gamma}^3 \\ \vdots \\ \widetilde{\gamma}^{M-3} \\ \widetilde{\gamma}^{M-2} \\ \widetilde{\gamma}^{M-1} \end{pmatrix}}_{=:r} \tag{8-4}$$

mit

$$\begin{aligned} \widetilde{\gamma}^1 &= (\gamma_{1,0} + \gamma_{0,1}, \gamma_{2,0}, \dots, \gamma_{M-2,0}, \gamma_{M-1,0} + \gamma_{M,1}), \\ \widetilde{\gamma}^j &= (\gamma_{0,j}, 0, \dots, 0, \gamma_{M,j}), \quad j = 2, \dots, M-2, \\ \widetilde{\gamma}^{M-1} &= (\gamma_{0,M-1} + \gamma_{1,M}, \gamma_{2,M}, \dots, \gamma_{M-2,M}, \gamma_{M-1,M} + \gamma_{M,M-1}). \end{aligned}$$

Es bleibt nun nachzuweisen, dass das Gleichungssystem (8-3) eine eindeutige Lösung besitzt. Dazu benötigen wir etwas Matrizentheorie.

8.3 Definition. Eine Matrix $A \in \mathbb{R}^{N,N}$ heißt nicht negativ oder monoton, falls $A_{ij} \geq 0$ für $i, j \in \{1, \dots, N\}$. Wir schreiben hierfür einfach $A \geq 0$.

Seien $A, B \in \mathbb{R}^{N,N}$. Dann ist $A \leq B \iff B - A \geq 0$.

8.4 Definition. $A \in \mathbb{R}^{N,N}$ heißt invers monoton, falls A^{-1} existiert und $A^{-1} \geq 0$ gilt.

8.5 Definition. A heißt L_0-Matrix, falls $A_{ij} \leq 0$ für $i, j \in \{1, \dots, N\}$ mit $i \neq j$.

8.6 Definition. A heißt M-Matrix, falls A eine invers monotone L_0-Matrix ist.

Bekanntermaßen gilt für $A \in \mathbb{R}^{N,N}$ die Äquivalenz

$$A \geq 0 \iff (u \leq v \Rightarrow Au \leq Av, \quad \forall u, v \in \mathbb{R}^N).$$

© Johannes Schropp 28. November 2008

8.7 Bemerkung. Die Matrix A des klassischen Differenzenverfahrens (vgl. Gleichungssystem (8-4)) ist offensichtlich eine L_0-Matrix. Um sicherzustellen, dass die Lösung u wohldefiniert ist, müssen wir die Invertierbarkeit von A nachweisen.

Dies geschieht mit

8.8 Satz (M-Kriterium). *Eine L_0-Matrix $A \in \mathbb{R}^{N,N}$ ist genau dann eine M-Matrix, wenn ein $e > 0$, $e \in \mathbb{R}^n$ existiert mit $Ae \geq 0$ und der Verbindungseigenschaft: Zu jedem $i_0 \in \{1, \ldots, N\}$ mit $(Ae)_{i_0} = 0$ gibt es eine Kette $i_0, i_1, \ldots, i_r \in \{1, \ldots, N\}$ mit $(Ae)_{i_r} > 0$ und $A_{i_{j-1}, i_j} \neq 0$ für alle $j \in \{1, \ldots, r\}$.*

8.9 Bezeichnung. Ein $e > 0$ mit $Ae \geq 0$ und der Verbindungseigenschaft heißt majorisierendes Element für A.

Wir betrachten jetzt wieder das Gleichungssystem $A^h u = r^h$ (vgl. (8-4)) des klassischen Differenzenverfahrens zu $-\Delta u = g$ in $\Omega = (0,1)^2$, $u = \gamma$ auf $\partial\Omega$.

8.10 Lemma. *Die Matrix $A^h \in \mathbb{R}^{(M-1)^2,(M-1)^2}$ des klassischen Differenzenverfahrens ist eine M-Matrix.*

Beweis: Offensichtlich ist A^h eine L_0-Matrix. Ferner gilt für $\mathbb{I} = (1, \ldots, 1)^T \in \mathbb{R}^{(M-1)^2}$

$$A^h \mathbb{I} = h^{-2} \begin{pmatrix} \delta^{(2)} \\ \delta^{(1)} \\ \vdots \\ \delta^{(1)} \\ \delta^{(2)} \end{pmatrix},$$

wobei

$$\delta^{(2)} = \begin{pmatrix} 2 \\ 1 \\ \vdots \\ 1 \\ 2 \end{pmatrix} \in \mathbb{R}^{M-1}, \quad \delta^{(1)} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \in \mathbb{R}^{M-1}.$$

Genauer gilt

$$\begin{aligned} A^h \mathbb{I}(x,y) &= 0 \text{ für } (x,y) \in \mathring{\Omega}_h, \\ A^h \mathbb{I}(x,y) &> 0 \text{ für } (x,y) \in \Omega_h \setminus \mathring{\Omega}_h \,. \end{aligned}$$

Wir müssen also zu jedem $(x_0, y_0) \in \mathring{\Omega}_h$ eine Kette $(x_0, y_0), \ldots, (x_r, y_r) \in \Omega_h$ finden mit $A_{(x_{i-1},y_{i-1}),(x_i,y_i)} \neq 0$ für $i = 1, \ldots, r$ und $(x_r, y_r) \in \Omega_h \setminus \mathring{\Omega}_h$. Gemäß Definition von A^h mittels B^h und C^h ist dies aber immer möglich. □

© Johannes Schropp 28. November 2008

8.11 Bemerkung. Da A^h eine M-Matrix ist, können wir das lineare Gleichungssystem (8-3) mit dem Gauß-Algorithmus ohne Pivotisierung auflösen. Dabei ist zu berücksichtigen, dass A^h eine Bandmatrix mit der Bandweite $2(M-1)+1 = 2M-1$ ist.

Die Elimination einer Bandmatrix der Dimension $(M-1)^2$ mit der Bandweite $2M-1$ erfordert etwa $\frac{1}{2}(M-1)^2(2M-1)^2 \sim 2M^4$ Multiplikationen. Der Aufwand steigt also mit M sehr stark an.

Wir behandeln nun das Problem eines krummlinigen Randes. Dazu setzen wir

$$\Omega_h = \Omega \cap \mathbb{R}^2_h$$

und ordnen jedem Punkt $(x, y) \in \Omega_h$ vier Nachbarpunkte $N_k = N_k(x, y, h) \in \overline{\Omega}$, $k = 1, 2, 3, 4$ zu.

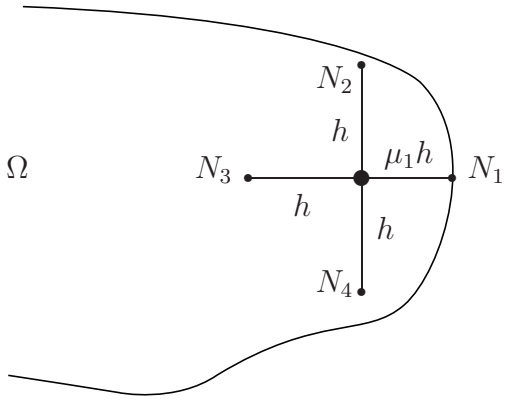


Abbildung 2: Nachbarpunkte

Im Folgenden sei

$$e^1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, e^2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, e^3 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, e^4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix},$$
$$\mu_k = \mu_k(x, y, h) = \sup\left\{\mu \in [0, 1] \,|\, (x, y) + the^k \in \Omega, \forall t \in [0, \mu]\right\},$$
$$N_k = N_k(x, y, h) = (x, y) + \mu_k(x, y, h)he^k, \quad k = 1, 2, 3, 4.$$

Die Menge der inneren Gitterpunkte ist

$$\mathring{\Omega}_h = \{(x, y) \in \Omega_h \,|\, N_k(x, y, h) \in \Omega, \quad k = 1, 2, 3, 4\}.$$

An einem inneren Gitterpunkt (x, y) gilt überdies $\mu_k(x, y, h) = 1$, $k = 1, 2, 3, 4$.

© Johannes Schropp 28. November 2008

Wir können also die Differentialgleichung $-\Delta u = g$ in Ω, $u = \gamma$ auf $\partial\Omega$ an inneren Gitterpunkten $(x,y) \in \overset{\circ}{\Omega}_h$ wieder durch

$$h^{-2}(-u(x-h,y)-u(x+h,y)-u(x,y-h)-u(x,y+h)+4u(x,y)) = g(x,y) \quad (8\text{-}5)$$

diskretisieren.

Die Menge

$$\Omega_h^R = \Omega_h \setminus \overset{\circ}{\Omega}_h = \{(x,y) \in \Omega_h \mid N_k(x,y,h) \in \partial\Omega \text{ für ein } k = 1,2,3,4\}$$

enthält die randnahen Punkte.

Zur Diskretisierung von $-\Delta u(x,y) = g(x,y)$, $(x,y) \in \Omega_h^R$ benötigen wir zunächst eine Formel für die gewöhnliche Ableitung v'' bei nicht äquidistanten Knoten.

8.12 Lemma. *Sei $v \in C^3([-a,a],\mathbb{R})$ für ein $a > 0$. Dann gilt für alle $\mu_0, \mu_1 \in (0,1]$ und $h \le a$*

$$\left| \frac{2}{\mu_0\mu_1(\mu_0+\mu_1)h^2} \Big\{ \mu_0 v(\mu_1 h) - (\mu_0+\mu_1)v(0) + \mu_1 v(-\mu_0 h) \Big\} - v''(0) \right| \le \frac{2}{3} h \cdot \max\{|v'''(x)| \mid |x| \le a\}.$$

Beweis: Taylorentwicklung liefert

$$\begin{aligned} \frac{1}{\mu_1 h}(v(\mu_1 h) - v(0)) &= v'(0) + \frac{1}{2}\mu_1 h v''(0) + \frac{1}{6}(\mu_1 h)^2 v'''(\xi), \\ \frac{1}{\mu_0 h}(v(-\mu_0 h) - v(0)) &= -v'(0) + \frac{1}{2}\mu_0 h v''(0) - \frac{1}{6}(\mu_0 h)^2 v'''(\eta) \end{aligned}$$

mit $\xi, \eta \in [-a,a]$.

Addition ergibt

$$\frac{1}{\mu_0\mu_1 h}\big(\mu_0 v(\mu_1 h) - (\mu_0+\mu_1)v(0) + \mu_1 v(-\mu_0 h)\big) = \frac{\mu_0+\mu_1}{2} h v''(0) + \frac{1}{6}\big((\mu_1 h)^2 v'''(\xi) - (\mu_0 h)^2 v'''(\eta)\big).$$

Division durch $\frac{(\mu_0+\mu_1)h}{2}$ liefert jetzt

$$\frac{2}{\mu_0\mu_1 h(\mu_0+\mu_1)h}\big(\mu_0 v(\mu_1 h) - (\mu_0+\mu_1)v(0) + \mu_1 v(-\mu_0 h)\big) - v''(0) = \frac{1}{3}\frac{1}{(\mu_0+\mu_1)h}\big((\mu_1 h)^2 v'''(\xi) - (\mu_0 h)^2 v'''(\eta)\big) =: R_h.$$

© Johannes Schropp 28. November 2008

Es folgt

$$\begin{aligned} |R_h| &\leq \frac{h}{3}\frac{1}{(\mu_0+\mu_1)}\left(\mu_1^2|v'''(\xi)|+\mu_0^2|v'''(\eta)|\right) \\ &\leq \frac{2h}{3}\max\{|v'''(x)|\,|\,|x|\leq a\}. \end{aligned}$$

□

Mit Hilfe von Lemma 8.12 können wir $-\Delta u(x,y) = g(x,y)$, $(x,y) \in \Omega_h^R$ ersetzen durch

$$\begin{aligned} g(x,y) = h^{-2}\Big\{ &-\frac{2}{\mu_1(\mu_1+\mu_3)}u(N_1) - \frac{2}{\mu_3(\mu_1+\mu_3)}u(N_3) - \frac{2}{\mu_2(\mu_2+\mu_4)}u(N_2) \\ &-\frac{2}{\mu_4(\mu_2+\mu_4)}u(N_4) + 2\left(\frac{1}{\mu_1\mu_3}+\frac{1}{\mu_2\mu_4}\right)u(x,y)\Big\}. \end{aligned} \tag{8-6}$$

Für jeden Nachbarn $N_k \in \partial\Omega$, $k \in \{1,2,3,4\}$ ist dabei der Wert $u(N_k) = \gamma(N_k)$ gemäß $u = \gamma$ auf $\partial\Omega$ einzusetzen.

Das hierdurch definierte Verfahren heißt klassisches Differenzenverfahren. Es liefert ein lineares Gleichungssystem

$$Au = r, \quad u \in \mathbb{R}^{\Omega_h}. \tag{8-7}$$

Will man (8-7) mit einem Bandalgorithmus lösen, so sind zur Aufstellung von A die Gitterpunkte durchzunummerieren.

Eine mögliche Nummerierung ist zeilenweise von links unten nach rechts oben, d.h. (x,y) kommt vor $(\widetilde{x},\widetilde{y})$, falls $y < \widetilde{y}$ oder $(x < \widetilde{x},\ y = \widetilde{y})$.

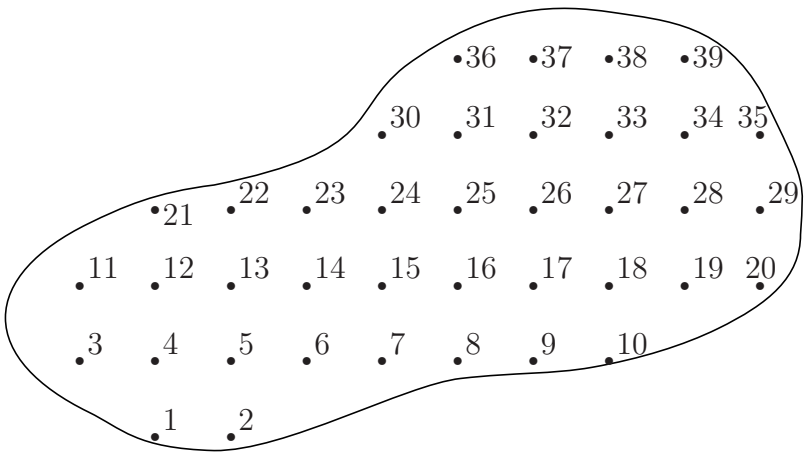


Abbildung 3: Nummerierung der Gitterpunkte

© Johannes Schropp 28. November 2008

Haben wir statt $-\Delta u(x,y) = g(x,y)$ in Ω die Aufgabe

$$-\Delta u(x,y) = g(x,y,u(x,y)), \quad (x,y) \in \Omega,$$

so ersetzen wir $g(x,y)$ in den Differenzengleichungen (8-5) und (8-6) einfach durch $g(x,y,u(x,y))$ und erhalten ein nichtlineares Gleichungssystem

$$Au = G(u), \quad u \in \mathbb{R}^{\Omega_h}$$

mit einem Diagonalfeld

$$(G(u))(x,y) = g(x,y,u(x,y)) + \text{Randterme}, \quad (x,y) \in \Omega_h.$$

b) Konsistenz und Stabilität des klassischen Differenzenverfahrens für die Dirichletsche Randwertaufgabe

Vorgelegt sei

$$\begin{aligned} -\Delta u(x,y) &= g(x,y), \quad (x,y) \in \Omega, \\ u(x,y) &= \gamma(x,y), \quad (x,y) \in \partial\Omega, \end{aligned} \tag{8-8}$$

wobei $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet ist.

Das klassische Differenzenverfahren auf $\Omega_h = \Omega \cap \mathbb{R}^2_h$ ist für alle $(x,y) \in \Omega_h$ durch die Formelzeile (8-6) gegeben. Da sich gemäß Lemma 8.12 für $(x,y) \in \Omega_h^R$ nur der Konsistenzfehler $O(h)$ ergibt, multiplizieren wir diese Gleichungen mit h durch und schreiben das dann entstehende Gleichungssystem in der Form

$$A^h u = r^h, \quad u \in \mathbb{R}^{\Omega_h} \tag{8-9}$$

bzw.

$$T^h u = A^h u - r^h = 0, \quad T^h : \mathbb{R}^{\Omega_h} \longrightarrow \mathbb{R}^{\Omega_h}.$$

Sei $h > 0$ und sei $\Omega_h = \Omega \cap \mathbb{R}^2_h$. Zu $u \in \mathbb{R}^{\Omega_h}$ sei

$$\|u\|_\infty = \max\{|u(x,y)| \,|\, (x,y) \in \Omega_h\}.$$

8.13 Definition. Sei $W \subset C^2(\overline{\Omega})$. Das numerische Modell $T^h(u) = A^h u - r^h = 0$, $u \in \mathbb{R}^{\Omega_h}$ heißt W-konsistent, falls jede Lösung $\overline{u} \in W$ der Randwertaufgabe (8-8)

$$\left\|T^h(\overline{u}_h)\right\|_\infty \longrightarrow 0 \quad \text{für} \quad h \longrightarrow 0$$

erfüllt. Das Verfahren heißt W-konsistent der Ordnung p, falls überdies

$$\left\|T^h(\overline{u}_h)\right\|_\infty = O(h^p) \quad \text{für} \quad h \longrightarrow 0$$

gilt. Dabei bezeichnet $\overline{u}_h$ die Restriktion von $\overline{u}$ auf Ω_h.

© Johannes Schropp 28. November 2008

8.14 Definition. Das Modell $T^h(u) = 0$ heißt W-konvergent, falls es zu jeder Lösung $\overline{u} \in W$ der Randwertaufgabe (8-8) so ein $h_0 > 0$ gibt, dass

$$T^h(u) = 0$$

für $0 < h \leq h_0$ eine Lösung $u^h \in \mathbb{R}^{\Omega_h}$ besitzt mit

$$\|\overline{u}_h - u^h\|_\infty \longrightarrow 0 \quad \text{für} \quad h \longrightarrow 0.$$

Gilt überdies $\|\overline{u}_h - u^h\|_\infty = O(h^p)$, so heißt das Modell W-konvergent der Ordnung p.

8.15 Definition. Das Modell $T^h(u) = 0$, $u \in \mathbb{R}^{\Omega_h}$ heißt stabil bezüglich h, falls es von h unabhängige Konstanten $h_0 > 0$ und $C > 0$ derart gibt, dass die Stabilitätsungleichung

$$\|u - v\|_\infty \leq C\|T^h(u) - T^h(v)\|_\infty$$

für alle $u, v \in \mathbb{R}^{\Omega_h}$, $0 < h \leq h_0$ erfüllt ist.

8.16 Bemerkung. Im linearen Fall ist dies gleichbedeutend zu

$$\|u\|_\infty \leq C\|Au\|_\infty, \quad \forall u \in \mathbb{R}^{\Omega_h}, \quad 0 < h \leq h_0.$$

8.17 Satz. *Sei das Modell W-konsistent oder W-konsistent der Ordnung p und stabil. Es existiere eine Lösung $u^h \in \mathbb{R}^{\Omega_h}$ von $T^h(u) = 0$. Dann ist das Modell auch W-konvergent bzw. W-konvergent der Ordnung p.*

Beweis: Wir setzen $u = u^h$ und $v = \overline{u}_h$ in die Stabilitätsungleichung ein und finden

$$\begin{aligned} \|u^h - \overline{u}_h\|_\infty &\leq C\|\underbrace{T^h(u^h)}_{=0} - T(\overline{u}_h)\|_\infty \\ &= C \cdot \|T^h(\overline{u}_h)\|_\infty \longrightarrow 0 \quad \text{für} \quad h \longrightarrow 0 \end{aligned}$$

bzw.

$$\|u^h - \overline{u}_h\|_\infty = C \cdot \|T^h(\overline{u}_h)\|_\infty = O(h^p) \quad \text{für} \quad h \longrightarrow 0.$$

□

8.18 Satz. *An jeder Lösung $\overline{u} \in C^4(\overline{\Omega})$ von (8-8) liegt Konsistenz des klassischen Differenzenverfahrens (8-9) der Ordnung 2 vor, d.h.*

$$\|A^h\overline{u}_h - r^h\|_\infty = \sup\{|(A^h\overline{u}_h)(x,y) - r^h(x,y)| \,|\, (x,y) \in \Omega_h\} = O(h^2).$$

(d.h. C^4-Konsistenz der Ordnung 2)

© Johannes Schropp 28. November 2008

Beweis: Für $(x,y) \in \mathring{\Omega}_h$ gilt mit $T^h(u) = A^h u - r^h$ nach Formel (8-2)

$$\begin{aligned}
|(A^h\overline{u}_h - r^h)(x,y)| &= \big|h^{-2}\big(-\overline{u}(x-h,y) - \overline{u}(x+h,y) - \overline{u}(x,y-h) \\
&\quad -\overline{u}(x,y+h) + 4\overline{u}(x,y)\big) - \underbrace{g(x,y)}_{=-\Delta\overline{u}(x,y)} \big| \\
&\le \big|h^{-2}(-\overline{u}(x-h,y) + 2\overline{u}(x,y) - \overline{u}(x+h,y)) + \overline{u}_{xx}(x,y)\big| \\
&+ \big|h^{-2}(-\overline{u}(x,y-h) + 2\overline{u}(x,y) - \overline{u}(x,y+h)) + \overline{u}_{yy}(x,y)\big| \\
&\le C\cdot h^2\big(\max\{|\overline{u}_{xxxx}(x+\xi h,y)|\,|\,|\xi|\le 1\} \\
&\qquad + \max\{|\overline{u}_{yyyy}(x,y+\eta h)|\,|\,|\eta|\le 1\}\big) \le \widetilde{C}h^2
\end{aligned}$$

mit einer von h und (x,y) unabhängigen Konstanten $\widetilde{C}$.

Für $(x,y) \in \Omega_h^R$ setzen wir

$$\begin{aligned} v(\xi) &= \overline{u}(x+\xi,y), \\ w(\eta) &= \overline{u}(x,y+\eta) \end{aligned}$$

und finden damit die Abschätzung

$$\begin{aligned}
|(A^h\overline{u}_h - r^h)(x,y)| &= h\Big|h^{-2}\Big\{\frac{-2}{\mu_1(\mu_1+\mu_3)}\overline{u}(N_1) - \frac{2}{\mu_3(\mu_1+\mu_3)}\overline{u}(N_3) \\
&\quad -\frac{2}{\mu_2(\mu_2+\mu_4)}\overline{u}(N_2) - \frac{2}{\mu_4(\mu_2+\mu_4)}\overline{u}(N_4) \\
&\quad +2\left(\frac{1}{\mu_1\mu_3}+\frac{1}{\mu_2\mu_4}\right)\overline{u}(x,y)\Big\} - \underbrace{g(x,y)}_{=-\Delta\overline{u}(x,y)}\Big|.
\end{aligned}$$

Unter Berücksichtigung von

$$\begin{aligned}
\overline{u}(N_1) &= v(\mu_1 h), \\
\overline{u}(N_3) &= v(-\mu_3 h), \quad v''(0) = \overline{u}_{xx}(x,y), \\
\overline{u}(N_2) &= w(\mu_2 h), \\
\overline{u}(N_4) &= w(-\mu_4 h), \quad w''(0) = \overline{u}_{yy}(x,y)
\end{aligned}$$

erhält man

$$\begin{aligned}
|(A^h\overline{u}_h - r^h)(x,y)| &\le h\Big|h^{-2}\Big\{\frac{-2}{\mu_1(\mu_1+\mu_3)}v(\mu_1 h) - \frac{2}{\mu_3(\mu_1+\mu_3)}v(-\mu_3 h) \\
&\quad +\frac{2}{\mu_1\mu_3}v(0)\Big\} + v''(0)\Big| \\
&\quad +h\Big|h^{-2}\Big\{\frac{-2}{\mu_2(\mu_2+\mu_4)}w(\mu_2 h) - \frac{2}{\mu_4(\mu_2+\mu_4)}w(-\mu_4 h)
\end{aligned}$$

© Johannes Schropp 28. November 2008

$$
\begin{aligned}
&\quad + \frac{2}{\mu_2\mu_4} w(0) \Big\} + w''(0) \Big| \\
&\leq \frac{2h^2}{3} \big(\max\{ |\overline{u}_{xxx}(x+\xi h, y)| \, -\mu_3 \leq \xi \leq \mu_1 \} \\
&\qquad \max\{ |\overline{u}_{yyy}(x, y+\eta h)| \, -\mu_4 \leq \eta \leq \mu_2 \} \big) \leq Ch^2.
\end{aligned}
$$

□

8.19 Satz. *Die Matrix A^h des klassischen Differenzeverfahrens ist eine M-Matrix. Wählt man ein Rechteck $(a,b)\times(c,d) \supset \overline{\Omega}$, so gilt für den positiven Vektor $e_h = e|_{\Omega_h} \in \mathbb{R}^{\Omega_h}$ mit*

$$e(x,y) = (x-a)(b-x) + (y-c)(d-y), \quad (x,y) \in \mathbb{R}^2$$

die Ungleichung

$$A^h e_h \geq \rho\, e_h$$

für $\rho := \frac{16}{(b-a)^2+(d-c)^2} > 0$, falls $h > 0$ hinreichend klein gewählt ist.

A^h erfüllt die Stabilitätsungleichung

$$\|u\|_\infty \leq \frac{1}{\rho} \|A^h u\|_\infty \quad \forall u \in \mathbb{R}^{\Omega_h}$$

für hinreichend kleine $h > 0$.

Beweis: Wir wissen, dass A^h eine L_0-Matrix ist.

Für $(x,y) \in \overset{\circ}{\Omega}_h$ ist die in (8-6) verwandte Differenzenformel exakt für Polynome bis zum Grad 3, d.h. für $e(x,y) = (x-a)(b-x) + (y-c)(d-y)$ gilt

$$
\begin{aligned}
(A^h e_h)(x,y) &= h^{-2}(-e(x-h,y) - e(x+h,y) - e(x,y-h) - e(x,y+h) + 4e(x,y)) \\
&= -(\Delta e)(x,y) \;=\; -(-2-2) \;=\; 4
\end{aligned}
$$

für alle $(x,y) \in \overset{\circ}{\Omega}_h$.

Für $(x,y) \in \Omega_h^R$ werden durch die Differenzenformel (8-6) noch Polynome bis zum Grad 2 exakt differenziert (vgl. Lemma 8.12).

Unter Beachtung der Formenzeile (8-6) sowie der Tatsache $u(N_k) = \gamma(N_k)$, falls $N_k \in \partial\Omega$, folgt also

$$(A^h e_h)(x,y) = h\Bigg[\underbrace{-\Delta e(x,y)}_{=-4} + h^{-2} \sum_{\substack{k=1 \\ N_k(x,y,h)\in\partial\Omega}}^{4} \frac{2}{\mu_k(\mu_k+\mu_{\widetilde{k}})} e(N_k) \Bigg]$$

mit $\widetilde{k} = (k+2) \mod 4$.

© Johannes Schropp 28. November 2008

Nach Voraussetzung $(x,y) \in \Omega_h^R$ ist die letztere Summe nicht leer und $\mu_k, \mu_{\widetilde{k}} \in (0,1]$, d.h. $0 < \mu_k(\mu_k + \mu_{\widetilde{k}}) \le 2$, $\frac{2}{\mu_k(\mu_k+\mu_{\widetilde{k}})} \ge 1$ und somit

$$(A^h e_h)(x,y) \ge h(4 + h^{-2} \underbrace{\min\{e(x,y) \,|\, (x,y) \in \overline{\Omega}\}}_{=:e_0}).$$

Da e stetig ist und $e(x,y) > 0$ für $(x,y) \in \overline{\Omega}$ gilt, folgt $e_0 > 0$.

Wir finden also

$$(A^h e_h)(x,y) \ge h^{-1} e_0$$

für $(x,y) \in \Omega_h^R$.

Insgesamt erhalten wir

$$A^h e_h \ge \min\left(4, \frac{e_0}{h}\right) \mathbb{I} > 0, \quad e_h > 0,$$

d.h. A^h ist eine M-Matrix.

Für $h \le \frac{e_0}{4}$ folgt

$$A^h e_h \ge 4\mathbb{I} \ge 4\frac{e_h}{\|e_h\|_\infty} \ge \frac{4}{\left(\frac{b-a}{2}\right)^2 + \left(\frac{d-c}{2}\right)^2} e_h =: \rho\, e_h$$

Da A^h eine M-Matrix ist, ergibt sich

$$\|(A^h)^{-1}\|_\infty = \|(A^h)^{-1}\mathbb{I}\|_\infty \le \frac{1}{4}\|e_h\|_\infty \le \frac{1}{\rho}.$$

Somit folgt die Stabilitätsungleichung

$$\|u\|_\infty = \|(A^h)^{-1}A^h u\|_\infty \le \|(A^h)^{-1}\|_\infty \|A^h u\|_\infty \le \frac{1}{\rho}\|A^h u\|_\infty. \tag{8-10}$$

□

Aus Satz 8.18 und Satz 8.19 erhalten wir

8.20 Korollar. *Das klassische Differenzenverfahren für die Dirichletsche Randwertaufgabe (8-8) in einem beschränkten Gebiet Ω ist $C^4(\overline{\Omega})$-konvergent der Ordnung 2 bzgl. $\|\cdot\|_\infty$.*

Beweis: Ist u^h die Lösung des klassischen Differenzenverfahrens, so gilt

$$\|\overline{u}_h - u^h\|_\infty \le \frac{1}{\rho}\|A^h\overline{u}_h - \underbrace{A^h u^h}_{=r^h}\|_\infty = \frac{1}{\rho}\|A^h\overline{u}_h - r^h\|_\infty = O(h^2).$$

□

© Johannes Schropp 28. November 2008

8.21 Bemerkung. Dass die Lösung $\overline{u}$ in $C^4(\overline{\Omega},\mathbb{R})$ liegt, ist für Gebiete mit „Ecken“ im Allgemeinen nicht erfüllt, z.B. im Fall $-\Delta u = 1$ in $\Omega = (0,1)^2$, $u = 0$ auf $\partial\Omega$. Wäre $\overline{u} \in C^4(\overline{\Omega},\mathbb{R})$ Lösung hiervon, so würde gelten

$$-\Delta u(0,0) = 1 = -u_{xx}(0,0) - u_{yy}(0,0) = 0,$$

da $u = 0$ auf $\partial\Omega$.

Wir betrachten jetzt kurz das klassische Differenzenverfahren für die nichtlineare Aufgabe

$$\begin{aligned} -\Delta u(x,y) &= g(x,y,u(x,y)) \text{ in } \Omega, \\ u(x,y) &= \gamma \text{ auf } \partial\Omega. \end{aligned} \tag{8-11}$$

Es lautet

$$\begin{aligned} g(x,y,u(x,y)) = h^{-2}\Bigg\{ & \frac{-2}{\mu_1(\mu_1+\mu_3)}u(N_1) - \frac{2}{\mu_3(\mu_1+\mu_3)}u(N_3) \\ & -\frac{2}{\mu_2(\mu_2+\mu_4)}u(N_2) - \frac{2}{\mu_4(\mu_2+\mu_4)}u(N_4) \\ & +2\left(\frac{1}{\mu_1\mu_3}+\frac{1}{\mu_2\mu_4}\right)u(x,y)\Bigg\}. \end{aligned}$$

Dabei beachten wir $u(N_k) = \gamma(N_k)$ für $(x,y) \in \Omega_h = \Omega \cap \mathbb{R}^2_h$, falls $N_k \in \partial\Omega$.

Dies liefert ein nichtlineares System der Form

$$T^h u = A^h u - G^h(u) = 0 \tag{8-12}$$

mit einem Diagonalfeld

$$G^h(u)(x,y) = g(x,y,u(x,y)) + \text{Randterme}.$$

Das Gleichungssystem (8-12) lässt sich nun z.B. mit dem Newtonverfahren lösen. Präziser: Man wähle ein u^0 und löse für $i = 0,1,2,\ldots$ das lineare Gleichungssystem

$$DT^h(u^i)d^i = -T^h(u^i)$$

und setze $u^{i+1} = u^i + d^i$, $i = 0,1,2,\ldots$.

Für den Fall $T^h(u) = A^h u - G^h(u)$ folgt

$$DT^h(u) = A^h - DG^h(u)$$

und mit $g_v(x,y,v) \le \mu < \rho = \frac{16}{(b-a)^2+(d-c)^2}$ folgt

$$DT^h(u)e_h = A^h e_h - \underbrace{DG^h(u)}_{\le \mu I^h} e_h \ge \rho e_h - \mu e_h = (\rho-\mu)e_h > 0,$$

© Johannes Schropp 28. November 2008

falls $h > 0$ hinreichend klein.

Da $DT^h(u)$ eine L_0-Matrix ist, ist e_h ein majorisierendes Element für $DT^h(u)$, d.h. $DT^h(u)$ ist M-Matrix und das Newtonverfahren ist in dieser Situation durchführbar.

Für ein beliebiges Gleichungssystem der Form $T(u) = 0$ erinnern wir an den lokalen Konvergenzsatz des Newtonverfahrens.

8.22 Satz (lokaler Konvergenzsatz, ohne Beweis). *Sei $U \subset \mathbb{R}^N$ offen und sei $T : U \longrightarrow \mathbb{R}^N$ zweimal stetig differenzierbar. Ferner existiere eine Lösung $\overline{u} \in U$ von $T(u) = 0$, und $DT(\overline{u})$ sei invertierbar. Dann gibt es eine Kugel*

$$K_\rho(\overline{u}) := \{u \in \mathbb{R}^N \mid \|u - \overline{u}\| < \rho\} \subset U, \quad \rho > 0$$

so, dass $T(u) = 0$ keine weitere Lösung in $K_\rho(\overline{u})$ besitzt. Für jeden Startwert $u^0 \in K_\rho(\overline{u})$ existiert die Newtonfolge

$$u^{n+1} = u^n - DT(u^n)^{-1}T(u^n),$$

liegt in $K_\rho(\overline{u})$ und konvergiert gegen $\overline{u}$.

Überdies gibt es ein $C > 0$ mit

$$\|u^{n+1} - \overline{u}\|_\infty \leq C\|u^n - \overline{u}\|_\infty^2$$

für alle $n \geq 0$, $u^0 \in K_\rho(\overline{u})$. Letztere Ungleichung bedeutet lokal quadratische Konvergenz.

8.23 Bemerkung. Eine Nullstelle $\overline{u} \in U$ von $T(u) = 0$ mit invertierbarem $DT(\overline{u})$ heißt regulär.

Typische Abbruchkriterien für Newtonverfahren sind

i) $\|T(u^i)\|_\infty \leq$ TOL, z.B. TOL $= 10^{-10}$,
oder/und
ii) $\|u^{i+1} - u^i\|_\infty \leq$ EPS, z.B. EPS $= 10^{-5}$.

8.24 Satz (ohne Beweis). *Vorgelegt sei (8-11), und es gelte $g \in C^1(\overline{\Omega} \times \mathbb{R}, \mathbb{R})$ sowie*

$$\frac{\partial g}{\partial u}(x, y, u) \leq \mu < \rho := \frac{16}{(b-a)^2 + (d-c)^2},$$

$\overline{\Omega} \subset (a, b)\times(c, d)$. Dann ist das klassische Differenzenverfahen (8-12) für die Aufgabe (8-11) $C^4(\overline{\Omega})$-konvergent der Ordnung 2.

Vor- und Nachteile der Differenzenverfahren lauten also:

- ⊕ Differenzenverfahren sind einfach aufzustellen, da lediglich Ableitungen durch Differenzenquotienten ersetzt werden.

© Johannes Schropp 28. November 2008

⊕ Die bei äquidistantem Gitter entstehenden Gleichungssysteme haben sehr viel Struktur und lassen sich effizient lösen.

⊖ Die Differenzengleichungen werden kompliziert bei nicht äquidistanten Stützstellen und komplexen Gebieten.

⊖ Wir brauchen Lösungen in $C^4(\overline{\Omega})$ um Konvergenz der Ordnung 2 zu sichern.

© Johannes Schropp 28. November 2008

9. Finite Elemente Methoden für elliptische Differentialgleichungen

a) Das Ritz'sche Verfahren

Vorgelegt sei

$$\begin{aligned} -\Delta u &= g \text{ in } \Omega, \quad \Omega \subset \mathbb{R}^2 \text{ beschränktes Gebiet,} \\ u &= \gamma \text{ auf } \partial\Omega, \quad \partial\Omega \text{ stückweise stetig differenzierbar.} \end{aligned} \tag{9-1}$$

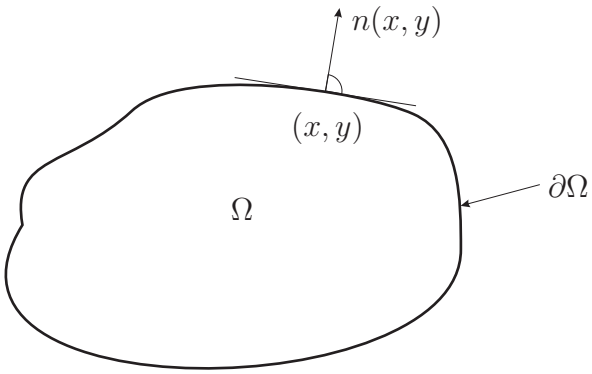


Abbildung 4: Beschränktes Gebiet mit stückweise glattem Rand

Also existiert die äußere Normale $n = n(x,y)$ für $(x,y) \in \partial\Omega$ mit Ausnahme endlich vieler Punkte. Ferner sei für jedes $\psi \in C(\partial\Omega, \mathbb{R})$ der Unterraum

$$D_\psi = \{u \in C(\overline{\Omega}, \mathbb{R}) \,|\, u = \psi \text{ auf } \partial\Omega\}$$

definiert.

Das allgemeine Ritz'sche Verfahren basiert auf einer Umformulierung der Randwertaufgabe (9-1) in ein Variationsproblem.

9.1 Satz. *Für eine Funktion $\overline{u} \in C^2(\overline{\Omega}) \cap D_\gamma$ sind paarweise äquivalent*

i) $\overline{u}$ löst die Randwertaufgabe (9-1);

ii) $\overline{u}$ ist stationärer Punkt des Funktionals

$$\begin{aligned} I : C^1(\overline{\Omega}) \cap D_\gamma &\longrightarrow \mathbb{R}, \\ u &\longmapsto \int_\Omega \frac{1}{2}|\nabla u|^2 - g \cdot u \, d(x,y); \end{aligned}$$

© Johannes Schropp 28. November 2008

iii) $u = \overline{u} \in C^1(\Omega) \cap D_\gamma$ *erfüllt*

$$\int_\Omega \nabla u \cdot \nabla v - g \cdot v \, d(x,y) = 0 \quad \forall v \in C^1(\overline{\Omega}) \cap D_0.$$

9.2 Bemerkung. Dass $\overline{u}$ stationärer Punkt des Funktionals I ist, bedeutet

$$\frac{\partial}{\partial \varepsilon} I(\overline{u} + \varepsilon v)\big|_{\varepsilon=0} = 0 \quad \forall v \in C^1(\overline{\Omega}) \cap D_0.$$

Man beachte:

$$\overline{u} \in D_\gamma, \quad v \in D_0, \quad \varepsilon \in \mathbb{R} \Longrightarrow \overline{u} + \varepsilon v \in D_\gamma.$$

9.3 Bemerkung. Man kann in *ii)* auch schreiben: Bei $u = \overline{u}$ nimmt das Funktional $I : C^1(\overline{\Omega}) \cap D_\gamma \longrightarrow \mathbb{R}$ sein Minimum an. Man bezeichnet *ii)* als Variationsproblem und *iii)* als Variationsgleichungen der Randwertaufgabe (9-1), da dort nur C^1-Lösungen gesucht werden.

9.4 Bemerkung. In *ii)* und *iii)* ist entscheidend, dass $u \in D_\gamma$ und $v \in D_0$ gilt.

Offensichtlich gilt

$$\nabla \cdot (v \nabla u) = \nabla v \cdot \nabla u + v \cdot \Delta u$$

für $v \in C^1(\Omega)$, $u \in C^2(\Omega)$.

Wir integrieren diese Beziehung und finden

$$\int_\Omega \nabla(v \nabla u) d(x,y) = \int_\Omega \nabla v \cdot \nabla u + v \Delta u d(x,y),$$

was unter Verwendung des Gauß'schen Satzes

$$\int_{\partial\Omega} (v \nabla u) \cdot n \, dS = \int_\Omega \nabla v \cdot \nabla u + v \Delta u d(x,y) \tag{9-2}$$

liefert, falls zusätzlich $u \in C^1(\overline{\Omega})$ gilt. Letztere Formelzeile heißt die erste Greensche Formel.

Beweis:

ii) $\Leftrightarrow$ *iii)* Für $v \in C^1(\overline{\Omega}) \cap D_0$ gilt

$$\begin{aligned} \frac{\partial}{\partial \varepsilon} I(\overline{u} + \varepsilon v) &= \frac{\partial}{\partial \varepsilon} \left\{ \int_\Omega \frac{1}{2} (\nabla \overline{u} + \varepsilon \nabla v) \cdot (\nabla \overline{u} + \varepsilon \nabla v) - g(\overline{u} + \varepsilon v) \, d(x,y) \right\} \\ &= \frac{\partial}{\partial \varepsilon} \left\{ \int_\Omega \frac{1}{2} \nabla \overline{u} \cdot \nabla \overline{u} + \varepsilon \cdot \nabla v \cdot \nabla \overline{u} \right. \end{aligned}$$

© Johannes Schropp 28. November 2008

$$
\begin{aligned}
&+ \quad \frac{1}{2}\varepsilon^2 \nabla v \cdot \nabla v - g \cdot \overline{u} - \varepsilon g \cdot v \, d(x,y) \Big\} \\
&= \int_\Omega \nabla v \cdot \nabla \overline{u} + \varepsilon \nabla v \cdot \nabla v - g \cdot v \, d(x,y).
\end{aligned}
$$

Somit bekommen wir

$$
\frac{\partial}{\partial \varepsilon} I(\overline{u} + \varepsilon v)\big|_{\varepsilon=0} = \int_\Omega \nabla v \cdot \nabla \overline{u} - g \cdot v \, d(x,y),
$$

d.h. *iii)* ist genau dann erfüllt, wenn $\overline{u}$ stationärer Punkt des Funtionals I ist.

i) $\Rightarrow$ *iii)* Wir multiplizieren die Randwertaufgabe (9-1) mit $v \in C^1(\overline{\Omega}) \cap D_0$ und integrieren über Ω:

$$
\begin{aligned}
\int_\Omega g \cdot v d(x,y) \quad &= \quad - \int_\Omega \Delta \overline{u} \cdot v d(x,y) \\
&\overset{(9-2)}{=} \int_\Omega \nabla \overline{u} \cdot \nabla v d(x,y) - \int_{\partial\Omega} (v \nabla \overline{u}) \cdot n \, dS.
\end{aligned}
$$

Wegen $v \in C^1(\overline{\Omega}) \cap D_0$ folgt $\int_{\partial\Omega} (v \cdot \nabla \overline{u}) \cdot n \, dS = 0$, und wir erhalten

$$
\int_\Omega g v \, d(x,y) = \int_\Omega \nabla \overline{u} \cdot \nabla v \, d(x,y)
$$

für alle $v \in C^1(\overline{\Omega}) \cap D_0$.

iii) $\Rightarrow$ *i)* Es gilt umgekehrt

$$
0 = \int_\Omega \nabla \overline{u} \cdot \nabla v - g \cdot v \, d(x,y) \overset{(9-2)}{=} \int_\Omega (-\Delta \overline{u} - g) \cdot v \, d(x,y)
$$

für alle $v \in C^1(\overline{\Omega}) \cap D_0$, woraus sich mit dem nachstehenden Lemma 9.5

$$
-\Delta \overline{u} - g = 0
$$

ergibt.

Somit haben wir den Satz bewiesen. □

9.5 Lemma (Fundamentallemma der Variationsrechnung (ohne Beweis)). *Es sei $z \in C(\overline{\Omega}, \mathbb{R})$, und es gelte*

$$
\int_\Omega z(x,y) v(x,y) \, d(x,y) = 0 \quad \forall v \in D_0.
$$

Dann ist $z \equiv 0$ in $\overline{\Omega}$.

© Johannes Schropp 28. November 2008

Das allgemeine Ritz'sche Verfahren zur Lösung der Randwertaufgabe (9-1) lautet jetzt: Wähle eine Funktion $u_0 \in C^1(\overline{\Omega}) \cap D_\gamma$ und Ansatzfunktionen $u_i \in C^1(\overline{\Omega}) \cap D_0$, $i = 1, \ldots, m$ und mache das Funktional

$$I : C^1(\overline{\Omega}) \cap D_\gamma \longrightarrow \mathbb{R}$$

stationär bezüglich $u \in u_0 + V$, $V = \text{span}\{u_1, \ldots, u_m\}$, d.h. finde ein $\widetilde{u} \in u_0 + V$ mit

$$\frac{\partial}{\partial \varepsilon} I(\widetilde{u} + \varepsilon v)\big|_{\varepsilon=0} = 0 \quad \forall v \in V.$$

Dies ist nach dem Beweis $ii) \Leftrightarrow iii)$ in Satz 9.1 äquivalent zu $\widetilde{u} \in u_0 + V$ erfüllt

$$\int_\Omega \nabla \widetilde{u} \cdot \nabla v - g \cdot v \, d(x, y) = 0 \quad \forall v \in V. \tag{9-3}$$

9.6 Definition. (9-3) heißen Galerkinsche Gleichungen zur Randwertaufgabe (9-1).

Wir suchen also $\widetilde{u}$ in der Form

$$\widetilde{u} = u_0 + \sum_{j=1}^{m} c_j u_j, \quad c_j \in \mathbb{R}, \quad j = 1, \ldots, m.$$

Es genügt nun, (9-3) auf der Basis $\{u_1, \ldots, u_m\}$ von V zu fordern, da (9-3) linear in V ist. Damit erhalten wir

$$\begin{aligned} 0 &= \int_\Omega \nabla \left(u_0 + \sum_{j=1}^{m} c_j u_j \right) \cdot \nabla u_i - g \cdot u_i \, d(x, y) \\ &= \sum_{j=1}^{m} c_j \int_\Omega \nabla u_j \cdot \nabla u_i \, d(x, y) + \int_\Omega (\nabla u_0 \cdot \nabla u_i - g u_i) \, d(x, y), \quad i = 1, \ldots, m. \end{aligned}$$

Wir bekommen also für $c = (c_1, \ldots, c_m) \in \mathbb{R}^m$ ein lineares Gleichungssystem

$$Ac = r, \quad A \in R^{m,m}, \quad r \in \mathbb{R}^m. \tag{9-4}$$

Hier ist

$$A_{ij} = \int_\Omega \nabla u_j \cdot \nabla u_i \, d(x, y), \quad 1 \le i, j \le m,$$

d.h. A ist eine symmetrische Matrix, und $r = (r_1, \ldots, r_m)$ ist definiert durch

$$r_i = -\int_\Omega (\nabla u_0 \cdot \nabla u_i - g u_i) \, d(x, y), \quad 1 \le i \le m.$$

Beim klassischen Ritz'schen Verfahren wählt man Ansatzfunktionen $u_i \in C^1(\overline{\Omega}) \cap D_0$, $i = 1, \ldots, m$, die im Allgemeinen auf dem gesamten Gebiet nicht verschwinden. A wird dann eine vollbesetzte Matrix.

© Johannes Schropp 28. November 2008

b) Finite Elemente Methoden

Die Finite Elemente Verfahren weichen von dem Ritz-Verfahren in zwei wesentlichen Punkten ab:

1.) geringere Glattheit der Ansatzfunktionen,

2.) Ansatzfunktionen mit lokalem Träger.

Zu 1.) Man unterteilt das Gebiet Ω in sogenannte finite Elemente durch eine Triangulierung.

Sei es der Einfachheit halber angenommen, dass Ω ein polygonales Gebiet ist, d.h. der Rand $\partial\Omega$ bestehe aus endlich vielen Geradenstücken.

Im allgemeinen Fall wird ein beliebiges beschränktes Gebiet durch Ω_{T_h} approximiert.

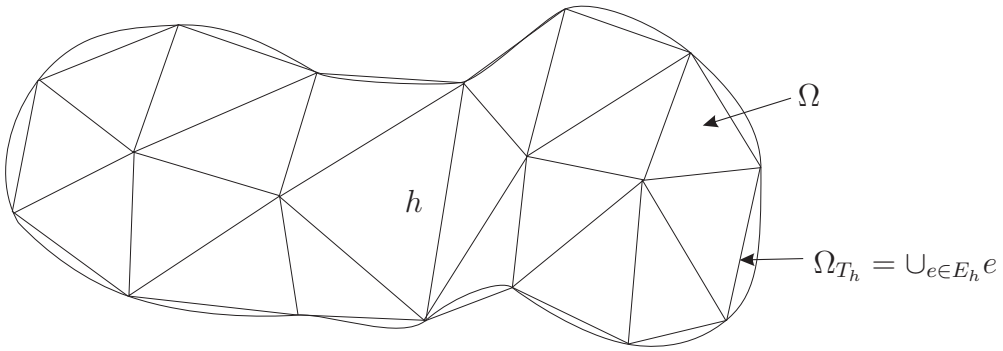


Abbildung 5: Triangulierung eines polygonalen Gebietes

E_h ist eine Menge von Dreiecken e. Je zwei Dreiecke $e, e' \in E_h$ haben entweder eine Seite, einen Eckpunkt oder nichts gemeinsam. h ist die dabei maximal auftretende Kantenlänge.

Man sucht eine Lösung u in $u_0 + V_{T_h}$, wobei

$$V_{T_h} = \{u \in C(\Omega_{T_h}) \,|\, u \text{ ist ein Polynom mit } \deg(u) \leq r \text{ in } x \text{ und } y \text{ auf jedem Dreieck } e \in E_h \text{ und } u = 0 \text{ auf } \partial\Omega_{T_h}\}.$$

Für $r = 1$ handelt es sich hierbei um „lineare finite Elemente“. Für $r = 2$ sprechen wir von „quadradischen finiten Elementen“, u.s.w.

Wir beschränken uns hier auf den Fall $r = 1$, d.h. u ist linear in x und y auf jedem Dreieck $e \in E_h$.

© Johannes Schropp 28. November 2008

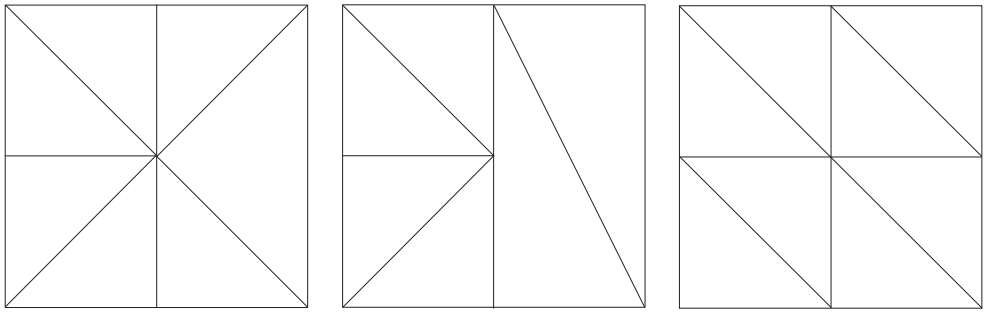


Abbildung 6: Beispiele für Triangulierungen

9.7 Bemerkung. Die Funktionen $u \in V_{T_h}$ gehören nicht zu $C^1(\overline{\Omega})$! Wir werden später aber sehen, dass die Variationsgleichung

$$\int_\Omega \nabla u \cdot \nabla v - g \cdot v \, d(x,y) = 0$$

für v aus einem geeigneten Sobolev-Raum gilt, welcher die Funktionen aus V_{T_h} umfasst.

Zu 2.) Benutze eine Basis von Ansatzfunktionen in V_{T_h}, welche jeweils im größten Teil des Gebietes Ω verschwinden. Dies impliziert, dass die entstehende Matrix dann dünn besetzt ist.

Vorgehensweise (lineare finite Elemente):

Seien P_i, $i = 1, \ldots, M$ die durchnummerierten Knoten der Triangulierung Ω_{T_h}, wobei die Knoten, die nicht auf $\partial\Omega$ liegen, gerade die P_i, $i = 1, \ldots, M$ mit $m < M$ seien.

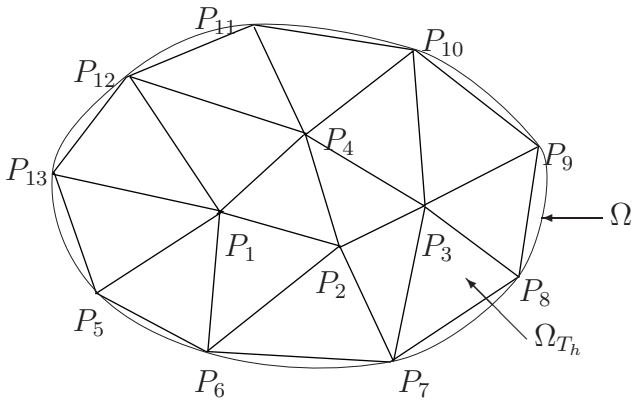


Abbildung 7: Nummerierung der Knoten, $M = 13$, $m = 4$

Definiere die Funktionen u_i, $i = 1, \ldots, M$ auf Ω_{T_h} durch

$$u_i(P_j) = \delta_{ij}, \quad 1 \le i, j \le M, \tag{9-5}$$

© Johannes Schropp 28. November 2008

u ist linear in x und y auf jedem $e \in E_h, u \in C(\Omega_{T_h})$.

Die u_i, $i = 1, \ldots, M$ heißen auch Formfunktionen. Es lässt sich zeigen, dass das Interpolationsproblem (9-5) eindeutig lösbar ist.

Struktur der Formfunktionen

Die Werte von u_i auf einer Dreiecksseite sind durch die Endwerte in den Knoten bestimmt. u ist daher auch stetig in Ω_{T_h} und verschwindet auf allen Dreiecken, bei denen der Knoten P_i nicht vorkommt.

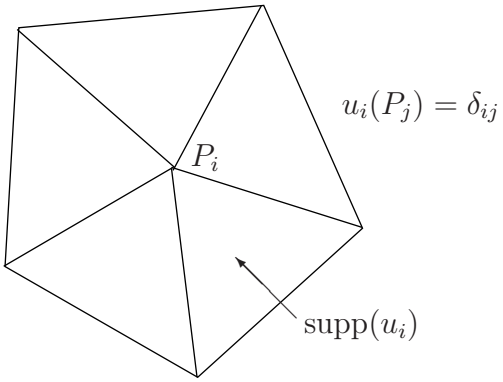


Abbildung 8: Träger $\mathrm{supp}(u_i)$ einer Formfunktion u_i

Wir setzen ferner

$$u_0 := \sum_{j=m+1}^{M} \gamma(P_j) u_j,$$

wobei $P_{m+1}, \ldots, P_M \in \partial\Omega$ zu beachten ist, und erhalten

$$u_0(P_i) = \sum_{j=m+1}^{M} \gamma(P_j) \underbrace{u_j(P_i)}_{=\delta_{ij}} = \gamma(P_i), \quad i = m+1, \ldots, M.$$

Der Ansatz erhält dann die Form

$$\widetilde{u} = u_0 + \sum_{j=1}^{m} c_j u_j = \sum_{j=1}^{M} c_j u_j \tag{9-6}$$

mit $c_j = \gamma(P_j)$ für $j = m+1, \ldots, M$.

Der Koeffizient c_i gibt dann gerade den Wert von $\widetilde{u}$ im Knoten P_i, $i = 1, \ldots, M$ an, denn

$$\widetilde{u}(P_i) = \sum_{j=1}^{M} c_j \underbrace{u_j(P_i)}_{=\delta_{ij}} = c_i, \quad i = 1, \ldots, M.$$

© Johannes Schropp 28. November 2008

Gesucht sind $c_1, \ldots, c_m$. Die Koeffizienten $c_{m+1}, \ldots, c_M$ sind schon bekannt.

Detaillierte Aufstellung des linearen Gleichungssystems (9-4)

9.8 Definition. Die Aufstellung des linearen Gleichungssystems (9-4) heißt Assemblierung.

Unser Gleichungssystem hat die Form $Ac = r$ mit

$$\begin{aligned} A_{ij} &= \int_\Omega \nabla u_j \cdot \nabla u_i \, d(x,y), \quad 1 \le i,j \le m, \\ r_i &= -\int_\Omega (\nabla u_0 \cdot \nabla u_i - g u_i) \, d(x,y), \quad 1 \le i \le m. \end{aligned}$$

Wir ersetzen Ω durch Ω_{T_h}, g durch

$$g_{T_h} = \sum_{j=1}^{M} g(P_j) u_j$$

und finden mit $u_0 = \sum_{j=m+1}^{M} \gamma(P_j) u_j$ dann

$$\begin{aligned} A_{ij} &= \int_{\Omega_{T_h}} \nabla u_i \cdot \nabla u_j \, d(x,y), \quad 1 \le i,j \le m, \\ r_i &= \int_{\Omega_{T_h}} g_{T_h} u_i - \nabla u_0 \cdot \nabla u_i \, d(x,y) \\ &= \sum_{j=1}^{M} g(P_j) \int_{\Omega_{T_h}} u_i u_j \, d(x,y) \\ &- \sum_{j=m+1}^{M} \gamma(P_j) \int_{\Omega_{T_h}} \nabla u_j \cdot \nabla u_i \, d(x,y), \quad i = 1, \ldots, m. \end{aligned}$$

Da das Integral über Ω_{T_h} als die Summe der Integrale über jeweiligen Dreiecken $e \in E_h$ dargestellt werden kann, d.h.

$$\int_{\Omega_{T_h}} f(x,y) \, d(x,y) = \sum_{e \in E_h} \int_e f(x,y) \, d(x,y),$$

bleiben damit die folgenden Integrale zu bestimmen

$$\int_e \nabla u_i \cdot \nabla u_j \, d(x,y), \quad \int_e u_i u_j \, d(x,y), \quad 1 \le i,j \le M. \tag{9-7}$$

Die Integrale in (9-7) lassen sich wie folgt berechnen.

© Johannes Schropp 28. November 2008

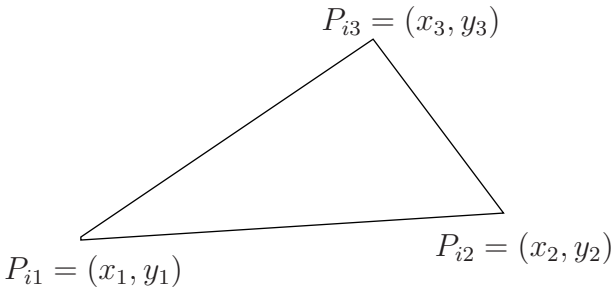


Abbildung 9: Dreieck e mit den Ecken P_{i1}, P_{i2}, P_{i3}

Sei $e \in E_h$ beliebig. e habe die Endpunkte P_{i1}, P_{i2}, P_{i3}.

Es sind dann die Integrale $S_{jk} = \int_e \nabla u_{ij} \cdot \nabla u_{ik}\, d(x,y)$, $k,j = 1,2,3$ von Null verschieden.

Die symmetrische Matrix $S^e = (S_{jk})_{1\le j,k\le 3} \in \mathbb{R}^{3,3}$ ist beim Aufbau der Gesamtmatrix A auf die Untermatrix

$$\begin{pmatrix} A_{i1,i1} & A_{i1,i2} & A_{i1,i3} \\ A_{i2,i1} & A_{i2,i2} & A_{i2,i3} \\ A_{i3,i1} & A_{i3,i2} & A_{i3,i3} \end{pmatrix}$$

aufzuaddieren. Man findet

$$S^e = \frac{1}{2 \cdot F_e} C_e C_e^T$$

mit

$$C_e = \begin{pmatrix} y_2 - y_3 & x_3 - x_2 \\ y_3 - y_1 & x_1 - x_3 \\ y_1 - y_2 & x_2 - x_1 \end{pmatrix} \in \mathbb{R}^{3,2}$$

und $F_e = |e| = (x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)$. Dabei ist F_e die Fläche des Dreiecks e mit den Eckpunkten P_{i1}, P_{i2}, P_{i3} ist.

Das zweite Integral in (9-7) ergibt sich als die Matrix $M^e \in \mathbb{R}^{3,3}$ mit

$$(M^e)_{jk} = \int_e u_{jk} \cdot u_{ik}\, d(x,y), \quad 1 \le j,k \le 3,$$

für welche

$$M^e = \frac{F_e}{24} \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix} \in \mathbb{R}^{3,3}$$

gilt.

c) Theoretische Grundlagen zur Finite Elemente Methode

© Johannes Schropp 28. November 2008

Wir möchten die im Abschnitt b) hergeleitete Variationsgleichung

$$\int_\Omega \nabla u \cdot \nabla v - g \cdot v \, d(x,y) = 0 \quad \forall v \in C^1(\overline{\Omega}) \cap D_0$$

auf einen geeigneten Funktionenraum verallgemeinern, welcher auch den Ansatzraum V_{T_h} der Finite Elemente Funktionen umfasst und die eindeutige Lösbarkeit der Variationsgleichung sicherstellt.

9.9 Definition. Es sei $\Omega \subset \mathbb{R}^2$ ein Gebiet. Wir definieren

$$L^2(\Omega) := \{u : \Omega \longrightarrow \mathbb{R} \,|\, |u|^2 \text{ Lebesgue-integrierbar}\}.$$

$u, v \in L^2(\Omega)$ sind gleich in $L^2(\Omega)$ genau dann, wenn $u(x) = v(x)$ für fast alle $x \in \Omega$ gilt.

Versehen mit dem Skalarprodukt

$$\langle u, v \rangle_0 = \int_\Omega u(x)v(x)dx$$

und der dadurch erzeugten Norm

$$\|u\|_{L^2} = \sqrt{\langle u, u \rangle} = \left(\int_\Omega |u(x)|^2 dx \right)^{1/2}$$

wird der Raum $(L^2(\Omega), \langle \cdot, \cdot \rangle_0)$ zum Hilbertraum.

9.10 Definition. Zu $u : \Omega \longrightarrow \mathbb{R}$ heißt

$$\text{supp}(u) = \overline{\{x \in \Omega \,|\, u(x) \neq 0\}}$$

Träger von u.

9.11 Definition. Der Raum

$$C_0^\infty(\Omega) = \{u \in C^\infty(\Omega) \,|\, \text{supp}(u) \text{ kompakt}\}$$

heißt Raum der Testfunktionen.

Dieser Raum ist nicht leer, denn der sogannte Friedrichs'sche Glättungskern $u_\varepsilon(\cdot - x_0)$ mit

$$u_\varepsilon(x) = \begin{cases} \exp\left(-\frac{\varepsilon^2}{\varepsilon^2 - \|x\|_2^2}\right), & \|x\|_2 < \varepsilon \\ 0, & \text{sonst} \end{cases} \qquad (\varepsilon > 0)$$

ist in $C_0^\infty(\Omega)$ für $x_0 \in \Omega$ und hinreichend kleine $\varepsilon > 0$.

Ferner benötigen wir noch den Begriff der schwachen Ableitung.

© Johannes Schropp 28. November 2008

9.12 Definition. Sei $u \in L^2(\Omega)$, und sei $\alpha \in \mathbb{N}_0^2$ ein Multiindex. u besitzt die schwache Ableitung $v := \partial^\alpha u \in L^2(\Omega)$, wenn für alle $\varphi \in C_0^\infty(\Omega)$ gilt

$$\int_\Omega v \cdot \varphi \, dx = (-1)^{|\alpha|} \int_\Omega u \partial^\alpha \varphi \, dx.$$

Offensichtlich besitzt $u \in C^k(\Omega)$ schwache Ableitungen $\partial^\alpha u$ für $|\alpha| \leq k$, und diese stimmen mit den klassischen Ableitungen überein.

9.13 Definition. Sei $m \in \mathbb{N}_0$. Dann heißt der Raum

$$H^m(\Omega) = \{u \in L^2(\Omega) \,|\, \partial^\alpha u \in L^2(\Omega) \quad \forall \alpha \in \mathbb{N}_0^2 : |\alpha| \leq m\}$$

Sobolev-Raum.

Der Sobolev-Raum $H^m(\Omega)$ versehen mit dem Skalarprodukt

$$\langle u, v \rangle_{H^m} = \sum_{|\alpha| \leq m} \int_\Omega \partial^\alpha u \cdot \partial^\alpha v \, dx = \sum_{|\alpha| \leq m} \langle \partial^\alpha u, \partial^\alpha v \rangle_0$$

ist ein Hilbert-Raum.

Es stellt sich nun die Frage, ob H^m-Funktionen stetig sind. Um diese Frage positiv beantworten zu können, benötigt man Voraussetzungen an den Rand $\partial\Omega$.

9.14 Definition. Sei $\Omega \subset \mathbb{R}^n$ beliebig. Ω besitzt die gewöhnliche Kegeleigenschaft, wenn es einen endlichen Kegel C so gibt, dass es zu jedem $x \in \Omega$ einen zu C kongruenten Kegel C_x gibt, dessen Spitze in x ist und der ganz in Ω liegt.

Kongruenz bedeutet dabei, dass C und C_x durch Kongruenzabbildung ineinander überführbar sind. Kongruenzabbildungen sind Parallelverschiebung, Drehung, Spiegelung und die Vernüpfung dieser Abbildungen.

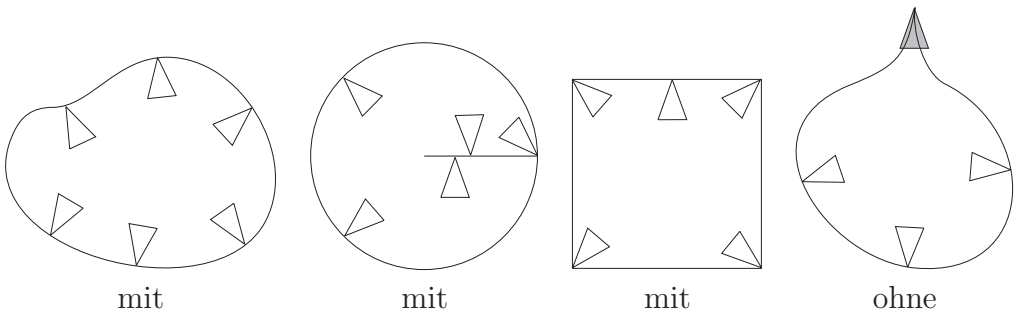


Abbildung 10: Beispiel für Gebiete mit und ohne Kegeleigenschaft

© Johannes Schropp 28. November 2008

9.15 Satz (Sobolevscher Einbettungssatz). *Sei $\Omega \subset \mathbb{R}^d$ ein Gebiet mit gewöhnlicher Kegeleigenschaft. Ist $m > \frac{d}{2}$, so gilt*

$$H^m(\Omega) \subset C(\overline{\Omega}),$$

und die Einbettungsabbildung $H^m(\Omega) \hookrightarrow C(\overline{\Omega})$ ist stetig.

Also ist für $d = 2, 3$ die Einbettung stetig, falls $m \geq 2$.

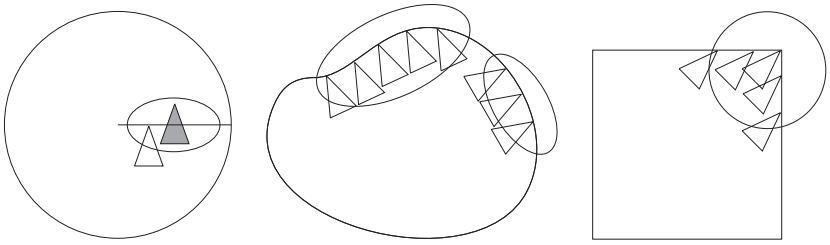
Ferner benötigen wir die Hilberträume

$$H_0^m(\Omega) = \{u \in H^m(\Omega) \,|\, u = 0 \text{ auf } \partial\Omega\}.$$

Da für Gebiete $\Omega \subset \mathbb{R}^2$ die Elemente aus $H_0^1(\Omega)$ nicht unbedingt stetig sein müssen, ist „$u = 0$ auf $\partial\Omega$“ möglicherweise nicht definiert. Der Ausdruck „$u = 0$ auf $\partial\Omega$“ muss in einem schwachen Sinne erklärt werden.

9.16 Definition. Ein Gebiet $G \in \mathbb{R}^n$ besitzt die strikte Kegeleigenschaft, falls es eine lokal (d.h. für alle Kompakta) endliche offene Überdeckung $\{\theta_i\}$ von ∂G mit zugehörigen Kegeln $\{K_i\}$ gibt mit

$$\forall x \in G \cap \theta_i : \quad x + K_i \subset G.$$



ohne mit mit

Abbildung 11: Beispiel für Gebiete mit der strikten und ohne strike Kegeleigenschaft

9.17 Satz. *Sei $\Omega \subset \mathbb{R}^2$ ein beschränktes Gebiet mit strikter Kegeleigenschaft. Dann gibt es eindeutig einen linearen stetigen Operator*

$$\Gamma : H^1(\Omega) \quad \longrightarrow \quad L^2(\partial\Omega)$$

mit $\Gamma(u) = u|_{\partial\Omega}$ für alle $u \in C^1(\overline{\Omega})$.

Man nennt $\Gamma(u)$ die verallgemeinerten Randwerte von u oder den Spuroperator. Häufig schreibt man „$u = 0$ auf $\partial\Omega$ anstelle von $\Gamma(u) = 0$.

Es gilt also

$$H_0^1(\Omega) = \{u \in H^1(\Omega) \,|\, \Gamma(u) = 0\}.$$

© Johannes Schropp 28. November 2008

Es lässt sich auch die Dichtheit von $C_0^\infty(\Omega)$ in $H_0^m(\Omega)$ zeigen.

Die Dirichlet'sche Aufgabe mit homogenen Randbedingungen

Nun wenden wir uns unserem Modellproblem zu. Vorgelegt sei die Aufgabe

$$\begin{aligned} -\Delta u &= g \text{ in } \Omega, \\ u &= 0 \text{ auf } \partial\Omega. \end{aligned} \tag{9-8}$$

$\Omega \subset \mathbb{R}^2$ sei dabei ein beschränktes Gebiet mit stückweise glattem Rand $\partial\Omega$ und der strikten Kegeleigenschaft. Es sei $g \in C(\overline{\Omega})$ und $u \in C^2(\Omega) \cap C(\overline{\Omega})$ eine klassische Lösung.

Multipliziert man (9-8) mit $v \in C_0^\infty(\Omega)$ und intergriert über Ω, so ergibt sich unter Benutzung des Divergenzsatzes

$$\begin{aligned} \int_\Omega g \cdot v \, dx &= -\int_\Omega \Delta u \cdot v \, dx \\ &= -\int_{\operatorname{supp} v} \Delta u \cdot v \, dx \\ &= \int_{\operatorname{supp} v} \nabla u \cdot \nabla v \, dx - \int_{\partial \operatorname{supp} v} \frac{\partial u}{\partial n} \cdot v \, dx \\ &= \int_{\operatorname{supp} v} \nabla u \cdot \nabla v \, dx \text{ für alle } v \in C_0^\infty(\Omega). \end{aligned} \tag{9-9}$$

Diese Gleichung gilt wegen der Dichtheit von $C_0^\infty(\Omega)$ und der Stetigkeit von $\langle g, \cdot \rangle - \langle \nabla u, \nabla \cdot \rangle$ in $H_0^1(\Omega)$ auch für alle $v \in H_0^1(\Omega)$.

Für $V = H_0^1(\Omega)$ setzen wir

$$\begin{aligned} a : V \times V &\longrightarrow \mathbb{R}, \\ a(u, v) &= \int_\Omega \nabla u \cdot \nabla v \, dx \end{aligned}$$

und

$$\begin{aligned} b : V &\longrightarrow \mathbb{R}, \\ b(v) &= \int_\Omega g \cdot v \, dx. \end{aligned}$$

Die schwache Formulierung von (9-8) lautet

$$a(u, v) = b(v), \quad \forall v \in V = H_0^1(\Omega). \tag{9-10}$$

9.18 Definition. $u \in V$ heißt schwache Lösung von (9-8), wenn u die Gleichung (9-10) erfüllt.

© Johannes Schropp 28. November 2008

Ist umgekehrt u eine schwache Lösung von (9-9) mit $u \in C^2(\Omega) \cap C(\overline{\Omega})$, so gilt

$$\int_\Omega (g + \Delta u) v \, dx = \int_\Omega (g \cdot v - \nabla u \cdot \nabla v) \, dx = 0$$

für alle $v \in C_0^\infty(\Omega)$.

Aus Dichtheitsgründen folgt

$$\int_\Omega \underbrace{(g + \Delta u)}_{\in L^2(\Omega)} v \, dx = 0 \quad \forall v \in L^2(\Omega)$$

und mit dem Fundamentallemma der Variationsrechnung (Satz von de la Vallée-Poisson) folgt

$$g + \Delta u = 0 \text{ fast überall in } \Omega.$$

Da $u \in H_0^1(\Omega)$, gilt $0 = \Gamma(u) = u|_{\partial\Omega}$.

9.19 Lemma. *Die Variationsgleichung $a(u, v) = b(v)$, $\forall v \in V$ hat die gleichen Lösungen $v \in V$ wie die Minimierungsaufgabe*

$$F(v) = \frac{1}{2} a(v, v) - b(v) = \int_\Omega \frac{1}{2} |\nabla v|^2 - g \cdot v \, dx \to \min \text{ für } v \in V.$$

Das Minimierungsproblem heißt Prinzip der minimalen potentiellen Energie oder das Dirichlet'sche Prinzip.

Es stellt sich nun die Frage nach den Bedingungen an a und b, unter welchen die Existenz und Eindeutigkeit von Lösungen von (9-10) gesichert werden.

9.20 Definition. Sei V ein reller Hilbertraum und $a : V \times V \to \mathbb{R}$ eine Bilinearform.

i) a heißt stetig auf V, falls ein $K > 0$ existiert mit

$$|a(u, v)| \leq K \cdot \|u\| \cdot \|v\|$$

für alle $u, v \in V$.

ii) a heißt koerziv (oder elliptisch) auf V, wenn es ein $\alpha > 0$ gibt mit

$$a(u, u) \geq \alpha \|u\|^2$$

für alle $u \in V$.

© Johannes Schropp 28. November 2008

9.21 Satz (Lax-Milgram). *Seien V ein reeller Hilbertraum, $a : V \times V \to \mathbb{R}$ eine stetige und koerzive Bilinearform sowie $b : V \to \mathbb{R}$ ein stetiges lineares Funktional. Dann existiert genau ein $u \in V$ mit*

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

Für diese Lösung gilt

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

Dabei ist

$$\|b\|_{V'} = \sup\left\{ \frac{|b(v)|}{\|v\|} \,\middle|\, v \in V, v \neq 0 \right\}.$$

Die Dirichlet'sche Aufgabe mit inhomogenen Randbedingungen

Wir betrachten das Problem

$$\begin{aligned} -\Delta u &= g \text{ in } \Omega, \\ u &= \gamma \text{ auf } \partial\Omega. \end{aligned} \tag{9-11}$$

$\Omega \subset \mathbb{R}^2$ sei wieder ein beschränktes Gebiet mit stückweise glattem Rand $\partial\Omega$ mit strikter Kegeleigenschaft. Ferner lasse sich $\gamma|_{\partial\Omega}$ zu einem zweimal stetig differenzierbaren $\gamma : \Omega \longrightarrow \mathbb{R}$ fortsetzen.

Wir suchen in diesem Fall eine Funktion u mit $u - \gamma \in V = H_0^1(\Omega)$. Dazu transformieren wir die Aufgabe (9-11) auf homogene Randbedingungen.

Sei $w = u - \gamma$ eine Lösung von

$$\begin{aligned} -\Delta w &= g + \Delta\gamma \text{ in } \Omega, \\ w &= 0 \text{ auf } \partial\Omega. \end{aligned}$$

Dann ist $u = w + \gamma$ eine Lösung von (9-11), denn

$$\begin{aligned} -\Delta u &= -\Delta(w + \gamma) = -\Delta w - \Delta\gamma = g \text{ in } \Omega, \\ u|_{\partial\Omega} &= \underbrace{w|_{\partial\Omega}}_{=0} + \gamma|_{\partial\Omega} = \gamma|_{\partial\Omega}. \end{aligned}$$

Die schwache Formulierung des Problems hat die Form

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

Somit lautet das Galerkin-Verfahren wie folgt. Sei $V_h \subset V$ ein endlich dimensionaler Unterraum von V. Gesucht ist ein $u_h \in V_h$ mit

$$a(u_h, v) = b(v) \quad \forall v \in V_h.$$

Man sieht nun, dass die Finite Elemente Methode wieder als ein Galerkin-Verfahren für einen Ansatzraum mit speziellen Eigenschaften interpretiert werden kann.

© Johannes Schropp 28. November 2008

d) Stabilität und Konvergenz der Finite Elemente Methode

Es sei im Folgenden $V = H_0^1(\Omega)$ und

$$a : V \times V \longrightarrow \mathbb{R}, \quad a(u,v) = \int_\Omega \nabla v \cdot \nabla u \, dx,$$
$$b : V \longrightarrow \mathbb{R}, \quad b(v) = \int_\Omega g \cdot v \, dx.$$

Die Variationsaufgabe besteht nun darin, dass ein $\overline{u} \in V$ gesucht wird mit

$$a(\overline{u}, v) = b(v) \quad \forall v \in V. \tag{9-12}$$

Das Galerkin-Verfahren lautet entsprechend: Finde ein $u_h \in V_h$ mit

$$a(u_h, v) = v(v) \quad \forall v \in V_h \tag{9-13}$$

für einen endlich dimensionalen Teilraum $V_h \subset V$ von V.

$e := \overline{u} - u_h$ bezeichne den dadurch entstehenden Fehler, wobei $\overline{u}$ und u_h die Gleichungen (9-12) bzw. (9-13) lösen.

Wir finden die Fehlergleichung

$$a(e, v) = 0 \quad \forall v \in V_h,$$

denn aus (9-12) folgt

$$a(\overline{u}, v) = b(v) \quad \forall v \in V_h \subset V$$

und aus (9-13) folgt

$$a(u_h, v) = b(v) \quad \forall v \in V_h.$$

Insgesamt erhalten wir

$$a(e, v) = a(\overline{u} - u_h, v) = b(v) - b(v) = 0 \quad \forall v \in V_h. \tag{9-14}$$

Ist a eine symmetrische positiv definite Bilinearform, so besagt (9-14), dass e und v für $v \in V_h$ orthogonal sind, d.h. $e \in V_h^\perp$.

(9-14) bedeutet die Orthogonalität des Fehlers auf dem Ansatzraum. Durch das Galerkin-Vefahren (9-13) wird also dasjenige Element $u_h \in V_h$ charakterisiert, welches bezüglich der Norm

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

den kleinsten Abstand zu $\overline{u} \in V$ hat.

© Johannes Schropp 28. November 2008

9.22 Lemma. *Sei $V_h \subset V$ ein Unterraum von V, a ein Skalarprodukt auf V und $\|u\|_a = \sqrt{a(u,u)}$ die davon erzeugte Norm. Dann gilt für $u_h \in V_h$:*

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

was mit

$$\|\overline{u} - u_h\|_a = \inf\{\|\overline{u} - v\| \,|\, v \in V_h\}.$$

äquivalent ist.

Beweis: Für ein festes $\overline{u} \in V$ sei $b(v) = a(\overline{u}, v)$, $v \in V_h$, d.h. b ist ein lineares Funktional auf V_h. Somit ist

$$\underbrace{a(\overline{u}, v)}_{=b(v)} = a(u_h, v) \quad v \in V_h$$

eine Variationsgleichung auf V_h. Diese hat nach Lemma 9.19 die gleichen Lösungen wie

$$F(u_h) = \inf\{F(v) \,|\, v \in V_h\}$$

mit $F(v) = \frac{1}{2}a(v,v) - b(v) = \frac{1}{2}a(v,v) - a(\overline{u}, v)$.

F hat die gleichen Minima wie das Funktional

$$\begin{aligned} G(v) &= (2F(v) + a(\overline{u}, \overline{u}))^{1/2} \\ &= (a(v,v) - 2a(\overline{u}, v) + a(\overline{u}, \overline{u}))^{1/2} \\ &= a(\overline{u} - v, \overline{u} - v)^{1/2} = \|\overline{u} - v\|_a. \end{aligned}$$

□

Wir setzen jetzt für a das Folgende voraus: Sei $\|\cdot\|$ eine Norm auf V bezüglich der gelte

i) $a : V \times V \longrightarrow \mathbb{R}$ ist stetig bezüglich $\|\cdot\|$, d.h. es existiert ein $M > 0$ mit

$$|a(u,v)| \leq M \cdot \|u\| \cdot \|v\| \quad \forall u, v \in V.$$

ii) a ist V-elliptisch, d.h. es existiert ein $\alpha > 0$ mit

$$a(u,u) \geq \alpha \|u\|^2 \quad \forall u \in V.$$

Ferner sei das lineare Funktional $b : V \longrightarrow \mathbb{R}$ stetig.

Dann sichert der Satz 9.21 von Lax und Milgram, dass die Variationsgleichung

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

genau eine Lösung $u \in V$ besitzt.

© Johannes Schropp 28. November 2008

In unserem Fall ist $a : V \times V \longrightarrow \mathbb{R}$ ein Skalarprodukt. Mit $\|\cdot\| = \|\cdot\|_a$ ergibt sich unter Verwendung der Cauchy-Schwarzschen Ungleichung die Gültigkeit von $i)$ mit $M = 1$ sowie von $ii)$ mit $\alpha = 1$.

Die V-Elliptizität impliziert die Stabilität der Galerkin-Approximation.

9.23 Lemma. *Die Lösung u_h der Galerkin-Gleichung $a(u, v) = b(v)$, $\forall v \in V_h$ ist stabil im Sinne der Abschätzung*

$$\|u_h\| \leq \frac{1}{\alpha}\|b\|_{V'}$$

für alle $h > 0$, wobei

$$\|b\|_{V'} = \sup\left\{\frac{|b(v)|}{\|v\|} \,\middle|\, v \in V, v \neq 0\right\}.$$

Beweis: Aus $a(u_h, v) = b(v)$ für alle $v \in V_h$ folgt

$$\begin{aligned} \alpha\|u_h\|^2 &\overset{ii)}{\leq} a(u_h, u_h) = b(u_h) \\ &\leq \frac{|b(u_h)|}{\|u_h\|} \cdot \|u_h\| \leq \|b\|_{V'} \cdot \|u_h\|. \end{aligned}$$

Also folgt mit $\alpha > 0$

$$\|u_h\| \leq \frac{\|b\|_{V'}}{\alpha}.$$

Somit gilt bis auf eine Konstante weiterhin die Approximationsaussage aus Lemma 9.22. □

9.24 Satz (Cea's Lemma). *Unter den Voraussetzungen $i)$ und $ii)$ an die Bilinearform a und der Stetigkeit des linearen Funktionals b gilt für den Fehler der Galerkin-Lösung u_h die Abschätzung*

$$\|\overline{u} - u_h\| \leq \frac{M}{\alpha} \cdot \inf\{\|\overline{u} - v\| \mid v \in V_h\}.$$

Beweis: Sei $v \in V_h$ beliebig. Aus der Fehlergleichung

$$a(\overline{u} - u_h, v) = 0 \quad v \in V_h \qquad \text{(Vgl. (9-14))}$$

folgt

$$a(\overline{u} - u_h, u_h - v) = 0,$$

da $u_h - v \in V_h$.

© Johannes Schropp 28. November 2008

Mit der V-Elliptizität erhalten wir dann

$$\begin{aligned}\alpha\|\overline{u}-u_h\|^2 &\leq a(\overline{u}-u_h,\overline{u}-u_h)\\ &= a(\overline{u}-u_h,\overline{u}-u_h)+\underbrace{a(\overline{u}-u_h,u_h-v)}_{=0}\\ &= a(\overline{u}-u_h,\overline{u}-v)\\ &\leq M\cdot\|\overline{u}-u_h\|\cdot\|\overline{u}-v\|\end{aligned}$$

für alle $v\in V_h$, d.h.

$$\|\overline{u}-u_h\|\leq\frac{M}{\alpha}\|\overline{u}-v\|\quad\forall v\in V_h.$$

Also folgt

$$\|\overline{u}-u_h\|\leq\frac{M}{\alpha}\cdot\inf\{\|\overline{u}-v\|\,|\,v\in V_h\}.$$

□

Es reicht also für eine asymptotische Fehlerdarstellung in h den Restapproximationsfehler von V_h, d.h. $\inf\{\|\overline{u}-v\|\,|\,v\in V_h\}$ abzuschätzen.

Betrachte nun die konkrete Variationsgleichung

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

mit $a(u,v)=\int_\Omega\nabla u\cdot\nabla v\,dx$, $b(v)=\int_\Omega g\cdot v\,dx$, welche die schwache Formulierung der Dirichlet'schen Randwertaufgabe darstellt. Dabei sei $V=H_0^1(\Omega)$.

Zu betrachten bleibt die Wahl der Normen. Offensichtlich kann als die Norm $\|\cdot\|=\|\cdot\|_a$

$$\|u\|_a=\left(\int_\Omega|\nabla u|^2\,dx\right)^{1/2}$$

genommen werden. Alternativ kann aber auch die durch das Skalarprodukt

$$\langle u,v\rangle_1=\sum_{|\alpha|\leq1}\langle\partial^\alpha u,\partial^\alpha v\rangle_0$$

induzierte Norm

$$\|u\|_1=\left(\int_\Omega|u(x)|^2\,dx+\int_\Omega|\nabla u(x)|^2\,dx\right)^{1/2},\quad u\in H_0^1(\Omega)$$

gewählt werden, falls die Bilinearform a bezüglich $\|\cdot\|_1$ stetig und V-elliptisch ist.

Für die Stetigkeit benutzen wir die Abschätzung

$$\|u\|_a\;=\;\left(\int|\nabla u(x)|^2\,dx\right)^{1/2}$$

© Johannes Schropp 28. November 2008

$$\leq \left(\int_\Omega |u(x)|^2 \, dx + \int_\Omega |\nabla u(x)|^2 \, dx \right)^{1/2} = \|u\|_1,$$

und somit folgt mit Cauchy-Schwarz

$$|a(u,v)| \leq \|u\|_a \cdot \|v\|_a \leq \|u\|_1 \cdot \|v\|_1 \quad \forall u, v \in H_0^1(\Omega),$$

d.h. die Bilinearform a ist auch stetig bezüglich $\|\cdot\|_1$ mit $M = 1$.

Die V-Elliptizität

$$a(u,u) \geq \alpha \|u\|_1^2, \quad \forall u \in V \quad (\alpha > 0)$$

gilt nicht im Allgemeinen für $V = H^1(\Omega)$. Sie gilt aber für $V = H_0^1(\Omega)$. Dazu benötigen wir den folgenden Satz.

9.25 Satz (Erste Poincarésche Abschätzung). *Sei Ω ein beschränktes Gebiet. Dann gibt es ein $C > 0$ mit*

$$\|u\|_0 \leq C \left(\int_\Omega |\nabla u(x)|^2 \, dx \right)^{1/2} \quad \forall u \in H_0^1(\Omega).$$

Mit dem Satz von Poincaré folgt

$$\|u\|_0 \leq C \|u\|_a$$

und damit

$$\|u\|_1^2 = \|u\|_0^2 + \|u\|_a^2 \leq (1 + C^2) \|u\|_a^2,$$

d.h.

$$\frac{1}{\sqrt{1+C^2}} \|u\|_1 \leq \|u\|_a, \quad u \in H_0^1(\Omega).$$

Dies liefert

$$a(u,u) = \|u\|_a^2 \geq \frac{1}{1+C^2} \|u\|_1^2,$$

und somit die V-Elliptizität bezüglich $\|\cdot\|_1$ mit $\alpha = \frac{1}{1+C^2}$.

Damit sind die Normen $\|\cdot\|_a$ und $\|\cdot\|_1$ auf $V = H_0^1(\Omega)$ äquivalent und erzeugen daher denselben Konvergenzbegriff.

Im Lemma 9.24 von Cea erhalten wir für $\|\cdot\| = \|\cdot\|_1$ die Abschätzung

$$\begin{aligned} \|\overline{u} - u_h\|_1 &\leq \frac{1}{\alpha} \cdot \inf\{\|\overline{u} - v\|_1 \mid v \in V_h\} \\ &= (1 + C^2) \cdot \inf\{\|\overline{u} - v\|_1 \mid v \in V_h\}. \end{aligned}$$

© Johannes Schropp 28. November 2008

Bis jetzt war unsere Abschätzung des Approximationsfehlers unabhängig von der konkreten Wahl von V_h. Jetzt möchten wir den Approximationsfehler der Finiten Elemente Methode analysieren. In diesem Fall ist $V_h = V_{T_h}$ mit

$$V_{T_h} = \{u \in C(\Omega_{T_h}) \,|\, u \text{ ist affin linear in } x \text{ und } y \\ \text{auf jedem Dreieck } e \in E_h \text{ und } u = 0 \text{ auf } \partial\Omega_{T_h}\}.$$

Es gilt

$$\begin{aligned} \|\overline{u} - u_h\|_1 &\leq (1 + C^2) \cdot \inf\{\|\overline{u} - v\|_1 \,|\, v \in V_{T_h}\} \\ &\leq (1 + C^2)\|\overline{u} - w\|_1 \end{aligned}$$

für jedes $w \in V_{T_h}$.

Wir konstruieren nun ein spezielles $w \in V_{T_h}$ durch Interpolation.

Sei Ω polygonal berandet mit strikter Kegeleigenschaft. Betrachte die Abbildung

$$\begin{aligned} I_h : \{u \in C(\overline{\Omega}) \,|\, u = 0 \text{ auf } \partial\Omega\} &\longrightarrow V_{T_h} \\ v &\longmapsto I_h(v) = \sum_{i=1}^{M} v(P_i)u_i \end{aligned}$$

mit den Formfunktionen u_i, $i = 1, \ldots, M$ zu den Knoten P_i der Triangulierung Ω_h.

$I_h(v)$ ist nur für $v \in H^2(\Omega)$ definiert, da $H^2(\Omega) \subset C(\overline{\Omega})$ nach dem Sobolevschen Einbettungssatz 9.15.

Abzuschätzen bleibt der Interpolationsfehler in der H^1-Norm. Für diesen Fehler kann man zeigen

$$\|I_h(v) - v\|_1 \leq C \cdot h \quad \forall v \in H^2(\Omega),$$

falls $(T_h)_{0<h<h_1}$ eine Familie von Triangulierungen von Ω ist, so dass für den jeweils maximalen Winkel $\gamma_{h,\max}$ in einem Dreieck $e \in T_h$ gilt

$$\gamma_{h,\max} \leq \gamma_{\max} < \pi$$

für $0 < h \leq h_0$. Letztere Bedingung heißt Maximalwinkelbedingung.

Liegt also die Lösung u der Variationsaufgabe in $H^2(\Omega) \cap H_0^1(\Omega)$, so dürfen wir $v = u$ wählen und erhalten für lineare Finite Elemente die Abschätzung

$$\|u - u_h\|_1 \leq (1 + C^2)\|u - I_h(u)\|_1 \leq \hat{C}h, \quad 0 < h \leq h_0.$$

Also ist die lineare Finite Elemente Methode konvergent der Ordnung 1 bezüglich $\|\cdot\|_1$ an Lösungen der Glattheit $H^2(\Omega) \cap H_0^1(\Omega)$.

Fehlerabschätzung in der L^2-Norm

© Johannes Schropp 28. November 2008

9.26 Lemma (Aubin-Nietzsche). *Besitzt das duale Problem*

$$a(v,z) = b(v) = \int_\Omega g \cdot v \, dx \quad \forall v \in V$$

für $g \in L^2(\Omega)$ eine Lösung z mit $\|z\|_2 := \|z\|_{H^2} \leq C\|f\|_0$, so gilt

$$\|u - u_h\|_0 \leq \widetilde{C} \cdot h \cdot \|u - u_h\|_1.$$

Somit folgt

$$\|\overline{u} - u_h\|_0 \leq \widetilde{C} \cdot h \cdot \|u - u_h\|_1 \leq \widetilde{C}\hat{C}h^2, \quad 0 < h \leq h_0,$$

d.h. wir haben Konvergenz der Ordnung 2 bezüglich $\|\cdot\|_0$.

9.27 Bemerkung. Höhere Konvergenzordnung in der H^1-Norm bzw. der L^2-Norm lassen sich mit polynomialen Finiten Elementen erreichen, falls die Lösung der Randwertaufgabe hinreichend oft schwach differenzierbar ist.

9.28 Bemerkung. Die Konvergenzaussage für lineare Finite Elemente bleibt auch richtig, falls der Rand $\partial\Omega$ sückweise stetig differenzierbar ist.

9.29 Bemerkung. Im Fall von höheren polynomialen Finiten Elementen ($r \geq 2$), benötigt man spezielle Randelemente, um höhere Konvergenzordnung bei stückweise glattem Rand $\partial\Omega$ sicherzustellen.

© Johannes Schropp 28. November 2008

10. Finite Differenzenverfahren für parabolische Differentialgleichungen

a) Das Prinzip der Linienmethode

Wir betrachten die nichtlineare Anfangswertaufgabe

$$\begin{aligned} u_t &= u_{xx} + f(u, u_x, x, t) \text{ in } \Omega = (0,1) \times (0,T), \\ u(x,0) &= u_0(x) \text{ für } 0 \le x \le 1, \\ u(0,t) &= \gamma_0(t),\ u(1,t) = \gamma_1(t) \text{ für } 0 \le t \le T. \end{aligned} \tag{10-1}$$

Das Prinzip der Linienmethode beruht darauf, zuerst eine Diskretisierung in der Ortsvariablen vorzunehmen. Die Zeitvariable t bleibt noch kontinuerlich.

Um diesen Schritt durchzuführen, betrachten wir zunächst die Randwertaufgabe

$$\begin{aligned} -w''(x) &= f(x, w(x), w'(x)) \text{ in } (0,1), \\ w(0) &= \gamma_0,\ w(1) = \gamma_1. \end{aligned} \tag{10-2}$$

Zu $\Delta x = \frac{1}{M} > 0$ ($M \in \mathbb{N}$) bekommen wir das Gitter $\Omega_{\Delta x} = \{j\Delta x \,|\, j = 0, \ldots, M\}$. Wir ersetzen $-w''(x)$ durch

$$-w''(x) \sim \frac{1}{\Delta x^2}(-w(x-\Delta x) + 2w(x) - w(x+\Delta x)) \tag{10-3}$$

und $w'(x)$ durch

$$w'(x) \sim \frac{1}{2\Delta x}(w(x+\Delta x) - w(x-\Delta x)). \tag{10-4}$$

Präziser gilt

$$\left| \frac{1}{\Delta x^2}(-w(x-\Delta x) + 2w(x) - w(x+\Delta x)) + w''(x) \right| = O(\Delta x^2),$$

falls $w \in C^4$ (vgl. Kapitel 8) bzw.

$$\left| \frac{1}{2\Delta x}(w(x+\Delta x) - w(x-\Delta x)) - w'(x) \right| = O(\Delta x^2)$$

für $w \in C^3$.

Führen wir jetzt den Ersetzungsprozess durch, so erhalten wir

$$\begin{aligned} w(0) &= \gamma_0, \\ \frac{1}{\Delta x^2}(-w(x-\Delta x) + 2w(x) - w(x+\Delta x)) &= \end{aligned}$$

© Johannes Schropp 28. November 2008

$$= f\left(x, w(x), \frac{1}{2\Delta x}(w(x+\Delta x) - w(x-\Delta x))\right), \qquad (10\text{-}5)$$
$$x = j\Delta x, j = 1, \ldots, M-1,$$
$$w(1) = \gamma_1.$$

Gesucht ist also ein Vektor $(w(0), w(\Delta x), \ldots, w(1-\Delta x), w(1))$, welcher (10-5) löst. Wir eliminieren $w(0)$, $w(1)$ aus dem System (10-5) und schreiben es in kompakter Form

$$\frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix}\begin{pmatrix} w(\Delta x) \\ w(2\Delta x) \\ w(3\Delta x) \\ \vdots \\ w(1-3\Delta x) \\ w(1-2\Delta x) \\ w(1-\Delta x) \end{pmatrix} = \qquad (10\text{-}6)$$
$$= \begin{pmatrix} f\left(\Delta x, w(\Delta x), \frac{1}{2\Delta x}(w(2\Delta x) - \gamma_0)\right) + \frac{\gamma_0}{\Delta x^2} \\ f\left(2\Delta x, w(2\Delta x), \frac{1}{2\Delta x}(w(3\Delta x) - w(\Delta x))\right) \\ \vdots \\ f\left(j\Delta x, w(j\Delta x), \frac{1}{2\Delta x}(w((j+1)\Delta x) - w((j-1)\Delta x))\right), j = 2, \ldots, M-2 \\ \vdots \\ f\left(1-2\Delta x, w(1-2\Delta x), \frac{1}{2\Delta x}(w(1-\Delta x) - w(1-3\Delta x))\right) \\ f\left(1-\Delta x, w(1-\Delta x), \frac{1}{2\Delta x}(\gamma_1 - w(1-2\Delta x))\right) + \frac{\gamma_1}{\Delta x^2} \end{pmatrix}.$$

Das Gleichungssystem (10-6) hat die Form

$$A^{\Delta x} w = G^{\Delta x}(w), \quad w \in \mathbb{R}^{\Omega_{\Delta x}} \qquad (10\text{-}7)$$

mit

$$A^{\Delta x} = \frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix},$$

$$G^{\Delta x}(w) = \begin{pmatrix} f\left(\Delta x, w(\Delta x), \frac{1}{2\Delta x}(w(2\Delta x) - \gamma_0)\right) + \frac{\gamma_0}{\Delta x^2} \\ f\left(j\Delta x, w(j\Delta x), \frac{1}{2\Delta x}(w((j+1)\Delta x) - w((j-1)\Delta x))\right), \\ j = 2, \ldots, M-2 \\ f\left(1-\Delta x, w(1-\Delta x), \frac{1}{2\Delta x}(\gamma_1 - w(1-2\Delta x))\right) + \frac{\gamma_1}{\Delta x^2} \end{pmatrix}.$$

© Johannes Schropp 28. November 2008

Insbesondere zeigt dies, dass

$$R^{\Delta x} = -A^{\Delta x} w + G^{\Delta x}(w)$$

eine Diskretisierung für den Operator A ist mit

$$\begin{aligned} A : D(A) \subset C^2((0,1)) \cap C([0,1]) &\longrightarrow C((0,1)), \\ w &\longmapsto w''(\cdot) + f(\cdot, w(\cdot), w'(\cdot)), \end{aligned}$$

wobei

$$D(A) := \{w \in C^2((0,1)) \cap C([0,1]) \,|\, w(0) = \gamma_0,\ w(1) = \gamma_1\}.$$

Wir kehren nun zu unserer Anfangsrandwertaufgabe (10-1) zurück. Gemäß des Prinzips der Linienmethode diskretisieren wir für beliebiges $t \in [0,T]$ zuerst die Raumvariable.

Wir suchen eine Gleichung für die Funktion

$$\begin{aligned} v(t) &= (u(\Delta x, t), u(2\Delta x, t), \ldots, u(1-2\Delta x, t), u(1-\Delta x, t)) \\ &= (v_1(t), v_2(t), \ldots, v_{M-2}(t), v_{M-1}(t)) \in \mathbb{R}^{M-1}, \quad 0 \le t \le T. \end{aligned}$$

Eine einfache Diskretisierung können wir jetzt in der Ortsvariablen durch das klassische Differenzenverfahren vornehmen.

Wir ersetzen den Ausdruck $u_{xx}(x,t) + f(u(x,t), u_x(x,t), u, t)$, $u(0,t) = \gamma_0(t)$, $u(1,t) = \gamma_1(t)$ für $t \in [0,T]$ durch

$$\begin{aligned} &\underbrace{-\frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix}}_{=A^{\Delta x}} \begin{pmatrix} v_1(t) \\ v_2(t) \\ v_3(t) \\ \vdots \\ v_{M-3}(t) \\ v_{M-2}(t) \\ v_{M-1}(t) \end{pmatrix} + \underbrace{\frac{1}{\Delta x^2}\begin{pmatrix} \gamma_0(t) \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \gamma_1(t) \end{pmatrix}}_{=r^{\Delta x}(t)} \\ &\quad + \underbrace{\begin{pmatrix} f\left(v_1(t), \frac{1}{2\Delta x}(v_2(t) - \gamma_0), \Delta x, t\right) \\ f\left(v_j(t), \frac{1}{2\Delta x}(v_{j+1}(t) - v_{j-1}(t)), j\Delta x, t\right), \\ j = 2, \ldots, M-2 \\ f\left(v_{M-1}(t), \frac{1}{2\Delta x}(\gamma_1 - v_{M-2}(t)), (M-1)\Delta x, t\right) \end{pmatrix}}_{=:H^{\Delta x}(v(t))} \\ &=: -A^{\Delta x} v(t) + r^{\Delta x}(t) + H^{\Delta x}(v(t)). \end{aligned}$$

Präziser wird $A(t)u(t,\cdot)$, $u(t,\cdot) \in D(A(t))$ durch

$$-A^{\Delta x} v(t) + r^{\Delta x}(t) + H^{\Delta x}(v(t))$$

© Johannes Schropp 28. November 2008

ersetzt, wobei der zeitabhängige Differentialoperator $A(t)$ wie folgt definiert ist:

$$\begin{aligned} A(t) : D(A(t)) \subset C^2((0,1)) \cap C([0,1]) &\longrightarrow C((0,1)), \\ u &\longmapsto u_{xx} + f(u, u_x, \cdot, t) \end{aligned}$$

mit dem Definitionsbereich

$$D(A(t)) := \{u \in C^2((0,1)) \cap C([0,1]) \,|\, u(0) = \gamma_0(t),\, u(1) = \gamma_1(t)\}.$$

Setzt man $v'(t) = (u_t(\Delta x, t), \ldots, u_t((M-1)\Delta x, t))$, so ergibt sich unter Beachtung von (10-1) die gewöhnliche Differentialgleichung

$$\begin{aligned} v'(t) &= -A^{\Delta x} v(t) + H^{\Delta x}(v(t)) + r^{\Delta x}(t) \\ &=: F_{\Delta x}(v(t), t), \quad 0 \le t \le T, \\ v(0) &= v^0 = (u_0(\Delta x), \ldots, u_0((M-1)\Delta x)) \end{aligned} \tag{10-8}$$

als Ersatz für die Anfangsrandwertaufgabe (10-1).

Der Übergang von (10-1) zu (10-8) wird als Semidiskretisierung bezeichnet. Zu einer vollständigen Diskretisierung von (10-1) kommen wir, wenn wir auf (10-8) eines der handelsüblichen Verfahren zur Lösung von Anfangswertproblemen anwenden. Bei der Durchführung dieser Verfahren sollte man sich jedoch die spezielle Gestalt von $F_{\Delta x}$ zu Nutze machen, denn das System (10-8) wird für $\Delta x \longrightarrow 0$ immer größer. Das einfachste Verfahren für (10-8) ist das Euler-Cauchy Verfahren.

Seien $\Delta t = \frac{T}{N} > 0$, v^0 vorgegeben. v^j approximiere $v(j\Delta t)$, $j = 0, \ldots, N$. Dann finden wir mit $t_j = j\Delta t$

$$\begin{aligned} v^{j+1} &= v^j + \Delta t \cdot F_{\Delta x}(v^j, t_j), \quad j = 0, \ldots, N-1, \\ v^0 &= (u_0(\Delta x), \ldots, u_0((M-1)\Delta x)) \end{aligned} \tag{10-9}$$

oder explizit mit

$$v^j = (u_1^j, \ldots, u_{M-1}^j), \quad u_i^j = u(x_i, t_j), \quad x_i = i\Delta x$$

folgt

$$\begin{aligned} \begin{pmatrix} u_1^{j+1} \\ \vdots \\ u_i^{j+1} \\ \vdots \\ u_{M-1}^{j+1} \end{pmatrix} = \begin{pmatrix} u_1^j \\ \vdots \\ u_i^j \\ \vdots \\ u_{M-1}^j \end{pmatrix} + \Delta t \Bigg\{ \frac{1}{\Delta x^2} \begin{pmatrix} -2u_1^j + u_2^j \\ \vdots \\ u_{i-1}^j - 2u_i^j + u_{i+1}^j \\ \vdots \\ u_{M-2}^j - 2u_{M-1}^j \end{pmatrix} \\ + \frac{1}{\Delta x^2} \begin{pmatrix} \gamma_0(t_j) \\ 0 \\ \vdots \\ 0 \\ \gamma_1(t_j) \end{pmatrix} + \begin{pmatrix} f\left(u_1^j, \frac{1}{2\Delta x}(u_2^j - \gamma_0(t_j)), x_1, t_j\right) \\ \vdots \\ f\left(u_i^j, \frac{1}{2\Delta x}(u_{i+1}^j - u_{i-1}^j), x_i, t_j\right) \\ \vdots \\ f\left(u_{M-1}^j, \frac{1}{2\Delta x}(\gamma_1(t_j) - u_{M-2}^j), x_{M-1}, t_j\right) \end{pmatrix} \Bigg\} \end{aligned} \tag{10-10}$$

© Johannes Schropp 28. November 2008

mit den Anfangsbedingungen

$$\begin{pmatrix} u_1^0 \\ \vdots \\ u_i^0 \\ \vdots \\ u_{M-1}^0 \end{pmatrix} = \begin{pmatrix} u_0(x_1) \\ \vdots \\ u_0(x_i) \\ \vdots \\ u_0(x_{M-1}) \end{pmatrix}. \tag{10-11}$$

(10-10), (10-11) heißen auch explizites Differenzenverfahren zur Anfangsrandwertaufgabe (10-1). Man erhält u_i^{j+1} aus u_{i-1}^j, u_i^j, u_{i+1}^j, was man durch folgendes Schema andeutet.

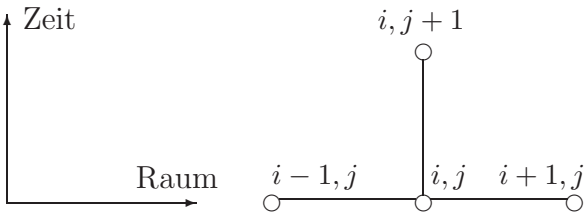


Abbildung 12: ,,Differenzenstern“

Wir betrachten als nächstes für das Liniensystem (10-8) das von einem Parameter $\vartheta \in [0,1]$ abhängige Verfahren

$$\begin{aligned} v_{j+1} &= v^j + \Delta t \left[\vartheta F_{\Delta x}(v^{j+1}, t_{j+1}) + (1-\vartheta) F_{\Delta x}(v^j, t_j)\right], \\ &\quad j = 0, \ldots, N-1, \\ v^0 &= (u_0(x_1), \ldots, u_0(x_{M-1})). \end{aligned} \tag{10-12}$$

Die Spezialfälle des ϑ-Verfahrens für das Liniensystem sind:
$\vartheta = 0$: Euler-Cauchy Verfahren,
$\vartheta = \frac{1}{2}$: Trapezenmethode oder sogenanntes ,,Crank-Nicholson Verfahren“,
$\vartheta = 1$: Implizites Euler-Cauchy Verfahren.

Für $\vartheta > 0$ erfordert (10-12) die Auflösung eines nichtlinearen Gleichungssystems und das Verfahren heißt implizit. Die Abbildung 13 stellt die zugehörigen Differenzensterne dar.

10.1 Beispiel (Chemische Reaktions-Transport-Gleichung). Die dazu gehörige Anfangsrandwertaufgabe lautet

$$\begin{aligned} u_t &= u_{xx} - k \cdot e_0(x) u, \quad 0 \le x \le 1,\ t \ge 0, \quad k \ge 0 \\ u(x,0) &= 1 - x, \quad 0 \le x \le 1 \\ u(0,t) &= 1,\ u(1,t) = 0, \quad e_0(x) \ge 0,\ 0 \le x \le 1. \end{aligned}$$

© Johannes Schropp 28. November 2008

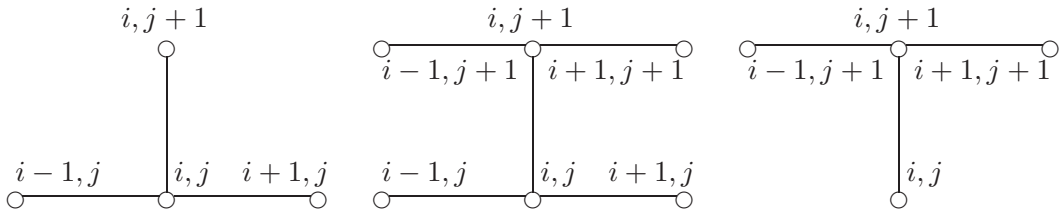


Abbildung 13: Differenzensterne für $\vartheta \in \{0, \frac{1}{2}, 1\}$

Wir finden für das Liniensystem

$$v'(t) = F_{\Delta x}(v(t), t),\ v(0) = v^0$$

mit

$$F_{\Delta x}(v,t) = -\frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix} v + \frac{1}{\Delta x^2}\begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$-k\begin{pmatrix} e_0(x_1)v_1 \\ e_0(x_2)v_2 \\ e_0(x_3)v_3 \\ \vdots \\ e_0(x_{M-3})v_{M-3} \\ e_0(x_{M-2})v_{M-2} \\ e_0(x_{M-1})v_{M-1} \end{pmatrix} = -B_{\Delta x}v + r^{\Delta x},$$

wobei $x_i = i\Delta x$, $i = 1, \dots, M-1$. Dabei ist

$$B_{\Delta x} = -\frac{1}{\Delta x^2}\begin{pmatrix} 2 & -1 & 0 & \dots & 0 & 0 & 0 \\ -1 & 2 & -1 & \dots & 0 & 0 & 0 \\ 0 & -1 & 2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 2 & -1 & 0 \\ 0 & 0 & 0 & \dots & -1 & 2 & -1 \\ 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix}$$

© Johannes Schropp 28. November 2008

$$
\begin{aligned}
& +k \cdot \operatorname{diag}(e_0(x_1), \ldots, e_0(x_{M-1})), \\
r^{\Delta x} &= \frac{1}{\Delta x^2}(1, 0, \ldots, 0, 0)^T.
\end{aligned}
$$

$B_{\Delta x}$ ist eine L_0-Matrix mit majorisierendem Element $\mathbb{I} = (1, 1, \ldots, 1, 1)^T$, d.h. eine M-Matrix, da $e_0(x) \geq 0$, $0 \leq x \leq 1$ und $h > 0$.

Diskretisierung der Anfangsrandwertaufgabe mit dem ϑ-Verfahren liefert dann die Iteration

$$
\begin{aligned}
v^{j+1} &= v^j + \Delta t \cdot \left[\vartheta(-B_{\Delta x} v^{j+1} + r) + (1-\vartheta)(-B_{\Delta x} v^j + r)\right], \\
& \quad j = 0, \ldots, N-1 \\
v^0 &= (u_0(x_1), \ldots, u_0(x_{M-1})),
\end{aligned}
$$

woraus sich

$$
\left(\frac{1}{\Delta t} I + \vartheta B_{\Delta x}\right) v^{j+1} = \left(\frac{1}{\Delta t} I - (1-\vartheta) B_{\Delta x}\right) v^j + r
$$

für $j = 0, \ldots, N-1$ ergibt.

Ferner ist $\frac{1}{\Delta t} + \vartheta B_{\Delta x}$ eine M-Matrix für $\vartheta \in [0, 1]$. In jedem Schritt ist also für $\vartheta > 0$ ein tridiagonales Gleichungssystem zu lösen.

Beobachtungen

a) $\vartheta = 0$:
Es ergeben sich nur für sehr kleine Zeitschritte gute Näherungen. Bei zu großem Δt treten starke Oszillationen in x-Richtung auf, die sich in t-Richtung verstärken.

b) $\vartheta \geq \frac{1}{2}$:
Die für $\vartheta = 0$ beobachteten Oszillationen treten in der numerischen Lösung nicht mehr auf. Der Mehraufwand (d.h. das Lösen eines linearen Gleichungssystems in jedem Zeitschritt) wird aber durch die impliziten Verfahren mehr als wettgemacht. Die impliziten Verfahren sind also für die Anfangsrandwertaufgabe sehr wichtig.

10.2 Bemerkung (Lösbarkeit von (10-12) im nichtlinearen Fall $f(u, x, t)$).
Sei $f \in C^1(\mathbb{R} \times [0, 1] \times [0, T], \mathbb{R})$, und es gelte

$$
\frac{\partial f}{\partial u}(u, x, t) < \mu
$$

für $u \in \mathbb{R}$, $0 \leq x \leq 1$ und $0 \leq t \leq T$. Dann lässt sich Folgendes zeigen: Das implizite ϑ-Verfahren ($\vartheta > 0$)

$$
v^{j+1} = v^j + \Delta t \cdot \left[\vartheta F_{\Delta x}(v^{j+1}, t_{j+1}) + (1-\vartheta) F_{\Delta x}(v^j, t_j)\right], \; j = 0, \ldots, N-1
$$

ist auf jedem Zeitlevel $t_j = j\Delta t$ für alle Δt mit $\vartheta \mu \Delta t < 1$ durchführbar.

© Johannes Schropp 28. November 2008

b) Konsistenz der Differenzenverfahren

Wir bringen die Differentialgleichung aus Abschnitt a), d.h. die vollständige Diskretisierung auf die Form

$$T^h(u) = 0, \quad u \in \mathbb{R}^{\Omega_h}, \quad T^h : \mathbb{R}^{\Omega_h} \longrightarrow \mathbb{R}^{\Omega_h}$$

mit der Familie von Gittern $\{\Omega_h\}_{0<h\le h_0}$. Wir betrachten hier die Anfangsrandwertaufgabe

$$\begin{aligned} u_t &= u_{xx} + f(u,x,t) \text{ in } \Omega = (0,1)\times(0,T), \\ u(x,0) &= u_0(x) \text{ für } 0 \le x \le 1, \\ u(0,t) &= \gamma_0(t),\ u(1,t) = \gamma_1(t) \text{ für } 0 \le t \le T. \end{aligned} \tag{10-13}$$

Dazu setzen wir $h = (\Delta x, \Delta t)$ und $\Delta x = \frac{1}{M}$, $\Delta t = \frac{T}{N}$ sowie

$$\Omega_h = \{(x_i, t_j) = (i\Delta x, j\Delta t) \,|\, i = 1,\dots,M-1,\ j = 0,\dots,N\}.$$

Für $u \in \mathbb{R}^{\Omega_h}$ gilt

$$\begin{aligned} u &= (u_1^0,\dots,u_{M-1}^0,\dots,u_1^1,\dots,u_{M-1}^1,u_1^N,\dots,u_{M-1}^N), \\ u_i^j &= u(x_i,t_j) = u(i\Delta x, j\Delta t), \quad i = 1,\dots,M-1,\ j = 0,\dots,N. \end{aligned}$$

Wir finden dann $T^h(u) = 0$ mit

$$(T^h(u))_i^j = \begin{cases} u_i^0 - u_0(x_i), & \begin{cases} j = 0 \\ i = 1,\dots,M-1 \end{cases} \\ \frac{1}{\Delta t}\left(u_i^j - u_i^{j-1}\right) & \\ -\vartheta\left[\frac{1}{\Delta x^2}\left(u_{i-1}^j - 2u_i^j + u_{i+1}^j\right) + f(u_i^j, x_i, t_j)\right] & \begin{cases} j = 1,\dots,N \\ i = 1,\dots,M-1 \end{cases} \\ -(1-\vartheta)\left[\frac{1}{\Delta x^2}\left(u_{i-1}^{j-1} - 2u_i^{j-1} + u_{i+1}^{j-1}\right) + f(u^{j-1}, x_i, t_{j-1})\right] & \end{cases} \tag{10-14}$$

Hierbei ist

$$u_0^j = \gamma_0(t_j), \quad u_0^{j-1} = \gamma_0(t_{j-1}), \quad u_M^j = \gamma_1(t_j), \quad u_M^{j-1} = \gamma_1(t_{j-1})$$

zu setzen.

Nach Kapitel 8 haben wir die Begriffe Konsistenz, Stabilität und Konvergenz für $T^h(u) = 0$ definiert. Wir untersuchen die Konsistenz für Anfangswertaufgabe (10-13).

Sei $\overline{u} \in C(\overline{\Omega})$ mit $u_t, u_x, u_{xx} \in C(\overline{\Omega})$ eine klassische Lösung von (10-13) und $\overline{u}|_h = \overline{u}|_{\Omega_h}$ bezeichne die Einschränkung von $\overline{u}$ auf Ω_h. Wir untersuchen den Konsistenzfehler $\|T^h(\overline{u}_h)\|_\infty$ in der Maximumsnorm.

© Johannes Schropp 28. November 2008

Sei $w \in \mathbb{R}^{\Omega_h}$ mit

$$\|w\|_\infty = \max\{|w_i^j| \mid i = 1, \ldots, M-1, \quad j = 0, \ldots, N\}.$$

Für $j = 0$, $i = 1, \ldots, M-1$ gilt

$$(T^h(\overline{u}_h))_i^0 = \underbrace{\overline{u}_i^0}_{=u_0(x_i)} - u_0(x_i) = 0.$$

Für $j = 1, \ldots, N$, $i = 1, \ldots, M-1$ erhalten wir

$$\begin{aligned}
(T^h(\overline{u}_h))_i^j &= \frac{1}{\Delta t}\left(\overline{u}_i^j - \overline{u}_i^{j-1}\right) \\
&- \vartheta\left[\frac{1}{\Delta x^2}\left(\overline{u}_{i-1}^j - 2\overline{u}_i^j + \overline{u}_{i+1}^j\right) + f(\overline{u}_i^j, x_i, t_j)\right] \\
&- (1-\vartheta)\left[\frac{1}{\Delta x^2}\left(\overline{u}_{i-1}^{j-1} - 2\overline{u}_i^{j-1} + \overline{u}_{i+1}^{j-1}\right) + f(\overline{u}_i^{j-1}, x_i, t_{j-1})\right] \\
&= \frac{1}{\Delta t}(\overline{u}_i^j - \overline{u}_i^{j-1}) - \vartheta(\overline{u}_t)_i^j - (1-\vartheta)(\overline{u}_t)_i^{j-1} \\
&+ \vartheta\Bigg\{\underbrace{(\overline{u}_{xx})_i^j + f(\overline{u}_i^j, x_i, t_j)}_{=(\overline{u}_t)_i^j} - \frac{1}{\Delta x^2}\left(\overline{u}_{i-1}^j - 2\overline{u}_i^j + \overline{u}_{i+1}^j\right) - f(\overline{u}_i^j, x_i, t_j)\Bigg\} \\
&+ (1-\vartheta)\Bigg\{\underbrace{(\overline{u}_{xx})_i^{j-1} + f(\overline{u}_i^{j-1}, x_i, t_j)}_{=(\overline{u}_t)_i^{j-1}} - \frac{1}{\Delta x^2}\left(\overline{u}_{i-1}^{j-1} - 2\overline{u}_i^{j-1} + \overline{u}_{i+1}^{j-1}\right) \\
&- f(\overline{u}_i^{j-1}, x_i, t_{j-1})\Bigg\} \\
&= \frac{1}{\Delta t}\left(\overline{u}_i^j - \overline{u}_i^{j-1}\right) - \vartheta(\overline{u}_t)_i^j - (1-\vartheta)(\overline{u}_t)_i^{j-1} \\
&+ \vartheta\Bigg\{(\overline{u}_{xx})_i^j - \frac{1}{\Delta x^2}\left(\overline{u}_{i-1}^j - 2\overline{u}_i^j + \overline{u}_{i+1}^j\right)\Bigg\} \\
&+ (1-\vartheta)\Bigg\{(\overline{u}_{xx})_i^{j-1} - \frac{1}{\Delta x^2}\left(\overline{u}_{i-1}^{j-1} - 2\overline{u}_i^{j-1} + \overline{u}_{i+1}^{j-1}\right)\Bigg\}.
\end{aligned}$$

Nach Kapitel 8 gilt für ein $C > 0$

$$\left|\frac{1}{\Delta x^2}\left(\overline{u}_{i-1}^j - 2\overline{u}_i^j + \overline{u}_{i+1}^j\right) - (\overline{u}_{xx})_i^j\right| \leq C \cdot \Delta x^2,$$

falls $\frac{\partial^\nu \overline{u}}{\partial x^\nu} \in C(\overline{\Omega})$, $\nu = 1, 2, 3, 4$.

Zur Abschätzung des übrigen Terms betrachten wir

$$\frac{1}{\Delta t}(w(t+\Delta t) - w(t)) - \vartheta w'(t+\Delta t) - (1-\vartheta)w'(t).$$

© Johannes Schropp 28. November 2008

Mit

$$\begin{aligned} w(t+\Delta t) &= w(t) + \Delta t w'(t) + \frac{\Delta t^2}{2} w''(t) + O(\Delta t^3), \\ w'(t+\Delta t) &= w'(t) + \Delta t w''(t) + O(\Delta t^2) \end{aligned}$$

für $w \in C^3([0,T])$ folgt

$$\begin{aligned} &\frac{1}{\Delta t}(w(t+\Delta t) - w(t)) - \vartheta w'(t+\Delta t) - (1-\vartheta) w'(t) \\ &\quad = w'(t) + \frac{\Delta t}{2} w''(t) + O(\Delta t^2) - \vartheta(w'(t) + \Delta t w''(t) + O(\Delta t^2)) - (1-\vartheta) w'(t) \\ &\quad = \Delta t w''(t) \left(\frac{1}{2} - \vartheta \right) + O(\Delta t^2), \end{aligned}$$

falls $w \in C^3([0,T])$.

Somit folgt

$$\left| \frac{1}{\Delta t}\left(\overline{u}_i^j - \overline{u}_i^{j-1}\right) - (1-\vartheta)(\overline{u}_t)_i^{j-1} \right| \le C \begin{cases} \Delta t, & \text{falls } \overline{u}, \overline{u}_t, \overline{u}_{tt} \in C(\overline{\Omega}), \\ \Delta t^2, & \text{falls } \overline{u}, \overline{u}_t, \overline{u}_{tt}, \overline{u}_{ttt} \in C(\overline{\Omega}) \text{ und } \vartheta = \frac{1}{2} \end{cases}$$

10.3 Satz. *Das ϑ-Verfahren für $\vartheta \in [0,1]$ zur Anfangsrandwertaufgabe (10-13) ist konsistent der Ordnung 1 in Δt und 2 in Δx bezgl. $\|\cdot\|_\infty$, d.h.*

$$\|T^h(\overline{u}_h)\|_\infty = O(\Delta t + \Delta x^2)$$

an jeder klassischen Lösung von (10-13) mit $\frac{\partial^\nu \overline{u}}{\partial t^\nu} \in C(\overline{\Omega})$, $\nu = 1,2$, $\frac{\partial^\nu \overline{u}}{\partial x^\nu} \in C(\overline{\Omega})$, $\nu = 0,1,2,3,4$.

Das Crank-Nicholson Verfahren ist sogar $O(\Delta t^2 + \Delta x^2)$ konsistent, falls zusätzlich $\frac{\partial^3 \overline{u}}{\partial t^3} \in C(\overline{\Omega})$ gilt.

Unsere Rechnungen zeigen keinen Unterschied im Konsistenzverhalten für das explizite ($\vartheta = 0$) und das implizite ($\vartheta > 0$) Verfahren. Die im Abschnitt a) beobachteten drastischen Unterschiede müssen daher mit der Stabilität zusammenhängen, die wir als nächstes diskutieren.

10.4 Bemerkung. Die Glattheitsbedingungen für $\overline{u}$ von Satz 10.3 sind oft in den Ecken $x = 0$, $t = 0$ bzw. $x = 1$, $t = 0$ nicht erfüllt. Beispielsweise erfordert $\overline{u} \in C(\overline{\Omega})$ für (10-13) die Verträglichkeitsbedingungen (auch Kompatibilitätsbedingungen genannt) $\gamma_0(0) = u_0(0)$, $\gamma_1(0) = u_0(1)$ und u_t, $u_{xx} \in C(\overline{\Omega})$ die weiteren Bedingungen

$$\begin{aligned} \gamma_0'(0) &= u_0''(0) + f(\gamma_0(0), 0, 0) \quad (t = 0, x = 0), \\ \gamma_1'(0) &= u_0''(1) + f(\gamma_1(0), 1, 0) \quad (t = 0, x = 1). \end{aligned}$$

Jedoch garantiert die Theorie der parabolischen Differentialgleichungen trotzdem die Glattheit von $\overline{u}(x,t)$ für $t > 0$, $0 \le x \le 1$, falls die Daten f, u_0, γ_0, γ_1 hinreichend glatt sind.

© Johannes Schropp 28. November 2008

c) Stabilität und Konvergenz der Differenzenverfahren

Wir erinnern an die Stabilitätsdefinition für ein diskretes Modell.

10.5 Definition. Es sei $T^h(u) = 0$, $T^h : \mathbb{R}^{\Omega_h} \longrightarrow \mathbb{R}^{\Omega_h}$ ein diskretes Modell. Sei der Raum $\mathbb{R}^{\Omega_h}$ mit einer Norm $\|\cdot\|$ versehen. Gibt es ein von h unabhängiges $C > 0$ mit

$$\|u - v\| \leq C\|T^h(u) - T^h(v)\|, \quad \forall u, v \in \mathbb{R}^{\Omega_h}, \quad 0 < h \leq h_0, \quad h = (\Delta x, \Delta t),$$

so heißt das Modell T^h stabil.

Wir analysieren hier die Stabilität der Wärmeleitungsgleichung

$$\begin{aligned} u_t &= u_{xx} \text{ in } (0,1) \times (0,T), \\ u(x,0) &= u_0(x), \quad 0 \leq x \leq 1, \\ u(0,t) &= \gamma_0,\ u(1,t) = \gamma_1 \text{ für } 0 \leq t \leq T. \end{aligned} \tag{10-15}$$

Mit den Vektoren $v^j = (u_1^j, \ldots, u_{M-1}^j)$, $j = 0, \ldots, N$, sowie den Schrittweiten $\Delta x = \frac{1}{M} > 0$, $\Delta t = \frac{T}{N} > 0$ hat das ϑ-Verfahren die Form

$$\begin{aligned} v^0 &= r^0 = (u_0(x_1), \ldots, u_0(x_{M-1})), \\ v^{j+1} &= v^j + \Delta t \left[\vartheta F_{\Delta x}(v^{j+1}, t_{j+1}) + (1-\vartheta) F_{\Delta x}(v^j, t_j)\right] \end{aligned}$$

mit

$$\begin{aligned} F_{\Delta x}(v,t) &= -\frac{1}{\Delta x^2} \begin{pmatrix} 2 & -1 & 0 & \ldots & 0 & 0 & 0 \\ -1 & 2 & -1 & \ldots & 0 & 0 & 0 \\ 0 & -1 & 2 & \ldots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 2 & -1 & 0 \\ 0 & 0 & 0 & \ldots & -1 & 2 & -1 \\ 0 & 0 & 0 & \ldots & 0 & -1 & 2 \end{pmatrix} v + \frac{1}{\Delta x^2} \begin{pmatrix} \gamma_0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \gamma_1 \end{pmatrix} \\ &=: -\Gamma v + r^1, \end{aligned}$$

d.h.

$$v^{j+1} = v^j + \Delta t \left[\vartheta(-\Gamma v^{j+1} + r^1) + (1-\vartheta)(-\Gamma v^j + r^1)\right], \quad v^0 = r^0.$$

Letztere Formelzeile ist mit

$$(I + \Delta t \vartheta I) v^{j+1} = (I - \Delta t(1-\vartheta)\Gamma) v^j + \Delta t r^1, \quad v^0 = r^0$$

und somit mit

$$\underbrace{\left(\frac{1}{\Delta t} I + \vartheta \Gamma\right)}_{=:A} v^{j+1} = \underbrace{\left(\frac{1}{\Delta t} I - (1-\vartheta)\Gamma\right)}_{=:B} v^j + r^1,\ j = 0, \ldots, N-1 \tag{10-16}$$

© Johannes Schropp 28. November 2008

$$v^0 = r^0$$

äquivalent.

In Matrixform lautet die Gleichung (10-16)

$$\begin{pmatrix} I & 0 & 0 & \dots & 0 & 0 & 0 \\ -B & A & 0 & \dots & 0 & 0 & 0 \\ 0 & -B & A & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \ddots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & A & 0 & 0 \\ 0 & 0 & 0 & \dots & -B & A & 0 \\ 0 & 0 & 0 & \dots & 0 & -B & A \end{pmatrix} \begin{pmatrix} v^0 \\ v^1 \\ v^2 \\ \vdots \\ v^{N-2} \\ v^{N-1} \\ v^N \end{pmatrix} - \begin{pmatrix} r^0 \\ r^1 \\ r^1 \\ \vdots \\ r^1 \\ r^1 \\ r^1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{10-17}$$

Wir müssen die Stabilität von Dreiecksblockmatrizen der obigen Form untersuchen.

10.6 Lemma. *Sei eine Norm $\|\cdot\|_*$ auf $\mathbb{R}^m$ gegeben. Die Matrizen $A(\Delta t), B(\Delta t) \in \mathbb{R}^{m,m}$ mögen von $\Delta t \in (0,T)$ abhängen. $A(\Delta t)$ sei invertierbar und erfülle*

$$\|A(\Delta t)^{-1}\|_* \le C_1 \cdot \Delta t, \quad \forall \Delta t \in (0,T]. \tag{10-18}$$

Ferner existiere ein $C_2 > 0$ mit

$$\|(A^{-1}B)^n(\Delta t)\|_* \le C_2 \quad \textit{für} \quad 0 \le n \cdot \Delta t \le T, n \in \mathbb{N}. \tag{10-19}$$

Dann gilt für die $(n+1)$-blockige Matrix

$$H(\Delta t) = \begin{pmatrix} I & 0 & 0 & \dots & 0 & 0 & 0 \\ -B & A & 0 & \dots & 0 & 0 & 0 \\ 0 & -B & A & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \ddots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & A & 0 & 0 \\ 0 & 0 & 0 & \dots & -B & A & 0 \\ 0 & 0 & 0 & \dots & 0 & -B & A \end{pmatrix} \tag{10-20}$$

mit $0 \le n\Delta t \le T$ die Stabilitätsungleichung

$$\|v\|_{*,\infty} \le C_2(1 + C_1 T) \cdot \|H(\Delta t)v\|_{*,\infty}, \quad \forall v \in \mathbb{R}^{m(n+1)}, \quad 0 \le n\Delta t \le T, \tag{10-21}$$

wobei

$$\|v\|_{*,\infty} = \|(v^0, v^1, \dots, v^n)\|_{*,\infty} = \max\{\|v^i\|_* \mid i = 0, \dots, n\}.$$

© Johannes Schropp 28. November 2008

10.7 Bemerkung. Setzt man $C(\Delta t) = (A^{-1}B)(\Delta t)$, so lautet (10-19)

$$\|C(\Delta t)^n\|_* \le C_2 \quad \text{für} \quad 0 \le n\Delta t \le T.$$

Diese Bedingung wird in der Literatur oft zur Definition der Stabilität eines Verfahrens der Form (10-17) herangezogen. Sie erweist sich auch als notwendig für ein konsistentes und konvergentes Verfahren (Lax'scher Äquivalenzsatz).

10.8 Bemerkung. Hinreichend für (10-19) ist

$$\|(A^{-1}B)(\Delta t)\|_* \le 1 + C_3\Delta t. \tag{10-22}$$

Dann folgt nämlich

$$\begin{aligned} \|(A^{-1}B)^n(\Delta t)\|_* &\le \|(A^{-1}B)(\Delta t)\|_*^n \le (1 + C_3\Delta t)^n \\ &\le \exp(C_3\Delta t)^n = \exp(C_3 n\Delta t) \\ &\le \exp(C_3 T) =: C_2 \quad \text{für} \quad 0 \le n\Delta t \le T. \end{aligned}$$

Beweis von Lemma 10.6: Die Gleichung

$$H(\Delta t)v = g = (g^0, g^1, \ldots, g^n)$$

bedeutet

$$v^0 = g^0, \quad Av^j - Bv^{j-1} = g^j, \quad j = 1, \ldots, n$$

oder

$$v^j = A^{-1}Bv^{j-1} + A^{-1}g^j, \quad j = 1, \ldots, n, \quad v^0 = g^0.$$

Wir zeigen nun für $j \le n$, $n\Delta t \le T$ die Darstellung

$$v^j = (A^{-1}B)^j g^0 + \sum_{k=1}^{j} (A^{-1}B)^{j-k} A^{-1} g^k.$$

Der Beweis wird durch Induktion über j erbracht:
$j = 0$: $\quad v^0 = g^0$
$j - 1 \to j$:

$$\begin{aligned} v^j &= A^{-1}Bv^{j-1} + A^{-1}g^j \\ &= (A^{-1}B)\left[(A^{-1}B)^{j-1}g_0 + \sum_{k=1}^{j-1}(A^{-1}B)^{j-1-k}A^{-1}g^k\right] + A^{-1}g^j \\ &= (A^{-1}B)^j g_0 + \sum_{k=1}^{j-1}(A^{-1}B)^{j-k}A^{-1}g^j + A^{-1}g^j \end{aligned}$$

© Johannes Schropp 28. November 2008

$$= (A^{-1}B)^j g_0 + \sum_{k=1}^{j} (A^{-1}B)^{j-k} A^{-1} g^k.$$

Somit folgt

$$\begin{aligned}
\|v^j\|_* &= \underbrace{\|(A^{-1}B)^j\|_*}_{\leq C_2} \cdot \|g^0\|_* + \sum_{k=1}^{j} \underbrace{\|(A^{-1}B)^{j-k}\|_*}_{\leq C_2} \cdot \underbrace{\|A^{-1}\|_*}_{\leq C_1 \Delta t} \|g^k\|_* \\
&\leq C_2 \|g^0\|_* + \sum_{k=1}^{j} C_2 \cdot C_1 \cdot \Delta t \cdot \|g^k\|_* \\
&\leq C_2 \|g^0\|_* + C_1 C_2 \underbrace{\sum_{k=1}^{j}}_{\leq j\Delta t \leq T} \max\{\|g^k\|_* \,|\, k = 1, \ldots, n\} \\
&\leq C_2 \left(\|g^0\|_* + C_1 T \cdot \max\{\|g^k\|_* \,|\, k = 1, \ldots, n\} \right) \\
&\leq C_2 (1 + C_1 T) \|g\|_{*,\infty} = C_2 (1 + C_1 T) \|H(\Delta t) v\|_{*,\infty}, \quad j = 0, \ldots, n.
\end{aligned}$$

Dies liefert

$$\|v\|_{*,\infty} \leq C_2 (1 + C_1 T) \|H(\Delta t) v\|_{*,\infty}.$$

□

Wir prüfen nun (10-18), (10-19) für das ϑ-Verfahren. Wir haben

$$\begin{aligned}
A &= \frac{1}{\Delta t} I + \vartheta \Gamma, \quad B = \frac{1}{\Delta t} I - (1 - \vartheta)\Gamma, \\
m &= M - 1, \quad \|\cdot\|_* = \|\cdot\|_\infty, \\
\Gamma &= \frac{1}{\Delta x^2} \begin{pmatrix}
2 & -1 & 0 & \ldots & 0 & 0 & 0 \\
-1 & 2 & -1 & \ldots & 0 & 0 & 0 \\
0 & -1 & 2 & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 2 & -1 & 0 \\
0 & 0 & 0 & \ldots & -1 & 2 & -1 \\
0 & 0 & 0 & \ldots & 0 & -1 & 2
\end{pmatrix}.
\end{aligned}$$

$A(\Delta t) = \frac{1}{\Delta t} I + \vartheta \Gamma$, $\vartheta \geq 0$ ist eine L_0-Matrix und

$$A(\Delta t)\mathbb{I} = \frac{1}{\Delta t}\mathbb{I} + \frac{\vartheta}{\Delta x^2} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} \geq \frac{1}{\Delta t}\mathbb{I} > 0.$$

© Johannes Schropp 28. November 2008

Also ist $A(\Delta t)$ eine M-Matrix, und es gilt

$$\|A^{-1}(\Delta t)\|_\infty = \|A^{-1}(\Delta t)\mathbb{I}\|_\infty \leq \|\Delta t\mathbb{I}\|_\infty = \Delta t.$$

(10-18) gilt also mit $C_1 = 1$.

Es gilt $B(\Delta t) = \frac{1}{\Delta t} I - (1-\vartheta)\Gamma$. Somit ist die Bedingung $B(\Delta t) \geq 0$ mit

$$\frac{1}{\Delta t} - (1-\vartheta)\frac{2}{\Delta x^2} \geq 0$$

äquivalent, da alle außerdiagonalen Elemente von $B(\Delta t)$ größer gleich Null sind. Daraus folgt

$$1 \geq 2(1-\vartheta)\frac{\Delta t}{\Delta x^2},$$

was mit

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)}. \tag{10-23}$$

gleichbedeutend ist.

Wir setzen nun (10-23) voraus und schreiben weiter:

$$\begin{aligned} (A(\Delta t) - B(\Delta t))\mathbb{I} &= \vartheta\Gamma\mathbb{I} + (1-\vartheta)\Gamma\mathbb{I} = \Gamma\mathbb{I} \geq 0 \quad \Longrightarrow \\ A(\Delta t)\mathbb{I} &\geq B(\Delta t)\mathbb{I} \quad \Longrightarrow \\ \mathbb{I} &\geq (A^{-1}B)(\Delta t)\mathbb{I}, \text{ da } A^{-1}(\Delta t) \geq 0. \end{aligned}$$

(10-23) stellt nun $B(\Delta t) \geq 0$ sicher, und wir haben $(A^{-1}B)(\Delta t) \geq 0$. Somit folgt

$$\|(A^{-1}B)(\Delta t)\|_\infty = \|(A^{-1}B)(\Delta t)\mathbb{I}\|_\infty \leq \|\mathbb{I}\|_\infty = 1.$$

Also sind die Bedingungen (10-22) und (10-21) mit $C_3 = 0$ bzw. $C_2 = 1$ erfüllt. Nach Lemma 10.6 erfüllt $H(\Delta t)$ die Stabilitätsungleichung (10-21) mit der Stabilitätskonstanten $C_2(1 + C_1 T) = 1 + T$, die sowohl von Δx als auch von Δt unabhängig ist. Da T^h bis auf inhomogene Terme mit $H(\Delta t)$ übereinstimmt, folgt die Stabilitätsungleichung mit $\|\cdot\| = \|\cdot\|_{*,\infty} = \|\cdot\|_\infty$.

10.1 Satz Unter der Bedingung

$$\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)}$$

ist das ϑ-Verfahren für die Wärmeleitungsgleichung bezgl. $\|\cdot\|_\infty$ auf $\mathbb{R}^{\Omega_h}$ stabil. Genauer gilt

$$\|u - v\|_\infty \leq (1+T)\|T^h(u) - T^h(v)\|_\infty, \quad \forall u, v \in \mathbb{R}^{\Omega_h}, \quad h = (\Delta x, \Delta t),$$

wobei $T^h : \mathbb{R}^{\Omega_h} \longrightarrow \mathbb{R}^{\Omega_h}$ wie in (10-14) definiert ist.

© Johannes Schropp 28. November 2008

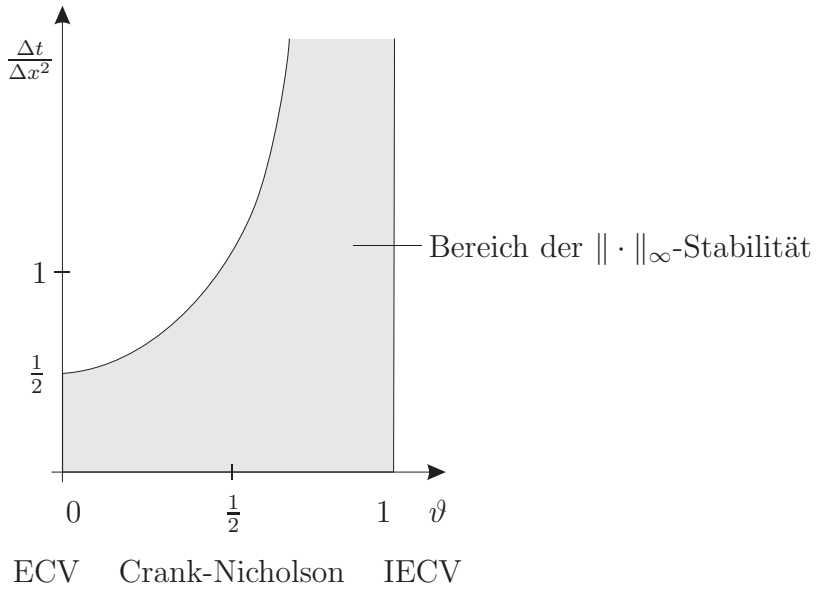


Abbildung 14: $\|\cdot\|_\infty$-Stabilität

10.9 Bemerkung. Die Restriktion $\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}$ für das explizite Verfahren und $\frac{\Delta t}{\Delta x^2} \leq \infty$ für das rein implizite Verfahren stimmen mit unseren Beobachtungen überein. Bei $\vartheta = \frac{1}{2}$ können wir jedoch $\frac{\Delta t}{\Delta x^2} \leq 1$ verletzen. Es lässt sich aber zeigen, dass bzgl. einer anderen Norm das ϑ-Verfahren für $\vartheta \in \left[\frac{1}{2}, 1\right]$ bedingungslos stabil ist. Diese Norm stellt eine Verallgemeinerung der $L^2((0,1))$-Norm auf die Gitterfunktionen dar.

10.10 Korollar. *Unter der Voraussetzung $\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-\vartheta)}$ ist das ϑ-Verfahren, $0 \leq \vartheta \leq 1$, bzgl. der Maximumsnorm konvergent der Ordnung 1 in Δt und 2 in Δx an der Lösung $\overline{u}$ der Wärmeleitungsgleichung, falls $\frac{\partial^\nu}{\partial t^\nu}\overline{u} \in C(\overline{\Omega})$, $\nu = 1, 2$, $\frac{\partial^\nu}{\partial x^\nu}\overline{u} \in C(\overline{\Omega})$, $\nu = 0, 1, 2, 3, 4$, $\Omega = (0,1) \times (0,T)$, d.h.*

$$\max\{|\overline{u}(x_i, t_j) - u^h(x_i, t_j)| \,|\, i = 1, \ldots, M-1,\ j = 0, \ldots, N\} \leq C \cdot (\Delta t + \Delta x^2),$$

wobei u^h die Lösung von $T^h(u) = 0$ bezeichnet. Für das Crank-Nicholson Verfahren gilt sogar

$$\max\{|\overline{u}(x_i, t_j) - u^h(x_i, t_j)| \,|\, i = 1, \ldots, M-1,\ j = 0, \ldots, N\} \leq C \cdot (\Delta t^2 + \Delta x^2),$$

falls zusätzlich $\frac{\partial^3}{\partial t^3}\overline{u} \in C(\overline{\Omega})$ gilt.

© Johannes Schropp 28. November 2008

11. Hyperbolische Differentialgleichungen

a) Differenzenverfahren für die Wellengleichung

Eines der bekanntesten Beispiele für hyperbolische Differentialgleichungen ist die Wellengleichung. Wir betrachten zunächst die analytische Lösung der Wellengleichung als reine Anfangswertaufgabe

$$\begin{aligned} u_{tt} &= c^2 u_{xx}, \quad x \in \mathbb{R},\, t \geq 0, \\ u(x,0) &= u_0(x), \quad u_t(x,0) = u_1(x),\, x \in \mathbb{R}. \end{aligned} \tag{11-1}$$

Dies modeliert einen unendlich ausgedehnten Stab oder eine unendliche Saite.

Mit dem Ansatz

$$u(x,t) = f_1(x+ct) + f_2(x-ct), \quad f_1, f_2 \in C^2(\mathbb{R})$$

findet man

$$\begin{aligned} u_{tt}(x,t) &= c^2 f_1''(x+ct) + (-c)^2 f_2''(x-ct) \\ &= c^2(f_1''(x+ct) + f_2''(x-ct)) = c^2 u_{xx}(x,t), \end{aligned}$$

d.h. $f_1(x+ct) + f_2(x-ct)$ genügt der Differentialgleichung (11-1).

Man kann nun auch die Anfangsbedingungen abgleichen und erhält

$$u(x,t) = \frac{1}{2}(u_0(x+ct) + u_0(x-ct)) + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(\xi)d\xi. \tag{11-2}$$

(11-2) ist als d'Alembert-Formel bekannt.

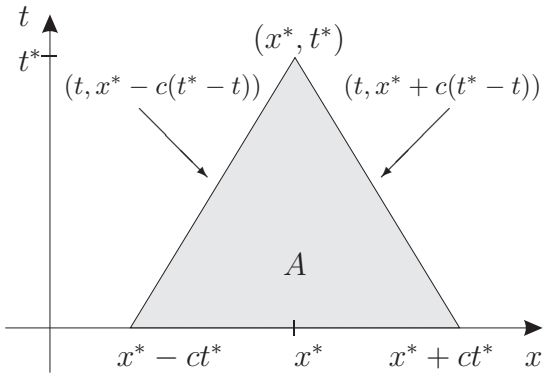


Abbildung 15: Abhängigkeitsbereich von (x^*, t^*)

Die auf der Abbildung (15) mit den Pfeilen gekennzeichneten Linien heißen Charakteristiken für (11-1).

© Johannes Schropp 28. November 2008

Gemäß der Lösungsdarstellung (11-2) hängt der Wert $u(x^*, t^*)$ von den Anfangsvorgaben im Intervall

$$I_{x^*,t^*} = [x^* - ct^*, x^* + ct^*]$$

ab. Dieses Intervall wird als Abhängigkeitsintervall des Punktes (x^*, t^*) bezeichnet. Das Dreieck A mit den Ecken $(x^* - ct^*, 0)$, $(x^* + ct^*, 0)$, (x^*, t^*) wird entsprechend das Abhängigkeitsdreieck von (x^*, t^*) genannt.

Fixieren wir umgekehrt ein festes $x^* \in \mathbb{R}$ für $t = 0$, so beeinflussen $u_0(x^*)$, $u_1(x^*)$ die Werte der Lösung u in

$$B := \{(x, t) \,|\, t \geq 0, x^* - ct \leq x \leq x^* + ct\}.$$

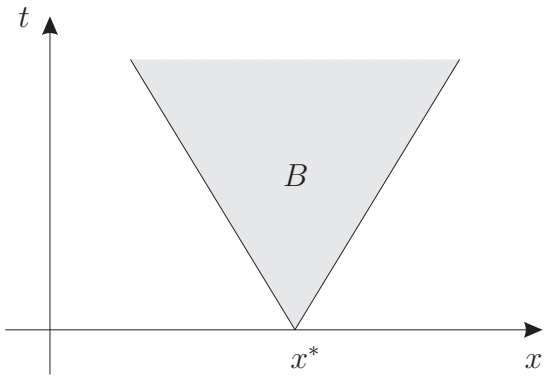


Abbildung 16: Einflussbereich von (x^*, t^*)

Wir betrachten nun die Anfangswertaufgabe

$$\begin{aligned} u_{tt} &= c^2 u_{xx}, \quad 0 \leq x \leq 1,\ t \geq 0, \\ u(x,0) &= u_0(x), \quad u_t(x,0) = u_1(x), \quad 0 \leq x \leq 1, \\ u(0,t) &= u(1,t) = 0, \quad t \geq 0. \end{aligned} \tag{11-3}$$

Ferner setzen wir die Verträglichkeitsbedingungen

$$u_0(0) = u_0(1) = 0, \quad u_1(0) = u_1(1) = 0$$

voraus.

Wir möchen nachweisen, dass die d'Alembertsche Lösungsformel (11-2) auch das Problem (11-3) löst. Dazu setzen wir u_0, u_1 2-periodisch folgenderweise auf ganz $\mathbb{R}$ fort:

$$\begin{aligned} u_i(1 + x) &= -u_i(1 - x), \quad 0 \leq x \leq 1, \\ u_i(x) &= u_i(x - 2n), \quad \text{falls } 2n \leq x \leq 2n + 2,\ n \in \mathbb{Z},\ i = 0, 1. \end{aligned}$$

Hiermit erreicht man

$$u_i(x) = -u_i(-x),$$

© Johannes Schropp 28. November 2008

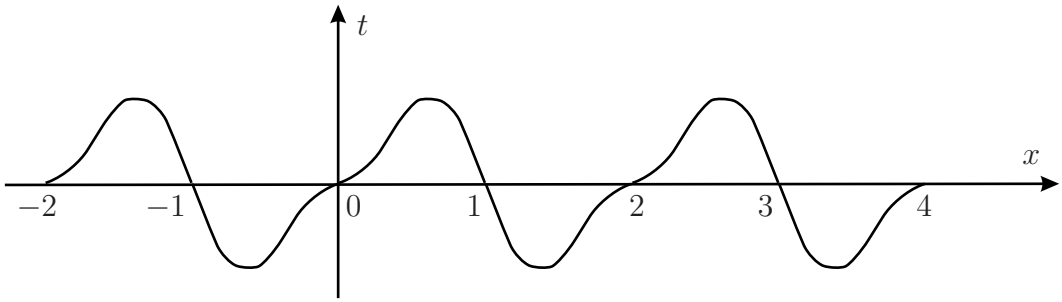


Abbildung 17: Periodische Fortsetzung

$$u_i(1+x) \;=\; -u_i(1-x), \quad x \in \mathbb{R},\, i = 0, 1,$$

d.h. die fortgesetzten Funktionen sind punktsymmetrisch zu $(0,0)$ und $(1,0)$.

Dann folgt für die Lösung der Wellengleichung (11-1) auf ganz $\mathbb{R}$ überdies

$$\begin{aligned} u(0,t) &= \frac{1}{2}\Big(\underbrace{u_0(ct)}_{=-u_0(-ct)} - u_0(-ct)\Big) + \frac{1}{2c}\int_{-ct}^{ct} u_1(\xi)d\xi \\ &= \frac{1}{2c}\left(\int_{-ct}^{0} -u_1(-\xi)d\xi + \int_0^{ct} u_1(\xi)d\xi\right). \end{aligned}$$

Mit der Transformation $\eta = -\xi$, $d\eta = -d\xi$ ergibt sich

$$u(0,t) = \frac{1}{2c}\left(\int_{ct}^{0} -u_1(\eta)(-1)d\eta + \int_0^{ct} u_1(\xi)d\xi\right) = 0.$$

Eine analoge Rechnung liefert $u(1,t) = 0$, d.h. u löst das Problem (11-3).

Allgemeiner betrachten wir nun

$$\begin{aligned} u_{tt} &= c^2 u_{xx} + f(u_x, u, x, t), \quad 0 \le x \le 1,\, 0 \le t \le T, \\ u(x,0) &= u_0(x), \quad u_t(x,0) \;=\; u_1(x), \quad 0 \le x \le 1, \\ u(0,t) &= \gamma_0(t), \quad u(1,t) = \gamma_1(t), \quad 0 \le t \le T. \end{aligned} \tag{11-4}$$

Wir führen auf $[0,1] \times [0,T]$ das Gitter

$$\begin{aligned} \Omega_h &= \{(i\Delta x, j\Delta t)\,|\, i = 1, \ldots, M-1,\, j = 0, \ldots, N\}, \\ \Delta x &= \frac{1}{M}, \quad \Delta t \;=\; \frac{T}{N}, \quad h \;=\; (\Delta x, \Delta t) \end{aligned}$$

ein.

Ohne den Umweg über die Linienmethode stellen wir direkt Differenzengleichungen für die Unbekannten $u_i^j = u(i\Delta x, j\Delta t)$, $i = 1, \ldots, M-1$, $j = 0, \ldots, N$ auf. Mit $x_i = i\Delta x$, $t_j = j\Delta t$ finden wir

$$\frac{1}{\Delta t^2}\left(u_i^{j+1} - 2u_i^j + u_i^{j-1}\right) \;=\; \frac{c^2}{\Delta x^2}\left(u_{i+1}^j - 2u_i^j + u_{i-1}^j\right)$$

© Johannes Schropp 28. November 2008

$$+ \quad f\left(\frac{1}{2\Delta x}\left(u_{i+1}^j - u_{i-1}^j\right), u_i^j, x_i, t_j\right) \tag{11-5}$$

für $i = 1, \ldots, M-1$ und $j = 1, \ldots, N$.

Für $j = 0$ setzen wir natürlich

$$u_i^0 = u_0(x_i), \quad i = 1, \ldots, M-1. \tag{11-6}$$

Zusätzlich setzen wir

$$u_0^j = \gamma_0(t_j), \quad u_M^j = \gamma_1(t_j), \quad j = 0, \ldots, N. \tag{11-7}$$

Uns fehlt noch aber eine Gleichung für u_i^1, z.B.

$$\frac{1}{\Delta t}\left(u_i^1 - u_i^0\right) = u_1(x_i), \quad i = 1, \ldots, M-1. \tag{11-8}$$

Während (11-5), (11-6), (11-7) bekanntermaßen die Konsistenzordnung $O(\Delta t^2 + \Delta x^2)$ haben, liegt für (11-8) nur $O(\Delta t)$ vor. Eine $O(\Delta t^2)$ Approximation gewinnt man aus der Entwicklung

$$\frac{u(x, \Delta t) - u(x, 0)}{\Delta t} = u_t(x, 0) + \frac{1}{2}\Delta t \cdot u_{tt}(x, 0) + O(\Delta t^2). \tag{11-9}$$

Für eine glatte Lösung $\overline{u}$ von (11-4) gilt nämlich

$$\begin{aligned} \overline{u}_{tt}(x, 0) &= c^2 \overline{u}_{xx}(x, 0) + f(\overline{u}_x, \overline{u}, x, 0) \\ &= c^2 u_0''(x) + f(u_0', u_0, x, 0). \end{aligned}$$

Man kann daher (11-8) durch

$$\frac{1}{\Delta t}\left(u_i^1 - u_i^0\right) = u_1(x_i) + \frac{\Delta t}{2}\left(c^2 u_0''(x_i) + f(u_0'(x_i), u_0(x_i), x_i, 0)\right) \tag{11-10}$$

für $i = 1, \ldots, M-1$ ersetzen.

Man bezeichnet (11-5)—(11-7), (11-8) oder (11-10) auch als explizites Verfahren.

Wir können aber (11-5) wiederum als Spezialfall eines von einem Parameter abhängigen Verfahrens ansehen. Setze dazu

$$\sigma_i^j = \frac{c^2}{\Delta x^2}\left(u_{i-1}^j - 2u_i^j + u_{i+1}^j\right) + f\left(\frac{1}{2\Delta x}\left(u_{i+1}^j - u_{i-1}^j\right), u_i^j, x_i, t_j\right) \tag{11-11}$$

und betrachte dann das im Allgemeinen implizite Verfahren

$$\frac{1}{\Delta t^2}\left(u_i^{j+1} - 2u_i^j + u_i^{j-1}\right) = \vartheta\sigma_i^{j+1} + (1 - 2\vartheta)\sigma_i^j + \vartheta\sigma_i^{j-1}, \quad \vartheta \in \left[0, \frac{1}{2}\right]. \tag{11-12}$$

© Johannes Schropp 28. November 2008

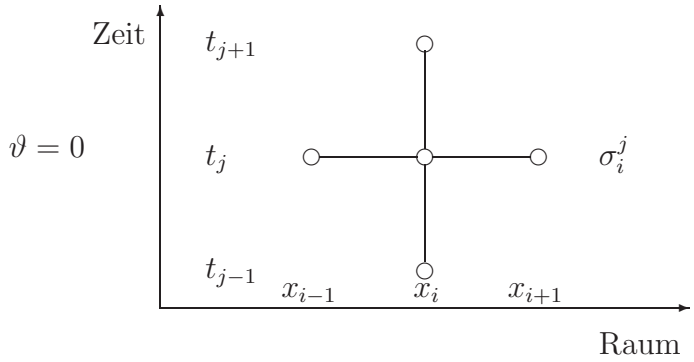


Abbildung 18: Differenzenstern für $\vartheta = 0$

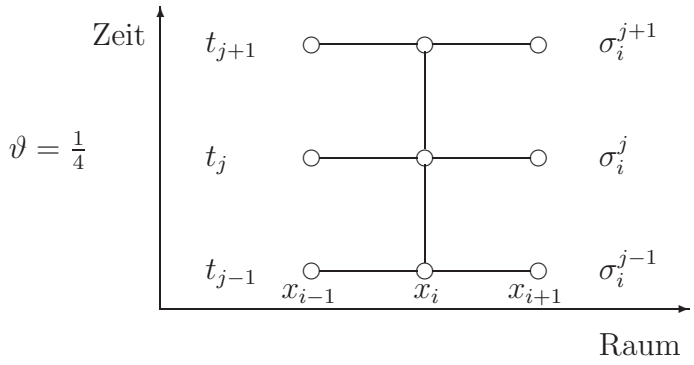


Abbildung 19: Differenzenstern für $\vartheta = \frac{1}{4}$

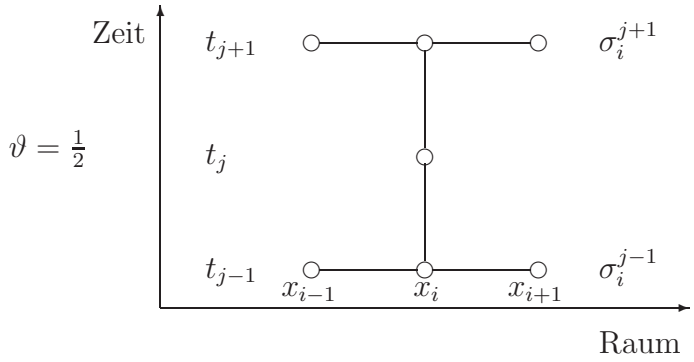


Abbildung 20: Differenzenstern für $\vartheta = \frac{1}{2}$

Die Anfangs- und Randbedingungen werden dabei wieder wie vorhin behandelt.

Auf den Abbildungen 18—20 werden die Differenzensterne für $\vartheta = 0, \frac{1}{4}, \frac{1}{2}$ graphisch dargestellt.

Zur Analyse des Konsistenzfehlers schreiben wir das numerische Verfahren wieder

© Johannes Schropp 28. November 2008

in der Form $T^h(u) = 0$, $T^h : \mathbb{R}^{\Omega_h} \longrightarrow \mathbb{R}^{\Omega_h}$ mit

$$(T^h(u))_i^j = \begin{pmatrix} u_i^0 - u_0(x_i),\ j = 0,\ i = 1,\dots,M-1 \\ \frac{1}{\Delta t}\left(u_i^1 - u_i^0\right) - u_1(x_i) - \frac{\Delta t}{2}\left[c^2 u_0''(x_i) + f(u_0'(x_i), u_0(x_i), x_i, 0)\right], \\ \qquad j = 1,\ i = 1,\dots,M-1 \\ \frac{1}{\Delta t^2}\left(u_i^j - 2u_i^{j-1} + u_i^{j-2}\right) - \left[\vartheta\sigma_i^j + (1-2\vartheta)\sigma_i^{j-1} + \vartheta\sigma_i^{j-2}\right], \\ \qquad j = 2,\dots,M,\ i = 1,\dots,M-1 \end{pmatrix}$$

und

$$\sigma_i^j = \frac{c^2}{\Delta x^2}\left(u_{i-1}^j - 2u_i^j + u_{i+1}^j\right) + f\left(\frac{1}{2\Delta x}\left(u_{i+1}^j - u_{i-1}^j\right), u_i^j, x_i, t_j\right).$$

11.1 Lemma. *An einer Lösung $\overline{u} \in C^4([0,1]\times[0,T])$ von (11-4) ist das parameterabhängige Verfahren $O(\Delta t^2 + \Delta x^2)$ konsistent, falls f einer Lipschitzbedingung bezüglich der Variablen u_x genügt.*

Beweis: Sei $\overline{u}_h$ die Restriktion von $\overline{u}$ auf Ω_h.

Für $j = 0$ gilt

$$T^h(\overline{u}_h)_i^0 = \overline{u}_i^0 - u_0(x_i) = 0, \quad i = 1,\dots,M-1.$$

Ist $j = 1$, so folgt

$$\begin{aligned} T^h(\overline{u}_h)_i^1 &= \frac{1}{\Delta t}\left(\overline{u}_i^1 - \overline{u}_i^0\right) - u_1(x_i) \\ &\quad \underbrace{- \frac{\Delta t}{2}\left[c^2 u_0''(x_i) + f(u_0'(x_i), u_0(x_i), x_i, 0)\right]}_{= c^2\overline{u}_{xx}(x_i,0) + f(\overline{u}_x(x_i,0), \overline{u}(x_i,0), x_i, 0) = \overline{u}_{tt}(x_i,0)} \\ &= \frac{\overline{u}(x_i,\Delta t) - \overline{u}(x_i,0)}{\Delta t} - \overline{u}_t(x_i,0) - \frac{\Delta t}{2}\overline{u}_{tt}(x_i,0) \\ &\overset{(11-9)}{=} O(\Delta t^2). \end{aligned}$$

Für $j = 2,\dots,M$ gehen wir wie folgt vor. Es sei

$$\overline{\sigma}_i^j = \frac{c^2}{\Delta x^2}\left(\overline{u}_{i-1}^j - 2\overline{u}_i^j + \overline{u}_{i+1}^j\right) + f\left(\frac{1}{2\Delta x}\left(\overline{u}_{i+1}^j - \overline{u}_{i-1}^j\right), \overline{u}_i^j, x_i, t_j\right).$$

Dann gilt nach Konstruktion

$$\begin{aligned} \overline{\sigma}_i^j &= c^2(\overline{u}_{xx})_i^j + f((\overline{u}_x)_i^j, \overline{u}_i^j, x_i, t_j) + O(\Delta x^2) \\ &= (\overline{u}_{tt})_i^j + O(\Delta x^2). \end{aligned}$$

Also folgt für den Konsistenzfehler

$$T^h(\overline{u}_h)_i^j = \frac{1}{\Delta t^2}\left(\overline{u}_i^j - 2\overline{u}_i^{j-1} + \overline{u}_i^{j-2}\right)$$

© Johannes Schropp 28. November 2008

$$
\begin{aligned}
- &\left[\vartheta(\overline{u}_{tt})_i^j + (1-2\vartheta)(\overline{u}_{tt})_i^{j-1} + \vartheta(\overline{u}_{tt})_i^{j-2}\right] + O(\Delta x^2)\\
= &\ (\overline{u}_{tt})_i^{j-1} + O(\Delta t^2) - \left[\vartheta(\overline{u}_{tt})_i^j + (1-2\vartheta)(\overline{u}_{tt})_i^{j-1} + \vartheta(\overline{u}_{tt})_i^{j-2}\right] + O(\Delta x^2)\\
= &\ -\vartheta \underbrace{\left[(\overline{u}_{tt})_i^j - 2(\overline{u}_{tt})_i^{j-1} + (\overline{u}_{tt})_i^{j-2}\right]}_{=O(\Delta t^2)} + O(\Delta t^2 + \Delta x^2)\\
= &\ O(\Delta t^2 + \Delta x^2),
\end{aligned}
$$

wobei hier

$$
\begin{aligned}
w(t-\Delta t) - 2w(t) + w(t+\Delta t) &= w(t) - \Delta t w'(t) + O(\Delta t^2)\\
&\quad - 2w(t) + w(t) + \Delta t w'(t) + O(\Delta t^2) = O(\Delta t^2 + \Delta x^2)
\end{aligned}
$$

für $w = \overline{u}_{tt}(x,\cdot)$ zu beachten ist. □

Im Falle $f = 0$ in (11-4), d.h. der linearen Wellengleichung, finden wir mit

$$
\begin{aligned}
v(t) &= (u(x_1,t),\ldots,u(x_{M-1},t)), \quad v^j = v(t_j) = v(j\Delta t),\\
\Gamma &= \frac{c^2}{\Delta x^2}\begin{pmatrix}
2 & -1 & 0 & \ldots & 0 & 0 & 0\\
-1 & 2 & -1 & \ldots & 0 & 0 & 0\\
0 & -1 & 2 & \ldots & 0 & 0 & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\
0 & 0 & 0 & \ldots & 2 & -1 & 0\\
0 & 0 & 0 & \ldots & -1 & 2 & -1\\
0 & 0 & 0 & \ldots & 0 & -1 & 2
\end{pmatrix},\\
r^j &= \frac{c^2}{\Delta x^2}\begin{pmatrix}
\gamma_0(t_j)\\ 0\\ 0\\ \vdots\\ 0\\ 0\\ \gamma_1(t_j)
\end{pmatrix}
\end{aligned}
$$

dann die Iteration

$$
\begin{aligned}
\frac{1}{\Delta t^2}\left(v^{j+1} - 2v^j + v^{j-1}\right) &= \vartheta\left(-\Gamma v^{j+1} + r^{j+1}\right) + (1-2\vartheta)(-\Gamma v^j + r^j)\\
&\quad + \vartheta(-\Gamma v^{j-1} + r^{j-1}), \quad j = 1,\ldots,N-1.
\end{aligned}
$$

Dies ist äquivalent zu

$$
\begin{aligned}
\left(\frac{1}{\Delta t^2} I + \vartheta\Gamma\right) v^{j+1} &= \frac{1}{\Delta t^2}\left(2v^j - v^{j-1}\right) + \vartheta r^{j+1} + (1-2\vartheta)(-\Gamma v^j + r^j)\\
&\quad + \vartheta(-\Gamma v^{j-1} + r^{j-1}), \quad j = 1,\ldots,N-1 \qquad (11\text{-}13)
\end{aligned}
$$

© Johannes Schropp 28. November 2008

mit den Startdaten

$$\begin{aligned} v^0 &= (u_0(x_1), \ldots, u_0(x_{M-1})), \\ \frac{1}{\Delta t}(v^1 - v^0) &= (u_1(x_1), \ldots, u_1(x_{M-1})) + \frac{\Delta t}{2} c^2 (u_0''(x_1), \ldots, u_0''(x_{M-1})). \end{aligned}$$

In jedem Fall erfordert (11-13) auf jedem Zeitlevel die Auflösung eines linearen Gleichungssystems mit der Matrix

$$A = \frac{1}{\Delta t^2} I + \vartheta \Gamma,$$

die offensichtlich eine M-Matrix ist.

b) Die Courant-Friedrichs-Levy Bedingung

Beim expliziten Differenzenverfahren ($\vartheta = 0$) hängt der Wert u_i^j, also die Näherung für $\overline{u}(x_i, t_j)$, von den Startwerten $u_{i-j}^0, u_{i-j+1}^0, \ldots, u_{i+j}^0$ ab. Man vergleiche hierzu folgenden Differenzenstern:

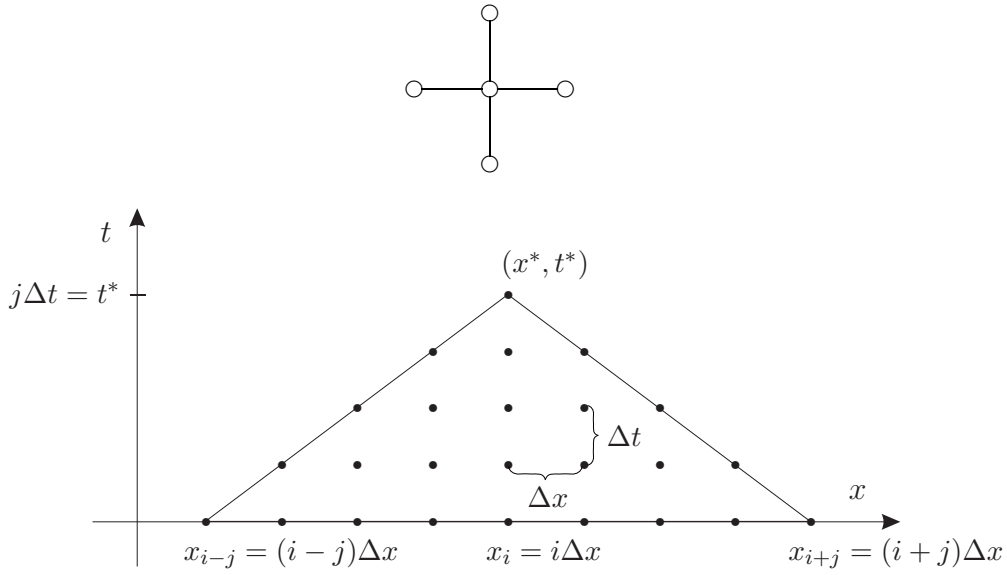


Abbildung 21: Numerisches Abhängigkeitsintervall

Man nennt $[(i-j)\Delta x, (i+j)\Delta x] = [x_{i-j}, x_{i+j}]$ das numerische Abhängigkeitsintervall des Punktes $(x^*, t^*) = (x_i, t_j)$.

Die Wellengleichung selbst liefert das Abhängigkeitsintervall $[x^* - ct^*, x^* + ct^*]$ für (x^*, t^*). Ist dieses Intervall größer als das numerische Abhängigkeitsintervall, so kann man keine Konvergenz des Verfahrens erwarten, denn Startwerte, die die exakte Lösung beeinflussen, werden für die numerische Lösung gar nicht berücksichtigt.

© Johannes Schropp 28. November 2008

Notwendig ist daher die Courant-Friedrichs-Levy Bedingung (auch CFL-Bedingung genannt).

Courant-Friedrich-Levy Bedingung:

Das numerische Abhängigkeitsintervall jedes Punktes $(x^*, t^*) = (x_i, t_j)$ umfasst das kontinuerliche Abhängigkeitsintervall.

In unserem Fall bedeutet dies $x_i - ct_j \geq x_{i-j}$ und $x_{i+j} \geq x_i + ct_j$.

Dies ist äquivalent zu

$$-ct_j \geq -x_j \text{ und } x_j \geq ct_j,$$

d.h.

$$c \cdot \Delta t \leq \Delta x,$$

wobei letzteres „CFL-Bedingung“ heißt.

Es lässt sich nun zeigen, dass die CFL-Bedingung im Falle der Wellengleichung

$$\begin{aligned} u_{tt} &= c^2 u_{xx}, \quad 0 \leq x \leq 1,\ 0 \leq t \leq T, \\ u(x,0) &= u_0(x), \quad u_t(x,0) = u_1(x), \quad 0 \leq x \leq 1, \\ u(0,t) &= \gamma_0(t), \quad u(1,t) = \gamma_1(t), \quad 0 \leq t \leq T \end{aligned} \tag{11-14}$$

für das explizite Differenzenverfahren ($\vartheta = 0$) eine Stabilitätsbedingung bezüglich der Norm $\|\cdot\|_{2,\infty}$ liefert. Dabei ist für $u \in \mathbb{R}^{\Omega_h}$

$$\|u\|_{2,\infty} = \max\left\{ \left| \sum_{i=1}^{M-1} (u_i^j)^2 \right|^{1/2} \,\middle|\, j = 0, \ldots, N \right\}.$$

11.2 Korollar. *Die Anfangsrandwertaufgabe (11-14) besitze eine Lösung*

$$\overline{u} \in C^4([0,1] \times [0,T]).$$

Dann ist das explizite Verfahren ($\vartheta = 0$) unter der CFL-Bedingung $\frac{\Delta t}{\Delta x} \leq \frac{1}{c}$ konvergent bezüglich $\|\cdot\|_{2,\infty}$ der Ordnung 2 in Δx und 2 in Δt.

© Johannes Schropp 28. November 2008

A. Anhang

a) Iterative Lösung großer Gleichungssysteme

Für die im Abschnitt 8.a) gefundenen ,,großen“ linearen Gleichungssysteme sind wegen des Speicherplatzes und Rechenaufwandes iterative Lösungsmethoden konkurrenzfähig.

Betrachte allgemein

$$Au = r, \quad r \in \mathbb{R}^N, \quad A \in \mathbb{R}^{N,N}. \tag{1-1}$$

A.1 Definition. $A = B - C$, $B, C \in \mathbb{R}^{N,N}$ heißt Zerlegung von A.

Sie heißt regulär, falls B invertierbar ist.

(1-1) ist äquivalent zu

$$Bu = Cu + r.$$

Dies legt die Iterationsvorschrift

$$Bu^{n+1} = Cu^n + r \Longleftrightarrow u^{n+1} = B^{-1}Cu^n + B^{-1}r, \quad u^0 \in \mathbb{R}^N, \quad n = 0, 1, 2, \ldots.$$

nahe.

A.2 Beispiele. Zu einer Matrix A mögen A_D, A_L und A_R die Diagonale, die strikte untere bzw. die strikte obere Dreieckmatrix bezeichnen. Präziser gilt

$$\begin{aligned} A_D &= \operatorname{diag}(A_{ii}, i = 1, \ldots, N), \\ (A_L)_{ij} &= \begin{cases} -A_{ij}, & i > j \\ 0, & \text{sonst} \end{cases}, \\ (A_R)_{ij} &= \begin{cases} -A_{ij}, & i < j \\ 0, & \text{sonst} \end{cases}. \end{aligned}$$

Mit dieser Zerlegung kann man A wie folgt darstellen:

$$A = A_D - A_L - A_R.$$

i) Jacobi-Verfahren: Es sei $B = A_D$, $C = A_L + A_R$. So folgt

$$A_D u^{n+1} = (A_L + A_R)u^n + r$$

oder

$$u^{n+1} = A_D^{-1}(A_L + A_R)u^n + A_D^{-1}r \tag{1-2}$$

© Johannes Schropp 28. November 2008

$$=: \quad J(A)u^n + A_D^{-1}r.$$

$J(A) = A_D^{-1}(A_L + A_R)$ heißt die Jacobi-Matrix von A.

ii) Gauss-Seidel Verfahren: Die Zerlegung $B = A_D - A_L$, $C = A_R$ führt zum sogenannten Gauss-Seidel Verfahren:

$$(A_D - A_L)u^{n+1} = A_R u^n + r$$

bzw.

$$u^{n+1} = (A_D - A_L)^{-1}A_R u^n + (A_D - A_L)^{-1}r. \tag{1-3}$$

$G(A) = (A_D - A_L)^{-1}A_R$ heißt die Gauss-Seidel Matrix von A.

Explizit lautet das Verfahren

$$A_{ii}u_i^{n+1} = -\sum_{j>1} A_{ij}u_j^n - \sum_{j<i} A_{ij}u_j^{n+1} + r_i, \quad i = 1, \ldots, N.$$

iii) Relaxationsverfahren: Sei $\omega > 0$ ein Parameter. Wir setzen $B = \frac{1}{\omega}A_D - A_L$, $C = \frac{1-\omega}{\omega}A_D + A_R$.

Das Relaxationsverfahren wird durch

$$\left(\frac{1}{\omega}A_D - A_L\right)u^{n+1} = \left(\frac{1-\omega}{\omega}A_D + A_R\right)u^n + r$$

bzw.

$$u^{n+1} = S_\omega(A)u^n + \left(\frac{1}{\omega}A_D - A_L\right)^{-1} r \tag{1-4}$$

gegeben, wobei

$$\begin{aligned} S_\omega &= \left(\frac{1}{\omega}(A_D - \omega A_L)\right)^{-1}\left(\frac{1-\omega}{\omega}A_D + A_R\right) \\ &= (A_D - \omega A_L)^{-1}((1-\omega)A_D + \omega A_R) \end{aligned}$$

die Relaxationsmatrix von A zum Parameter ω bezeichnet.

Man unterscheidet folgende Fälle:

- $\omega \in (0, 1)$: Unterrelaxation
- $w = 1$: Gauss-Seidel
- $\omega > 1$: Überrelaxation

© Johannes Schropp 28. November 2008

(1-4) heißt für $\omega > 1$ auch SOR[1]-Verfahren.

Die Durchführung ist ebenso einfach wie beim Gauss-Seidel Verfahren.

Wir diskutieren im Folgenden die optimale Wahl des Relaxationsparamaters ω. Diese orientiert sich an Spektralradius der Matrix $S_\omega(A)$.

A.3 Definition. Sei $T \in \mathbb{R}^{N,N}$. Dann heißt

$$\sigma(T) = \max\{|\lambda| \,|\, \lambda \in \mathbb{C} \quad \text{EW von } T\}$$

der Spektralradius von T.

A.4 Satz. *Sei $\|\cdot\|$ ein Vektorraum auf $\mathbb{R}^N$ und*

$$\|T\| = \sup\{\|Tu\| \,|\, \|u\| = 1\}$$

die zugeordnete Matrixnorm. Dann gilt für jede Matrix $T \in \mathbb{R}^N$

$$\sigma(T) = \lim_{n\to\infty} \|T^n\|^{\frac{1}{n}} = \inf\{\|T^n\|^{\frac{1}{n}}, n \in \mathbb{N}\}.$$

A.5 Satz. *Sei $T \in \mathbb{R}^{N,N}$, so sind gleichwertig*

i) $T^n \longrightarrow 0$ *für* $n \longrightarrow \infty$

ii) $\sigma(T) < 1$

iii) Es gibt eine Vektornorm $\|\cdot\|_*$ auf $\mathbb{R}^n$ mit $\|T\|_* < 1$

iv) Das Iterationsverfahren $u^{n+1} = Tu^n + r$ konvergiert für jedes $r \in \mathbb{R}^N$ und jeden Startvektor $u^0 \in \mathbb{R}^N$ gegen die eindeutige Lösung von $u = Tu + r$

Beweis:

i) $\Rightarrow$ *ii)* Da $\{T_n\}$ gegen 0 konvergiert, gibt es ein n_0 mit $\|T^{n_0}\|_\infty < 1$. Für dieses gilt $\|T^{n_0}\|_\infty^{1/n_0} < 1$. Somit folgt

$$\sigma(T) = \inf\{\|T^n\|_\infty^{1/n} \,|\, n \in \mathbb{N}\} \le \|T^{n_0}\|_\infty^{1/n_0} < 1.$$

ii) $\Rightarrow$ *iii)* Wähle ρ mit $\sigma(T) < \rho < 1$. Nach Satz A.4 existiert ein $n_0 \in \mathbb{N}$ mit

$$\|T^n\|_\infty^{1/n} \le \rho \quad \forall n \ge n_0.$$

[1] successive overrelaxation

© Johannes Schropp 28. November 2008

Setze $\|u\|_* = \sum_{i=0}^{n_0} \rho^{-i} \|T^i u\|_\infty$, $u \in \mathbb{R}^n$.
Sei $u \in \mathbb{R}^n$ mit $\|u\|_* = 1$, so gilt

$$\begin{aligned}
\|Tu\|_* &= \sum_{i=0}^{n_0} \rho^{-i} \|T^{i+1}u\|_\infty \\
&= \rho \sum_{i=0}^{n_0-1} \rho^{-(i+1)} \|T^{i+1}u\|_\infty + \rho^{-n_0} \|T^{n_0+1}u\|_\infty \\
&\leq \rho \sum_{i=0}^{n_0-1} \rho^{-(i+1)} \|T^{i+1}u\|_\infty + \rho^{-n_0} \|T^{n_0+1}\|_\infty \|u\|_\infty \\
&\leq \rho \sum_{i=0}^{n_0-1} \rho^{-(i+1)} \|T^{i+1}u\|_\infty + \rho^{-n_0} \rho^{n_0+1} \|u\|_\infty \\
&= \rho \sum_{i=1}^{n_0} \rho^{-i} \|T^i u\|_\infty + \rho \|u\|_\infty \\
&= \rho \left(\sum_{i=1}^{n_0} \rho^{-i} \|T^i u\|_\infty + \|u\|_\infty \right) \\
&= \rho \sum_{i=0}^{n_0} \rho^{-i} \|T^i u\|_\infty = \rho \|u\|_* = \rho,
\end{aligned}$$

d.h. $\|T\|_* \leq \rho < 1$.

iii) $\Rightarrow$ *iv)* Auf T ist der Banachsche Kontraktionssatz mit $L = \|T\|_* < 1$ anwendbar, denn

$$\|Tu - r - (Tv - r)\|_* = \|T(u - v)\| \leq \|T\|_* \|u - v\|_*.$$

iv) $\Rightarrow$ *i)* Wende *iv)* mit $r = 0$ an. Dies liefert $u^n \longrightarrow 0$ für $n \longrightarrow \infty$ für jede Folge $u^{n+1} = Tu^n$ und beliebiges $u^0 \in \mathbb{R}^N$. Damit gilt $u^n = T^n u^0 \longrightarrow 0$ für $n \longrightarrow \infty$ und alle $u_0 \in \mathbb{R}^N$, d.h. $T^n \longrightarrow 0$ für $n \longrightarrow \infty$.

□

Unsere obigen Beispiele ordnen sich *iv)* unter mit

$$T = \begin{cases} J(A), & \text{Jacobi-Verfahren,} \\ G(A), & \text{Gauss-Seidel Verfahren,} \\ S_\omega(A), & \text{SOR-Verfahren.} \end{cases}$$

Sei nun eine der Bedingungen von Satz A.5 erfüllt.

Betrachte die Iteration

$$u^{n+1} = Tu^n + r, \quad n = 0, 1, 2, \ldots, \quad u^0 \in \mathbb{R}^N.$$

© Johannes Schropp 28. November 2008

$\overline{u}$ löse dabei die Gleichung $u = Tu + r$. So ist der Fehler im n-ten Schritt durch $e^n = \overline{u} - u^n$ gegeben.

Ferner gilt

$$e^{n+1} = \overline{u} - u^{n+1} = T\overline{u} + r - (Tu^n + r) = T(\overline{u} - u^n) = Te^n, \quad n = 0, 1, 2, \ldots,$$

d.h. $e^n = T^n e^0$ mit $e^0 = \overline{u} - u^0$.

Hieraus erhalten wir für eine Vektornorm $\|\cdot\|$:

$$\|e^n\| \leq \|T^n\| \|e^0\|. \tag{1-5}$$

Setzen wir

$$\rho_n = \left(\frac{\|e^n\|}{\|e^0\|} \right)^{\frac{1}{n}},$$

so folgt mit (1-5)

$$\|e^n\| = \frac{\|e^n\|}{\|e^0\|} \|e^0\| = \rho_n^n \|e^0\| \leq \|T^n\| \|e^0\|.$$

Daher ist $\rho_n \leq \|T^n\|^{1/n}$.

Für große n schätzt also wegen $\lim_{n\to\infty} \|T^n\|^{1/n} = \sigma(T)$ der Spektralradius $\sigma(T)$ die durchschnittliche Verkleinerung des Fehlers pro Iterationsschritt ab. Für $\sigma(T) \approx 1$ wird die Konvergenz sehr langsam.

A.6 Bemerkung (Die e-Norm). Seien $u, e \in \mathbb{R}^n$, $e > 0$. Wir definieren die durch e gewichtete Vektornorm

$$\|u\|_e = \max \left\{ \frac{|u_i|}{e_i} \,\middle|\, i = 1, \ldots, N \right\}$$

sowie die zugeordnete Matrixnorm

$$\|T\|_e = \sup\{\|Tu\|_e \,|\, u \in \mathbb{R}^N, \|u\|_e = 1\}.$$

Für $T \in \mathbb{R}^{N,N}$, $T \geq 0$ gilt $\|T\|_e = \|Te\|_e$.

A.7 Satz. *Sei $A \in \mathbb{R}^{N,N}$ eine M-Matrix. Dann sind das Jacobi- und das Gauss-Seidel Verfahren zur Lösung von $Au = r$ global konvergent. Insbesondere gilt für jeden Vektor $e > 0$ mit $Ae > 0$*

$$\|G(A)\|_e \leq \|J(A)\|_e < 1. \tag{1-6}$$

Beweis: Wegen

$$u^{n+1} = G(A)u^n + (A_D - A_L)^{-1} r$$

© Johannes Schropp 28. November 2008

bzw.

$$u^{n+1} = J(A)u^n + A_D^{-1}r$$

genügt es, nur (1-6) zu zeigen, da alle verbleibenden Aussagen mit Satz A.5 folgen.

Sei $e > 0$ mit $Ae > 0$ gegeben (ein solches e existiert, da A eine M-Matrix ist). Dann gilt

$$0 < Ae = A_D e - A_L e - A_R e,$$

wobei $A_L, A_R \geq 0$ und A_D eine Diagonalmatrix mit positiven Diagonalelementen ist, da sich aus

$$0 < (Ae)_i = A_{ii}e_i + \sum_{j \neq i} \underbrace{A_{ij}}_{\leq 0} e_j \leq A_{ii}e_i$$

die Abschätzung $A_{ii} > 0$, $i = 1, \ldots, N$ ableiten lässt.

Es folgt somit

$$A_D e > (A_L + A_R)e$$

und

$$J(A)e = A_D^{-1}(A_L + A_R)e < e.$$

Da $J(A) \geq 0$, erhalten wir

$$\|J(A)\|_e = \|J(A)e\|_e < \|e\|_e = 1.$$

Wegen $(A_D - A_L)e \geq Ae > 0$ ist auch $A_D - A_L$ eine M-Matrix und $G(A) = (A_D - A_L)^{-1}A_R \geq 0$. Schließlich gilt

$$\begin{aligned}
J(A)e &= A_D^{-1}(A_L + A_R)e = (A_D - A_L)^{-1}(A_D - A_L)A_D^{-1}(A_L + A_R)e \\
&= (A_D - A_L)^{-1}\left\{A_R + A_L - A_L A_D^{-1}(A_L + A_R)\right\}e \\
&= G(A)e + (A_D - A_L)^{-1}\left\{A_L(I - A_D^{-1}(A_L + A_R))\right\} \\
&= G(A)e + \underbrace{(A_D - A_L)^{-1}\left\{A_L(I - J(A))\right\}e}_{\geq 0, \text{ denn } e - J(A)e > 0} \\
&\geq G(A)e,
\end{aligned}$$

was nun mit $\|G(A)\|_e \leq \|J(A)\|_e$ äquivalent ist. □

Als Modellproblem für unsere Verfahren betrachten wir die Randwertaufgabe

$$\begin{aligned}
-\Delta u &= g \text{ in } \Omega = (0,1)^2, \\
u &= \gamma \text{ auf } \partial\Omega.
\end{aligned}$$

Sei $A^h u = r \in \mathbb{R}^{\Omega_h}$, $h = \frac{1}{M}$, $\Omega_h = \{(ih, jh) \,|\, 1 \leq i, j \leq M - 1\}$ das klassische Differenzenverfahren hierzu.

© Johannes Schropp 28. November 2008

Die Matrix $A^h \in \mathbb{R}^{\Omega_h,\Omega_h}$ besitzt die Eigenvektoren $u^{kl} \in \Omega_h$

$$(u^{kl})_{ij} = \sin(ik\pi h)\sin(jl\pi h), \quad 1 \le i,j \le M-1.$$

Man rechnet leicht nach

$$A^h u^{kl} = \lambda_{kl} u^{kl}$$

mit

$$\lambda_{kl} = \frac{2}{h^2}\left[2 - \cos(k\pi h) - \cos(l\pi h)\right] > 0.$$

Also folgt

$$\begin{aligned}
A_D^h u^{kl} &= \frac{4}{h^2} u^{kl} = (A_L^h + A_R^h)u^{kl} + \lambda_{kl} u^{kl}, \\
J(A^h)u^{kl} &= \left(A_D^h\right)^{-1}\left(A_L^h + A_R^h\right) u^{kl} \\
&= \frac{h^2}{4}\left(A_D^h u^{kl} - \lambda_{kl} u^{kl}\right) \\
&= \frac{h^2}{4}\left(\frac{4}{h^2} - \lambda_{kl} u^{kl}\right) = \underbrace{\left(1 - \frac{\lambda_{kl} h^2}{4}\right)}_{=:\mu_{kl}} u^{kl}
\end{aligned}$$

mit

$$\begin{aligned}
\mu_{kl} &= 1 - \frac{1}{2}\left[2 - \cos(k\pi h) - \cos(l\pi h)\right] \\
&= \frac{1}{2}\left[\cos(k\pi h) + \cos(l\pi h)\right], \quad 1 \le k,l \le M-1.
\end{aligned}$$

Hieraus erhalten wir

$$\begin{aligned}
\sigma(J(A^h)) &= \max\{|\mu_{kl}| \,|\, 1 \le k,l \le M-1\} \\
&= \mu_{11} = \cos(\pi h) = 1 - \frac{1}{2}\pi^2 h^2 + O(h^4).
\end{aligned}$$

Das Jacobi-Verfahren ist also sehr langsam.

Wir untersuchen jetzt lineare Gleichungssysteme $Au = r$, für die die Matrix A die folgende Blockdiagonalform hat:

$$A = \begin{pmatrix}
B_1 & C_1 & 0 & \dots & 0 & 0 & 0 \\
A_2 & B_2 & C_2 & \dots & 0 & 0 & 0 \\
\dots & \dots & \dots & \vdots & \dots & \dots & \dots \\
0 & 0 & 0 & \dots & A_{p-1} & B_{p-1} & C_{p-1} \\
0 & 0 & 0 & \dots & 0 & A_p & B_p
\end{pmatrix},$$

© Johannes Schropp 28. November 2008

wobei B_i, $i = 1, \ldots, p$ Diagonalmatrizen der Größe N_i mit $\sum_{i=1}^{p} N_i = N$ sind. Eine solche Matrix heißt T-Matrix.

Das klassische Differenzenverfahren für

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

für beschränkte Gebiete Ω mit Schrittweite h liefert ein Gleichungssystem mit einer T-Matrix, wenn die Gitterpunkte z.B. wie folgt durchnummeriert werden:

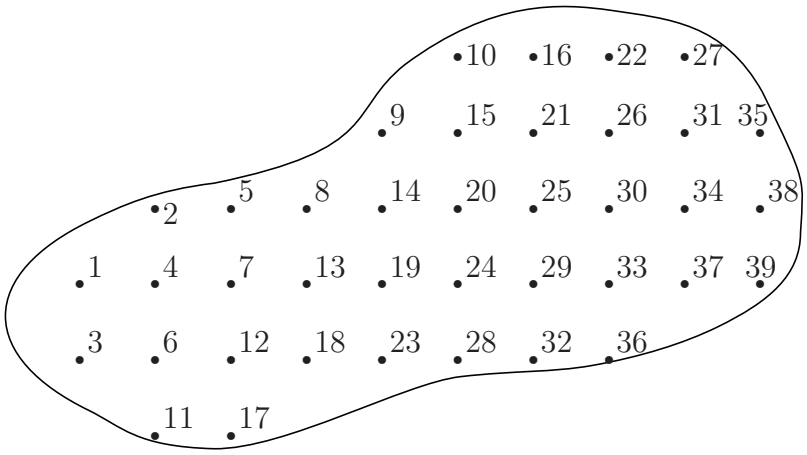


Abbildung 22: Diagonale Nummerierung, $p = 10$, $\Omega_h = \Omega \cap \mathbb{R}^2_h$

p ist die Anzahl der Diagonalen in Ω_h, die Größe von B_i ist die Anzahl der Gitterpunkte in der i-ten Diagonalen. B_i wird diagonal, da jeder Gitterpunkt auf der i-ten Diagonalen keinen Nachbarn in der i-ten Diagonalen hat, sondern jeweils höchstens zwei Nachbarn in der $(i-1)$-ten und $(i+1)$-ten Diagonalen.

Für T-Matrizen gilt der folgende

A.8 Satz (Householder). *Sei $A \in \mathbb{R}^{N,N}$ eine T-Matrix mit $A_{ii} \neq 0$, $i = 1, \ldots, N$. Dann gilt:*

a) $\sigma(G(A)) = \sigma(J(A))^2$

b) Hat $J(A)$ nur reelle Eigenwerte und ist $\sigma(J(A)) < 1$, so folgt mit

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \sigma(J(A))^2}} \in (1, 2)$$

die Relation

$$1 > \sigma(S_\omega(A)) > \sigma(S_{\omega_{opt}}(A)) = \omega_{opt} - 1$$

für alle $\omega \in (0, 2)$, $\omega \neq \omega_{opt}$.

© Johannes Schropp 28. November 2008

Aus dem letzteren Satz können wir folgende Konsequenzen ziehen. Wegen $A_R = A_L^T$ und somit $J(A) = J(A)^T$ ist die Matrix $J(A) = A_D^{-1}(A_L + A_R)$ für unser Modellproblem symmetrisch, d.h. sie besitzt nur reelle Eigenwerte.

Ferner gilt

$$\sigma(G(A)) = \cos \pi h = 1 - \frac{1}{2}h^2\pi^2 + O(h^4).$$

Es liefert also für das Gauss-Seidel Verfahren

$$\sigma(G(A)) = \sigma(J(A))^2 = \left(1 - \frac{1}{2}h^2\pi^2 + O(h^4)\right)^2 = 1 - \pi^2h^2 + O(h^4).$$

Dies ist zwar etwas besser als beim Jacobi-Verfahren, aber immer noch sehr langsam.

Für das SOR-Verfahren erhalten wir

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \sigma(J(A))^2}} = \frac{2}{1 + \sqrt{1 - \cos(\pi h)^2}} = \frac{2}{\sin(\pi h)}$$

sowie

$$\sigma(S_{\omega_{\text{opt}}}(A)) = \omega_{\text{opt}} - 1 = \frac{2 - (1 + \sin(\pi h))}{1 + \sin(\pi h)} = \frac{1 - \sin(\pi h)}{1 + \sin(\pi h)} = 1 - 2\pi h + O(h^2).$$

Obwohl dies immer noch nahe bei 1 liegt, ist es aber sehr viel besser als beim Gauss-Seidel Verfahren.

Betrachte nun

$$\begin{aligned} u^{n+1} &= Tu^n + r, \quad n = 0, 1, 2, \ldots, \\ e^n &= \overline{u} - u^n, \end{aligned}$$

wobei $\overline{u}$ die Gleichung $u = Tu + r$ löse.

Will man $\|e^n\| \leq \varepsilon \|e^0\|$ für vorgegebenes $\varepsilon > 0$ erreichen, so benötigt man wegen

$$\|e^n\| \leq \|T^n\| \|e^0\|$$

natürlich $\|T^n\| \leq \varepsilon$ bzw. $\|T^n\|^{1/n} \leq \varepsilon^{1/n}$.

Mit $\|T^n\|^{1/n} \approx \sigma(T)$ für große n werden wir auf

$$\sigma(T) \leq \varepsilon^{1/n} \text{ oder } \ln(\sigma(T)) \leq \frac{1}{n} \ln(\varepsilon)$$

bzw.

$$n \geq \frac{\ln(\varepsilon)}{\ln(\sigma(T))}$$

© Johannes Schropp 28. November 2008

geführt.

Die Größe $N = \left\lceil \frac{\ln(\varepsilon)}{\ln(\sigma(T))} \right\rceil$ gibt also etwa die Zahl der erforderlichen Iterationen an, um den Fehler um den Faktor ε zu verkleinern.

Für unser Modellproblem gilt

$$\begin{aligned} \ln \sigma(S_{\omega_{\text{opt}}}(A)) &\approx \ln(1 - 2\pi h) \approx -2\pi h, \\ \ln \sigma(J(A)) &\approx \ln\left(1 - \frac{1}{2}\pi^2 h^2\right) \approx -\frac{1}{2}\pi^2 h^2, \\ \ln \sigma(G(A)) &\approx 2 \ln(\sigma(J(A))) \approx -\pi^2 h^2. \end{aligned}$$

Man beachte dabei $\ln(1 - \alpha) = -\alpha + O(\alpha^2)$.

Die folgende Tabelle enthält neben den geschätzten Iterationszahlen und dem Gesamtaufwand an Multiplikationen, welcher sich daraus ergibt, dass für den Fall $\Omega = (0,1)^2$, $N = (M-1)^2$, $A \in \mathbb{R}^{(M-1)^2,(M-1)^2}$ ein Gauss-Seidel und ein Jacobi-Schritt jeweils etwa M^2 Multiplikationen, ein Relaxationsschritt aber etwa $3M^2$ benötigt.

ε		Jacobi	Gauss-Seidel	SOR, $\omega = \omega_{\text{opt}}$
10^{-3}	Iterationen	$1,4M^2$	$0,7M^2$	$1,1M$
	Multiplikationen	$1,4M^4$	$0,7M^4$	$3,3M^3$
10^{-6}	Iterationen	$2,8M^2$	$1,4M^2$	$2,2M$
	Multiplikationen	$2,8M^4$	$1,4M^4$	$6,6M^3$
10^{-9}	Iterationen	$4,2M^2$	$2,1M^2$	$3,3M$
	Multiplikationen	$4,2M^4$	$2,1M^4$	$9,9M^3$

© Johannes Schropp 28. November 2008