Differenzenverfahren für Partielle Differentialgleichungen

Wolf Hofmann

25. August 2005

Inhaltsverzeichnis

Vorbemerkung: Dieses Skript führt in die Theorie der Differenzen-Verfahren für Partielle Differntialgleichungen ein. Behandelt werden sollen hierbei die Wärmeleitungsgleichung, die Poissongleichung und die Wellengleichung. Die ersten Überlegungen behandeln den räumlich eindimensionalen Fall, der schon (fast) alle auftretenden Schwierigkeiten enthält. Später folgt eine Ausdehnung auf den räumlich zweidimensionalen Fall. Die Übertragung auf höhere Dimensionen ist dann fast nur Technik.
Es wird sich zeigen, daß zur numerischen Lösung der auftretenden Probleme eine ganze Menge Numerische Lineare Algebra nötig ist. Sie wird, soweit nicht aus der Vorlesung Numerik I+II für Studienanfänger bekannt, ebenfalls behandelt. Dies betrifft u.a. insbesondere eine relativ ausführliche Darstellung des Mehrgitter-Verfahrens, welches *das* Verfahren ist, das der numerischen Lösung der auftretenden linearen Gleichungssysteme angemessen ist.
Das Skript hat seinen Ursprung in einer Vorlesung über Numerische Behandlung von Partiellen Differentialgleichungen, die im Rahmen einer Gastvorlesung von Prof. Dr. G. Stoyan von der Elte-Universität Budapest gehalten worden ist. Aus seiner professionellen Erfahrung stammen auch viele Hinweise auf die Anwendbarkeit oder Nichtanwendbarkeit der verschiedenen Verfahren oder ihrer Varianten. Solche Hinweise sind in Lehrbüchern leider nur sehr sehr selten zu finden.
Dieses Skript enthält a) eine notwendigerweise beschränkte und subjektive Stoffauswahl aus dem Gebiet der Partiellen Differentialgleichungen (PDG) und b) – mit einiger Wahrscheinlichkeit – auch eine Reihe von Fehlern. Aus beiden Gründen ist es ungeeignet, ein Lehrbuch zu ersetzen. Sein Zweck ist es, den Hörer vom Zwang des Mitschreibens zu befreien. Es entbindet ihn nicht von der Notwendigkeit, den Stoff in Lehrbüchern nachzulesen und zu vertiefen und sich mit der notwendigen Referenzliteratur vertraut zu machen, die es ihm gestattet, Stoffgebiete nachzulesen, die nicht in der Vorlesung behandelt wurden.

Kapitel I

Parabolische Differentialgleichungen

§ 1 Die Wärmeleitungsgleichung

Unser Ziel ist die numerische Behandlung der allgmeinen Wärmeleitungsgleichung mittels Differenzenverfahren, die (eindimensional) wie folgt aussieht:

$$c_p \, \rho \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k \frac{\partial u}{\partial x}\right) - v \frac{\partial u}{\partial x} - qu + f$$

u	$\widehat{=}$	Temperatur
c_p	$\widehat{=}$	Wärmekapazität
ρ	$\widehat{=}$	Dichte
k	$\widehat{=}$	Wärmeleitfähigkeit
v	$\widehat{=}$	Geschwindigkeit
q	$\widehat{=}$	Abbaurate
f	$\widehat{=}$	Quellterm

Üblicherweise sind:

$c_p, \rho > 0$	häufig konstant, möglicherweise abhängig von u
$v = v(x, t)$	gegeben
$f = f(x, t)$	oft nichtlinear
$k = k(x, t)$	oft stückweise konstant (z.B. beim Übergang von einem Medium in ein anderes)

Nutzanwendung: Meteorologie, Luftverschmutzung, Bodenverschmutzung (Grundwasser).

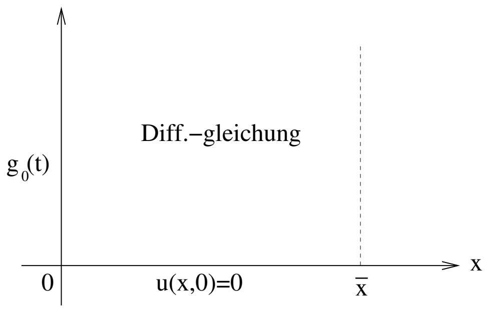
Ein typisches **Beispiel**: Aus einem Container (Tanker) strömt Gas aus.

$$\frac{\partial u}{\partial t} = D \frac{\partial^2 u}{\partial x^2} - v \frac{\partial u}{\partial x} - qu$$

Dabei sind

$u \hat{=}$ Gaskonzentration
$D \hat{=}$ Diffusionskoeffizient
$v \hat{=}$ Windgeschwindigkeit
$q \hat{=}$ Abbauglied

Container im Nullpunkt: Randbedingungen in $x = 0 : g_0(t) \hat{=}$ ausströmendes Gas
Null-Anfangsbedingungen



Parameter in der Differentialgleichung:
v, meßbar
D, schwierig, Vorwissen nötig (was ist im Tank drin? Diffusion in der Luft?)
q, abhängig von den chemischen Eigenschaften des Gases

Randbedingungen in $\bar{x}$? künstliche?
beliebt Neumann : $\frac{\partial u}{\partial x}(\bar{x}) = 0$
(primitiv, klappt aber oft).

Problem z.B. Tschernobyl: q war nicht bekannt, man wußte nicht, was drinnen war.

Ein weiteres Anwendungsbeispiel: Das Börsenverhalten von Aktien (Modell von Black–Scholes 1973) zur Unterstützung von Kauf und Verkauf kann durch eine parabolische Differentialgleichung modelliert werden.

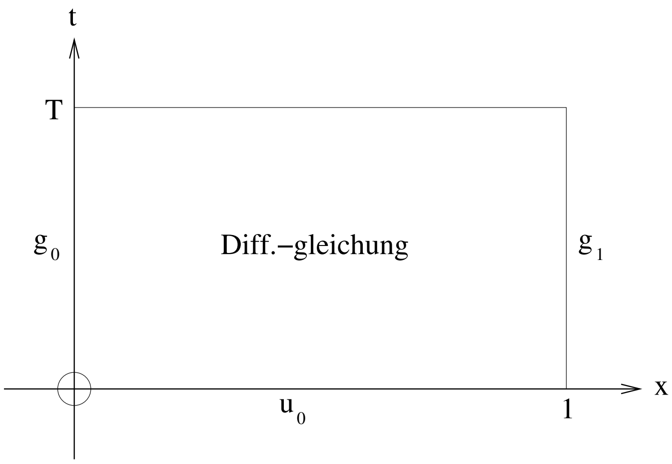
Wir beschänken uns in unseren Untersuchungen zunächst auf das einfachste Beispiel:

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f, \quad 0 < x < 1, \quad 0 < t < T$$

$$t = 0 : \; u(x, 0) = u_0(x), \quad 0 \leq x \leq 1$$

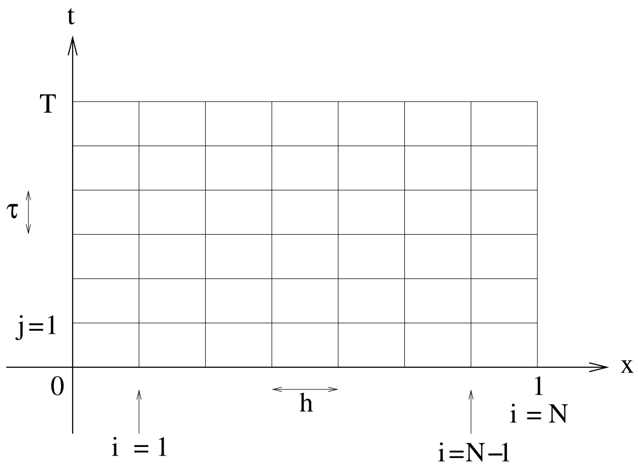
$$t > 0 : \; u(0, t) = g_0(t), \quad 0 \leq x \leq 1$$

$$u(1, t) = g_1(t).$$


t
T
g 0
Diff.−gleichung
g 1
x
u 0
1

§ 2 Diskretisierung (einfachster Fall)

Wir überziehen das Gebiet, in dem die Lösung berechnet werden soll, mit einem (nicht notwendigerweise quadratischen) Gitter. τ bzw. h sind Zeit- bzw. Ortsschrittweite.



$$\tau = T/m, \quad m \geq 1$$
$$h = 1/N, \quad N \geq 2 \qquad \text{(damit mindestens ein innerer Punkt existiert)}$$

und definieren eine

Gitterfunktion $u_i^j = u(x_i, t_j), \;\; x_i = i \cdot h, \;\; t_j = j \cdot \tau; \quad i, j = 0, 1, 2, \ldots$

und bezeichnen ihre Approximation mit

$$y_i^j \approx u(x_i, t_j), \;\; x_i = i \cdot h, \;\; t_j = j \cdot \tau; \quad i, j = 0, 1, 2, \ldots$$

Eine *Zeitschicht* umfaßt alle Gitterpunkte für ein festes t (alle auf einer Linie).

Ersetzt man die Zeitableitung durch den

vorwärtsgenommenen Differenzenquotienten

$$\frac{\partial u(x_i, t_j)}{\partial t} \approx \frac{y_i^{j+1} - y_i^j}{\tau}, \tag{2.1}$$

so kann man die Güte dieser Approximation abschätzen durch die Taylorentwicklung in Zeitrichtung (wir unterdrücken das x-Argument)

$$u(t_{j+1}) = u(t_j) + \tau \dot{u}(t_j) + \frac{\tau^2}{2} \ddot{u}(t_j) + 0(\tau^3) \quad \text{(Voraussetzung: } u \in C^2)$$

und erhält

$$\frac{u(t_{j+1}) - u(t_j)}{\tau} = \dot{u}(t_j) + \frac{\tau}{2} \ddot{u}(t_j) + 0(\tau^2), \tag{2.2}$$

eine (schlechte) Approximation der 1. Ordnung (lineare Konsistenzordnung).

Hier und im Folgenden bezeichnen wir mit $\dot{u}$ Ableitungen nach der Zeit- und mit u' Ableitungen nach der Ortsvariablen (entsprechend für höhere Ableitungen mit mehr Punkten bzw. Strichen).

Eine bessere Approximationseigenschaft erhält man durch den

Zentrale Differenzenquotienten:

1.Ableitung:

Wir betrachten den Differenzenquotienten (2.1) als eine Approximation für die Ableitung auf der Zeitstufe $t_{j+1/2}$.
Wir betrachten die Taylorentwicklung in den Punkten $t_{j+\frac{1}{2}\pm\frac{1}{2}}$ an der Stelle $t_{j+\frac{1}{2}}$ (und unterdrücken die Ortsabhängigkeit in der Bezeichnung).

$$u(t_{j+\frac{1}{2}\pm\frac{1}{2}}) = u(t_{j+\frac{1}{2}}) \pm \frac{\tau}{2}\dot{u}(t_{j+\frac{1}{2}}) + \frac{1}{2!}\left(\frac{\tau}{2}\right)^2 \ddot{u}(t_{j+\frac{1}{2}}) \pm \frac{\tau^3}{48}\dddot{u}_\pm \quad (\text{Vor: } u \in C^3).$$

Dabei werden mit $\dddot{u}_\pm = \dddot{u}(\xi_\pm)$ Zwischenstellen bezeichnet.
Subtrahiert man die Darstellungen für t_{j+1} und t_j , so folgt

$$u(t_{j+1}) - u(t_j) = \tau\dot{u}(t_{j+\frac{1}{2}}) + \frac{\tau^3}{24}(\dddot{u}_+ + \dddot{u}_-).$$

$\dddot{u}$ ist stetig und nach dem Zwischenwertsatz existiert dann eine Zwischenstelle z , so daß $2\dddot{u}(z) = \dddot{u}(\xi_+) + \dddot{u}(\xi_-)$. Damit folgt

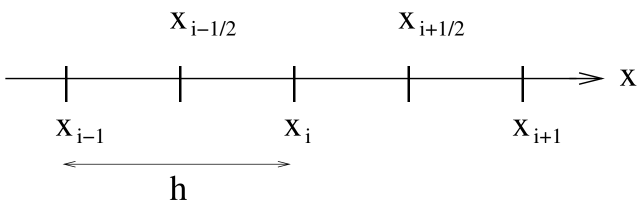
$$\text{(2.3)} \qquad \frac{u(t_{j+1}) - u(t_j)}{\tau} = \dot{u}(t_{j+\frac{1}{2}}) + \frac{\tau^2}{24}(\dddot{u}(\xi_+) + (\dddot{u}(\xi_-)) = \dot{u}(t_{j+\frac{1}{2}}) + \frac{\tau^2}{24}\dddot{u}(z)$$

also eine Approximation 2. Ordnung.

2. Ableitung

Zur Approximation der 2.ten Ableitung in Ortsrichtung verwenden wir den zentralen Differenzenquotienten in x_i , gebildet aus den zentralen Differenzenquotienten für die ersten Ableitungen in den Punkten $x_{i-\frac{1}{2}}$ und $x_{i+\frac{1}{2}}$.

Zentraler Differenzenquotient für u''



$$\text{(2.4)} \qquad \frac{\partial^2 u(x_i, t_j)}{\partial x^2} \approx \frac{\frac{y^j_{i+1} - y^j_i}{h} - \frac{y^j_i - y^j_{i-1}}{h}}{h} = \frac{y^j_{i+1} - 2y^j_i + y^i_{i-1}}{h^2}$$

Um die Approximationsgüte abzuschätzen benutzen wir wieder die Taylorentwicklung: (die Indizes bezeichnen die x_i-Werte, das t-Argument wird unterdrückt.)

$$u_{i\pm 1} = u_i \pm h u_i' + \frac{h^2}{2} u_i'' \pm \frac{h^3}{6} u_i''' + \frac{h^4}{24} u_i^{(4')} \pm \frac{h^5}{5!} u_i^{(5')} + \frac{h^6}{6!} u_\pm^{(6')}$$

$$\Longrightarrow$$

$$u_{i+1} + u_{i-1} = 2u_i + h^2 u_i'' + \frac{h^4}{12} u_i^{(4')} + 2\frac{h^6}{6!} u_\pm^{(6')}$$

$\pm$ als Indizes bezeichnen Zwischenstellen.

Damit folgt

$$\text{(2.5)} \quad \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = \begin{cases} u_i'' + \frac{h^2}{12} u_i^{(4')} + \frac{h^4}{720} (u_+^{(6')} + u^{(6')})_- & \text{falls} \quad u \in C^6 \\ u_i'' + \frac{h^2}{24} (u_+^{(4')} + u_-^{(4')}) & \text{falls} \quad u \in C^4 \end{cases}$$

Die Indizes „+“ und „−“ bezeichnen die Werte an Zwischenstellen $x \pm \vartheta_\pm h, |\vartheta_\pm| \leq 1$.

$$= \begin{cases} u_i'' + \frac{h^2}{12} u_i^{(4')} + \frac{h^4}{360} u^{(6')}(z) & \text{falls} \quad u \in C^6 \\ u_i'' + \frac{h^2}{12} u^{(4')}(z) & \text{falls} \quad u \in C^4 \end{cases}$$

jeweils an einer Zwischenstelle z,

also Approximationen 2. Ordnung.

Oft werden folgende **Bezeichnungen** benutzt.

$$\text{(2.6)} \quad y_i^j \begin{cases} \text{Indizes oben für die Zeitschicht} \\ \text{Indizes unten für die Ortsschicht} \end{cases}$$

oder indexlos

(2.7)

$\boldsymbol{y}$ für eine Zeitschicht (z.B. t_j)
$\hat{\boldsymbol{y}}$ für die folgende Zeitschicht (z.B. t_{j+1})
$\bar{\boldsymbol{y}}$ für die vorhergehende Zeitschicht (also t_{j-1})

Die entsprechenden Differenzenquotienten für eine beliebige, aber feste Zeitschicht t_j, werden wie folgt abgekürzt:

vorwärtsgenommener Differenzenquotient:

$$\dot{u}(x_i, t_j) \approx \frac{y_i^{j+1} - y_i^j}{\tau} =: y_{t,i}^j \qquad \frac{\hat{\boldsymbol{y}} - \boldsymbol{y}}{\tau} \approx \frac{\partial u}{\partial t}$$

rückwärtsgenommener Differenzenquotient:

$$\dot{u}(x_i, t_j) \approx \frac{y_i^j - y_i^{j-1}}{\tau} =: y_{\bar{t},i}^j \qquad \frac{\boldsymbol{y} - \bar{\boldsymbol{y}}}{\tau} \approx \frac{\partial u}{\partial t}$$

Entsprechend wird bezeichnet auf der Zeitschicht t_j mit

$$\begin{array}{lcl} y^j_{x,i} & \hat{=} & \text{vorwärts genommener Differenzenquotient in } x\text{–Richtung im Punkt } x_i \\ y^j_{\bar{x},i} & \hat{=} & \text{rückwärts genommener Differenzenquotient in } x\text{–Richtung im Punkt } x_i \end{array}$$

und entsprechend der zentrale Differenzenquotient 2. Ordnung

$$\frac{\partial^2 u(x_i,t_j)}{\partial x^2} \approx \frac{y^j_{i+1} - 2y^j_i + y^j_{i-1}}{h^2} =: y^j_{\bar{x}x,i}$$

bzw. ohne Auszeichnung der speziellen Schichten

$$\frac{\partial^2 u}{\partial x^2} \approx \boldsymbol{y}_{\bar{x}x}$$

Man beachte, daß $\boldsymbol{y}_{\bar{x}x}$ in $\bar{x}$ und x symmetrisch ist: $\boldsymbol{y}_{\bar{x}x} = \boldsymbol{y}_{x\bar{x}}$. Im Sinne der Hintereinanderausführung gilt

$$(y_{x,i})_{\bar{x},\,i} = (y_{\bar{x},i})_{x,\,i}\,, \quad \text{denn}$$

$$(y_{x,i})_{\bar{x},\,i} = \left(\frac{y_{i+1}-y_i}{h}\right)_{\bar{x},\,i} = \left(\frac{\frac{y_{i+1}-y_i}{h} - \frac{y_i-y_{i-1}}{h}}{h}\right) = \left(\frac{y_i-y_{i-1}}{h}\right)_{x,\,i} = (y_{\bar{x},\,i})_{x,\,i}\,.$$

Mit diesen Bezeichnungen erhalten wir für die näherungsweise Berechnung der Lösung von

$$\begin{aligned} \dot{u} &= u'' + f, \quad 0 < x < 1, \quad 0 < t < T \\ u(x,0) &= u_0(x) \\ u(0,t) &= g_0(t), \\ u(1,t) &= g_1(t) \end{aligned}$$

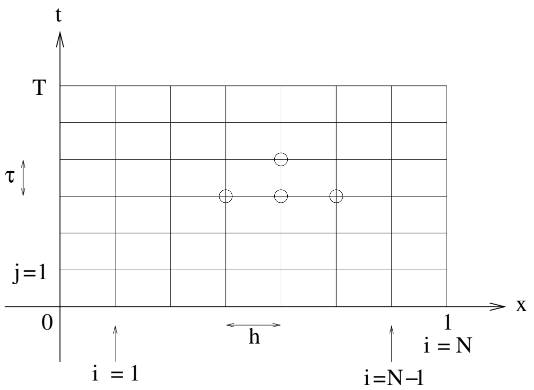
das **explizite Differenzenschema**

$$\begin{array}{lcl} y^j_{t,\,i} & = & y^j_{\bar{x}x,\,i} + f^j_i, \quad i = 1,\ldots,N-1, \;\; j = 0,1,\ldots, \\ y^0_i & = & u_0(x_i), \quad i = 0,1,\ldots,N \\ \left.\begin{array}{l} y^{j+1}_0 = g_0(t_{j+1}) \\ y^{j+1}_N = g_1(t_{j+1}) \end{array}\right\} & & j \geq 0 \end{array} \tag{2.8}$$

unter den Minimalvoraussetzungen

$$\begin{array}{rll} u \in C^4 & \text{bzgl. Ort} & \text{vgl. (2.5)} \\ \in C^2 & \text{bzgl. Zeit} & \text{vgl. (2.2).} \end{array}$$

Aus jeweils 3 Punkten einer Zeitschicht wird ein Punkt der nächst höheren Zeitschicht berechnet (vgl. Abb.).



Dies ist ein Verfahren, das im allgemeinen nicht viel taugt. Daß das so ist, wird durch Übungen belegt. Warum das so ist und wie man das verbessern kann, soll im folgenden untersucht werden.

Schreibt man die Differenzengleichung in Matrixgestalt, so erhält man für die Zeitschichten $j \geq 0$ (beachte: dies sind $N-1$ Gleichungen für jede Zeitschicht)

(2.9)
$$\begin{pmatrix} \dfrac{y_1^{j+1}-y_1^j}{\tau} \\ \vdots \\ \vdots \\ \vdots \\ \dfrac{y_{N-1}^{j+1}-y_{N-1}^j}{\tau} \end{pmatrix} = \underbrace{\begin{pmatrix} \frac{1}{h^2} & -\frac{2}{h^2} & \frac{1}{h^2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & & & \frac{1}{h^2} & -\frac{2}{h^2} & \frac{1}{h^2} \end{pmatrix}}_{N+1} \begin{pmatrix} y_0^j \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ y_N^j \end{pmatrix} + \begin{pmatrix} f_1^j \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ f_{N-1}^j \end{pmatrix}$$

In diesen Gleichungen werden die zu den Randwerten gehörigen Terme (die mit y_0^j, y_N^j) mit den Werten f_i^j zu einem Vektor $\boldsymbol{\varphi}^j$ zusammengefaßt.
Mit den Bezeichnungen

(2.10) $$\boldsymbol{y}^j = (y_1^j, \ldots, y_{N-1}^j)^T \quad \text{(nur die unbekannten Werte)}$$

und der (symmetrischen) $(N-1) \times (N-1)$ Matrix (man beachte die Vorzeichen)

(2.11) $$\boldsymbol{A}_h^0 := \frac{1}{h^2} \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ & & & & -1 & 2 \end{pmatrix} =: \frac{1}{h^2} \quad \text{tridiag}(-1, 2, -1)$$

erhält man für (2.9) die Darstellung

$$
(2.12)\quad \left\{
\begin{array}{l}
\boldsymbol{y}_t^j := \dfrac{\boldsymbol{y}^{j+1} - \boldsymbol{y}^j}{\tau} = -\boldsymbol{A}_h^0 \boldsymbol{y}^j + \boldsymbol{\varphi}^j, \quad \boldsymbol{\varphi}^j = \begin{pmatrix} \dfrac{1}{h^2} y_0^j & + & f_1^j \\ & & f_2^j \\ & & \vdots \\ & & f_{N-2}^j \\ \dfrac{1}{h^2} y_N^j & + & f_{N-1}^j \end{pmatrix}, \quad j \geq 0 \\
\text{oder (indexlos auch bzgl. der Zeitschicht)} \\
\boldsymbol{y}_t := \dfrac{1}{\tau}(\hat{\boldsymbol{y}} - \boldsymbol{y}) = -\boldsymbol{A}_h^0 \boldsymbol{y} + \boldsymbol{\varphi}
\end{array}
\right.
$$

Es wird sich zeigen, daß $\boldsymbol{A}_h^0$ von ausschlaggebender Bedeutung für dieses, und später zu betrachtende Verfahren ist, weshalb wie zunächst Eigenschaften dieser Matrix studieren. Dies bedingt auch einen Exkurs in die Numerische Lineare Algebra.

§ 3 Hilfmittel aus der linearen Algebra

Wir beginnen mit letzterem.

Satz 3.1 Neumann'sche Reihe
Sei $\boldsymbol{C} \in \mathbb{C}^{n\times n}, \|\boldsymbol{C}\| < 1$ (bezüglich einer Matrixnorm)
$\Rightarrow$

$$\exists \ (\boldsymbol{I}-\boldsymbol{C})^{-1} = \sum_{\nu=0}^{\infty} \boldsymbol{C}^{\nu}$$

Beweis:
1) Die Reihe konvergiert (absolut): $\|\sum \boldsymbol{C}^\nu\| \le \sum \|\boldsymbol{C}^\nu\| \le \sum \|\boldsymbol{C}\|^\nu, \quad \|\boldsymbol{C}\| < 1$.
2) $(\boldsymbol{I}-\boldsymbol{C}) \sum\limits_{\nu=0}^{\infty} \boldsymbol{C}^\nu = \sum\limits_{\nu=0}^{\infty} (\boldsymbol{I}-\boldsymbol{C})\boldsymbol{C}^\nu \underset{\substack{\uparrow \\ \text{abs.Kvg.}}}{=} \sum\limits_{\nu=0}^{\infty} \boldsymbol{C}^\nu - \sum\limits_{\nu=1}^{\infty} \boldsymbol{C}^\nu = \boldsymbol{C}^0 = \boldsymbol{I}$.
Entpsrechend für die rechtsseitige Inverse.

Satz 3.2
Sei $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ hermite'sch (d.h. $\boldsymbol{A} = \bar{\boldsymbol{A}}^T =: \boldsymbol{A}^*$)
$\Longrightarrow$

(3.1) Alle Eigenwerte $\lambda_i(\boldsymbol{A})$ von $\boldsymbol{A}$ sind reell.
(3.2) $\boldsymbol{A}$ besitzt ein orthonormales System von n Eigenvektoren (Basis).
(3.3) Extremaleigenschaft des *Rayleigh–Quotienten.*
Für alle $\boldsymbol{x} \in \mathbb{C}^n, \boldsymbol{x} \neq 0$ bzw. $\boldsymbol{y} \in \mathbb{C}^n$, $\|\boldsymbol{y}\|_2 = 1$ gilt
$\lambda_{\min}(\boldsymbol{A}) \le \dfrac{\bar{\boldsymbol{x}}^T \boldsymbol{A}\boldsymbol{x}}{\bar{\boldsymbol{x}}^T\boldsymbol{x}} = \bar{\boldsymbol{y}}^T\boldsymbol{A}\boldsymbol{y} \le \lambda_{\max}(\boldsymbol{A})$.
Die Grenzen werden für die Eigenvektoren zu $\lambda_{\min}$ bzw. $\lambda_{\max}$ angenommen.

Beweis (3.1): Die quadratische Form $\bar{\boldsymbol{x}}^T\boldsymbol{A}\boldsymbol{x}$ ist reell $\forall \boldsymbol{x} \in \mathbb{C}^n$:

$$\bar{\boldsymbol{x}}^T\boldsymbol{A}\boldsymbol{x} = \bar{\boldsymbol{x}}^T\bar{\boldsymbol{A}}^T\boldsymbol{x} = (\bar{\boldsymbol{A}}\bar{\boldsymbol{x}})^T\boldsymbol{x} = \boldsymbol{x}^T\bar{\boldsymbol{A}}\bar{\boldsymbol{x}} = \overline{\bar{\boldsymbol{x}}^T\boldsymbol{A}\boldsymbol{x}}, \quad \Longrightarrow$$

(3.4) $$\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \quad \Longrightarrow \quad \underbrace{\bar{\boldsymbol{x}}^T\boldsymbol{A}\boldsymbol{x}}_{\in\mathbb{R}} = \lambda \underbrace{\bar{\boldsymbol{x}}^T\boldsymbol{x}}_{>0 \text{ für } \boldsymbol{x}\neq 0} \quad \Longrightarrow \quad \lambda \in \mathbb{R}.$$

Beweis (3.2): Eine hermite'sche Matrix ist diagonalisierbar, besitzt also n linear unabhängige Eigenvektoren (Fischer: Lineare Algebra).

Wir zeigen zunächst: Eigenvektoren zu verschiedenen Eigenwerten sind orthogonal bezüglich des inneren Produkts $(\boldsymbol{x},\boldsymbol{y}) = \sum\limits_{i=1}^{n} x_i\bar{y}_i = \boldsymbol{y}^*\boldsymbol{x}, \quad \boldsymbol{y}^* = \bar{\boldsymbol{y}}^T$.

$$\begin{aligned} \boldsymbol{A}\boldsymbol{x}^1 = \lambda_1\boldsymbol{x}^1 \quad &\Longrightarrow \quad (\boldsymbol{A}\boldsymbol{x}^1,\boldsymbol{x}^2) = \lambda_1(\boldsymbol{x}^1,\boldsymbol{x}^2) \\ \boldsymbol{A}\boldsymbol{x}^2 = \lambda_2\boldsymbol{x}^2 \quad &\overset{(3.1)}{\Longrightarrow} \quad (\boldsymbol{x}^1,\boldsymbol{A}\boldsymbol{x}^2) = \lambda_2(\boldsymbol{x}^1,\boldsymbol{x}^2). \end{aligned}$$

Durch Subtraktion

$$(3.5) \qquad (\boldsymbol{A}\boldsymbol{x}^1, \boldsymbol{x}^2) - (\boldsymbol{x}^1, \boldsymbol{A}\boldsymbol{x}^2) = (\lambda_1 - \lambda_2)(\boldsymbol{x}^1, \boldsymbol{x}^2)$$

Für jede quadratische Matrix gilt (vgl. vorseitige Definition des inneren Produkts)

$$(3.6) \qquad (\boldsymbol{A}\boldsymbol{x}^1, \boldsymbol{x}^2) = \bar{\boldsymbol{x}}^{2^T} \boldsymbol{A}\boldsymbol{x}^1 = (\boldsymbol{A}^T \bar{\boldsymbol{x}}^2)^T \boldsymbol{x}^1 = \overline{(\bar{\boldsymbol{A}}^T \boldsymbol{x}^2)}^T \boldsymbol{x}^1 = (\boldsymbol{x}^1, \bar{\boldsymbol{A}}^T \boldsymbol{x}^2).$$

Falls $\bar{\boldsymbol{A}}^T = \boldsymbol{A}$, folgt damit aus (3.5): $(\boldsymbol{x}^1, \boldsymbol{x}^2) = 0$ wegen $\lambda_1 \neq \lambda_2$.

Da Eigenvektoren zum gleichen Eigenwert orthogonalisiert werden können (Verfahren von Ehrhardt–Schmidt), existiert ein orthogonales System von n Eigenvektoren $\boldsymbol{x}^\mu$ zu den Eigenwerten λ_μ, das normiert werden kann: $((\boldsymbol{x}^\nu, \boldsymbol{x}^\mu) = \delta_{\nu\mu})$.

Beweis (3.3): $\forall \boldsymbol{x} \in \mathbb{C}^n, \|\boldsymbol{x}\|_2 = 1 \quad \exists!$ Darstellung $\boldsymbol{x} = \sum_{\nu=1}^{n} \alpha_\nu \boldsymbol{x}^\nu, \quad \sum_{\nu=1}^{n} |\alpha_\nu|^2 = 1$

Damit folgt:

$$(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}) = \left(\sum \alpha_\nu \lambda_\nu \boldsymbol{x}^\nu, \sum \alpha_\nu \boldsymbol{x}^\nu\right) = \sum |\alpha_\nu|^2 \lambda_\nu \quad \begin{array}{l} \leq \lambda_{\max}(\boldsymbol{A}) \\ \geq \lambda_{\min}(\boldsymbol{A}). \end{array}$$

■

Aus diesem Satz erhalten wir

Folgerung 3.3

(3.7) $\boldsymbol{A} \in \mathbb{C}^{n\times n}, \boldsymbol{A} \geq 0 \quad$ (*positiv semidefinit*: $\bar{\boldsymbol{x}}^T \boldsymbol{A}\boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}) \geq 0 \quad \forall \boldsymbol{x} \in \mathbb{C}^n$)
$\Longrightarrow$ alle Eigenwerte sind ≥ 0.

(3.8) Ist $\boldsymbol{A}$ hermite'sch, so gilt sogar

$$\boldsymbol{A} \text{ positiv} \quad \left\{ \begin{array}{ll} \text{definit} & (\boldsymbol{A} > 0) \\ \text{semidefinit} & (\boldsymbol{A} \geq 0) \end{array} \right.$$

$$\Longleftrightarrow \quad \text{Alle Eigenwerte von } \boldsymbol{A} \text{ sind } \left\{ \begin{array}{l} > \\ \geq \end{array} \right. 0.$$

(3.9) $\boldsymbol{A} \in \mathbb{C}^{n\times n}$: Die Spektralnorm $\|\boldsymbol{A}\|_2 = \|\boldsymbol{A}\|_S = +\sqrt{\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{A})}$ ist der euklidischen Vektornorm $\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^*\boldsymbol{x}}$ zugeordnet.

(3.10) Ist $\boldsymbol{A} = \bar{\boldsymbol{A}}^T$, (hermitesch), so gilt

$$\|\boldsymbol{A}\|_S = \max_i |\lambda_i(\boldsymbol{A})|$$

Beweis (3.7) folgt aus (3.4).

Beweis (3.8) folgt aus (3.3). (Extremaleigenschaft des Rayleigh–Quotienten)

Beweis (3.9)

$$\|\boldsymbol{A}\|_2 = \sup_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = \sup_{\boldsymbol{x}\neq 0} \sqrt{\frac{(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})}} \overset{(3.6)}{=} \sup_{\boldsymbol{x}\neq 0} \sqrt{\frac{(\boldsymbol{x}, \bar{\boldsymbol{A}}^T \boldsymbol{A}\boldsymbol{x})}{(\boldsymbol{x}, \boldsymbol{x})}} \overset{(3.3)}{=} \sqrt{\lambda_{\max}(\bar{\boldsymbol{A}}^T \boldsymbol{A})},$$

denn $\bar{\boldsymbol{A}}^T\boldsymbol{A}$ ist hermite'sch.

Beweis (3.10)
Aus (3.9) folgt für $\boldsymbol{A} = \boldsymbol{A}^T : \boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}^2$, also $\|\boldsymbol{A}\|_2 = \sqrt{\lambda_{\max}(\boldsymbol{A}^2)}$ und $\boldsymbol{A}^2 \geq 0$, denn

$$\boldsymbol{A}^2\boldsymbol{y} = \mu\boldsymbol{y} \quad\Longrightarrow\quad \mu(\boldsymbol{y},\boldsymbol{y}) = (\boldsymbol{A}^2\boldsymbol{y},\boldsymbol{y}) = (\boldsymbol{A}\boldsymbol{y},\boldsymbol{A}\boldsymbol{y}) \geq 0 \quad\Longrightarrow\quad \mu \geq 0 \quad\Longrightarrow\quad \boldsymbol{A}^2 \geq 0.$$

$$\text{Weiterhin}\quad \boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x} \quad\Longrightarrow\quad \boldsymbol{A}^2\boldsymbol{x} = \boldsymbol{A}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}(\lambda\boldsymbol{x}) = \lambda^2\boldsymbol{x},$$

also

(3.11) Ist λ Eigenwert von $\boldsymbol{A}$, so ist λ^2 Eigenwert von $\boldsymbol{A}^2$ zum gleichen EV.

Wir zeigen weiter

(3.12) Hat $\boldsymbol{A}^2$ nur nichtnegative Eigenwerte μ_i, so ist $\sqrt{\mu_i}$ oder $-\sqrt{\mu_i}$ Eigenwert von $\boldsymbol{A}$,

denn (charakt. Polynom)

$$\begin{aligned} 0 = \det(\boldsymbol{A}^2 - \mu\boldsymbol{I}) &= \det\{(\boldsymbol{A} + \sqrt{\mu}\,\boldsymbol{I})(\boldsymbol{A} - \sqrt{\mu}\,\boldsymbol{I})\} \\ &= \det(\boldsymbol{A} + \sqrt{\mu}\,\boldsymbol{I})\det(\boldsymbol{A} - \sqrt{\mu}\,\boldsymbol{I}), \end{aligned}$$

und einer der Faktoren muß verschwinden.

Damit gilt $\quad \lambda_{\max}(\boldsymbol{A}^2) = \left(\max_i |\lambda_i(\boldsymbol{A})|\right)^2 \overset{(3.9)}{\Longrightarrow} (3.10)$. ■

Eigenschaften von $\boldsymbol{A}_h^0$

Satz 3.4
$\boldsymbol{A}_h^0 > 0$ (d.h. *positiv definit*: $(\boldsymbol{A}_h^0\boldsymbol{x},\boldsymbol{x}) > 0 \quad \forall\boldsymbol{x} \neq 0$). (vgl. (2.11))

Zum Beweis benutzen wir ein gewichtetes Skalarprodukt

Definition 3.5 **gewichtetes Skalarprodukt** $(\ ,\ )_{(0,h)}$

$$(\boldsymbol{y},\boldsymbol{v})_{(0,h)} := \sum_{i=1}^{N-1} y_i\bar{v}_i h, \quad h > 0.$$

Bedeutung der Indizierung:

$0 \mathrel{\hat{=}}$ es werden keine Ableitungen benutzt,
$h \mathrel{\hat{=}}$ Ortsschrittweite.

Beweis Satz 3.4

Wir zeigen $(\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})_{(0,h)} > 0 \quad \forall \boldsymbol{y} \neq 0$. (Beachte dazu: $(\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})_{(0,h)} = (\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})h$.)

Mit $\boldsymbol{y} = (y_1,\ldots,y_{N-1})^T$, entsprechend den Dimensionen von $\boldsymbol{A}_h^0$, ist ("auf jeder Zeitschicht")

$$y_{\bar{x}x,\,i} = \frac{y_{i+1}-2y_i+y_{i-1}}{h^2} = (y_{x,\,i} - y_{\bar{x},\,i})\frac{1}{h}.$$

Wir setzen ergänzend fest: $y_0 = y_N = 0$, (entspricht Nullrandbedingung)

Dann gilt $(\boldsymbol{A}_h^0\boldsymbol{y})_i = -\boldsymbol{y}_{\bar{x}x,\,i}, \quad i = 1,\ldots,N-1$ (vgl. (2.9)-(2.12)) und somit

$$\begin{aligned}
(\boldsymbol{A}_h^0,\boldsymbol{y},\boldsymbol{y})_{(0,h)} &= -\sum_{i=1}^{N-1} y_{\bar{x}x,i}\bar{y}_i h, \quad \text{und mit } y_{\bar{x}x,i} = (y_{x,i} - y_{\bar{x},i})\frac{1}{h} \\
&= -\sum_{i=1}^{N-1} y_{x,i}\bar{y}_i + \sum_{i=1}^{N-1} y_{\bar{x},i}\bar{y}_i \\
&\quad \text{(das ist eine Art diskreter partieller Integration)} \\
&= -\sum_{i=1}^{N-1} y_{x,i}\bar{y}_i + \sum_{i=0}^{N-2} y_{x,i}\bar{y}_{i+1}, \quad \text{da } y_{\bar{x},i} = \frac{y_i - y_{i-1}}{h} = y_{x,i-1} \\
&= -\sum_{i=0}^{N-1} y_{x,i}\bar{y}_i + \sum_{i=0}^{N-1} y_{x,i}\bar{y}_{i+1}, \quad \text{da } y_0 = y_N = 0 \\
&= \sum_{i=0}^{N-1} (\bar{y}_{i+1} - \bar{y}_i) y_{x,i} = \sum_{i=0}^{N-1} \frac{(\bar{y}_{i+1} - \bar{y}_i)}{h}\, h\, y_{x,i}
\end{aligned}$$

$$(3.13)\quad (\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})_{(0,h)} = \sum_{i=0}^{N-1} (y_{x,i})^2 h \quad \Longrightarrow \quad \boldsymbol{A}_h^0 \geq 0 \quad \text{(positiv semidefinit)}$$

Wegen

$$\begin{aligned}
(\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})_{(0,h)} = 0 \quad &\Longleftrightarrow \quad y_{x,\,i} = 0, \;\; i = 0,\ldots N-1 \\
&\Longleftrightarrow \quad y_{i+1} - y_i = 0, \;\; i = 0,\ldots N-1 \quad \Longleftrightarrow \quad y_i = 0\;\forall\, i \quad \text{wegen } y_0 = y_N = 0 \\
&\Longleftrightarrow \quad \boldsymbol{y} = 0
\end{aligned}$$

$$\Longrightarrow \boldsymbol{A}_h^0 > 0. \blacksquare$$

Folgerung 3.6
$\boldsymbol{A}_h^0 > \boldsymbol{0} \quad \Longrightarrow \quad \boldsymbol{A}_h^0$ ist invertierbar

denn alle Eigenwerte von $\boldsymbol{A}_h^0$ sind $> 0 \quad \Longrightarrow \quad$ die Diagonale der Jordannormalform $J(\boldsymbol{A})$ enthält nur positive Elemente $\quad \Longrightarrow \quad \exists J(\boldsymbol{A})^{-1} \Longrightarrow \exists \boldsymbol{A}^{-1}$.

Als nächstes untersuchen wir die Eigenwerte von $\boldsymbol{A}_h^0$. Aus ihrer Kenntnis, bzw. aus der Kenntnis von Schranken für die Eigenwerte kann man u.a. Abschätzungen für die Norm von $\boldsymbol{A}$ und $\boldsymbol{A}^{-1}$ ableiten (vgl. (3.10)).

Wir stellen eine Analogie zu der kontinuierlichen Aufgabe her.

Eigenwerte von $\boldsymbol{A}_h^0$

Die Motivation zur Berechnung der Eigenwerte folgt aus der kontinuierlichen Aufgabe.

Hilfssatz: Die Eigenwertaufgabe

$$-u'' =: Lu = \lambda u, \quad x \in (0,1), \quad u(0) = u(1) = 0$$

$$\begin{array}{rlcll} \text{hat die Eigenfunktion} & y^k(x) & = & \sin(k\pi x), & k \in \boldsymbol{Z} \\ \text{zu den Eigenwerten} & \lambda_k & = & k^2\pi^2, & k \in \boldsymbol{Z} \end{array}$$

Beweis: Durch Nachrechnen ■

Das diskrete Analogon zur obigen Eigenwertaufgabe lautet

$$-y_{\bar{x}x,i} =: (\boldsymbol{A}_h^0 \boldsymbol{y})_i = \lambda y_i, \quad i = 1, \ldots, N-1, \quad y_0 = y_N = 0 \tag{3.14}$$

Aus Analogiegründen zur kontinuierlichen Aufgabe führen wir die Vektoren $\tilde{\boldsymbol{y}} := (0, \underbrace{y_1, \ldots, y_{N-1}}_{\boldsymbol{y}}, 0)^T = (0, \boldsymbol{y}, 0)^T \in \mathbb{R}^{n+1}$ ein und betrachten mit der Matrix

$$\widetilde{\boldsymbol{A}}_h^0 := \begin{pmatrix} 1 & & & & & \\ -\frac{1}{h^2} & \frac{2}{h^2} & -\frac{1}{h^2} & & \boldsymbol{0} & \\ & -\frac{1}{h^2} & \ddots & \ddots & & \\ & & \ddots & \ddots & -\frac{1}{h^2} & \\ & \boldsymbol{0} & & -\frac{1}{h^2} & \frac{2}{h^2} & -\frac{1}{h^2} \\ & & & & & 1 \end{pmatrix} = \begin{pmatrix} 1 & \cdots & & \\ -\frac{1}{h^2} & \vdots & & \vdots \\ & \vdots & \boldsymbol{A}_h^0 & \vdots \\ & \vdots & \cdots & \vdots & -\frac{1}{h^2} \\ & & & & 1 \end{pmatrix}$$

die Eigenwertaufgabe

$$\widetilde{\boldsymbol{A}}_h^0 \tilde{\boldsymbol{y}} = \lambda \tilde{\boldsymbol{y}}. \tag{3.15}$$

Durch Untersuchung dieser Eigenwertaufgabe beweisen wir

Lemma 3.7

$\boldsymbol{A}_h^0$ hat die Eigenvektoren

$$\boldsymbol{y}^h = (y^h_{1,\ldots} y^h_{N-1})^T, \quad y_i^h = \sin(k\pi x_i), \quad x_i = i \cdot h = i \cdot \tfrac{1}{N}, \quad i = 1, \ldots, N-1,$$

zu den Eigenwerten

$$\lambda_k^h = \frac{4}{h^2} \sin^2\left(\frac{k\pi h}{2}\right), \quad k = 1, \ldots, N-1, \quad 0 < \lambda_1 < \ldots < \lambda_{N-1}.$$

Beweis:
Mit $x_0 = 0 \cdot h$, $x_N = Nh = 1$ ist $y_0^k := \sin(k\pi x_0) = 0$, $y_N^k := \sin(k\pi x_N) = 0$.
Wir rechnen (komponentenweise) nach, daß die Vektoren

$$\tilde{\boldsymbol{y}}^k := (\tilde{y}_0^k, \ \ldots, \tilde{y}_N^k)^T = (0, \boldsymbol{y}^k, 0)^T, k = 1, \ldots, N-1$$

Eigenvektoren von $\widetilde{\boldsymbol{A}}_h^0$ sind zu den Eigenwerten λ_k^h, $k = 1, 2, \ldots, N-1$.

Die erste und letzte Zeile von (3.15) lauten $0 = \lambda \cdot 0$.

Für $i = 1, \ldots, N-1$ gilt (wir unterdrücken die oberen Indizes $k = 1, \ldots, N-1$)

$$\begin{aligned} y_{i+1} + y_{i-1} &= \sin(k\pi x_{i+1}) + \sin(k\pi x_{i-1}), \quad x_{i\pm1} = x_i \pm h, \quad h = \frac{1}{N} \\ &= \sin(k\pi x_i + k\pi h) + \sin(k\pi x_i - k\pi h) \quad \text{(Additionstheoreme)} \\ &= 2\underbrace{\sin(k\pi x_i)}_{y_i}\cos(k\pi h) = 2y_i\cos(k\pi h). \end{aligned}$$

$$\Longrightarrow$$

$$\begin{aligned} y_{i+1} - 2y_i + y_{i-1} &= 2(\cos(k\pi h) - 1)y_i, \qquad \text{mit} \quad \cos 2\varphi = 1 - 2\sin^2\varphi \\ &= -4\left(\sin^2\frac{k\pi h}{2}\right)y_i \qquad \text{Division durch } -h^2 \text{ liefert} \\ \frac{-y_{i+1} + 2y_i - y_{i-1}}{h^2} &= -y_{\bar{x}x,i} = \frac{4}{h^2}\left(\sin^2\frac{k\pi h}{2}\right)y_i, \quad i = 1, \ldots, N-1. \end{aligned}$$

Wegen $y_0 = y_N = 0$ ist dieses Gleichungssystem nichts anderes als

$$\boldsymbol{A}_h^0\boldsymbol{y}^k = \lambda_k^h\boldsymbol{y}^k.$$

Wir haben also Eigenvektoren und Eigenwerte von $\boldsymbol{A}_h^0$ gefunden.
Wegen

$$k = 1, \ldots, N-1, \ \ h = \frac{1}{N} \qquad \Longrightarrow \qquad kh\frac{\pi}{2} \le (N-1)\frac{1}{N} \cdot \frac{\pi}{2} < \frac{\pi}{2}$$

sind die Eigenwerte mit k monoton wachsend (sin wächst monoton in $[0, \frac{\pi}{2}]$). ■

Wegen $\lambda_{\min} = \frac{4}{h^2}\sin^2(\frac{\pi}{2N}) > 0$ folgt aus (3.8) nochmals die Invertierbarkeit von $\boldsymbol{A}_h^0$.

Vergleich der Eigenwerte von kontinuierlicher und diskreter Aufgabe

Für kleine x ist $\sin x \approx x$ eine gute Approximation. Deshalb gilt für kleine Werte von $kh = k \cdot \frac{1}{N}$

$$\lambda_k^h = \frac{4}{h^2}\sin^2(\frac{k\pi h}{2}) \approx \frac{4}{h^2}(\frac{k\pi h}{2})^2 = k^2\pi^2 \quad \text{(gute Näherung)}$$

Für große k , z.B. $k = N-1$, kann von Approximation, selbst für sehr kleines $h = 1/N$, keine Rede sein, wie man unmittelbar ersieht:

$$\lambda_{N-1}^h = \frac{4}{h^2}\left(\sin\frac{(N-1)\pi h}{2}\right)^2 \overset{h=\frac{1}{N}}{=} 4N^2\left(\sin(\frac{(N-1)}{N}\cdot\frac{\pi}{2})\right)^2 \overset{N \text{ groß}}{\approx} 4N^2,$$

aber

$$\lambda_{N-1} = (N-1)^2\pi^2.$$

Eigenwertschranken

Unmittelbar folgt aus Lemma 3.7 mit $h = \frac{1}{N}$

$$\lambda_{\max}(\boldsymbol{A}_h^0) = \lambda_{N-1}^h = \frac{4}{h^2}\sin^2\left(\frac{N-1}{N}\,\frac{\pi}{2}\right) \leq \frac{4}{h^2}. \tag{3.16}$$

Zur Abschätzung der Eigenwerte nach unten benutzen wir $N \geq 2$, also $h \leq \frac{1}{2}$ (es gibt mindestens einen inneren Punkt in $[0,1]$).

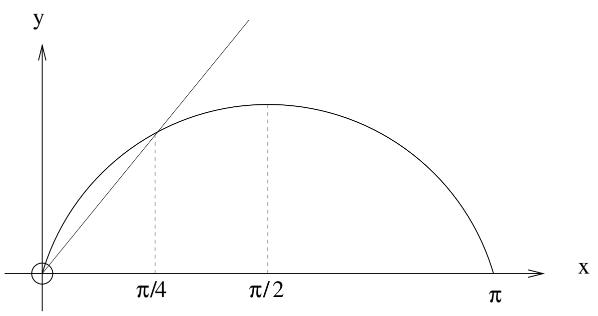
Nun gilt für $0 \leq x \leq \dfrac{\pi}{4}$

$y = \dfrac{\sin\pi/4}{\pi/4}\cdot x \leq \sin x.$

Wegen $\sin\dfrac{\pi}{4} = \sqrt{2}/2$ folgt

$\dfrac{\sqrt{2}/2}{\pi/4}x \leq \sin x,$ bzw.

$\dfrac{2\sqrt{2}}{\pi}x \leq \sin x$



Damit folgt für $x = \dfrac{\pi h}{2}$ (wegen $h \leq \frac{1}{2}$)

$$\lambda_{\min}(\boldsymbol{A}_h^0) = \lambda_1^h = \frac{4}{h^2}\sin^2\frac{\pi h}{2} \geq \frac{4}{h^2}\,\frac{8}{\pi^2}\underbrace{\frac{\pi^2 h^2}{4}}_{x^2} = 8, \tag{3.17}$$

eine von h unabhängige untere Schranke.

Skalarprodukte, Normen und Abschätzungen

Definition 3.8 Skalarprodukte und Normen

(3.18) $(\boldsymbol{y},\boldsymbol{v})_{(0,h)} := \sum\limits_{i=1}^{N-1} y_i\,\bar{v}_i\,h$ ist ein Skalarprodukt (vgl. 3.5), $\quad h=\frac{1}{N}$.

Die zugehörige Vektornorm ergibt sich aus

(3.19) $\|\boldsymbol{y}\|^2_{(0,h)} = (\boldsymbol{y},\boldsymbol{y})_{(0,h)}$.

Wegen $\boldsymbol{A}_h^0 > 0$ (pos. definit, (Satz 3.4)), ist $(\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{x})$ ein Skalarprodukt mit

(3.20) $(\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})_{(0,h)} \overset{(3.13)}{=} \sum\limits_{i=0}^{N-1}(\boldsymbol{y}_{x,i})^2 h, \quad (y_0^j = y_N^j = 0\ \forall j)$.

Es erzeugt die Norm

(3.21) $\|\boldsymbol{y}\|^2_{(1,h)} = \sum\limits_{i=0}^{N-1}(\boldsymbol{y}_{x,i})^2 h = (\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})_{(0,h)} =: \|\boldsymbol{y}\|^2_{\boldsymbol{A}_h^0}, \quad (y_0^j = y_N^j = 0 \quad \forall j)$

liefert eine Vektornorm (energetische Norm).

Der Index $(1,h)$ in (3.21) weist auf eine Ableitung in der Definition hin.

Beachte: Die Darstellung (3.20) (bzw. (3.13)) gilt nur, wenn man $y_0 = y_N = 0$ setzt.

Mit (3.18), (3.20) erhält man

$$\frac{(\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})_{(0,h)}}{(\boldsymbol{y},\boldsymbol{y})_{(0,h)}} = \frac{(\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})}{(\boldsymbol{y},\boldsymbol{y})}$$

und mit Satz 3.2,(3.3) (Eigenschaft des Rayleigh-Quotienten) und (3.16), (3.17) die Abschätzungen

(3.22) $$8 \le \lambda_{min}(\boldsymbol{A}_h^0) \le \frac{(\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})_{(0,h)}}{(\boldsymbol{y},\boldsymbol{y})_{(0,h)}} \le \lambda_{max}(\boldsymbol{A}_h^0) \le \frac{4}{h^2}\,.$$

Hieraus folgen wegen

$$\|\boldsymbol{y}\|^2_{(1,h)} = \frac{(\boldsymbol{A}_h^0\boldsymbol{y},\boldsymbol{y})_{(0,h)}}{\|\boldsymbol{y}\|^2_{(0,h)}} \cdot \|\boldsymbol{y}\|^2_{(0,h)}$$

die Normvergleiche

(3.23) $$8\,\|\boldsymbol{y}\|^2_{(0,h)} \le \|\boldsymbol{y}\|^2_{(1,h)} \le \frac{4}{h^2}\|\boldsymbol{y}\|^2_{(0,h)}$$

und mit $\|\boldsymbol{A}_h^0\|_S = \lambda_{max}(\boldsymbol{A}_h^0)$ weiterhin

$$8 \le \|\boldsymbol{A}_h^0\|_S \le \frac{4}{h^2}\,.$$

Bemerkung: Solche Normabschätzungen kann man auch ohne Rückgriff auf die Eigenwertaussagen erhalten (vgl. (3.28)). Sie fallen dann etwas schlechter aus, sind aber auch bei Matrizen anwendbar, deren Eigenwerte man nicht kennt. (vgl. den § über die Behandlung der Differentialgleichung $u_\tau = \frac{\partial}{\partial t}\left(k(x)\frac{\partial u}{\partial t}\right)$. Wir werden (später) Fehlerabschätzungen bzgl. dieser Skalarproduktnormen erhalten.
In praktischen Anwendungen möchte man gerne Abschätzungen bzgl. der Maximumnorm. Solche Abschätzungen erhält man oft nur mit Hilfe der Abschätzungen über Skalarprodukte. Wir zeigen, ohne Benutzung von Eigenwerten, die für die Anwendung wichtige Abschätzung

$$\|\boldsymbol{y}\|_\infty \le \frac{1}{2}\|\boldsymbol{y}\|_{(1,h)} \qquad (\boldsymbol{y} \in \mathbb{R}^{N-1}) \tag{3.24}$$

Beweis: Unter Benutzung von $y_0 = y_N = 0$ gilt

$$\begin{aligned} y_i &= h\sum_{j=0}^{i-1}\frac{(y_{j+1}-y_j)}{h} = \sum_{j=0}^{i-1} y_{x,j}\cdot h \quad \text{(diskretes Analogon zu } y(x) = \underbrace{y(0)}_{=0} + \int_0^x y(\xi)d\xi) \\ &= \sum_{j=0}^{i-1} y_{x,j}\sqrt{h}\sqrt{h} \overset{\text{CSU}}{\le} \sqrt{\sum_{j=0}^{i-1}(y_{x,j})^2 h \sum_{j=0}^{i-1} h} \quad \text{und mit} \quad \sum_{j=0}^{i-1} h = x_i \end{aligned}$$

$$y_i^2 \le x_i \sum_{j=0}^{i-1}(y_{x,j})^2 h. \tag{3.25}$$

Analog erhält man durch „rückwärtsintegrieren“: $y_i = \sum_{j=i}^{N-1}\frac{(y_{j+1}-y_j)}{h}$, die Identität

$$y_i = -\sum_{j=i}^{N-1} y_{x,j}h = \sum_{j=i}^{N-1}(-y_{x,j})\sqrt{h}\sqrt{h}$$

woraus auf gleiche Weise wegen $\sum_{j=i}^{N-1} h = (N-i)h = (1-x_i)$ mit der CSU folgt

$$y_i^2 \le (1-x_i)\sum_{j=i}^{N-1}(y_{x,j})^2 h. \tag{3.26}$$

Multipliziere (3.25) mit $(1-x_i)$ und (3.26) mit x_i und addiere dann (= Konvexkombination: $y_i^2(1-x_i) + y_i^2 \cdot x_i$) $\Longrightarrow$

$$y_i^2 \le x_i(1-x_i)\sum_{j=0}^{N-1}(y_{x,j})^2 h = x_i(1-x_i)\|\boldsymbol{y}\|_{(1,h)}^2. \tag{3.27}$$

Unter Beachtung von $\max\limits_{x\in[0,1]} x(1-x) = \frac{1}{4}$ folgt hieraus

$$\|\boldsymbol{y}\|_\infty \le \frac{1}{2}\|\boldsymbol{y}\|_{(1,h)}.$$

■

Zur späteren Verwendung leiten wir, <u>ohne Rückgriff auf Eigenwerte</u>, einen weiteren Vergleich her

$$\|\boldsymbol{y}\|^2_{(0,h)} \overset{!}{\leq} \frac{1}{6}\sum_{j=0}^{N-1}(\boldsymbol{y}_{x,i})^2 h = \frac{1}{6}\|\boldsymbol{y}\|^2_{(1,h)} \overset{(3.21)}{=} \frac{1}{6}\|\boldsymbol{y}\|^2_{\boldsymbol{A}^0_h}. \tag{3.28}$$

Beweis: Multipliziere (3.27) mit h und summiere: $\sum\limits_{1}^{N-1}$ $\Longrightarrow$

$$\sum_{i=1}^{N-1} y_i^2 h = \|\boldsymbol{y}\|^2_{(0,h)} \leq \|\boldsymbol{y}\|^2_{(1,h)} \sum_{i=1}^{N-1} x_i(1-x_i)h.$$

Vermutung: $\sum\limits_{i=1}^{N-1} x_i(1-x_i)h \approx \frac{1}{6}$ wegen $\int\limits_0^1 x(1-x)dx = \frac{1}{6}$.

Nun ist

$$\begin{aligned}
\sum_{i=1}^{N-1} x_i(1-x_i)h &= h^3\sum_{i=1}^{N-1} i(N-i) = h^3\left[N\frac{N(N-1)}{2} - \frac{N(N-1)}{2}\cdot\frac{(2N-1)}{3}\right] \\
&= h^3\frac{N(N-1)}{2}\left[N - \frac{2N-1}{3}\right] \quad \text{und mit } hN = 1 \\
&= h^2\frac{(N-1)}{6}[N+1] \\
&= h^2\frac{(N^2-1)}{6} = \frac{1-h^2}{6} < \frac{1}{6}.
\end{aligned}$$

■

Bemerkung: Die 8 aus (3.23) hat sich zu einer 6 verschlechtert.

§ 4 Stabilität (und „bessere“ Verfahren)

Stabilität eines Verfahrens bedeutet, daß bei beschränkten Anfangs- und Randwerten die fortlaufend errechneten Werte beschränkt bleiben. Eine genaue Definition folgt noch. Wir beweisen zunächst ein einfaches Stabilitätsergebnis, das zeigt, daß das beschriebene explizite Verfahren verbesserungsbedürftig ist.

Wir untersuchen das explizite Verfahren (2.8) für den Spezialfall $f = 0$ und Nullrandwerte $y_0^j = y_N^j = 0 \ \forall_j$. Es läßt sich schreiben als

$$\boldsymbol{y}_t^j = -A_h^0\, \boldsymbol{y}^j, \quad j \geq 0, \quad \boldsymbol{y}^0 = u_0, \tag{4.1}$$

und zeigen:

Notwendig und hinreichend dafür, daß alle $\boldsymbol{y}^j$ gemäß (4.1) beschränkt sind, ist

$$\frac{\tau}{h^2} \leq \frac{1}{2} \left(\cos^2 \frac{\pi h}{2}\right)^{-1}. \tag{4.2}$$

Bemerkungen:

1. Für die zu (4.1) gehörige kontinuierliche Aufgabe

   $$\dot{u} = u'', \quad u(x,0) = u_0(x), \quad 0 < x < 1, \quad u(0,t) = u(1,t) = 0$$

   gilt das Randmaximumprinzip, aus dem folgt, daß die Lösung $u(x,t)$ für $0 \leq x \leq 1$ und für alle t beschränkt ist durch $\max\limits_{x \in [0,1]} |u_0(x)|$. Deshalb ist die Beschränktheitsforderung notwendig für ein vernünftiges Verfahren.

2. Natürlich werden wir (4.2) verallgemeinern.

3. Für kleine h wächst $\cos^2 \frac{\pi h}{2} \xrightarrow{h \to 0} 1$, weshalb als hinreichende Bedingung üblicherweise $\frac{\tau}{h^2} \leq \frac{1}{2}$ genannt wird.

4. Diese Bedingung verlangt auf Grund der hohen Anzahl von Zeitschritten einen großen Rechenaufwand.

Beweis: (4.2)

$\boldsymbol{A}_h^0$ hat ein Orthonormalsystem (ONS) von $N-1$ Eigenvektoren (vgl. (3.2)).

$$\boldsymbol{A}_h^0\ \boldsymbol{v}^k = \lambda_h^h\, \boldsymbol{v}^k, \quad k = 1, \ldots, N-1, \quad \lambda_k^h > 0$$

Man kann also auf jeder Zeitschicht die durch (4.1) berechneten Vektoren darstellen durch (Fourier-Zerlegung)

$$\boldsymbol{y}^j = \sum_{k=1}^{N-1} c_k^j \boldsymbol{v}^k \tag{4.3}$$

Einsetzen dieser Darstellung in das Differenzenschema ergibt

$$-\boldsymbol{A}_h^0 \sum_{k=1}^{N-1} c_k^j \boldsymbol{v}^k = \frac{\boldsymbol{y}^{j+1} - \boldsymbol{y}^j}{\tau} = \sum_{k=1}^{N-1} \frac{c_k^{j+1} - c_k^j}{\tau} \boldsymbol{v}^k,$$

Koeffizientenbergleich liefert

$$\begin{aligned} \frac{c_k^{j+1} - c_k^j}{\tau} &= -\lambda_k^h \; c_k^j \\ c_k^{j+1} &= \left(1 - \tau\lambda_k^h\right) c_k^j \qquad \text{(Rekursionsformel)} \\ &= \left(1 - \tau\lambda_k^h\right)^2 c_k^{j-1} \\ &\vdots \\ c_k^{j+1} &= \left(1 - \tau\lambda_k^h\right)^{j+1} c_k^0. \end{aligned} \tag{4.4}$$

Gemäß (4.3) sind alle $\boldsymbol{y}^j$ genau dann beschränkt, wenn dies auch für alle Koeffizienten $c_k^j, \quad k = 1, \ldots, N-1, \quad j \geq 0$ gilt. Notwendig und hinreichend dafür ist (vgl. (4.4)):

$$|1 - \tau\lambda_k^h| \leq 1,$$

bzw.

$$-1 \leq 1 - \tau\lambda_k^h \leq 1 \quad \forall k.$$

Die linke Ungleichung besagt $\tau\lambda_k^h \leq 2 \quad \forall k$. Wegen $\tau > 0, \quad \lambda_k^h > 0$ ist die rechte Ungleichung immer erfüllt.

Wir müssen, da die λ_k^h monoton geordnet sind, diese Ungleichung also für den maximalen Eigenwert von $\boldsymbol{A}_h^0$ fordern, also (vgl. Lemma 3.7)

$$\tau \frac{4}{h^2} \sin^2 \frac{(N-1)\pi h}{2} \leq 2. \tag{4.5}$$

Nun ist

$$\frac{(N-1)\pi h}{2} = \frac{N\pi h}{2} - \frac{\pi h}{2} = \frac{\pi}{2} - \frac{\pi h}{2}$$

und

$$\sin \frac{(N-1)\pi h}{2} = \sin\left(\frac{\pi}{2} - \frac{\pi h}{2}\right) = -\sin\left(\frac{\pi h}{2} - \frac{\pi}{2}\right) = \cos\frac{\pi h}{2}. \quad \text{(Phasenverschiebung)}$$

Insgesamt liefert (4.5)

$$\frac{\tau}{h^2} \leq \frac{1}{2}\left(\cos^2\frac{\pi h}{2}\right)^{-1}.$$

■

Ist also τ in der Größenordnung h^2, so gibt es Ärger. Daß dieses Resultat nicht nur theoretisch sondern auch numerisch Ärger bereitet, zeigen die Übungen. Die Forderung $\frac{\tau}{h^2} \leq \frac{1}{2}$ ist sehr einschneidend. Man braucht „bessere“ Bedingungen (d.h. bessere Verfahren).

Motivation für neue Verfahren:

Der vorwärtsgenommene Differenzenquotient beim expliziten Verfahren hat nur eine lineare Konsistenzordnung (vgl. (2.2)). Derselbe Differenzenquotient als Näherung für die Ableitung auf der Zeitschicht $t_{j+\frac{1}{2}}$ hat quadratische Konsistenzordnung. Daher rührt der Vorschlag, das ganze Verfahren für die Zeitschicht $t_{j+\frac{1}{2}}$ zu formulieren, d.h.

Mittelung der Werte auf alter und neuer Zeitschicht

Wir untersuchen das Verfahren

$$\begin{aligned} \boldsymbol{y}_t = -\boldsymbol{A}_h^0 \boldsymbol{y}^\sigma \quad &+ \quad \boldsymbol{\varphi} \\ \boldsymbol{y}^\sigma \quad &:= \quad \sigma \boldsymbol{y}^{j+1} + (1-\sigma)\boldsymbol{y}^j \;\; \forall_j, \quad \sigma \in [0,1]. \end{aligned} \tag{4.6}$$

Bezeichnung: σ als Exponent bedeutet immer nur eine Mittelung, nie eine Potenz.

Bemerkung: $\boldsymbol{\varphi}$ enthält nur additive, bekannte Werte auf jeder Zeitschicht (Randwerte und f-Werte, vgl. (2.12)). Wir entscheiden später auf welcher Zeitschicht wir $\boldsymbol{\varphi}$ betrachten. Naheliegend ist $t_{j+\sigma}$.

Durch Taylorentwicklung untersuchen wir zunächst die Auswirkung der Ersetzung von $\boldsymbol{y}$ durch $\boldsymbol{y}^\sigma$. Wir entwickeln u^σ (für eine differenzierbare Funktion u) an der Stelle $t_{j+\frac{1}{2}}$ (die Indizes + und – bezeichnen wieder Funktionswerte an einer Zwischenstelle)

$$\begin{aligned} u^\sigma &= \sigma\left(u(t_{j+\frac{1}{2}}) + \frac{\tau}{2}\dot{u} + \frac{\tau^2}{8}\ddot{u} + \frac{\tau^3}{8\cdot 6}\dddot{u}_+\right) + (1-\sigma)\left(u(t_{j+\frac{1}{2}}) - \frac{\tau}{2}\dot{u} + \frac{\tau^2}{8}\ddot{u} - \frac{\tau^3}{8.6}\dddot{u}_-\right) \\ &= u(t_{j+\frac{1}{2}}) + \frac{\tau}{2}(2\sigma - 1)\,\dot{u} + \frac{\tau^2}{8}\ddot{u} + O(\tau^3) \end{aligned}$$

$\Longrightarrow$

Für $\sigma = \frac{1}{2}$ wird $u(t_{j+\frac{1}{2}})$ durch u^σ von 2. Ordnung approximiert. Da die Approximation der Ableitung 2. Ordnung in Zeitrichtung (durch den zentralen Differenzenquotienten) auch von 2. Ordnung war (vgl.(2.3)), ist keine Einbuße der Approximationsordnung zu befürchten.

In Abhängigkeit von der Wahl von σ erhält man aus (4.6) folgende Verfahren:

$$\sigma = \begin{cases} 0 & : \text{ explizites Verfahren: IndexEuler-Verfahren} \\ \frac{1}{2} & : \text{ Crank-Nicolson Schema (implizit)} \\ 1 & : \text{ implizites Diff.-Schema: implizites Euler Verfahren.} \end{cases}$$

Umformung von (4.6):

$$\begin{aligned} \boldsymbol{y}^\sigma &= \sigma\hat{\boldsymbol{y}} + (1-\sigma)\boldsymbol{y} \\ &= \sigma(\hat{\boldsymbol{y}} - \boldsymbol{y}) + \boldsymbol{y} \\ &= \sigma\tau\;\boldsymbol{y}_t + \boldsymbol{y} \end{aligned}$$

Einsetzen im (4.6) :

$$\boldsymbol{y}_t + \boldsymbol{A}_h^0(\sigma\tau\boldsymbol{y}_t + \boldsymbol{y}) \;=\; \boldsymbol{\varphi} \qquad \text{bzw.}$$

$$\underbrace{(\boldsymbol{I} + \sigma\tau\boldsymbol{A}_h^0)}_{=:\boldsymbol{B}}\boldsymbol{y}_t + \boldsymbol{A}_h^0\boldsymbol{y} \;=\; \boldsymbol{\varphi} \tag{4.7}$$

Mit $\boldsymbol{B} = (I + \sigma\tau\boldsymbol{A}_h^0)$ ordnet sich dieses Verfahren ein in die Klasse der Verfahren

$$\boxed{\boldsymbol{B}\boldsymbol{y}_t + \boldsymbol{A}\boldsymbol{y} = \boldsymbol{\varphi} \quad \text{mit Matrizen} \quad \boldsymbol{A}, \boldsymbol{B}} \tag{4.8}$$

Normalform des 2-Schicht-Verfahren

die wir im folgenden untersuchen, zunächst unter der Voraussetzungen

$$\boldsymbol{B} > 0 \quad \text{(d.h. pos. def.)}, \quad \boldsymbol{B} = \boldsymbol{B}^T$$
$$\boldsymbol{A} = \boldsymbol{A}^T, \quad \boldsymbol{A} > 0.$$

Aus $\boldsymbol{B} > 0$ folgt $\boldsymbol{B}$ ist invertierbar. Wünschenswert ist natürlich : $\boldsymbol{B}$ leicht invertierbar.

Beachte: (4.8) ist i. allg. implizit.

Wir untersuchen nun getrennt
a) die Stabilität der Verfahren (4.8) bzgl. der Anfangswerte (AWe), wobei $\boldsymbol{\varphi} \equiv 0$ gesetzt wird (d.h: rechte Seite der Differentialgleichung und Randwerte =0), und
b) bezüglich der „rechten Seite $\boldsymbol{\varphi}$", wobei $y_0 = 0$, $y_0^j = y_N^j = 0$ (Nullanfangs- und Randwerte). Die Stabilität bzgl. Anfangswerten und rechter Seite erhält man dann durch Superposition.

Definition 4.1 Stabilität
Das Differenzenschema $\boldsymbol{B}\boldsymbol{y}_t + \boldsymbol{A}\boldsymbol{y} = \boldsymbol{\varphi}$ heißt

$$\text{stabil bzgl. der} \quad \begin{cases} \text{Anfangswerte} \\ \text{rechten Seite} \end{cases}$$

falls Abschätzungen der folgenden Art gelten

$$||\boldsymbol{y}^j||_{(a)} \leq \begin{cases} M_1||\boldsymbol{y}^0||_{(b)} & \text{wobei} \quad \boldsymbol{\varphi} \equiv 0 \quad \text{(insbesondere} \quad y_0^j = y_N^j = 0) \\ M_2||\boldsymbol{\varphi}||_{(c)} & \text{wobei} \quad \boldsymbol{y}^0 = 0 \quad \text{(üblicherweise} \quad y_0^j = y_N^j = 0) \end{cases}$$

mit **vernünftigen** Normen $||\ ||_{(a)}, ||\ ||_{(b)}, ||\ ||_{(c)}$ und Konstanten $M_1, M_2 > 0$ die unabhängig von der Diskretisierung (d.h. von τ, h)) sind.

Bemerkungen:

1. Die Normen sollen in dem Sinn **vernünftig** sein, daß sie für $h \to 0$ und/oder $\tau \to 0$ weder gegen Null noch gegen ∞ gehen. Die L_2-Norm (für Zeilenverktoren) ist deshalb nicht zulässig, wohl aber z.B.

$$
\begin{array}{lll}
||\boldsymbol{y}||^2_{(0,h)} = \sum\limits_{i=0}^{N-1} y_i^2 \, h & \overset{h\to 0}{\longrightarrow} & \int\limits_0^1 u^2(x)\,dx \\
||\boldsymbol{y}||^2_{(1,h)} = \sum\limits_{i=0}^{N-1} (y_{x,i})^2 \, h & \underset{\text{beachte: } y_0=0}{\overset{h\to 0}{\underset{\uparrow}{\longrightarrow}}} & \int\limits_0^1 u'^2(x)\,dx \qquad (y_0 = 0, \text{ vgl. } (3.21))
\end{array}
$$

$\sqrt{\int\limits_0^1 u'^2(x)\,dx}$ ist, zusammen mit der Forderung $u(0) = 0$ auch eine Norm für differenzierbare Funktionen.
Die Stabilitätsdefinition macht klar, warum die Normen aus Definition 3.8 eingeführt werden mußten.

2. Stabilität sichert, daß die Lösung bei beschränkten Anfangswerten, Randwerten und rechter Seite beschränkt bleibt (Randmaximumprinzip) und liefert die stetige Abhängigkeit der Lösung von den Anfangswerten und der rechten Seite.

3. Verwunderlich mag erscheinen, daß man immer nur Nullrandwerte betrachtet. Der Grund dafür ist, wie wir sehen werden, daß die Randwerte für Konvergenzbetrachtungen keine Rolle spielen (vgl. § 5). Bei Konvergenzaussagen zeigt man, daß der Fehler *exakte Lösung - approximierte Lösung* gegen Null geht. Auf dem Rand werden immer die exakten Daten vorgegeben, weshalb dort der Fehler immer gleich Null ist.

Wir zeigen zunächst

Satz 4.2 Stabilität bzgl. Anfangswerten und rechter Seite
Für das Differenzensschema

(4.9) $\quad \boldsymbol{B}\boldsymbol{y}_t + \boldsymbol{A}\boldsymbol{y} = \boldsymbol{\varphi}, \quad \boldsymbol{y} \in \mathbb{R}^{n-1}$ sei $\boldsymbol{A}^T = \boldsymbol{A} > 0,\ \boldsymbol{B} > 0$ und $y_0^j = y_N^j = 0\ \forall j.$

Dann gilt

a) $\boldsymbol{B} - \frac{\tau}{2}\boldsymbol{A} \geq 0 \iff$ (4.9) ist stabil bzgl. der Anfangswerte mit $M_1 = 1$ und
(pos. semidef.) $\qquad ||\boldsymbol{y}^j||_{(1,h)} \leq ||\boldsymbol{y}^{j-1}||_{(1,h)} \leq \ldots \leq ||\boldsymbol{y}^0||_{(1,h)}.$

b) $\boldsymbol{B} - \frac{\tau}{2}\boldsymbol{A} - \frac{\varepsilon}{2}\boldsymbol{I} \geq 0$ für ein $\varepsilon > 0 \implies$
(4.9) ist stabil bzgl. Anfangswerten und rechter Seite und es gilt

$$
\begin{aligned}
||\boldsymbol{y}^j||_{(1,h)} &\leq ||\boldsymbol{y}^0||_{(1,h)} + \frac{1}{\sqrt{\varepsilon}}||\boldsymbol{\varphi}||_{(b)}, \quad \text{wobei} \\
||\boldsymbol{\varphi}||_{(b)} &= \left(\sum_{k=0}^{j-1} \tau \ ||\boldsymbol{\varphi}^k||^2_{(0,h)}\right)^{1/2}
\end{aligned}
$$

Bemerkung: $\sqrt{\tau \sum\limits_{r=0}^{j} ||\boldsymbol{\varphi}^\nu||^2_{(0,h)}} =: ||\boldsymbol{\varphi}||_{(b)}$ ist eine vernünftige Norm, denn

$$||\boldsymbol{\varphi}^\nu||^2_{(0,h)} = \sum_{i=1}^{N-1} |\varphi_i^\nu|^2 h \quad \overset{h\to 0}{\longrightarrow} \quad \int\limits_0^1 \varphi^2(x,t_\nu)\,dx \quad \text{und mit} \quad j\cdot\tau = T$$

$$\tau \sum_{\nu=0}^{j} ||\boldsymbol{\varphi}^\nu||^2_{(0,h)} \quad \overset{\substack{\tau\to 0\\ h\to 0}}{\longrightarrow} \quad \int\limits_0^T \int\limits_0^1 \varphi^2(x,t)\,dx\,dt.$$

Also gilt $||\boldsymbol{\varphi}||_{(b)} \approx \sqrt{\int_0^T \int_0^1 \varphi(x,t)^2\,dx\,dt}$. j zählt die Zeitschichten, deshalb kann $||\boldsymbol{\varphi}||_{(b)}$ mit j wachsen.
In dem $\boldsymbol{\varphi}^j$ aus (2.12) sind, je nach Komponente, f_i^j- Werte enthalten und Randwerte. Da wir nur Nullrandwerte betrachten, gilt für die Norm genauer

$$||\boldsymbol{\varphi}^\nu||^2_{(0,h)} \overset{\substack{\tau\to 0\\ h\to 0}}{\longrightarrow} \int\limits_0^T \int\limits_0^1 f^2(x,t)\,dx\,dt.$$

Beweis (Ideen):
(4.9) wird zu einer Gleichung umgeformt in der $||\hat{\boldsymbol{y}}||$ und $||\boldsymbol{y}||$ auftreten und Ausdrücke, die vorzeichenmäßig beherrschbar sind (Skalarprodukt-Multiplikation). Danach werden Stabilität in Bezug auf Anfangswerte und rechten Seiten getrennt untersucht. Das allgemeine Ergebnis folgt durch Superposition.

Multipliziere (4.9) bzgl. $(.)_{(0,h)}$ mit $2\tau\boldsymbol{y}_t \qquad \Longrightarrow$

$$2\tau(\boldsymbol{B}\boldsymbol{y}_t,\boldsymbol{y}_t)_{(0,h)} + 2\tau\left(\boldsymbol{A}\boldsymbol{y}, \frac{\hat{\boldsymbol{y}}-\boldsymbol{y}}{\tau}\right)_{(0,h)} = 2\tau(\boldsymbol{\varphi},\boldsymbol{y}_t)_{(0,h)} \quad \text{bzw.}$$

$$2\tau\underbrace{(\boldsymbol{B}\boldsymbol{y}_t,\boldsymbol{y}_t)_{(0,h)}}_{>0} + 2\left(\boldsymbol{A}\boldsymbol{y},\hat{\boldsymbol{y}}-\boldsymbol{y}\right)_{(0,h)} = 2\tau(\boldsymbol{\varphi},\boldsymbol{y}_t)_{(0,h)}$$

Idee: Von $\boldsymbol{B}$ kann man noch etwas subtrahieren, ohne daß die Vorzeichenbedingung verloren geht (Gershgorin), und damit den 2. Summanden umformen. In (4.7) ist $\boldsymbol{B} = \boldsymbol{I} + \sigma\tau\boldsymbol{A}_h^0$. Subtrahiert man $\frac{\tau}{2}\boldsymbol{A}_h^0$, erhält man

$$\boldsymbol{B} - \frac{\tau}{2}\boldsymbol{A}_h^0 = \boldsymbol{I} + \underbrace{(\sigma - \frac{1}{2})}_{=0 \text{ für } \sigma=\frac{1}{2}}\,\tau\boldsymbol{A}_h^0$$

Im allgemeinen Fall von Satz 4.2 subtrahieren wir $\frac{\tau}{2}\boldsymbol{A}$ und erhalten

$$(4.10) \qquad 2\tau\left((\boldsymbol{B} - \frac{\tau}{2}\boldsymbol{A})\boldsymbol{y}_t,\boldsymbol{y}_t\right)_{(0,h)} + \underbrace{\tau^2(\boldsymbol{A}\boldsymbol{y}_t,\boldsymbol{y}_t)_{(0,h)} + 2(\boldsymbol{A}\boldsymbol{y},\hat{\boldsymbol{y}}-\boldsymbol{y})_{(0,h)}}_{=:\,d} = 2\tau(\boldsymbol{\varphi},\boldsymbol{y}_t)$$

Umformung:

$$\begin{aligned} d &= (\boldsymbol{A}(\hat{\boldsymbol{y}}-\boldsymbol{y}),\hat{\boldsymbol{y}}-\boldsymbol{y})_{(0,h)} + 2(\boldsymbol{A}\boldsymbol{y},\hat{\boldsymbol{y}}-\boldsymbol{y})_{(0,h)} \\ &= (\boldsymbol{A}(\hat{\boldsymbol{y}}+\boldsymbol{y}),\hat{\boldsymbol{y}}-\boldsymbol{y})_{(0,h)} \\ &= (\boldsymbol{A}\hat{\boldsymbol{y}},\hat{\boldsymbol{y}})_{(0,h)} - (\boldsymbol{A}\boldsymbol{y},\boldsymbol{y})_{(0,h)} + \underbrace{(\boldsymbol{A}\boldsymbol{y},\hat{\boldsymbol{y}})_{(0,h)} - (\boldsymbol{A}\hat{\boldsymbol{y}},\boldsymbol{y})_{(0,h)}}_{=0 \text{ wegen } \boldsymbol{A}=\boldsymbol{A}^T,\ \hat{\boldsymbol{y}},\boldsymbol{y} \text{ reell}} \\ &= ||\hat{\boldsymbol{y}}||^2_{(1,h)} - ||\boldsymbol{y}||^2_{(1,h)} \end{aligned}$$

Damit folgt aus (4.10) die sog. *energetische Identität*

$$2\tau\Big((\boldsymbol{B}-\frac{\tau}{2}\boldsymbol{A})\boldsymbol{y}_t,\boldsymbol{y}_t\Big)_{(0,h)} + ||\hat{\boldsymbol{y}}||^2_{(1,h)} - ||\boldsymbol{y}||^2_{(1,h)} = 2\tau(\boldsymbol{\varphi},\boldsymbol{y}_t)_{(0,h)}. \tag{4.11}$$

Aus dieser Gleichung werden wir den Beweis ableiten.

Beweis a): Stabilität bzgl. der Anfangswerte $(\boldsymbol{\varphi}\equiv 0)$
„$\Longrightarrow$“
Aus (4.11) folgt sofort: Ist $\boldsymbol{B}-\frac{\tau}{2}\boldsymbol{A}\geq 0$ (positive Semidefinitheit), so gilt

$$||\hat{\boldsymbol{y}}||^2_{(1,h)} =: ||\boldsymbol{y}^{j+1}||^2_{(1,h)} \leq ||\boldsymbol{y}^j||^2_{(1,h)} \leq \cdots \leq ||\boldsymbol{y}^0||^2_{(1,h)} \tag{4.12}$$

also Stabilität bzgl. der Anfangswerte mit $M_1 = 1$.
In Anbetracht des Randmaximumprinzips ist $M_1 = 1$ auch die richtige Konstante, wenn $||\ ||_{(a)} = ||\ ||_{(b)}$ (vgl. Definition 4.1).

„$\Longleftarrow$“
Mit $\boldsymbol{\varphi}\equiv 0$ und der Stabilität (mit $M=1$): $||\hat{\boldsymbol{y}}||^2_{(1,h)} \leq ||\boldsymbol{y}||^2_{(1,h)}$ folgt aus (4.11)

$$\Big((\boldsymbol{B}-\frac{\tau}{2}\boldsymbol{A})\boldsymbol{y}_t,\boldsymbol{y}_t\Big) \geq 0.$$

$\boldsymbol{A}$ und $\boldsymbol{B}$ sind invertierbar. Verfahren (4.9) und $\boldsymbol{\varphi}\equiv 0$ liefern $\boldsymbol{y}_t = -\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{y}$. Mit beliebigen $\boldsymbol{y}$ durchläuft auch $\boldsymbol{y}_t$ den ganzen Raum. Also gilt

$$\Big((\boldsymbol{B}-\frac{\tau}{2}\boldsymbol{A})\boldsymbol{y},\boldsymbol{y}\Big) \geq 0. \tag{4.13}$$

Das bedeutet $\boldsymbol{B}-\frac{\tau}{2}\boldsymbol{A}\geq 0$.

Beweis b) Die Stabilität bzgl. der Anfangswerte folgt aus a).

Stabilität bzgl. der rechten Seite: $(\boldsymbol{y}^0 = 0,\ y^j_0 = y^j_N = 0)$

Wir schätzen die rechte Seite von (4.11) ab mit der CSU

$$(u,v) \leq ||u||\ ||v||, \quad ||\ || = \sqrt{(\ ,\ )}$$

und der ε-Ungleichung:

$$2ab \le \varepsilon a^2 + \frac{1}{\varepsilon} b^2 \text{ für } a, b, \varepsilon > 0.$$

Sie folgt aus $(\varepsilon a - b)^2 = \varepsilon^2 a^2 - 2\varepsilon ab + b^2 \ge 0$.

Somit folgt

$$2\tau(\boldsymbol{\varphi}^j, \boldsymbol{y}_t^j)_{(0,h)} \overset{\text{CSU}}{\le} 2\tau||\boldsymbol{\varphi}^j||_{(0,h)}||\boldsymbol{y}_t^j||_{(0,h)} \overset{\varepsilon-\text{Ungleichung}}{\le} \frac{\tau}{\varepsilon}||\boldsymbol{\varphi}^j||^2_{(0,h)} + \varepsilon\tau||\boldsymbol{y}_t||^2_{(0,h)}.$$

Dies wird in (4.11) eingetragen, der letzte Summand wird unter Beachtung von $||\boldsymbol{y}_t||^2_{(0,h)} = (\boldsymbol{I}\boldsymbol{y}_t, \boldsymbol{y}_t)_{(0,h)}$ auf die linke Seite gebracht. So folgt

$$2\tau\left((\boldsymbol{B} - \frac{\tau}{2}\boldsymbol{A} - \frac{\varepsilon}{2}\boldsymbol{I})\boldsymbol{y}_t, \boldsymbol{y}_t\right)_{(0,h)} + ||\hat{\boldsymbol{y}}||^2_{(1,h)} - ||\boldsymbol{y}||^2_{(1,h)} \le \frac{\tau}{\varepsilon}||\boldsymbol{\varphi}^j||^2_{(0,h)}$$

Die Voraussetzung $\boldsymbol{B} - \frac{\tau}{2}\boldsymbol{A} - \frac{\varepsilon}{2}\boldsymbol{I} \ge 0$ liefert somit

$$||\boldsymbol{y}^{j+1}||^2_{(1,h)} \le ||\boldsymbol{y}^j||^2_{(1,h)} + \frac{\tau}{\varepsilon}||\boldsymbol{\varphi}^j||^2_{(0,h)}.$$

Wiederholte Anwendung dieses Schrittes ergibt (wegen $\boldsymbol{y}^0 = 0$)

$$||\boldsymbol{y}^{j+1}||^2_{(1,h)} \le \underbrace{\|\boldsymbol{y}^0\|^2_{(1,h)}}_{=0} + \frac{\tau}{\varepsilon}\sum_{\nu=0}^{j}||\boldsymbol{\varphi}^\nu||^2_{(0,h)} = \frac{\tau}{\varepsilon}\sum_{\nu=0}^{j}||\boldsymbol{\varphi}^\nu||^2_{(0,h)}. \tag{4.14}$$

Damit liefert (4.14) die Behauptung b) im Fall $\boldsymbol{y}^0 \equiv 0$.

Nun kann man die Lösung $\boldsymbol{y}$ der Aufgabe mit Nullrandwerten, Anfangswerten und rechter Seite durch Superposition erhalten $\boldsymbol{y} = \boldsymbol{y}_A + \boldsymbol{y}_R$, wobei

$\boldsymbol{y}_A \;\hat{=}\;$ Lösung der Aufgabe mit Anfangswerten, Nullrandwerten und rechter Seite $= 0$
$\boldsymbol{y}_R \;\hat{=}\;$ Lösung der Aufgabe mit Null-Anfangs- und - Randwerten, aber mit rechter Seite

Mit Hilfe der Dreieckungsgleichung. $||\boldsymbol{y}||_{(1,h)} \le ||\boldsymbol{y}_A||_{(1,h)} + ||\boldsymbol{y}_R||_{(1,h)}$ folgt die Behauptung aus (4.13) und (4.14). ■

Bemerkungen

1. Dies ist die Theorie von Samarskij (in Birkhäuser gibt es 2 dicke Bücher von ihm).
2. Die Voraussetzung $\boldsymbol{B} = \boldsymbol{B}^T$ (vgl. nach (4.8)) wurde bisher nicht benötigt.
3. Der Satz zeigt, warum die Normvergleiche (3.22), (3.24), (3.28) wichtig sind. Dadurch ist es möglich die Aussagen des Satzes auch auf andere Normen zu übertragen.

Wir wenden Satz 4.2 an auf das Verfahren

$$(4.15) \qquad \underbrace{(\boldsymbol{I} + \sigma\tau\boldsymbol{A}_h^0)}_{B}\boldsymbol{y}_t + \boldsymbol{A}_h^0\boldsymbol{y} = \boldsymbol{\varphi}, \quad (\boldsymbol{A}_h^0)^T = \boldsymbol{A}_h^0,\ y_0^j = y_N^j = 0\ \forall j$$

und zeigen, daß man σ, τ und ε-Werte angeben kann,welche die Stabilität sichern. Wir beginnen mit der **Stabilität bzgl. der Anfangswerte**.

Für Skalarproduktnormen und zugeordnete Matrixnormen gilt für beliebige $\boldsymbol{A} \in \mathbb{R}^n$ mit der CSU

$$(4.16) \qquad (\boldsymbol{A}\boldsymbol{y}, \boldsymbol{y}) \le ||\boldsymbol{A}\boldsymbol{y}||\,||\boldsymbol{y}|| \le ||\boldsymbol{A}||\,||\boldsymbol{y}||^2 = ||\boldsymbol{A}||(\boldsymbol{y}, \boldsymbol{y}) \quad \text{bzw.} \frac{\boldsymbol{A}}{||\boldsymbol{A}||} \le \boldsymbol{I}\,.$$

Damit verschärfen wir die Bedingung $\boldsymbol{I} + \sigma\tau\boldsymbol{A}_h^0 - \frac{\tau}{2}\boldsymbol{A}_h^0 \ge 0$ aus Satz 4.2 a) zu

$$\frac{\boldsymbol{A}_h^0}{||\boldsymbol{A}_h^0||} + \sigma\tau\boldsymbol{A}_h^0 - \frac{\tau}{2}\boldsymbol{A}_h^0 \ge 0$$

und mit der Spektralnorm wegen $||\boldsymbol{A}_h^0|| \le \frac{4}{h^2}$ nochmals zu

$$\frac{h^2}{4}\boldsymbol{A}_h^0 + \sigma\tau\boldsymbol{A}_h^0 - \frac{\tau}{2}\boldsymbol{A}_h^0 \ge 0 \quad \text{bzw.} \quad (\frac{h^2}{4} + \tau(\sigma - \frac{1}{2}))\boldsymbol{A}_h^0 \ge 0.$$

Dies ist wegen $\boldsymbol{A}_h^0 > 0$ richtig, falls $\frac{h^2}{4} + \tau(\sigma - \frac{1}{2}) \ge 0$ bzw.

$$(4.17) \qquad \sigma \ge \frac{1}{2} - \frac{h^2}{4\tau}$$

Bedeutung

1. Das explizite Verfahren $(\sigma = 0)$ ist stabil bzgl. der Anfangswerte, falls
$$0 \ge \frac{1}{2} - \frac{h^2}{4\tau}, \quad \text{bzw.} \quad \frac{\tau}{h^2} \le \frac{1}{2}.$$
Dies war für den Spezialfall $\boldsymbol{\varphi} = 0$ schon bewiesen.

2. Das implizite Vefahren ist für $\sigma \ge \frac{1}{2}$ (insbesondere also Crank-Nicolson) *unbedingt stabil* bzgl. der Anfangswerte, d.h. stabil ohne Bedingungen an σ und τ.

Stabilität bzgl. der rechten Seite verlangt für ein $\varepsilon > 0$

$$\underbrace{\boldsymbol{I} + \sigma\tau\boldsymbol{A}_h^0}_{\boldsymbol{B}} - \frac{\tau}{2}\boldsymbol{A}_h^0 - \frac{\varepsilon}{2}\boldsymbol{I} \ge 0.$$

Mit (4.16): $\frac{\boldsymbol{A}}{||\boldsymbol{A}||} \le \boldsymbol{I}$, und damit

$$\boldsymbol{I} = \frac{\varepsilon}{2}\boldsymbol{I} + (1 - \frac{\varepsilon}{2})\boldsymbol{I} \ge \frac{\varepsilon}{2}\boldsymbol{I} + (1 - \frac{\varepsilon}{2})\frac{\boldsymbol{A}_h^0}{||\boldsymbol{A}_h^0||}$$

verschärfen wir die Bedingung unter der Voraussetzung $1-\frac{\varepsilon}{2}\geq 0$ zu

$$\frac{\varepsilon}{2}\boldsymbol{I}+(1-\frac{\varepsilon}{2})\frac{\boldsymbol{A}_h^0}{||\boldsymbol{A}_h^0||}+\sigma\tau\boldsymbol{A}_h^0-\frac{\tau}{2}\boldsymbol{A}_h^0-\frac{\varepsilon}{2}\boldsymbol{I}\overset{!}{\geq} 0 \quad \text{bzw.}$$

$$\left(\frac{1-\frac{\varepsilon}{2}}{||\boldsymbol{A}_h^0||}+\tau(\sigma-\frac{1}{2})\right)\boldsymbol{A}_h^0\geq 0.$$

Wegen $||\boldsymbol{A}_h^0||\leq\frac{4}{h^2}$ und $\boldsymbol{A}_h^0>0$ ist diese Bedingung sicher erfüllt, falls

$$(1-\frac{\varepsilon}{2})\frac{h^2}{4}+\tau(\sigma-\frac{1}{2})\geq 0 \quad \text{bzw.}$$

$$\sigma\geq\frac{1}{2}-\frac{h^2}{4\tau}(1-\frac{\varepsilon}{2}). \tag{4.18}$$

Insgesamt erhalten wir also

Folgerung 4.3
Für das Verfahren

$$(\boldsymbol{I}+\sigma\tau\boldsymbol{A}_h^0)\boldsymbol{y}_t+\boldsymbol{A}_h^0\boldsymbol{y}=\boldsymbol{\varphi},\quad (\boldsymbol{A}_h^0)^T=\boldsymbol{A}_h^0,\ y_0^j=y_N^j=0\ \forall j$$

gilt mit den Normen, die in Satz (4.2) verwandt wurden

a) Ist $\quad\sigma\geq\frac{1}{2}-\frac{h^2}{4\tau}\left(1-\frac{\varepsilon}{2}\right),\quad 0<\varepsilon\leq 2,$

so ist das (implizite) Verfahren *stabil* bzgl. Anfangswerten und rechter Seite (vgl. (4.18)), bzw. *bedingt stabil*, d.h. in Abhängigkeit von der Wahl der Schrittweiten.

b) Ist $\quad\sigma\geq\frac{1}{2},\quad$ ($\varepsilon\leq 2$ ist keine echte Einschränkung)

so ist das (implizite) Verfahren *unbedingt stabil* bzgl. Anfangswerten und rechter Seite
(d.h. ohne Bedingungen an die Schrittweiten τ, h).

c) Ist $\quad\sigma\geq\frac{1}{2}-\frac{h^2}{4\tau}\quad$ (vgl. (4.17))

(also $\varepsilon=0$ in a)), so ist das Verfahren *bedingt stabil* bzgl. der Anfangswerte.

d) Durch Verschärfung von c) zu $\sigma=0$ folgt insbesondere die schon bekannte Bedingung

Ist $\quad\frac{\tau}{h^2}\leq\frac{1}{2},$

(also $\sigma=0,\ \varepsilon=0$ in a)), so ist das (explizite) Verfahren bedingt stabil bzgl. der Anfangswerte.

Bermerkung:
$\varepsilon > 0$ kann in Voraussetzung a) beliebig klein sein. Daher liegt die Vermutung nahe, daß die Voraussetzung aus c) nicht nur die Stabilität bzgl. der Anfangswerte, sondern auch bzgl. der rechten Seite liefern könnte. Dies wird sich in der Tat bestätigen.

§ 5 Approximations- und Verfahrensfehler

Wir untersuchen das Verfahren (4.6), (4.7): $\boldsymbol{y}_t + \boldsymbol{A}_h^0 \boldsymbol{y}^{(\sigma)} = \boldsymbol{\varphi}$ in der Gestalt

$$(5.1) \qquad y^j_{t,i} - y^{j(\sigma)}_{\bar{x}x,i} - \varphi = 0, \quad i = 1, \ldots, N-1, \quad \boldsymbol{y}^{(\sigma)} = \sigma\hat{\boldsymbol{y}} + (1-\sigma)\boldsymbol{y}, \ \ \sigma \in [0,1].$$

Wir behalten uns eine genaue Definition von φ noch vor. φ ergab sich aus einer Diskretisierung des Differentialgleichungsproblems (vgl. etwa (2.12).). Auch andere Diskretisierungen sind denkbar. Es ist nur darauf zu achten, daß die Diskretisierung für $h, \tau \to 0$ gegen die Differentialgleichung konvergiert, und daß φ so gewählt wird, daß der Stabilitätsbegriff (vgl. dazu die Norm $\|\boldsymbol{\varphi}\|_{(b)}$) nicht darunter leidet.

Man beachte, daß gemäß der Definition von $\boldsymbol{y}^{(\sigma)}$ für jede Ableitung $\partial_x^\alpha (\alpha \hat{=} \text{Multiindex})$, und auch für jede entsprechende Diskretisierung d_x^α von ∂_x^α gilt

$$(\partial^\alpha u)^{(\sigma)} = \partial^\alpha \left(u^{(\sigma)}\right), \ \ (d^\alpha \boldsymbol{u})^{(\sigma)} = d^\alpha \left(\boldsymbol{u}^{(\sigma)}\right).$$

Sei u die Lösung der kontinuierlichen Aufgabe ($\hat{=}$ exakte Lösung) und $\boldsymbol{u} = (u_i^j)$ die durch ihre Restriktion auf das Gitter entstandene Gitterfunktion.

Wir bezeichnen mit $\boldsymbol{\psi}$ (vgl. (5.2)) den Approximationsfehler des Differenzenschemas.

$$(5.2) \qquad \psi_i^j := u^j_{t,i} - u^{j(\sigma)}_{\bar{x}x,i} - \varphi_i - \underbrace{\left(y^j_{t,i} - y^{j(\sigma)}_{\bar{x}x,i} - \varphi_i\right)}_{=0 \text{ laut Verfahren}}.$$

Da $\boldsymbol{u}$ und $\boldsymbol{y}$ die gleichen Randwerte haben, kann man ohne Einschränkung von **Nullrandwerten** ausgehen.

Definition 5.1
Sei u die exakte Lösung, so heißt die Gitterfunktion

$$\boldsymbol{\psi} := \boldsymbol{u}_t - \boldsymbol{u}^{(\sigma)}_{\bar{x}x} - \boldsymbol{\varphi} \quad (\text{Œ} \quad \text{Nullrandwerte})$$

Approximationsfehler (*Diskretisierungsfehler*, *truncation error*) des Differenzenschemas

$$\boldsymbol{y}_t - \boldsymbol{y}^{(\sigma)}_{\bar{x}x} - \boldsymbol{\varphi} = \boldsymbol{0}.$$

Zur Abschätzung des Approximationsfehlers benutzen wir (im Punkt x_i) die Taylorentwicklungen (2.3) und (2.5) bzgl. der Zeit an der Stelle $t^{j+\frac{1}{2}}$ (die Indizes z bezeichnen wieder geeignete Zwischenstellen) und erhalten (Raumindizes unterdrücken, φ wird zunächst nur mitgeführt):

$$\begin{aligned}
\psi^j =& \dot{u}^{j+\frac{1}{2}} + \frac{\tau^2}{24}\dddot{u}_z - (u'' + \frac{h^2}{12}u^{(4')} + \frac{h^4}{360}u_z^{(6')})^{(\sigma)} - \varphi \\
=& \dot{u}^{j+\frac{1}{2}} + \frac{\tau^2}{24}\dddot{u}_z - \left((u'' + \frac{h^2}{12}u^{(4')})^{(\sigma)} + (\frac{h^4}{360}u_z^{(6')})^{(\sigma)}\right) - \varphi \\
=& \dot{u}^{j+\frac{1}{2}} + \frac{\tau^2}{24}\dddot{u}_z \\
&- \left(\sigma(u''^{j+1} + \frac{h^2}{12}u^{(4'),j+1}) + (1-\sigma)(u''^{j} + \frac{h^2}{12}u^{(4'),j})\right) - \varphi + O(h^4)
\end{aligned}$$

Zusammenfassend folgt unter Beachtung von $\frac{\tau^2}{2}\ddot{u}_z = O(\tau^2)$

$$\psi^j = \dot{u}^{j+\frac{1}{2}} - \big[\underbrace{\sigma u''^{j+1} + (1-\sigma)u''^{j}}_{u''(\sigma)} + \frac{h^2}{12}(\underbrace{\sigma u^{(4'),j+1} + (1-\sigma)u^{(4'),j})}_{u^{(4')(\sigma)}})\big] - \varphi + O(\tau^2) + O(h^4)$$

(5.3) $$\psi^j = \dot{u}^{j+\frac{1}{2}} - u''^{(\sigma)} - \frac{h^2}{12}u^{(4')(\sigma)} - \varphi + O(\tau^2) + O(h^4).$$

Wir berechnen $u''^{(\sigma)}$ und $u^{(4')(\sigma)}$ durch Taylorentwicklung an $t^{j+\frac{1}{2}}$ bis auf Größenordnungen von $O(\tau^2)$ und $O(h^4)$. Der besseren Lesbarkeit halber schreiben wir - ausnahmsweise - die Zeitindizes unten:

$$\begin{aligned}
u''_{j+1} &= u''_{j+\frac{1}{2}} + \frac{\tau}{2}\dot{u}''_{j+\frac{1}{2}} + \frac{1}{2}\left(\frac{\tau}{2}\right)^2 \ddot{u}_z'' \\
u''_{j} &= u''_{j+\frac{1}{2}} - \frac{\tau}{2}\dot{u}''_{j+\frac{1}{2}} + \frac{1}{2}\left(\frac{\tau}{2}\right)^2 \ddot{u}_z'' \\
\Longrightarrow u''^{(\sigma)} &= u''_{j+\frac{1}{2}} + \frac{\tau}{2}(2\sigma-1)\dot{u}''_{j+\frac{1}{2}} + O(\tau^2) \\
u^{(4')}_{j+1} &= u^{(4')}_{j+\frac{1}{2}} + \frac{\tau}{2}\dot{u}^{(4')}_{j+\frac{1}{2}} + O(\tau^2) \\
u^{(4')}_{j} &= u^{(4')}_{j+\frac{1}{2}} - \frac{\tau}{2}\dot{u}^{(4')}_{j+\frac{1}{2}} + O(\tau^2) \\
\Longrightarrow u^{(4')(\sigma)} &= u^{(4')}_{j+\frac{1}{2}} + \frac{\tau}{2}(2\sigma-1)\dot{u}^{(4')}_{j+\frac{1}{2}} + O(\tau^2)
\end{aligned}$$

In (5.3) ist einzutragen $\frac{h^2}{12}u^{(4')(\sigma)}$. Für den 2. Summanden S_2 von $u^{(4')(\sigma)}$ folgt somit

$$S_2 := \frac{h^2}{12}\cdot\frac{\tau}{2}(2\sigma-1)\dot{u}^{(4')}_{j+\frac{1}{2}}\,.$$

Wegen $(\tau-h^2)^2 = \tau^2 - 2\tau h^2 + h^4 \geq 0$ ist $\tau h^2 \leq \frac{1}{2}(\tau^2+h^4) \implies S_2 \leq O(\tau^2+h^4)$.

Eintragen dieser Ergebnisse in (5.3) liefert

$$\begin{aligned}
\psi &= \dot u_{j+\frac12} - \left(u''_{j+\frac12} + \frac{\tau}{2}(2\sigma-1)\dot u''_{j+\frac12}\right) - \frac{h^2}{12}u^{(4')}_{j+\frac12} - \varphi + O(\tau^2+h^4)\\
&= \underbrace{\dot u_{j+\frac12} - u''_{j+\frac12} - f_{j+\frac12}}_{=0 \text{ laut Dgl.}} - \varphi + f_{j+\frac12} - \frac{\tau}{2}(2\sigma-1)\dot u''_{j+\frac12} - \frac{h^2}{12}u^{(4')}_{j+\frac12} + O(\tau^2+h^4)
\end{aligned}$$

$$\psi = -\varphi + f_{j+\frac12} - \frac{\tau}{2}(2\sigma-1)\dot u''_{j+\frac12} - \frac{h^2}{12}u^{(4')}_{j+\frac12} + O(\tau^2+h^4) \tag{5.4}$$

Bemerkung zur Schreibweise: Im Zusammenhang mit Differenzenverfahren schreibt man oft zur Vereinfachung der Schreibweise z.B. (wie oben geschehen)

$$O(\tau^2) + O(h^4) + O(\tau^2+h^4) = O(\tau^2+h^4),$$

was auf Grund der Definition des Landausymbols nicht korrekt ist. Die linke Schreibweise bedeutet: Es gibt Terme mit $O(\tau^2)$, $O(h^4)$ und solche mit $O(\tau^2+h^4)$. Da man bei der Konvergenzbehandlung üblicherweise τ und h <u>gleichzeitig</u> gegen Null gehen läßt, haben beide Schreibweisen den gleichen Effekt.

Setzt man in (5.4) $\varphi := f_{j+\frac12}$, so folgt hieraus, falls $u \in C^4$ bzgl. des Ortes und $u \in C^3$ bzgl. der Zeit:

$$\left\{\begin{aligned} \psi &= O(\tau+h^2), \text{ für beliebiges } \sigma\\ \psi &= O(\tau^2+h^2), \text{ für } \sigma = \tfrac12 \end{aligned}\right\} \quad \text{bzw. } \psi = O\left(\tau(\sigma-\tfrac12)+\tau^2+h^2\right)$$

Dieses Ergebnis läßt sich jedoch noch verbessern.

Aus der Differentialgleichung $\dot u - u'' - f = 0$ folgt $\dot u'' - u^{(4')} - f'' = 0$ oder

$$u^{(4')}_{j+\frac12} = \dot u''_{j+\frac12} - f''_{j+\frac12}$$

Eintragen in (5.4) ergibt

$$\underbrace{\psi = -\varphi + f_{j+\frac12} + \frac{h^2}{12}\, f''_{j+\frac12}}_{\substack{\overset{!}{=}0\\ \text{durch Wahl von } \varphi}} - \underbrace{\left[\frac{\tau}{2}(2\sigma-1) + \frac{h^2}{12}\right]}_{\substack{\overset{!}{=}0\\ \text{durch Wahl von } \sigma.}} \quad \dot u''_{j+\frac12} + O(\tau^2+h^4)$$

Insgesamt erhalten wir damit

Satz 5.2
Für den Approximationsfehler $\boldsymbol{\psi}$ des Verfahrens $\boldsymbol{y}_t - \boldsymbol{y}_{\bar{x}x}^{(\sigma)} - \boldsymbol{\varphi} = \mathbf{0}$ gilt

a) falls $\boldsymbol{\varphi} = \boldsymbol{f}^{j+\frac{1}{2}}$, $u \in C^4$ bzgl. des Ortes und $u \in C^3$ bzgl. der Zeit

$$\left.\begin{cases} \boldsymbol{\psi} = O(\tau + h^2), & \text{für beliebiges } \sigma \\ \boldsymbol{\psi} = O(\tau^2 + h^2), & \text{für } \sigma = \frac{1}{2} \end{cases}\right\} \quad \text{bzw. } \boldsymbol{\psi} = O\left(\tau(\sigma - \tfrac{1}{2}) + \tau^2 + h^2\right)$$

b) falls $\boldsymbol{\varphi} = \boldsymbol{f}^{j+\frac{1}{2}} + \dfrac{h^2}{12}\boldsymbol{f}''^{\,j+\frac{1}{2}}$, $\sigma = \dfrac{1}{2} - \dfrac{h^2}{12\tau}$, und
$u \in C^6$ bzgl. des Ortes und $u \in C^4$ bzgl. der Zeit gilt $\boldsymbol{\psi} = O(\tau^2 + h^4)$

Bemerkungen:

1. Die Numerik zeigt im Fall b) in der Tat eine deutlich schnellere Konvergenz.
2. In der Praxis will man f'' nicht gerne berechnen. Das ist oft zu fehleranfällig, wenn es denn überhaupt möglich ist. Man approximiert deshalb $\boldsymbol{\varphi}$ wie folgt durch:

$$(5.5) \qquad \left(f + \frac{h^2}{12}f''\right)_i^{j+\frac{1}{2}} \approx f_i^{j+\frac{1}{2}} + \frac{1}{12}\left(f_{i+1} - 2f_i + f_{i-1}\right)^{j+\frac{1}{2}} = \frac{1}{12}\left(f_{i+1} + 10f_i + f_{i-1}\right)^{j+\frac{1}{2}} \approx \varphi_i^j$$

   und erhält als Geschenk im allgemeinen die Folgerung

$$\boldsymbol{f} \geq \mathbf{0} \quad \Longrightarrow \quad \boldsymbol{\varphi} \geq \mathbf{0}.$$

3. **Beachte:**

   Die in der Stabilitätsdefinition benutzte Norm lautet nun

$$\|\boldsymbol{\varphi}\|_b^2 = \sum\sum \tau\, h |f_i^{j+\frac{1}{2}} + \frac{h^2}{12} f''^{\,j+\frac{1}{2}}|^2.$$

   Mit $\tau, h \to 0$ strebt auch diese Norm gegen $\int\limits_0^T \int\limits_0^1 f^2(x,t)\,dx\,dt$, sodaß die Stabilität ungefährdet ist.

   Nach Folgerung (4.3) war $\sigma \geq \dfrac{1}{2} - \dfrac{h^2}{4\tau}\left(1 - \dfrac{\varepsilon}{2}\right)$ hinreichend für Stabilität bzgl. der Anfangswerte und der rechten Seite. Die Forderung in Voraussetzung b) des vorigen Satzes lautet $\sigma = \dfrac{1}{2} - \dfrac{h^2}{12\tau}$. Für $\varepsilon \leq \dfrac{4}{3}$ gilt nun

$$\frac{1}{2} - \frac{h^2}{12\tau} \geq \frac{1}{2} - \frac{h^2}{4\tau}\left(1 - \frac{\varepsilon}{2}\right),$$

d.h unter der Voraussetzung $\varepsilon \leq \dfrac{4}{3}$ ist die Stabilität bzgl. der Anfangswerte und der rechten Seite (gemeint ist die Abhängigkeit von f) gewährleistet. Wenn wir später zeigen können, daß in Voraussetzung a) aus Folgerung 4.3 auch $\varepsilon = 0$ zugelassen ist, so ist dies nochmals eine Verschärfung des obigen Ergebnisses.

Definition 5.3 Verfahrensfehler des Schemas $\boldsymbol{y}_t - \boldsymbol{y}^{(\sigma)} - \boldsymbol{\varphi} = \boldsymbol{0}$

Sei $u(x,t)$ die exakte Lösung des kontinuierlichen Problems, (beliebige Anfangswerte, Randwerte und rechte Seite).

$\boldsymbol{u}$ die zugehörige Gitterfunktion

$\boldsymbol{y}$ die Gitterfunktion der Lösungsvektoren des Differenzenschemas

$$\boldsymbol{y}_t - \boldsymbol{y}_{\bar{x}x}^{(\sigma)} - \boldsymbol{\varphi} = \boldsymbol{0}.$$

Dann definieren wir als *Verfahrensfehler* die Gitterfunktion $\boldsymbol{z} := \boldsymbol{u} - \boldsymbol{y}$.

Bemerkung: Da für die Approximation als Rand- und Anfangswerte die exakten Werte vorgegeben werden, hat $\boldsymbol{z}$ Nullanfangs- und Randwerte.

Zur Abschätzung des Verfahrensfehlers (und damit zu Konvergenzaussagen) konstruieren wir ein Differenzschema für $\boldsymbol{z}$.

$$\begin{aligned} \boldsymbol{z}_t - \boldsymbol{z}_{\bar{x}x}^{(\sigma)} &= \boldsymbol{u}_t - \boldsymbol{u}_{\bar{x}x}^{(\sigma)} - \underbrace{\left[\boldsymbol{y}_t - (\boldsymbol{y}_{\bar{x}x})^{(\sigma)}\right]}_{= \boldsymbol{\varphi} \text{ laut Diffenzenschema}} \\ &= \boldsymbol{u}_t - \boldsymbol{u}_{\bar{x}x}^{(\sigma)} - \boldsymbol{\varphi} = \boldsymbol{\psi} \qquad \text{(vgl. Definition 5.1)} \end{aligned}$$

$\boldsymbol{z}$ erfüllt also ein **Differenzschema mit Nullrandwerten und Nullanfangswerten** und dem Approximationsfehler als rechte Seite.

$$\boldsymbol{z}_t - \boldsymbol{z}_{\bar{x}x}^{(\sigma)} - \boldsymbol{\psi} = \boldsymbol{0}.$$

Mit Hilfe des Stabilitätssatzes 4.2 b)) (Stabilität bzgl. der rechten Seite), der Folgerung 4.3 und des Satzes 5.2 erhält man also die Konvergenzaussage

Satz 5.4 Konvergenz
Gegeben sei das Differenzenschema $\boldsymbol{y}_t - \boldsymbol{y}_{\bar{x}x}^{(\sigma)} - \boldsymbol{\varphi} = \boldsymbol{0}$ mit beliebigen Anfangswerten und beliebigen Dirichlet-Randbedingungen.
Sei $\boldsymbol{z} = \boldsymbol{u} - \boldsymbol{y}$ der Verfahrensfehler und $\boldsymbol{\psi} = \boldsymbol{u}_t - \boldsymbol{u}_{\bar{x}x}^{(\sigma)} - \boldsymbol{\varphi}$ der Approximationsfehler, so gilt für $\sigma \geq \dfrac{1}{2} - \dfrac{h^2}{4\tau}\left(1 - \dfrac{\varepsilon}{2}\right)$:

$$\|\boldsymbol{z}^j\|_{A_h^0} \leq \left(\frac{1}{\varepsilon} \sum_{\nu=0}^{j-1} \tau \|\boldsymbol{\psi}^\nu\|_{(0,h)}^2 \right)^{\frac{1}{2}} \leq \sqrt{\frac{j\tau}{\varepsilon}} \max_{\nu \leq j-1} \|\boldsymbol{\psi}^\nu\|_{(0,h)}$$

und für $T \geq j \cdot \tau$ folgt also:

Falls $u \in C^3$ bzgl. Zeit, $u \in C^4$ bzgl. Ort gilt

$$\|\boldsymbol{z}^j\|_{A_h^0} \leq \begin{cases} O(\tau + h^2) & \text{für} \quad \sigma \neq \frac{1}{2} \\ O(\tau^2 + h^2) & \text{für} \quad \sigma = \frac{1}{2}, \ \boldsymbol{\varphi} = \boldsymbol{f}^{j+\frac{1}{2}}, \\ & \qquad (\text{stabil für } \varepsilon \leq 2) \end{cases}$$

Falls $u \in C^4$ bzgl. Zeit, $u \in C^6$ bzgl. Ort gilt

$$\|\boldsymbol{z}^j\|_{A_h^0} \leq \begin{cases} O(\tau^2 + h^4) & \text{für} \quad \sigma = \frac{1}{2} - \frac{h^2}{12}, \ \boldsymbol{\varphi} = \left(\boldsymbol{f} + \frac{h^2}{12}\boldsymbol{f}''\right)^{j+\frac{1}{2}}, \\ & \qquad (\text{stabil für } \varepsilon \leq \frac{4}{3}). \end{cases}$$

Bemerkungen:

1. Wesentlich für Konvergenz ist Stabilität bzgl. der rechten Seite

2. Beliebt (insbesondere bei Ingenieuren) sind Fehlerabschätzungen in der Maximumnorm (vgl. (3.24): $\|\boldsymbol{y}\|_\infty \leq \frac{1}{2}\|\boldsymbol{y}\|_{(1,h)}$). Eine Abschätzung des Approximationsfehlers $\|\boldsymbol{\psi}^k\|_{(0,h)}$ nach oben durch die Maximumsnorm ist zwar herleitbar, doch ist der Approximationsfehler $\|\boldsymbol{\psi}^k\|_{(0,h)}$ nur bei speziellen Beispielen auswertbar.

   In der Praxis ist man deshalb mit Konvergenz zufrieden und prüft die Genauigkeit durch Schrittweitenhalbierungen oder man testet das Verfahren (und natürlich damit auch die Programmierung) an geeigneten Testbeispielen.

Bemerkungen zur Konstruktion von Testbeispielen mit exakten Lösungen

Betrachte $r := \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2}$.

Ersetze u durch ein Polynom $p(x,t)$, berechne $g := \dfrac{\partial p}{\partial t} - \dfrac{\partial^2 p}{\partial x^2}$.

Dann ist p die exakte Lösung der Aufgabe $\dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial t^2} = g$ mit den Rand- und Anfangswerten $p|_\Gamma$ und $p|_{t=0}$.

Hinweis: Solche Konstruktionen, ebenfalls unter Verwendung von Polynomen, funktionieren auch bei nicht linearen Aufgaben der Art

$$c(u)\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k(u)\frac{\partial u}{\partial x}\right) + f(x) = 0.$$

Bevor wir auf solche Gleichungen eingehen, betrachten wir ein schnelles numerisches Verfahren zur Lösung der bisher behandelten Aufgaben.

§ 6 Spezielle Lösungsverfahren für lineare Gleichungssysteme

Der, das $\left\{\begin{array}{l}\text{verkürzte Gauß - Algorithmus}\\[1ex] \text{Tridiagonalalgorithmus}\\[1ex] \text{Thomas - Algorithmus}\\[1ex] \text{Double sweep - Verfahren}\end{array}\right.$

zur rechenzeitsparenden Auflösung von tridiagonalen Gleichungssystemen.

Wir leiten zuerst das Verfahren her und geben dann leicht nachprüfbare hinreichende Bedingungen für seine Durchführbarkeit an. Vorgelegt sei das tridiagonale Gleichungssystem:

(6.1)

$$\begin{array}{ccccccccl}
b_0y_0 & -c_0y_1 & & & & & = & f_0 & i=0\\
-a_1y_0 & +b_1y_1 & \ddots & & & & = & f_1 & \\
 & \ddots & \ddots & \ddots & & & & & \\
 & & -a_iy_{i-1} & +b_iy_i & -c_iy_{i+1} & & = & f_i & i=1,\ldots,n-1\\
 & & & \ddots & \ddots & \ddots & & & \\
 & & & & -a_ny_{n-1} & +b_ny_n & = & f_n & i=n.
\end{array}$$

Man zeigt schnell, daß das GEV eine LU-Zerlegung liefert (Lower, Upper): $\boldsymbol{A} = \boldsymbol{LU}$ mit Matrizen (Diagonale + untere oder obere Nebendiagonale)

$$\boldsymbol{L} = \begin{bmatrix} \ddots & & & \mathbf{0} \\ \ddots & \ddots & & \\ & \ddots & \ddots & \\ \mathbf{0} & & \ddots & \ddots \end{bmatrix}, \qquad \boldsymbol{U} = \begin{bmatrix} \ddots & \ddots & & \mathbf{0} \\ & \ddots & \ddots & \\ & & \ddots & \ddots \\ \mathbf{0} & & & \ddots \end{bmatrix}$$

Die Gleichung $\boldsymbol{LUy} = \boldsymbol{f}$ wird dann gelöst durch

1) $\boldsymbol{Lv} = \boldsymbol{f}$

2) $\boldsymbol{Uy} = \boldsymbol{v}$

Das System 2) hat dann – mit $\boldsymbol{U} = (u_{ij})$ – (abgesehen von der letzten Gleichung) die Gestalt

$$u_{ii}y_i + u_{i,i+1}y_{i+1} = v_i \qquad i = 0, \dots, n-1$$

$$(6.2) \qquad \begin{aligned} y_i &= \underbrace{-\frac{u_{i,i+1}}{u_{ii}}}\ y_{i+1} + \underbrace{\frac{v_i}{u_{ii}}} \\ &= \quad \alpha_{i+1} \quad y_{i+1} + \beta_{i+1} \end{aligned}$$

Wir leiten Rekursionsformeln zur Berechnung der $\alpha_{i+1}, \beta_{i+1}$ her und berechnen y_n. Dann können aus (6.2) die Werte y_i berechnet werden.

Wir setzen y_{i-1} gemäß (6.2) in die Zeilen $i = 1, \dots, n-1$ von (6.1) ein. $\Longrightarrow$

$$-a_i(\alpha_i y_i + \beta_i) + b_i y_i - c_i y_{i+1} = f_i \qquad i = 1, \dots, n-1$$

und erhalten durch Auflösen nach y_i die Rekursionsformel

$$(6.3) \qquad y_i = \frac{c_i}{b_i - \alpha_i a_i}\, y_{i+1} + \frac{f_i + a_i\beta_i}{b_i - \alpha_i a_i}, \quad i = 1, \dots, n-1, \qquad \underline{b_i - \alpha_i a_i \neq 0}$$

Aus dem Vergleich von (6.2) und (6.3) erhalten wir Rekursionsformeln für α_i, β_i:

$$(6.4) \qquad \alpha_{i+1} = \frac{c_i}{b_i - \alpha_i \alpha_i}, \quad \beta_{i+1} = \frac{f_i + \beta_i a_i}{b_i - \alpha_i a_i}, \quad i = 1, \dots, n-1, \quad b_i - \alpha_i a_i \neq 0$$

Aus der 0.ten Zeile von (6.1)folgt

$$b_0 y_0 - c_0 y_1 = f_0 \quad \Longrightarrow \quad y_0 = \frac{c_0}{b_0} y_1 + \frac{f_0}{b_0}, \qquad \underline{b_0 \neq 0}$$

Aus dem Vergleich mit (6.2) also

$$(6.5) \qquad \alpha_1 = \frac{c_0}{b_0}, \quad \beta_1 = \frac{f_0}{b_0}, \quad b_0 \neq 0.$$

Definieren wir $a_0 := 0$, so ist (6.5) in (6.4) enthalten für $i = 0$, und die Gleichung (6.3) gilt ebenfalls für $i = 0$.

Den Anfangswert y_n für (6.3) erhält man aus der letzten Zeile von (6.1), wenn man y_{n-1} aus (6.3), bzw. (6.2) in die letzte Zeile von (6.1) einsetzt.

$$-a_n(\alpha_n y_n + \beta_n) + b_n y_n = f_n \quad \Longrightarrow$$

$$(6.6) \qquad y_n = \frac{f_n + \beta_n a_n}{b_n - \alpha_n a_n} =: \beta_{n+1}, \qquad \underline{b_n - \alpha_n a_n \neq 0}$$

d.h. (6.6) ist auch der 2. Formel von (6.4) für β_{i+1} enthalten, wenn man dort $i = n$ zuläßt.

Insgesamt lautet das *Tridiagonalverfahren* also:

(I)	$\alpha_1 = \dfrac{c_0}{b_0}, \quad \beta = \dfrac{f_0}{b_0}, \quad b_0 \neq 0 \quad$ (vgl. (6.5)
	Berechne gemäß (6.4), (6.6)
(II)	$\alpha_{i+1} = \dfrac{c_i}{b_i - \alpha_i a_i}, \quad i = 1, \dots, n-1, \quad b_i - \alpha_i a_i \neq 0 \quad i = 1, \dots, n$
(III)	$\beta_{i+1} = \dfrac{f_i + \beta_i a_i}{b_i - \alpha_i a_i}, \quad i = 1, \dots, n, \quad b_i - \alpha_i a_i \neq 0 \quad i = 1, \dots, n$
	und gemäß (6.3)
(IV)	$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad i = 0, 1, \dots, n-1 \quad y_n = \beta_{n+1} = \dfrac{f_n + \beta_n a_n}{b_n - \alpha_n a_n}$

Bevor wir hinreichende Bedingungen für die Durchführbarkeit angeben, machen wir eine kurze Aufwandsbetrachtung.

Bezeichnet man als eine Operation den Aufwand für 1 Addition + Multiplikation oder für 1 Division, so benötigt man für

(I) 2 Operationen

(II) $2(n-1)$ Operationen (je 1 für den Nenner und eine für die Division

(III) $2n$ Operationen (je 1 für den Zähler und eine für den Bruch,
(der Nenner ist aus (II) bekannt),

(IV) n Operationen,

insgesamt also ca. 5 mal (Dimension des Systems–1). Beachte: Die Dimension des Systems ist $(n+1) \times (n+1)$.

Satz 6.1
Das Tridiagonalverfahren für $\boldsymbol{A} = \text{tridiag}(a_i, b_i, c_i), \quad a_0 := 0 =: c_n,$ ist durchführbar (d.h. $\exists A^{-1}$), falls

(1) $b_0 \neq 0$

(2) $|b_i| \geq |a_i| + |c_i|, \quad i = 0, \dots, n$

(3) $|b_i| > |a_i|, \quad i = 1, \dots, n$

Die Forderung (3) kann ersetzt werden durch

(3′) $|b_0| > |c_0| > 0$ und $|a_i| > 0$ für $i = 1, 2, \dots, n$

oder

(3″) $\exists j$ sodaß: (2) mit "$>$" gilt, $\quad |c_i| > 0, i = 1, \dots, j-1, \quad |a_i| > 0 \; \forall\, i > j.$

Bemerkung: Das Beispiel $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$ zeigt, daß die Bedingungen für die Invertierbarkeit von A nur hinreichend sind. $(3), (3'), (3'')$ versagen.

Beweis zunächst unter Benutzung von (3):

Zeige: Alle Nenner sind $\neq 0$ im Rahmen eines Induktionsbeweises für die Aussage $|\alpha_i| \leq 1 \quad \forall i$

$$\alpha_1 = \frac{c_0}{b_0} \quad \overset{(1)(2)}{\Longrightarrow} \quad |\alpha_1| \leq 1.$$

Sei also $|\alpha_i| \leq 1$, dann gilt

$$|b_i - \alpha_i a_i| \geq |b_i| - |\alpha_i||a_i| \geq |b_i| - |a_i| \underset{\uparrow}{\overset{(3)}{>}} 0, \quad i \geq 1 \tag{6.7}$$

Damit sind alle $\alpha_{i+1}, \beta_{i+1}$ aus (II) und (III) erklärt und es gilt

$$|\alpha_{i+1}| = \frac{|c_i|}{|b_i - \alpha_i a_i|} \leq \frac{|c_i|}{|b_i| - |a_i|} \overset{(2)}{\leq} 1.$$

Das Verfahren ist durchführbar.

Ersetzt man (3) durch $(3')$, so ist $|\alpha_1| = |\frac{c_0}{b_0}| < 1$ und man erhält induktiv $|\alpha_i| < 1 \forall\, i$, denn mit $|\alpha_i| < 1,$ für ein $i \leq n$, $|a_j| > 0$ für $j = 1, ..., n$, erhält man im Induktionsschritt

$$|b_i - \alpha_i a_i| \geq |b_i| - |\alpha_i||a_i| \underset{\uparrow}{\overset{(3')}{>}} |b_i| - |a_i| \overset{(2)}{\geq} |c_i| \geq 0 \quad \Longrightarrow \quad |\alpha_{i+1}| < 1 \quad \forall_i \geq 1$$

und alle Nenner sind $\neq 0$.

Benutzt man $(3'')$, so zeigt man $|\alpha_i| \leq 1$ (induktiv) für $i = 1, \ldots, n$ gemäß

$$\alpha_1 = \frac{c_0}{b_0}, \quad 0 \leq |\alpha_1| \leq 1,$$

und im Induktionsschritt

$$|b_i - \alpha_i a_i| \geq |b_i| - |\alpha_i||a_i| \geq |b_i| - |a_i| \geq |c_i| \underset{\uparrow}{>} 0 \text{ für } i = 1, \ldots, j-1$$

$$\Longrightarrow \quad |\alpha_{i+1}| \leq 1,\ i \leq j.$$

Für das j gilt:

$$|b_j - \alpha_j a_j| \geq |b_j| - |\alpha_j||a_j| \geq |b_j| - |a_j| \underset{\uparrow}{>} |c_j| \geq 0 \quad \Longrightarrow \quad |\alpha_{j+1}| < 1$$

und mit $|\alpha_{j+1}| < 1, |a_{j+1}| > 0$ (vgl. 3'))

$$|b_{j+1} - \alpha_{j+1} a_j| \geq |b_{j+1}| - |\alpha_{j+1}||a_{j+1}| \underset{\uparrow}{>} |b_{j+1}| - |a_{j+1}| \overset{(2)}{\geq} |c_j| \geq 0 \Longrightarrow |\alpha_{j+2}| < 1$$

usw. ∎

Bemerkung zu Satz 6.1
In der Praxis tritt üblicherweise folgender Fall auf:

$$\begin{aligned} &b_i > 0 \quad i = 0, \dots, n \\ &a_i < 0, \;\; i = 1, \dots, n \\ &c_i < 0, \;\; i = 0, \dots, n-1 \\ &b_i + a_i + c_i \geq 0 \quad \forall\, i \quad (b_0 = c_n = 0) \\ &\qquad\qquad > 0 \text{ für mindestens ein } i. \end{aligned}$$

Hierfür ist das Tridiagonalverfahren durchführbar (vgl. (3')).

Wir zeigen nun, daß man für spezielle Matrizen (M-Matrizen), die bei der Diskretisierung parabolischer Probleme entstehen, Invertierbarkeit und Normabschätzungen (in der Maximumnorm!) ohne die Kenntnis der Eigenwerte erhalten kann.

Definition 6.2
$A \in \mathbb{R}^{n\times n}$ heißt *M-Matrix* wenn

$$\begin{aligned} &a_{ij} \leq 0, \text{ für } i \neq j \text{ und} \\ &\exists\, \boldsymbol{p} \in \mathbb{R}^n, \; \boldsymbol{p} > 0 \text{ (komponentenweise), sodaß } \boldsymbol{Ap} > \boldsymbol{0} \text{ (komponentenweise).} \end{aligned}$$

Satz 6.3
Ist $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ eine M-Matrix, so gilt

a) $\exists \boldsymbol{A}^{-1}$ und $\boldsymbol{A}^{-1} \geq \boldsymbol{0}$ (elementweise)

b) $\displaystyle ||\boldsymbol{A}^{-1}||_\infty \leq \frac{||\boldsymbol{p}||_\infty}{\min\limits_i (\boldsymbol{Ap})_i}$, $\boldsymbol{p}$ gemäß Definition 6.2 $\displaystyle \left(||\boldsymbol{A}||_\infty := \max_{i=1,\dots,n} \sum_{k=1}^n |a_{ik}| \right)$.

c) Ist $\boldsymbol{Ax} = \boldsymbol{b}$, so gilt schärfer $\displaystyle ||\boldsymbol{x}||_\infty \leq \frac{||\boldsymbol{p}||_\infty}{\min\limits_{i:\, b_i \neq 0} (\boldsymbol{Ap})_i} ||\boldsymbol{b}||_\infty$

Beweis a): Wir zeigen zeigen $\exists\, \boldsymbol{D}^{-1}$, zerlegen

(6.8) $$\boldsymbol{A} = \boldsymbol{D} - \boldsymbol{B} = \boldsymbol{D}(\boldsymbol{I} - \underbrace{\boldsymbol{D}^{-1}\boldsymbol{B}}_{\boldsymbol{C}}) =: \boldsymbol{D}(\boldsymbol{I} - \boldsymbol{C})$$

und zeigen: $\exists\, (\boldsymbol{I} - \boldsymbol{C})^{-1}$. Hieraus folgt die Behauptung a).

Wegen $\boldsymbol{p} > 0$, $\boldsymbol{Ap} > 0$, $a_{ij} \leq 0, i \neq j$ folgt $a_{ii} > 0 \forall\, i$, sonst W!, denn

$$(\boldsymbol{Ap})_i = \sum_{j=1}^{i-1} p_j a_{i,j} + p_{ii} a_{i,i} + \sum_{j=i+1}^{n} p_j p_j a_{i,j} > 0, \; a_0 = a_n = 0 \quad \xRightarrow{a_{ij} \leq 0,\, i \neq j} \quad a_{ii} > 0.$$

$\Longrightarrow \quad \boldsymbol{D} = \operatorname{diag}(\boldsymbol{A}) > 0$ (elementweise) ist invertierbar.

Wegen $\boldsymbol{B} \geq \mathbf{0}$, $\boldsymbol{D}^{-1} > \mathbf{0}$ (jeweils elementweise), folgt $\boldsymbol{C} := \boldsymbol{D}^{-1}\boldsymbol{B} \geq \mathbf{0}$ (elementweise).

Multiplikation von $\mathbf{0} < \boldsymbol{A}\boldsymbol{p} = \boldsymbol{D}(\boldsymbol{I} - \boldsymbol{C})\boldsymbol{p}$, (vgl. (6.8)), mit $\boldsymbol{D}^{-1}$ von links liefert

$$\mathbf{0} < \boldsymbol{D}^{-1}\boldsymbol{A}\boldsymbol{p} = \boldsymbol{p} - \boldsymbol{C}\boldsymbol{p} \quad \text{und mit} \quad \boldsymbol{C}\boldsymbol{p} \geq \mathbf{0}$$

$$\mathbf{0} \leq \boldsymbol{C}\boldsymbol{p} < \boldsymbol{p}. \tag{6.9}$$

Wegen $\boldsymbol{p} > \mathbf{0}$ ist $\boldsymbol{P} := \operatorname{diag}(p_i)$ regulär und $\exists\, \boldsymbol{P}^{-1} \geq \mathbf{0}$ (elementweise).

$$\begin{aligned} \Longrightarrow \quad & ||\boldsymbol{x}||_p := ||\boldsymbol{P}^{-1}\boldsymbol{x}||_\infty && \text{ist eine Vektornorm,} \\ & ||\boldsymbol{A}||_p := ||\boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P}||_\infty && \text{ist die zugeordnete Matrixnorm.} \end{aligned}$$

Wir zeigen

$$||\boldsymbol{C}||_p < 1.$$

Mit $\boldsymbol{e} = (1, 1, \ldots, 1)^T > \mathbf{0}$ gilt $\boldsymbol{P}\boldsymbol{e} = \boldsymbol{p}$

$$\overset{(6.9)}{\Longrightarrow} \qquad \mathbf{0} \leq \boldsymbol{C}\boldsymbol{P}\boldsymbol{e} < \boldsymbol{P}\boldsymbol{e}.$$

Multiplikation mit $\boldsymbol{P}^{-1}$ von links ergibt

$$\mathbf{0} \leq (\boldsymbol{P}^{-1}\boldsymbol{C}\boldsymbol{P})\boldsymbol{e} < \boldsymbol{e} \qquad \text{(komponentenweise),}$$

d.h. in jeder Zeile der Matrix $\boldsymbol{P}^{-1}\boldsymbol{C}\boldsymbol{P}$ ist die Betragssumme der Elemente < 1, d.h.

$$||\boldsymbol{C}||_p = ||\boldsymbol{P}^{-1}\boldsymbol{C}\boldsymbol{P}||_\infty < 1 \quad \overset{\text{Satz3.1}}{\Longrightarrow} \quad \exists(\boldsymbol{I} - \boldsymbol{C})^{-1} = \sum_{\nu=0}^{\infty} \boldsymbol{C}^\nu \geq \mathbf{0} \qquad \text{(elementweise)}$$

$$\overset{(6.8)}{\Longrightarrow} \qquad \exists \boldsymbol{A}^{-1} = \underbrace{(\boldsymbol{I} - \boldsymbol{C})^{-1}}_{\geq \mathbf{0}} \underbrace{\boldsymbol{D}^{-1}}_{\geq \mathbf{0}} \geq \mathbf{0} \qquad \text{(elementweise).}$$

Beweis c)

Für $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ und $\boldsymbol{A}^{-1} \geq \mathbf{0}$ (elementweise) suchen wir für $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ ein Abschätzung

$$||\boldsymbol{x}||_\infty \leq k||\boldsymbol{b}||_\infty$$

Bemerkung: Ist k unabhängig von $\boldsymbol{b}$, so gilt laut Definition der Matrixnorm

$$||\boldsymbol{A}^{-1}||_\infty \leq k,$$

denn

$$||\boldsymbol{x}||_\infty = \|\boldsymbol{A}^{-1}\boldsymbol{b}\|_\infty \leq \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{x}\| \quad \text{und} \quad \|\boldsymbol{A}^{-1}\|_\infty = \inf_{\boldsymbol{b} \neq \mathbf{0}} \{k; \|\boldsymbol{A}^{-1}\boldsymbol{b}\|_\infty \leq k\|\boldsymbol{b}\|_\infty \ \forall\, \boldsymbol{b} \in \mathbb{R}^n\}.$$

Ziel: Für einen Vektor $\boldsymbol{p} > \mathbf{0}$ mit $(\boldsymbol{A}\boldsymbol{p})_i > 0\ \forall_i$ bestimmen wir ein $m > 0$ so, daß

$$\boldsymbol{A}(m\boldsymbol{p} \pm \boldsymbol{x}) = m\boldsymbol{A}\boldsymbol{p} \pm \boldsymbol{A}\boldsymbol{x} = m\boldsymbol{A}\boldsymbol{p} \pm \boldsymbol{b} \overset{!}{\geq} 0 \tag{6.10}$$

Es genügt $m > 0$ so zu bestimmen, daß (6.10) für die Komponenten i mit $b_i \neq 0$ erfüllt ist, denn für die anderen Komponenten gilt (6.10) ohnehin. Fordere also

$$m(\boldsymbol{Ap})_i \geq \pm b_i \quad \forall_i \quad \text{mit} \quad b_i \neq 0.$$

Der „schlimmste" Fall liegt vor, wenn links das Minimum, rechts das Maximum angenommen wird. Wir definieren deshalb m durch die Forderung

$$m \min_{i:\,\boldsymbol{b}\neq\boldsymbol{0}} (\boldsymbol{Ap})_i = ||\boldsymbol{b}||_\infty \quad \text{bzw.} \quad m := \frac{||\boldsymbol{b}||_\infty}{\min\limits_{i:\,\boldsymbol{b}\neq\boldsymbol{0}} (\boldsymbol{Ap})_i}.$$

Dann gilt für alle Komponenten $m\boldsymbol{Ap} \geq \pm\boldsymbol{b}$ und wegen $\boldsymbol{A}^{-1} \geq \boldsymbol{0}$, $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$, folgt

$$m\boldsymbol{p} \geq \pm\boldsymbol{x} \qquad \text{bzw. wegen } \boldsymbol{p} > \boldsymbol{0} \qquad |x_i| \leq m\, p_i$$

$$\Longrightarrow \quad ||\boldsymbol{x}||_\infty \leq m||\boldsymbol{p}||_\infty = \frac{||\boldsymbol{p}||_\infty}{\min\limits_{i:\,b_i\neq 0} (\boldsymbol{Ap})_i} ||b||_\infty, \qquad \text{also Behauptung c).}$$

Beweis b)
Wird insbesondere m unabhängig von $\boldsymbol{b}$ bestimmt, d.h. unabhängig von den Nullkomponenten von $\boldsymbol{b}$, so folgt aus der letzten Abschätzung

$$||\boldsymbol{x}||_\infty \leq \frac{||\boldsymbol{p}||_\infty}{\min\limits_{i=1,\dots,n} (\boldsymbol{Ap})_i} ||\boldsymbol{b}|| \ \forall \boldsymbol{b}, \qquad \text{also Behauptung b).}$$

■

Im Anschluß an Satz 6.1 und der Bemerkung von Satz 6.1 zeigen wir

Satz 6.4
Für $\boldsymbol{A} = \text{tridiag}(-a_i, b_i, -c_i)$ sei

$$\begin{aligned} a_i, b_i, c_i \ &> \ 0\ \forall_i \quad \text{außer } a_0 = c_n = 0 \\ b_i \ &\geq \ a_i + c_i \ \ \forall_i \\ &> \ \text{für mindestens ein } i \end{aligned}$$

$$\Longrightarrow \qquad \boldsymbol{A} \ \text{ ist eine } M-\text{Matrix.}$$

Bemerkungen:

1. Solche Matrizen entstehen bei der Diskretisierung eines parabolischen Problems.
2. Satz 6.1 zeigt, daß für diese Matrizen das Tridiagonalverfahren durchführbar ist, daß also die Werte auf der neuen Zeitschicht berechenbar sind.

3. Satz 6.4 liefert dann Abschätzungen für die Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$.

4. „$\Longleftarrow$“ gilt nicht. Beispiel: $\boldsymbol{A} = \boldsymbol{I}$ ist M-Matrix ($\boldsymbol{p} = \boldsymbol{e} = (1, ..., 1)^T$) und $\boldsymbol{A} = \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$ mit $\boldsymbol{p} = (1, 0.25)^T$ ebenfalls.

Beweis:

1. Gemäß der Definition der M-Matrizen (vgl. Definition 6.2) ist ein Vektor $\boldsymbol{p} > 0$ (komponentenweise) zu finden mit $\boldsymbol{Ap} > 0$.

   Mit $\boldsymbol{e} = (1, \dots, 1)^T$ gilt $\boldsymbol{Ae} \geq 0$ (komponentenweise, da nach Voraussetzung $b_i \geq a_i + c_i$).

   Für alle $\varepsilon > 0$ gilt deshalb

   $$(\varepsilon \boldsymbol{I} + \boldsymbol{A})\boldsymbol{e} \geq \varepsilon \boldsymbol{e} > \boldsymbol{0}, \quad \varepsilon \boldsymbol{I} + \boldsymbol{A} \text{ ist eine } M\text{-Matrix mit } \boldsymbol{p} = \boldsymbol{e} \overset{\text{Satz 6.3}}{\Longrightarrow}$$

   (6.11) $$\exists\, (\varepsilon \boldsymbol{I} + \boldsymbol{A})^{-1} \geq \boldsymbol{0} \quad \text{(elementweise)}$$

   Satz 6.1 mit (3') zeigt $\exists\, \boldsymbol{A}^{-1}$.

   Deshalb kann man im (6.11) den Grenzübergang $\varepsilon \to 0$ machen. (Die Existenz von $\boldsymbol{A}^{-1}$ bleibt ja erhalten) $\Longrightarrow$

   (6.12) $$\boldsymbol{A}^{-1} \geq \boldsymbol{0} \qquad \text{(elementweise)}$$

2. Wir konstruieren nun das gesuchte $\boldsymbol{p} > 0$. Für $\boldsymbol{e} = (1, \dots, 1)^T$ existiert (vgl. (6.12)), die Lösung von

   $$\boldsymbol{Ay} = \boldsymbol{e} \quad \overset{(6.12)}{\Longrightarrow} \quad \boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{e} \geq \boldsymbol{0} \quad \text{(elementweise)}.$$

   Sei $\boldsymbol{A}^{-1} = (\alpha_{ij})$. Annahme: $\exists i : \; y_i = 0$ d.h. $\sum\limits_{j=1}^{n} \alpha_{ij} \cdot 1 = 0$

   $\overset{(6.12)}{\Longrightarrow} \alpha_{ij} = 0 \quad \forall j,$ also existiert $\boldsymbol{A}^{-1}$ nicht W!

   Also kann $\boldsymbol{p} = \boldsymbol{y}$ gewählt werden. ■

Damit erhalten wir, insbesondere aus Beweisteil 2) die Charakterisierung:

Satz 6.5

Für $\boldsymbol{A} \in R^{n \times n}$ gilt

$$\boldsymbol{A} \text{ ist } M\text{-Matrix} \quad \Longleftrightarrow \quad \exists \boldsymbol{A}^{-1} \geq \boldsymbol{0} \text{ (elementweise)}.$$

Beweis: „$\Longrightarrow$“ liefert sofort Satz 6.3 a).

„$\Longleftarrow$“ liefert Beweisteil 2) des vorigen Satzes, der keinen Gebrauch von der Tridiagonalform machte. ■

§ 7 Die Gleichung $u_t = \frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right) + f$

In der Anwendung ist k oft stückweise stetig. Wie kann $\frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right) =: R$ geeignet diskretisiert werden? Physikalisch stellt dieser Ausdruck eines Wärmestrom dar, wenn u die Temperatur ist. Deshalb sollte er bei der Diskretisierung nicht durch Ausdifferenzieren ($R = ku_{\bar{x}x} + k_{x^0}u_{x^0}$, wobei etwa $\boldsymbol{y}_{x^0} = \frac{\boldsymbol{y}_{i+1}-\boldsymbol{y}_{i-1}}{2} = \frac{1}{2}\left(\boldsymbol{y}_x + \boldsymbol{y}_{\bar{x}}\right)$ – zentral wegen besserer Approximationsordnung –) in nicht interpretierbare Summanden zerlegt werden. Würde man dies trotzdem tun, so ergäbe sich für R eine nicht symmetrische Matrix. Frau(Man) kann das leicht nachrechnen. Dies ist auch aus mathematischen Gründen ungünstig, da der Operator $Lu = \frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right)$ bei geeigneten Randbedingungen selbstadjungiert ist, wie eine kurzer Rechnung zeigt.
Selbstadjungiert (das ist für Operatoren die Verallgemeinerung des Begriffs symmetrisch, bzw. hermite'sch bei Matrizen) bedeutet (z.B. für $x \in [0,1]$

$$(Lu, v) = (u, Lv), \quad (u,v) = \int_0^1 u(x)v(x)dx;$$

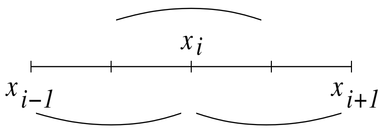
damit

$$\begin{aligned}(Lu,v) = \int_0^1 (ku')'v\,dx =& [ku'v]_0^1 - \int_0^1 ku'v'\,dx \\ =& \underbrace{[ku'v]_0^1 - [kv'u]_0^1}_{=0 \text{ bei entsprechenden Randwerten}} + \int_0^1 (kv')'v\,dx = \int_0^1 u(x)u(x)dx\end{aligned}$$

In der Tat zeigen auch numerische Beispiele, daß eine nichtsymmetrische Diskretisierung Konvergenz des Verfahrens gegen falsche Werte liefern kann. Deshalb wird der Wärmestrom als Ganzes diskretisiert. Da wir beim Wärmestrom nur mit Ortsableitungen zu tun haben, unterdrücken wir in der Bezeichnung die Ortsabhängigkeit. Dabei benutzen wir, so weit als möglich, zentrale Differenzenquotienten der besseren Approximationseigenschaften wegen.
Mit $y_i \approx u(x_i)$ approximieren wir

$$\left(k\frac{\partial u}{\partial x}\right)_{i+\frac{1}{2}} \approx k_{i+\frac{1}{2}}\frac{y_{i+1}-y_i}{h} = k_{i+\frac{1}{2}}y_{x,i},$$

$$(ku')_i' \approx \frac{1}{h}(k_{i+\frac{1}{2}}y_{x,i} - k_{i-\frac{1}{2}}y_{\bar{x},i}) =: (k \cdot y_x)_{\bar{x},i} \qquad \text{(abkürzende Bezeichnung)}$$



$$(k \cdot y_x)_{\bar{x},i} = \frac{1}{h}\left(k_{i+\frac{1}{2}}\frac{y_{i+1}-y_i}{h} - k_{i-\frac{1}{2}}\frac{y_i - y_{i-1}}{2}\right) \tag{7.1}$$

$$= \frac{1}{h^2}\left(k_{i+\frac{1}{2}}y_{i+1} - \frac{1}{h^2}\left(k_{i+\frac{1}{2}} + k_{i-\frac{1}{2}}\right)y_i + \frac{1}{h^2}k_{i-\frac{1}{2}}y_{i-1}\right) =: -\left(\boldsymbol{A}_h(k)\boldsymbol{y}\right)_i \tag{7.2}$$

Als Diskretisierungsmatrix für $A_h(k)$ erhalten wir – nachdem die Randwerte aus der Matrix eliminiert wurden (d.h. “der rechten Seite zugeschlagen”) – eine tridiagonale, symmetrische Matrix

$$A_h(k) = \frac{1}{h^2}\begin{pmatrix} k_{\frac{1}{2}}+k_{\frac{3}{2}} & -k_{\frac{3}{2}} & & \\ -k_{\frac{3}{2}} & k_{\frac{3}{2}}+k_{\frac{5}{2}} & -k_{\frac{5}{2}} & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots \end{pmatrix}$$

Wir untersuchen

1. die Approximationseigenschaften, d.h. den Diskretisierungsfehler der Diskretisierung,
2. die Stabilität des resultierenden Verfahrens,
3. die Konvergenzeigenschaften.

Diskretisierungsfehler

Für $u \in C^4$, $k \in C^3$ (bzgl. Ort) untersuchen wir die Diskretisierung von (7.1).

$$u_{i+\frac{1}{2}\pm\frac{1}{2}} = u_{i+\frac{1}{2}} \pm \frac{h}{2}u'_{i+\frac{1}{2}} + \frac{h^2}{8}u''_{i+\frac{1}{2}} \pm \frac{h^3}{48}u'''_{i+\frac{1}{2}} + \frac{1}{4!}\left(\frac{h}{2}\right)^4 u''''_{\pm}$$

liefert

$$u_{x,i} = \frac{u_{i+1} - u_i}{h} = u'_{i+\frac{1}{2}} + \frac{h^2}{24}u'''_{i+\frac{1}{2}} + O(h^3).$$

Damit erhält man

$$\begin{aligned} k_{i+\frac{1}{2}}u_{x,i} &= (ku')_{i+\frac{1}{2}} + \tfrac{h^2}{24}(ku''')_{i+\frac{1}{2}} + O(h^3), \\ k_{i-\frac{1}{2}}u_{\bar{x},i} &= (ku')_{i-\frac{1}{2}} + \frac{h^2}{24}(ku''')_{i-\frac{1}{2}} + O(h^3). \end{aligned}$$

Damit folgt (vgl. die Bezeichnung in (7.1))

$$\begin{aligned} (ku_x)_{\bar{x},i} &= \frac{1}{h}\left(k_{i+\frac{1}{2}}u_{x,i} - k_{i-\frac{1}{2}}u_{\bar{x},i}\right) \\ &= \frac{1}{h}\Big[(ku')_{i+\frac{1}{2}} - (ku')_{i-\frac{1}{2}} + \underbrace{\frac{h}{24}\left\{(ku''')_{i+\frac{1}{2}} - (ku''')_{i-\frac{1}{2}}\right\}}_{\frac{h^2}{24}(ku''')'_z = O(h^2)}\Big] + O(h^2). \end{aligned}$$

Die ersten beiden Summanden der rechten Seite müssen an der Stelle x_i entwickelt werden.

$$(ku')_{i\pm\frac{1}{2}} = (ku')_i \pm \frac{h}{2}(ku')'_i + \frac{1}{2!}\left(\frac{h}{2}\right)^2 (ku')''_i \pm \frac{h^3}{48}(ku')'''_{z\pm}\,.$$

Damit folgt

$$(ku_x)_{\bar{x},i} = (ku')'_i + \underbrace{\frac{h^2}{48}\left[(ku')'''_{z_+} - (ku')'''_{z_-}\right]}_{O(h^2)} + O(h^2) \quad \text{bzw.}$$

(7.3) $$(ku_x)_{\bar{x},i} = (ku')'_i + O(h^2) \quad \text{für } u \in C^2,\ k \in C^3 \text{ (bzgl. Ort)}$$

Wir untersuchen nun gleich **das gewichtete Verfahren**

(7.4) $$\boldsymbol{y}_t = (k\boldsymbol{y}_x)^{(\sigma)}_{\bar{x}} + \boldsymbol{\varphi}, \quad \boldsymbol{y}^{(\sigma)} = \sigma\hat{\boldsymbol{y}} + (1-\sigma)\boldsymbol{y} = \sigma\tau\boldsymbol{y}_t + \boldsymbol{y}$$

Beachte: $k = k(x)$ ist zeitlich konstant, deshalb kann man numerische Differentiation bzgl. x und Mittelwertbildung mittels σ vertauschen. Auf Grund des Matrixdarstellung (7.2) folgt aus der Linearität in $\boldsymbol{y}$ aus (7.4)

$$\begin{aligned} \boldsymbol{y}_t &= -\boldsymbol{A}_h(k)[\sigma\hat{\boldsymbol{y}} + (1-\sigma)\boldsymbol{y}] + \boldsymbol{\varphi} \\ &= -\boldsymbol{A}_h(k)[\sigma\tau\boldsymbol{y}_t + \boldsymbol{y}] + \boldsymbol{\varphi} \qquad \text{oder} \end{aligned}$$

(7.5) $$(\boldsymbol{I} + \sigma\tau\boldsymbol{A}_h(k))\boldsymbol{y}_t + \boldsymbol{A}_h(k)\boldsymbol{y} = \boldsymbol{\varphi}.$$

Dies ist eine Gestalt, die unter die Verfahrensklasse

$$\boldsymbol{B}\boldsymbol{y}_t + \boldsymbol{A}\boldsymbol{y} = \boldsymbol{\varphi}, \quad \boldsymbol{B} = \boldsymbol{I} + \sigma\tau\boldsymbol{A}, \quad \text{mit} \quad \boldsymbol{A} = \boldsymbol{A}_h(k)$$

fällt, für die Stabilitätssätze gelten. Wir müssen die Voraussetzungen prüfen. $\boldsymbol{A}$ symmetrisch ist klar (vgl. (7.2)). Wir zeigen zunächst

$\boldsymbol{A}_h(k)$ ist positiv definit. (Beweis analog zu (3.13))

$$\begin{aligned} (\boldsymbol{A}_h(k)\boldsymbol{y}, \boldsymbol{y})_{(0,h)} &= \sum_{i=1}^{N-1} (\boldsymbol{A}_h(k)\boldsymbol{y})_i y_i h \overset{(7.1)}{=} -\sum_{i=1}^{N-1} (ky_x)_{\bar{x},i} y_i h \\ &\overset{(7.1)}{=} -\sum_{i=1}^{N-1} k_{i+\frac{1}{2}} y_{x,i} y_i + \sum_{i=1}^{N-1} k_{i-\frac{1}{2}} y_{x,i-1} y_i \quad \text{und mit} \quad y_0 = y_n = 0 \\ &= -\sum_{i=0}^{N-1} k_{i+\frac{1}{2}} y_{x,i} y_i + \sum_{i=1}^{N} k_{i-\frac{1}{2}} y_{x,i-1} y_i \\ &= -\sum_{i=1}^{N} k_{i-\frac{1}{2}} y_{x,i-1} y_{i-1} + \sum_{i=1}^{N} k_{i-\frac{1}{2}} y_{x,i-1} y_i \\ &= \sum_{i=1}^{N} k_{i-\frac{1}{2}} y_{x,i-1}(y_i - y_{i-1}) \end{aligned}$$

(7.6) $$= \sum_{i=1}^{N} k_{i-\frac{1}{2}} (y_{x,i-1})^2 h = \sum_{i=0}^{N-1} k_{i+\frac{1}{2}} (y_{x,i})^2 h > 0$$

$$= 0 \text{ nur für } \boldsymbol{y} \equiv 0, \ \text{da } y_0 = y_N = 0$$

Unter der Voraussetzung $0 < c_0 \le k(x) \le c_1 \quad \forall x \in [0,1]$ folgt somit aus (7.6) (vgl. (3.8))

(7.7) $$\boldsymbol{A}_h(k) > 0 \text{ und } (\boldsymbol{A}_h(k)\boldsymbol{y},\boldsymbol{y})_{(0,h)} \ge c_0||\boldsymbol{y}||^2_{(1,h)} =: c_0||\boldsymbol{y}||^2_{\boldsymbol{A}_h^0}$$

Wir können also eine Vektornorm definieren durch

(7.8) $$||\boldsymbol{y}||^2_{\boldsymbol{A}_h(k)} = ||\boldsymbol{y}||^2_{(1,h,k)} = (\boldsymbol{A}_h(k)\boldsymbol{y},\boldsymbol{y})_{(0,h)} := \sum_{i=0}^{N-1} k_{j+\frac{1}{2}}(y_{x,i})^2 k,\ y_0 = y_N = 0.$$

Abschätzungen von $\boldsymbol{A}_h(k)$ nach oben und unten (im Sinne $(\ ,\ )_{(0,h)}$).
Unter Benutzung von (3.23): $||\boldsymbol{y}||^2_{(1,h)} \ge 8||\boldsymbol{y}||^2_{(0,h)}$ folgt aus (7.7)

$$(\boldsymbol{A}_h(k)\boldsymbol{y},\boldsymbol{y})_{(0,h)} \ge c_0||\boldsymbol{y}||^2_{(1,h)} \ge 8c_0||\boldsymbol{y}||^2_{(0,h)},$$

also

(7.9) $$\boldsymbol{A}_h(k) \ge 8c_0\boldsymbol{I} \quad \text{(im Sinne positiv semidefinit).}$$

Abschätzung nach oben

$$(\boldsymbol{A}_h(k)\boldsymbol{y},\boldsymbol{y})_{(0,h)} \overset{(7.6)}{=} \sum_{i=0}^{N-1} k_{i+\frac{1}{2}}(y_{x,i})^2 h \le c_1||\boldsymbol{y}||^2_{(1,h)} \overset{(3.23)}{\le} \frac{4c_1}{h^2}||\boldsymbol{y}||^2_{(0,h)} = \frac{4c_1}{h^2}(\boldsymbol{y},\boldsymbol{y})_{(0,h)}\,,$$

also insgesamt (im Sinne positiv semidefinit)

(7.10) $$8c_0\boldsymbol{I} \le \boldsymbol{A}_h(k) \le \frac{4c_1}{h^2}\boldsymbol{I}$$

Stabilität:
Die allgemeine Stabilitätsbedingung für das Verfahren 7.5 lautet gemäß Satz 4.2 b)

$$\boldsymbol{B} = \boldsymbol{I} + \sigma\tau\boldsymbol{A}_h(k) \ge \frac{\tau}{2}\boldsymbol{A}_h(k) + \frac{\varepsilon}{2}\boldsymbol{I} \quad \text{bzw.}$$
$$(1-\frac{\varepsilon}{2})\boldsymbol{I} + \sigma\tau\boldsymbol{A}_h(k) - \frac{\tau}{2}\boldsymbol{A}_h(k) \ge \boldsymbol{0}$$

Mit Hilfe von 7.10 verschärfen wir sie (mit $\varepsilon \le 2$) zu

$$(1-\frac{\varepsilon}{2})\frac{h^2}{4c_1}\boldsymbol{A}_h(k) + \sigma\tau\boldsymbol{A}_h(k) - \frac{\tau}{2}\boldsymbol{A}_h(k) \ge \boldsymbol{0}$$

bzw.

$$\left[(1-\frac{\varepsilon}{2})\frac{h^2}{4c_1} + \tau(\sigma-\frac{1}{2})\right]\boldsymbol{A}_h(k) \ge 0,$$

Wegen $\boldsymbol{A}_h(k) \ge 0$ ist dies erfüllt, wenn

$$(1-\frac{\varepsilon}{2})\frac{h^2}{4c_1} + \tau(\sigma-\frac{1}{2}) \ge 0 \qquad \text{bzw.}$$

(7.11) $$\sigma \geq \frac{1}{2} - \frac{h^2}{4c_1\tau}(1-\frac{\varepsilon}{2}) \qquad \text{vgl. Folgerung 4.3 b)}$$

Für das explizite Verfahren $(\sigma = 0)$ bedeutet dies

$$\tau \leq \frac{h^2}{2c_1}(1-\frac{\varepsilon}{2}).$$

In der Praxis kann c_1 sehr groß sein $\Longrightarrow$ fatale Auswirkungen auf $\tau \quad \Longrightarrow$.

Dringende Empfehlung: Kein explizites Verfahren bei variablen Koeffizienten $\boldsymbol{k}(\boldsymbol{x})$.

Konvergenz:
Analog zum Vorgehen in Definition 5.3 sei

$$\boldsymbol{z} = \boldsymbol{u} - \boldsymbol{y} \qquad (\boldsymbol{u} \text{ die Gitterfunktion zur exakten Lösung } u)$$

Anfangs- und Randwerte sind = Null : $\boldsymbol{z}_0 = 0,\ z_0^j = z_N^j = 0\ \forall j \geq 0$.

Für den Approximationsfehler $\boldsymbol{\psi} = \boldsymbol{u}_t - (k\boldsymbol{u}_x)_{\bar{x}}^{(\sigma)} - \boldsymbol{\varphi}$ zeigt man durch Taylorentwicklung wie früher

$$\boldsymbol{\varphi} = O\left(\tau(\sigma - \frac{1}{2}) + \tau^2 + h^2\right)$$

Der Anteil $\tau(\sigma - \frac{1}{2}) + \tau^2$ kommt aus der Zeitdiskretisierung, ist also unverändert gegenüber dem einfachen Verfahren. Der „Anteil h^2 “ wurde in (7.3) hergeleitet. Unter Benutzung der Vektornorm (7.8): $||\boldsymbol{y}||_{\boldsymbol{A}_h(k)} = ||\boldsymbol{y}||_{(1,h,k)} \leq \sqrt{c_1}||\boldsymbol{y}||_{(1,h)}$ gilt analog zu Satz 5.4 somit die

Konvergenzabschätzung

Für $T \geq j\tau$ gilt mit $\sigma \geq \frac{1}{2} - \frac{h^2}{4c_1\tau}(1-\frac{\varepsilon}{2})$

$$\frac{1}{\sqrt{c_1}}||\boldsymbol{z}^j||_{\boldsymbol{A}_h(k)} \leq ||\boldsymbol{z}^j||_{(1,h)} \leq ||\boldsymbol{z}^j||_{(1,h)} = \sqrt{\frac{T}{\varepsilon}}\, O\left(\tau(\sigma - \frac{1}{2}) + \tau^2 + h^2\right),$$

und damit auch

$$||\boldsymbol{z}^j||_{\boldsymbol{A}_h(k)} = \sqrt{\frac{T}{\varepsilon}}\, O\left(\tau(\sigma - \frac{1}{2}) + \tau^2 + h^2\right).$$

Beachte: $\theta < \varepsilon \leq 2$ und $\sigma = \frac{1}{2}$ sind möglich.

§ 8 Die allgemeine 1-dimensionale Wärmeleitungsgleichung

$$
\text{(8.1)} \qquad c\rho_p \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k\,\frac{\partial u}{\partial x}\right) - v\,\frac{\partial u}{\partial x} - qu + f
$$

mit variablen Koeffizienten $c, \rho_p, k, v, d > 0$

Vorbemerkung und Überblick:

Man kann Stabilitätsbedingungen und Konvergenz auch zeigen, wenn die Diskretisierungsmatrix $\boldsymbol{A}$ der rechten Seite von (8.1), abgesehen von f, eine M-Matrix ist (vgl. dazu den nächsten Paragraphen). Vorteil: Mit der M-Matrizentheorie vermeidet man Oszillationen, die bei der L_2-Theorie (Skalarprodukttheorie) auftauchen.
Nachteil: Man erhält oft schlechtere Konvergenzordnungen, muß gelegentlich empfindliche Einschränkungen bzgl. der Schrittweiten hinnehmen, und die M-Matrizentheorie ist nicht immer anwendbar.

Wir führen in diesem Paragraphen die L_2-Theorie weiter, geben jedoch gelegentlich schon Hinweise auf die M-Matrizentheorie, die wir im nächsten Paragraphen untersuchen.

Wir betrachten zunächst nur eine x-Abhängigkeit der Koeffizienten. Bei der Diskretisierung des Konvektionsanteils $v\,\dfrac{\partial u}{\partial x}$, wird sich zeigen, dass wir bei $v \neq 0$ die Symmetrie der Diskretisierungsmatrix verlieren. Der Stabilitätssatz (4.2) muß abgeändert werden, liefert dann aber nur noch die Stabilität bzgl. der Anfangswerte.

Bei zeitabhängigen Koeffizienten beschränken wir uns auf die Untersuchung des Diffusionsanteils $\dfrac{\partial}{\partial x}\left(k\,\dfrac{\partial u}{\partial x}\right)$. Wir stellen fest, dass die Diskretisierungsmatrix nun zeitabhängig wird. Deshalb ist der Stabilitätssatz (4.2) nicht mehr anwendbar. Ein neuer Stabilitätssatz muß bewiesen werden, der allerdings auch nur die Stabilität bzgl. der Anfangswerte liefert. Man rufe sich ins Gedächtnis zurück, dass der Konvergenzssatz (5.4) wesentlich auf der Stabilität bzgl. der rechten Seite beruht. Wir beweisen deshalb zum Abschluß des Paragraphen, dass man die Stabilität bzgl. der rechten Seite aus der Stabilität bzgl. der Anfangswerte folgern kann.

Wir behandeln im Folgenden die einzelnen Summanden der Differentialgleichung (8.1) getrennt.

1) Die Abbaurate q $(d > 0)$ wird in $\boldsymbol{A}_h$ integriert.

Sie liefert bei der Diskretisierung eine Diagonalmatrix $\boldsymbol{Q} = \operatorname{diag}(q_i)$. Diese wird dem Diskretisierungsterm

$$
\frac{\partial}{\partial x}\left(k(x)\,\frac{\partial u}{\partial x}\right) \approx -\boldsymbol{A}_h(k)\boldsymbol{y}
$$

zugeschlagen. Man erhält dann die symmetrische Matrix

$$
\boldsymbol{A} = \boldsymbol{A}_h(k) + \boldsymbol{Q}.
$$

Dadurch wird die positive Definitheit von $\boldsymbol{A}_h(k)$ sogar gestärkt. Ist $\boldsymbol{A}_h(k)$ eine M-Matrix, so auch $\boldsymbol{A}_h(k)+\boldsymbol{Q}$. Kein Problem.

2) Der Term $c\rho_p =: \kappa \geq \delta_0 > 0 \;\; \forall\, x,t$. (Wärmekapazität und Dichte)

Er liefert bei der Diskretisierung $\kappa u_t \approx \boldsymbol{\mathcal{K}} \boldsymbol{I}\, \boldsymbol{y}_t$ mit einer Diagonalmatrix $\boldsymbol{\mathcal{K}} = \operatorname{diag}(\kappa_i)$.
Im Stabilitätssatz 4.2 hat dann $\boldsymbol{B}$ die Gestalt

$$\boldsymbol{B} = (\boldsymbol{\mathcal{K}}\, \boldsymbol{I} + \sigma\, \tau\, \boldsymbol{A})$$

(der Abbauterm ist in $\boldsymbol{A}$ enthalten, vgl. oben).
Aus $(\boldsymbol{A}\boldsymbol{x},\boldsymbol{x}) > 0$ folgt also $(\boldsymbol{B}\boldsymbol{x},\boldsymbol{x}) > 0$.

Die Stabilitätsbedingung (vgl. Satz 4.2) lautet (setze $\boldsymbol{A} := \boldsymbol{A}_h(k)$ bzw. $\boldsymbol{A}_h(k)+\boldsymbol{D}$)

$$\boldsymbol{B} := \boldsymbol{\mathcal{K}}\boldsymbol{I} + \sigma\, \tau\, \boldsymbol{A} \overset{!}{\geq} \frac{\tau}{2}\boldsymbol{A} + \frac{\varepsilon}{2}\boldsymbol{I} \quad \text{für ein } \epsilon > 0\,.$$

Mit der Bedingung $\kappa \geq \delta_0 > 0$ muß man fordern

$$\boldsymbol{B} \geq \delta_0\, \boldsymbol{I} + \sigma\, \tau\, \boldsymbol{A} \overset{!}{\geq} \frac{\tau}{2}\boldsymbol{A} + \frac{\varepsilon}{2}\boldsymbol{I}$$

und mit der Verschärfung $\boldsymbol{I} \geq \dfrac{\boldsymbol{A}}{\|\boldsymbol{A}\|}$ (vgl. (4.16) sogar

$$\left(\delta_0 - \frac{\varepsilon}{2}\right) \frac{\boldsymbol{A}}{\|\boldsymbol{A}\|} + \sigma\, \tau\, \boldsymbol{A} - \frac{\tau}{2}\boldsymbol{A} \geq 0,$$

was durch

$$\frac{\delta_0 - \dfrac{\varepsilon}{2}}{\|\boldsymbol{A}\|} + \sigma\tau - \frac{\tau}{2} \geq 0 \quad \text{bzw.}$$

$$\sigma \geq \frac{1}{2} - \frac{\delta_0 - \frac{\varepsilon}{2}}{\tau\, \|\boldsymbol{A}\|} \tag{8.2}$$

garantiert wird.
Beachte: $\sigma = \dfrac{1}{2}$ **ist möglich.**

Für das explizite Verfahren $(\sigma = 0)$ bedeutet dies

$$\tau \leq \frac{2\left(\delta_0 - \frac{\varepsilon}{2}\right)}{\|\boldsymbol{A}\|}$$

Dies ist fatal, wenn δ_0 klein ist, was realistisch ist.

$\Longrightarrow$

Kein explizites Verfahren für kleines $\kappa = c\rho_p$
Verfahren mit $\sigma \geq \dfrac{1}{2}$ (vgl. (8.2)) sind nicht berührt!

<u>3) Der Konvektionsterm $v\,\dfrac{\partial u}{\partial x}$</u> wird in $\boldsymbol{A}$ integriert.

Die Primitivdiskretisierung (der einfacheren Schreibweise wegen sei $k = k_0 =$konst.)

$$
\begin{aligned}
(-\boldsymbol{A}\boldsymbol{y})_i := \left(k_0\,\frac{\partial^2 u}{\partial x^2} - v\,\frac{\partial u}{\partial x}\right)_i \;&\approx\; k_0\, y_{\overline{x}x,\,i} - v_i\,\frac{y_{i+1}-y_{i-1}}{2h} && \text{zentraler Differenzenquot. wegen Diskret.-Ordnung}\\
&=\; \frac{k_0}{h^2}(y_{i+1}-2y_i+y_{i-1}) - \frac{v_i}{2h}(y_{i+1}-y_{i-1})\\
&=\; \left(\frac{k_0}{h^2}+\frac{v_i}{2h}\right) y_{i-1} - \frac{2k_0}{h^2} y_i + \left(\frac{k_0}{h^2}-\frac{v_i}{2h}\right) y_{i+1}.
\end{aligned}
$$

ist von 2.ter Ordnung. Man erhält, nachdem die Randwerte aus der Matrix eliminiert worden sind, folgende tridiagonale Matrix

$$
(8.3) \qquad \boldsymbol{A} = \begin{pmatrix}
\ddots & & \ddots & & \mathbf{0} \\
\ddots & \ddots & & \ddots & \\
-\frac{k_0}{h^2}-\frac{v_i}{2h} & & \frac{2k_0}{h^2} & & -\frac{k_0}{h^2}+\frac{v_i}{2h} \\
 & \ddots & & \ddots & \ddots \\
\mathbf{0} & & \ddots & & \ddots
\end{pmatrix}
$$

Falls $v \neq 0$, ist diese Matrix nicht symmetrisch (Satz (4.2) gilt nicht). Ein neuer Stabilitätssatz muß bewiesen werden. Eine Ortsabhängigkeit von k würde die Symmetrie nicht stören (vgl. (7.3)).
Wir haben schon darauf hingewiesen, daß, zumindest in Randnähe, die Stabilität bzgl. der L_2-Theorie Oszillationen nicht vermeiden kann. Wir werden sehen, daß die Stabilitätstheorie , die sich auf $M-$ Matrizen stützt (vgl. nächstes Kapitel), oszillationsfrei arbeitet. Wir geben deshalb schon hier einen

Hinweis auf die Anwendung der $\boldsymbol{M}-$ Matrizentheorie.

Zur Anwendung dieser Theorie benötigt man zumindest die Vorzeichenverteilung

$$
\begin{aligned}
-\frac{k_0}{h^2} \pm \frac{v}{2h} \;&\leq\; 0, \quad \text{bzw.}\\
(*) \qquad\qquad h \;&\leq\; \frac{2k_0}{|v|} \quad \text{(Bedingung an Ortsschrittweite)}
\end{aligned}
$$

Die Bedingung $(*)$ kann - für kleines k_0 - in der Anwendung Probleme bereiten. Dies liegt nicht so sehr an der absoluten Größe von k_0, wie folgendes Anwendungsbeispiel $-$ auch schon 1-dimensional $-$ zeigt:

In der Praxis normiert man Längen üblicherweise auf $0 \leq x \leq 1$.
Werden zum Beispiel Verschmutzungsprobleme in einem Fluß gerechnet (Chemieunfälle), so muß man die Flußlänge L auf 1 normieren, d.h. Einführen einer Ortsvariablen $x = Lx'$, $0 \leq x' \leq 1$. Wird diese Transformation in die Differentialgleichung einge-

setzt, so folgt für konstantes v (der Einfachheit halber)

$$\tilde{u}(x') = u(x), \quad \frac{\partial u}{\partial x} = \frac{\partial \tilde{u}}{\partial x'} \cdot \frac{1}{L}, \quad \frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 \tilde{u}}{\partial x'^2} \cdot \frac{1}{L^2}$$

$$\frac{\partial \tilde{u}}{\partial t} = k \frac{\partial^2 \tilde{u}}{\partial x'^2} \cdot \frac{1}{L^2} + \frac{v}{L} \frac{\partial \tilde{u}}{\partial x'}$$

d.h. es werden ersetzt $v \to \frac{v}{L}$ und $k \to \frac{k}{L^2}$

d.h. obige Begingung geht über in

$$\frac{2k_0}{|v|} \to \frac{2k_0}{Lv} \quad \text{also} \quad \underline{\underline{h \le \frac{2k_0}{Lv}}}$$

Bei Flußlängen von ca. 1000 km hat dies katastrophale Folgen, selbst bei großen und schnellen Rechnern, denn für Rechnungen mit so kleinen Schrittweiten ist das Modell zu ungenau (man bräuchte u.a. die Information (Anfangswerte) für die sehr kleinen Schrittweiten).
Noch problematischer wird es, falls Nebenflüsse eingeschlossen werden (Nebenrohre in der Industrie). Das h ist für die Praxis zu klein.

Abhilfe: Statt des zentralen Differenzenquotienten $\boldsymbol{y}_{\mathring{x}}$ für $\frac{\partial u}{\partial x}$ könnte man einen einseitigen Differenzenquotieten $\boldsymbol{y}_{\overline{x}}$ (rückwärtsgenommen) wählen.

physikalische Begründung: Läuft die Strömung von links nach rechts $(v > 0)$, so enthält die ankommende Strömung wertvollere Informationen, als die abfließende (stromaufwärts schauen).

mathematische Begründung: Bei einer einseitigen Approximation für die x-Ableitung gilt für die Matrix $(\boldsymbol{A}_h\, \boldsymbol{y})_i := -k\, y_{\overline{x}x,i} + (v y_{\overline{x}})_i$

$$\begin{aligned}(\boldsymbol{A}_h\, \boldsymbol{y})_i &= -\frac{k_0}{h^2}(y_{i+1} - 2y_i + y_{i-1}) + \frac{v_i}{2h}(y_i - y_{i-1}) \\ &= -\frac{k}{h^2}\, y_{i+1} + \left(\frac{2k}{h^2} + \frac{v}{h}\right) y_i - \left(\frac{k}{h^2} + \frac{v}{h}\right) y_{i-1}\end{aligned}$$

$$(8.4) \qquad \boldsymbol{A}_h = \begin{pmatrix} \ddots & & \ddots & & \ddots & \\ & -\frac{k}{h^2} - \frac{v}{h} & & \frac{2k}{h^2} + \frac{v}{h} & & -\frac{k}{h^2} \\ & \ddots & & & \ddots & & \ddots \end{pmatrix}$$

und man erkennt sofort:

Vorteil: Für $v \ge 0$ liegt eine M-Matix vor (vgl. Satz (6.4)), auch für variable Koeffizienten, denn
1. Die Vorzeichenbedingung ist erfüllt
2. Die strenge Diagonaldominanz gilt in der ersten und letzten Zeile

Die Matrix ist also auch invertierbar.

Nachteil: $(\boldsymbol{A}_h\,\boldsymbol{y})_i \approx -L\,u + \underset{\substack{\uparrow\\ \text{von einsei-}\\ \text{tiger Approx.}}}{O(h)} \quad , \quad L\,u = -k\,\dfrac{\partial^2 u}{\partial x^2} + v\,\dfrac{\partial u}{\partial x}$

d.h., man hat Diskretisierungsordnung (Konvergenzgeschwindigkeit) für Oszillationsfreiheit geopfert, wie wir noch zeigen werden).

Beachte: Verfahren, die ihre Stabilität aus der L_2-Theorie beziehen (d.h. aus Normen, die von Skalarprodukten herrühren), können Oszillationen nicht verhindern, wenn es keine Abschätzung in der Maximumnorm gibt. Sie werden in der Praxis daher oft (obwohl von besserer Papierform bzgl. der Konvergenz) gegenüber Verfahren niedrigerer Ordnung, die Oszillationen ausschließen, hintenan gestellt.

Trotzdem führen wir hier auch die zugehörige L_2-Theorie vor, denn oft rechnet man in Randnähe bei Vorhandensein von Oszillationen mit oszillationsfreien Verfahren und steigt auf schneller konvergente Verfahren um in einem gewissen Abstand von der Anfangswertgeraden.

Erweiterung des Stabilitätssatzes (4.2)

Der Stabilitätssatz 4.2 für das Verfahren $\boldsymbol{B}\,y_t + \boldsymbol{A}\,y = \boldsymbol{\varphi}$ wurde hergeleitet aus der energetischen Identität (4.10), zu deren Beweis die Symmetrie von $\boldsymbol{A}$ benötigt wurde, die im allgemeinen Fall nicht notwendig vorliegt (vgl. (8.3) für $v \neq 0$, und (8.4)). Wir schreiben deshalb unser Verfahren um, können dann allerdings nur noch die Stabilität bzgl. der Anfangswerte zeigen aber nicht bzgl. der rechten Seite, was für die Konvergenz nötig war. Man (bzw. Samarskij) kann jedoch einen Satz beweisen, der die Stabilität der rechten Seite zurückführt auf die Stabilität bzgl. der Anfangswerte. Damit ist unser weiteres Vorgehen vorgezeichnet.
Im parabolischen Fall existiert unter der Voraussetzung $y_0 = y_N = 0$ üblicherweise $\boldsymbol{A}^{-1}$, entweder durch den Nachweis, daß A^{-1} eine M-Matrix ist, vgl. (8.3), (8.4), oder mittels der rellen positiven Definitheit von $\boldsymbol{A}$ (vgl. (8.7),(8.12)).
Dann kann man unser Verfahren mit einem geeigneten $\boldsymbol{A}$, (vgl. die vorigen Abschnitte, aber $\kappa = \text{const}.$, $v = \text{const}.$, vgl. dazu auch (8.11)-(8.13))

$$(\kappa \boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A})\,\boldsymbol{y}_t + \boldsymbol{A}\,\boldsymbol{y} = \boldsymbol{\varphi}$$

von links mit $\boldsymbol{A}^{-1}$ multiplizieren und erhält

$$\text{(8.5)} \qquad \underbrace{(\kappa \boldsymbol{A}^{-1} + \sigma\,\tau\,\boldsymbol{I})}_{\widetilde{B}}\,\boldsymbol{y}_t + \underbrace{\boldsymbol{I}}_{\widetilde{\boldsymbol{A}}}\,\boldsymbol{y} = \underbrace{\boldsymbol{A}^{-1}\,\varphi}_{\widetilde{\varphi}}$$

also das Verfahren

$$\text{(8.6)} \qquad \widetilde{\boldsymbol{B}}\,\boldsymbol{y}_t + \widetilde{\boldsymbol{A}}\,\boldsymbol{y} = \widetilde{\boldsymbol{\varphi}} \quad \text{mit } \widetilde{\boldsymbol{B}} = \kappa \boldsymbol{A}^{-1} + \sigma\,\tau\,\boldsymbol{I}, \quad \widetilde{\boldsymbol{A}} = \boldsymbol{I}, \quad \widetilde{\boldsymbol{\varphi}} = \boldsymbol{A}^{-1}\,\boldsymbol{\varphi}\,.$$

$\widetilde{\boldsymbol{A}}$ ist nun wieder symmetrisch. Man kann also vorgehen wie bei Satz (4.2). Es zeigt sich allerdings, daß man für $\boldsymbol{A}$, bzw. $\boldsymbol{A}^{-1}$, und damit auch für $\boldsymbol{B}$ die positive Definitheit nicht mehr immer nachweisen kann (z.B. für $v \neq 0$). Man muß sie abschwächen

zur reellen positiven Definitheit (kein Rückschluß auf die Eigenwerte möglich)
(8.7)

$$(\boldsymbol{A}\boldsymbol{v},\boldsymbol{v})_{(0,h)}\left\{\begin{matrix}\geq\\>\end{matrix}\right. 0 \quad \forall \boldsymbol{v}\in\mathbb{R}^n, \boldsymbol{v}\neq\boldsymbol{0} \quad \overset{\text{def}}{\Longleftrightarrow} \quad \boldsymbol{A} \text{ reell positiv}\left\{\begin{matrix}\text{semidefinit } (\overset{r}{\geq})\\ \text{definit } (\overset{r}{>})\end{matrix}\right.$$

(es wird keine Symmetrie gefordert)

Nun kann der Beweis von Satz (4.2) wörtlich mit dieser Abschwächung durchgeführt werden, wenn die Invertierbarkeit von $\boldsymbol{A}$ und $\boldsymbol{B}$ gesichert sind. Die folgte aus der positiven Definitheit. Wir zeigen für beliebiges $\boldsymbol{C}\in\mathbb{R}^{n\times n}$:

(8.8) $\quad \boldsymbol{C}$ reell positiv definit $\quad\Longrightarrow\quad \exists\, \boldsymbol{C}^{-1}$ und $\boldsymbol{C}^{-1}$ reell positiv definit.

Indirekt: Existiert $\boldsymbol{C}^{-1}$ nicht, so hat $\boldsymbol{C}$ hat einen Eigenwert 0 und dazu einen reellen Eigenvektor $\boldsymbol{x}\neq\boldsymbol{0}$ $\Rightarrow (\boldsymbol{C}\boldsymbol{x},\boldsymbol{x})=(\boldsymbol{0},\boldsymbol{x})=0$, ein Widerspruch zur reellen positiven Definitheit.
Für invertierbares $\boldsymbol{C}\overset{r}{\geq}0$ gilt:

$$(\boldsymbol{C}\boldsymbol{y},\boldsymbol{y})_{(0,h)} \underset{\underset{\boldsymbol{C}\boldsymbol{y}=\boldsymbol{v}}{\uparrow}}{=} (\boldsymbol{v},\boldsymbol{C}^{-1}\boldsymbol{v})_{(0,h)} \underset{\underset{\text{reell}}{\uparrow}}{=} (\boldsymbol{C}^{-1}\boldsymbol{v},\boldsymbol{v}) \overset{r}{\geq} 0.$$

Nun kann der Beweis von Satz 4.2 b) wörtlich wiederholt werden mit den abgeschwächten Definitheitsbedingungen

$$\widetilde{\boldsymbol{B}}=\kappa\boldsymbol{A}^{-1}+\sigma\,\tau\,\boldsymbol{I}\overset{r}{>}\boldsymbol{0},\ \widetilde{\boldsymbol{A}}^T=\widetilde{\boldsymbol{A}}=\boldsymbol{I}\overset{r}{>}\boldsymbol{0} \text{ und } \widetilde{\boldsymbol{B}}\overset{r}{\geq}\frac{\tau}{2}\widetilde{\boldsymbol{A}}.$$

und man erhält (für beliebige aufeinanderfolgende Zeitschichten $\hat{\boldsymbol{y}}$, $\boldsymbol{y}$) die energetische Identität in der Gestalt

(8.9) $$2\tau((\tilde{\boldsymbol{B}}-\frac{\tau}{2}\boldsymbol{I})\boldsymbol{y}_t,\boldsymbol{y}_t)+\|\hat{\boldsymbol{y}}\|^2_{(0,h)}-\|\boldsymbol{y}\|^2_{(0,h)}=\boldsymbol{A}^{-1}\boldsymbol{\varphi}.$$

Wegen $\widetilde{\boldsymbol{B}}=\boldsymbol{A}^{-1}+\sigma\,\tau\,\boldsymbol{I}$, $\quad\widetilde{\boldsymbol{A}}=\boldsymbol{I}$ reduzieren sich die Voraussetzungen zu
(8.10)

$$(\widetilde{\boldsymbol{B}}\boldsymbol{v},\boldsymbol{v})_{(0,h)}\overset{r}{>}0,\ \ (\kappa\boldsymbol{A}^{-1}\boldsymbol{v},\boldsymbol{v})_{(0,h)}+(\sigma\,\tau\,\boldsymbol{v},\boldsymbol{v})_{(0,h)}\overset{r}{\geq}\left(\frac{\tau}{2}\boldsymbol{v},\boldsymbol{v}\right)_{(0,h)} \quad \forall\boldsymbol{v}\in\mathbb{R}^n,\boldsymbol{v}\neq 0.$$

Falls $\boldsymbol{A}^{-1}\overset{r}{\geq}0$ ist letzteres erfüllt, wenn $\sigma\,\tau\geq\frac{\tau}{2}$ bzw. $\sigma\geq\frac{1}{2}$.
Setzt man $\boldsymbol{A}^{-1}\overset{r}{>}0$, voraus, so ist auch $\boldsymbol{A}^{-1}\overset{r}{\geq}0$ (vgl. (8.8)) und es folgt (beachte $\boldsymbol{I}\overset{r}{>}0$) : $(\kappa\boldsymbol{A}^{-1}\boldsymbol{x},\boldsymbol{x})+(\sigma\,\tau\,\boldsymbol{I}\boldsymbol{x},\boldsymbol{x})=((\boldsymbol{A}^{-1}+\sigma\,\tau\,\boldsymbol{I})\boldsymbol{x},\boldsymbol{x})>0$, also $\widetilde{\boldsymbol{B}}\overset{r}{>}\boldsymbol{0}$ und nach (8.8) auch die Invertierbarkeit. Dann liefert Satz 4.2 b)

$$\|\boldsymbol{y}^{j+1}\|_{(0,h)}\leq\|\boldsymbol{y}^j\|_{(0,h)}$$

für beliebige aufeinanderfolgende Zeitschichten $\boldsymbol{y}^{j+1}$, $\boldsymbol{y}^j$ und wegen $\widetilde{\boldsymbol{A}}=\boldsymbol{I}$ gilt sogar

$$\|\boldsymbol{y}^{j+1}\|_2\leq\|\boldsymbol{y}^j\|_2.$$

Dies beweist

Satz 8.1
Das Verfahren

$$\begin{aligned} (\kappa \boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A})\boldsymbol{y}_t + \boldsymbol{A}\boldsymbol{y} &= \boldsymbol{\varphi} \\ \text{bzw.}\quad (\kappa \boldsymbol{A}^{-1} + \sigma\,\tau\,\boldsymbol{I})\boldsymbol{y}_t + \boldsymbol{I}\boldsymbol{y} &= \boldsymbol{A}^{-1}\boldsymbol{\varphi} \end{aligned}$$

(mit $\kappa = \text{const}\,.$) ist stabil bzgl. der Anfangswerte (also $y_0^j = y_N^j = 0\ \forall j,\ \ \boldsymbol{\varphi} \equiv \boldsymbol{0}$), falls

(i) $(\boldsymbol{A}\boldsymbol{y}, \boldsymbol{y})_{(0,h)} > 0\ \forall\, \boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{y} \neq \boldsymbol{0}$ $\quad$ ($\boldsymbol{A}$ reell positiv definit) und

(ii) $\sigma \geq \frac{1}{2}$

Bemerkungen Fordert man in die stärkere Bedingung (Stabilität bzgl. der rechten Seite)

$$\widetilde{\boldsymbol{B}} \geq \frac{\tau}{2}\,\widetilde{\boldsymbol{A}} + \frac{\varepsilon}{2}\,\boldsymbol{I}$$

so führt dies auf obigem Wege zu der Bedingung $\sigma \geq \dfrac{1}{2} + \dfrac{\varepsilon}{2\tau}$. Wegen $\tau \to 0$ kann man hieraus keine Stabilitätsbedingung bzgl. der rechten Seite gewinnen. Man kann nicht ohne weiteres $\varepsilon \to 0$ gehen lassen, da sonst die Stabilität bezl. der rechten Seite ebenfalls verloren geht, vgl. dazu Satz (4.2) b), wo ε im Nenner steht.

Wir zeigen zunächst, dass (i) für die Matrizen (8.3) und (8.4) erfüllt ist.

Reelle pos. Definith. für $(\boldsymbol{A}\,\boldsymbol{y})_i := -(k y_x)_{\overline{x},i} + v y_{\overset{\circ}{x}},\ v = \text{const},\ y_0 = y_N = 0$

$$= \frac{1}{h}\left(k_{i+\frac{1}{2}}\frac{y_{i+1} - y_i}{h} - k_{i-\frac{1}{2}}\frac{y_i - y_{i-1}}{2}\right) + v\frac{y_{i+1} - y_i}{h}.$$

(vgl. (7.2), (8.3))

$$\begin{aligned} (\boldsymbol{A}\boldsymbol{y}, \boldsymbol{y})_{(0,h)} &= \sum_{i=1}^{N-1} (\boldsymbol{A}\boldsymbol{y})_i y_i h \\ &= -\sum_{i=1}^{N-1} (k y_x)_{\overline{x},i}\, y_i h + \sum_{i=1}^{N-1} v(y_{\overset{\circ}{x},i}) y_i h \quad \text{und mit } y_0^j = y_N^i = 0\ \forall j \qquad (8.11) \end{aligned}$$

Bekannt ist (vgl. (8.11)):

$$\sum_{i=1}^{N} (k y_x)_{\overline{x},i} = \sum_{i=1}^{N} \underbrace{k_{i-\frac{1}{2}}}_{\geq c_0} (y_{x,i-1})^2 h \geq c_0 \|\boldsymbol{y}\|^2_{(1,h)}.$$

Wir formen die zweite Summe um für v:=const.

$$\begin{aligned}\sum_{i=1}^{N-1} v(y_{\overset{\circ}{x},i})y_i h &= v\sum_{i=1}^{N-1} \frac{y_{i+1}-y_{i-1}}{2}\, y_i = \frac{v}{2}\left(\sum_{i=1}^{N-1} y_{i+1}\, y_i - \sum_{i=1}^{N-1} y_i\, y_{i-1}\right)\\ &= \frac{v}{2}\left(\sum_{i=2}^{N} y_i\, y_{i-1} - \sum_{i=1}^{N-1} y_i\, y_{i-1}\right)\\ &= \frac{v}{2}\Big(\underset{\overset{\uparrow}{=0}}{y_N}\, y_{N-1} - y_1\, \underset{\overset{\uparrow}{=0}}{y_0}\Big), \quad \text{klappt nicht für variables } v.!\end{aligned}$$

Es bleibt

$$(8.12) \qquad (\boldsymbol{Ay},\boldsymbol{y})_{(0,h)} = \sum_{i=1}^{N} k_{i-\frac{1}{2}}\,(y_{x,i-1})^2 h \overset{(7.7)}{\geq} c_0\|\boldsymbol{y}\|^2_{(1,h)} > 0 \quad \text{für } k \geq c_0.$$

Fazit: Der Konvektionsterm mit konstanter Geschwindigkeit berührt die reelle positive Definitheit nicht.

Reelle pos. Definitheit für $(\boldsymbol{A}_h y)_i := -(ky_x)_{\overline{x},i} + vy_{\overline{x}},\ v = \text{const},\ y_0 = y_N = 0$ (vgl. (8.4))

Analog zum vorigen brauchen wir nur den Konvektionsterm zu untersuchen, also den Term (vgl. (8.11))

$$v\sum_{i=1}^{N-1}(y_i - y_{i-1})y_i$$

Nun ist

$$\begin{aligned}(y_i - y_{i-1})y_i &= y_i^2 - y_i\, y_{i-1}, \quad \text{wegen} \quad 2ab = -(a-b)^2 + a^2 + b^2\\ &= y_i^2 - \frac{1}{2}\left[y_i^2 + y_{i-1}^2 - (y_i - y_{i-1})^2\right]\\ &= y_i^2 + \frac{1}{2}\left[-y_i^2 - y_{i-1}^2 + (y_{\overline{x},i})^2\cdot h^2\right]\\ &= \frac{1}{2}\left[y_i^2 - y_{i-1}^2 + (y_{\overline{x},i})^2\, h^2\right]\\ \sum_{i=1}^{N-1}(y_i - y_{i-1})\, y_i &= \sum_{i=1}^{N-1}\frac{1}{2}\left[y_i^2 - y_{i-1}^2 + (y_{\overline{x},i})^2\, h^2\right] = \frac{1}{2}\left(y_{N-1}^2 + \sum_{i=1}^{N-1}(y_{\overline{x},i})^2\, h^2\right),\\ &\quad \text{und wegen } y_N = 0,\ y_{N-1}^2 = (y_{n-1} - y_N)^2 = h^2 y_{\overline{x},N}\\ &= \frac{1}{2}\sum_{i=1}^{N}(y_{\overline{x},i})^2\, h^2 \overset{(3.21)}{=} \frac{h}{2}\,\|\boldsymbol{y}\|^2_{(1,h)} \xRightarrow{\text{insgesamt}}\end{aligned}$$

$$(8.13) \qquad (\boldsymbol{A}_h\boldsymbol{y},\boldsymbol{y})_{(0,h)} \geq \|\boldsymbol{y}\|^2_{(1,h)} + \frac{vh}{2}\,\|\boldsymbol{y}\|^2_{(1,h)} = \left(c_0 + \frac{vh}{2}\right)\|y\|^2_{(1,h)} > 0 \quad \forall \boldsymbol{y} \neq \boldsymbol{0}$$

Also reell positiv definit.
Fazit: wie oben!

Zusammenfassend haben wir also folgendes (Zwischen-) Ergebnis

Satz 8.2
Für den Differentialoperator

$$Lu := \kappa \frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left(k(x)\frac{\partial u}{\partial x}\right) + v\frac{\partial u}{\partial x} + q(x)\,u$$

seien $\quad 0 < c_0 \le k(x) \le c_1, \quad q(x) > 0,\ \kappa = \text{const},\ v = \text{const},\ \kappa, v > 0\,.$

Dann liefern die Diskretisierungen

$(*)\quad (\boldsymbol{Ay})_i := -(ky_x)_{\overline{x},i} + v\,y_{\overset{\circ}{x},i}$

$(**)\quad (\boldsymbol{Ay})_i := -(ky_x)_{\overline{x},i} + v\,y_{\overline{x},i}$

reell positiv definite Matrizen.

Mit diesen Diskretisierungen ist das implizite Verfahren

$$(\kappa\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A})y_t + \boldsymbol{A}y = \boldsymbol{\varphi}, \quad \sigma \ge \frac{1}{2}$$

für die Aufgabe $Lu = f$ mit Rand- und Anfangswerten stabil bzgl. der Anfangswerte, und von der Konvergenzordnung $\mathrm{O}\,(\tau^2 + h^2)$ für $(*)$ und $\mathrm{O}\,(\tau^2 + h)$ für $(**)$.

Die Wärmeleitung mit zeitabhängigem Diffusionskoefizienten

Der Einfachheit halber verzichten wir (zunächst) auf Konvektionsterm, Abbaurate und rechte Seite und betrachten die Aufgabe

(8.14)
$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k(x,t)\frac{\partial u}{\partial t}\right), \quad 0 < x < 1,\ t > 0$$
$$u(x,0) = u_0(x),\ 0 \le x \le 1$$
$$u(0,t) = g_0(t),\ u(1,t) = g_1(t),\ t > 0.$$

Für die Aufgabe (8.14) betrachten wir für die Punkte $x_i,\ i = 1, ..., N-1,$ die Approximation (vgl. (7.3))

$$y^j_{t,i} = \frac{1}{h}\left[\left(k^{j+\frac{1}{2}}_{i+\frac{1}{2}}\,y_x\right)_{i+\frac{1}{2}} - \left(k^{j+\frac{1}{2}}_{i-\frac{1}{2}}\,y_{\bar{x},i}\right)_{i-\frac{1}{2}}\right]^{(\sigma)}$$

(8.15) Hierbei wird für jede Zeitschicht j der Diffusionskoeffizient k auf die Schicht $j + \frac{1}{2}$ gesetzt (im Hinblick auf Crank- Nicolson) und unabhängig von der Wahl von σ festgehalten. Die Wirkung von σ beschränkt sich auf die y-Werte.

Wir erhalten somit (vgl. (7.2)) für jede Zeitschicht t^j eine **symmetrische Matrix** $\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)$ (zum Aussehen vgl. (7.3) mit k auf der Zeitschicht $t^{j+\frac{1}{2}}$), deren **positive Definitheit** wie in (7.6) nachgerechnet wird.

Wir können wieder die zeitliche Mittelwertbildung durch σ mit der numerischen Differentiation bzgl. x vertauschen und unser Verfahren erhält die Gestalt

$$\boldsymbol{y}_t^j = -\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)\,\boldsymbol{y}^\sigma \quad \text{bzw.}$$

$$\begin{aligned} \left(\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)\right)\boldsymbol{y}_t^j + \boldsymbol{A}_h^{(j+\frac{1}{2})}(k)\,\boldsymbol{y}^j = 0. \\ \underbrace{\left((\boldsymbol{A}_h^{(j+\frac{1}{2})}(k))^{-1} + \sigma\tau\boldsymbol{I}\right)}_{\tilde{\boldsymbol{B}}^j}\boldsymbol{y}^j + \boldsymbol{I}\boldsymbol{y}^j = 0. \end{aligned} \tag{8.16}$$

Natürlich ist $\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)$, und damit auch die Inverse, reell positiv definit als Folge der positiven Definitheit. Wie zu Beginn des Paragraphen kann man die Abbaurate für beliebiges $q(x,t)$ in $\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)$ integrieren, ein konstantes $\kappa > 0$ einschließen und einen Geschwindigkeitsterm $v\frac{\partial u}{\partial x}$ mit konstantem v berücksichtigen. Für die zu (8.3),(8.4) analogen Matrizen zeigt man wie in (8.11)-(8.13) die reell positive Definitheit. Dann kann man wieder Satz 4.2 analog zu (8.9) anwenden. Dabei spielt es keine Rolle, daß $\tilde{\boldsymbol{B}}$ nun zeitabhängig ist, wenn nur für jede Zeitschicht die zu (8.10) analogen Voraussetzungen für $\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)$ erfüllt sind. Damit erhalten wir die zu Satz 8.1 analoge Aussage

Satz 8.3
Es seien

(i) $(\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)\boldsymbol{y},\boldsymbol{y})_{(0,h)} > 0 \;\; \forall \boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{y} \neq \boldsymbol{0}$ ($\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)$ reell positiv definit) und

(ii) $\sigma \geq \frac{1}{2}$.

Dann ist das Verfahren

$$\begin{aligned} (\kappa\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A}_h^{(j+\frac{1}{2})}(k))\boldsymbol{y}_t + \boldsymbol{A}_h^{(j+\frac{1}{2})}(k)\boldsymbol{y} &= \boldsymbol{\varphi} \\ \text{bzw.} \quad \left(\kappa\left(\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)\right)^{-1} + \sigma\,\tau\,\boldsymbol{I}\right)\boldsymbol{y}_t + \boldsymbol{I}\boldsymbol{y} &= \left(\boldsymbol{A}_h^{(j+\frac{1}{2})}(k)\right)^{-1}\boldsymbol{\varphi} \end{aligned}$$

(mit $\kappa = \text{const.}$) ist stabil bzgl. der Anfangswerte (also $y_0^j = y_N^j = 0 \;\forall j,\;\; \boldsymbol{\varphi} \equiv \boldsymbol{0}$).

Die Stabilität bzgl. der rechten Seite

wird zurückgeführt auf die Stabilität bzgl. der Anfangswerte. Stabilität bzgl. Anfangswerten und rechter Seite wird wieder durch Superposition gezeigt. Da die Stabilität bzgl. der Anfangswerte bekannt ist, genügt es die Stabilität bzgl. der rechten Seite für eine Aufgabe mit Nullanfangswerten zu betrachten.

Wir behandeln die Aufgabe

$$u_t - \frac{\partial}{\partial x}\left(k(x,t)\,\frac{\partial u}{\partial x}\right) + v(x,t)\,\frac{\partial u}{\partial x} + q(x,t) = f(x,t), \quad u(0,t) = u(1,t) = 0. \tag{8.17}$$

und zu ihrer Lösung ein Differenzenschema der Form (vgl. (8.5), (8.6))

$$\widetilde{\boldsymbol{B}}^j\, \boldsymbol{y}_t^j + \widetilde{\boldsymbol{A}}\, \boldsymbol{y}^j = \widetilde{\boldsymbol{\varphi}}^j, \quad \text{mit} \tag{8.18}$$

$$\widetilde{\boldsymbol{B}}^j = (\boldsymbol{A}^j)^{-1} + \sigma\,\tau\,\boldsymbol{I},\ \ \widetilde{\boldsymbol{A}} = \boldsymbol{I},\ \ \widetilde{\boldsymbol{\varphi}} = (\boldsymbol{A}^j)^{-1}\,\boldsymbol{\varphi},\ \ (\boldsymbol{A}^j\,\boldsymbol{y})^{(\sigma)} = \boldsymbol{A}^j\,\boldsymbol{y}^{(\sigma)},\ \ \sigma \ge \frac{1}{2}\,.$$

Die Diskretisierungen aus § 7, § 8, welche die Gestalt von $\boldsymbol{A}$ bestimmen, sind zugelassen (also die Matrizen (7.2),(8.3),(8.4),(8.14)). Welche Zeitschichten für $\boldsymbol{\varphi}$ zugelassen werden ist ohne Belang. Wir treffen Voraussetzungen, welche die Stabilität bzgl. der Anfangswerte sichern (vgl. dazu die Sätze 4.2 und 8.1 und 8.4).

Der einfacheren Schreibweise wegen unterdrücken wir im folgenden Satz bei den Matrizen die Indizes und Argumente, die auf Zeit- bzw. Ortsabhängigkeit hinweisen.

Satz 8.4 (Samarskij)
Für das Differenzenschema

$$\frac{1}{\tau}\,\widetilde{\boldsymbol{B}}(\boldsymbol{y}^j - \boldsymbol{y}^{j-1}) + \widetilde{\boldsymbol{A}}\,\boldsymbol{y}^{j-1} = \widetilde{\boldsymbol{\varphi}}^{j-1}, \tag{8.19}$$

$$\widetilde{\boldsymbol{B}} = \boldsymbol{A}^{-1} + \sigma\,\tau\,\boldsymbol{I},\ \ \widetilde{\boldsymbol{A}} = \boldsymbol{I},\ \ \widetilde{\boldsymbol{\varphi}} = \boldsymbol{A}^{-1}\,\boldsymbol{\varphi},\ \ \sigma \ge \frac{1}{2},\ \ y_0^j = y_N^j = 0\ \forall j$$

seien folgende Voraussetzungen erfüllt

$$\widetilde{\boldsymbol{A}} > 0\ \ \text{(im Sinne reell positiv definit)}, \qquad \sigma \ge \frac{1}{2}\,.$$

Dann ist (8.19) stabil bzgl. Anfangswerten und rechter Seite:

$$\|\boldsymbol{y}^j\|_{(0,h)} \le \|\boldsymbol{y}^0\|_{(0,h)} + \sum_{k^1}^{j} \tau\,\|\boldsymbol{\varphi}^{k-1}\|_{(0,h)}\,. \tag{8.20}$$

Beweis
Die Stabilität bzgl. Anfangswerten (bereits bekannt) und rechter Seite erhält man durch Superposition. Es genügt also die Stabilität bzgl. der rechten Seite zu beweisen für $\boldsymbol{y}^0 = \boldsymbol{0}\,,\ y_0^j = y_N^j = 0\ \forall j\,.$

Für die Lösung von (8.19) machen wir den Ansatz für $m > j \;\; \forall j$ mit $j\tau \le T$:

$$\boldsymbol{y}^j = \sum_{k=1}^{m} \boldsymbol{y}^j_{(k)},$$

die $\boldsymbol{y}^j_{(k)}$ seien Lösungen von

$$\text{(8.21)} \qquad \frac{1}{\tau}\widetilde{\boldsymbol{B}}\left(\boldsymbol{y}^j_{(k)} - \boldsymbol{y}^{j-1}_{(k)}\right) + \widetilde{\boldsymbol{A}}\,\boldsymbol{y}^{j-1}_{(k)} = \delta_{kj}\,\widetilde{\boldsymbol{\varphi}}^{k-1}, \quad k = 1,\ldots,m, \;\; \boldsymbol{y}^0_{(k)} = \boldsymbol{0}$$

mit

$$\text{(8.22)} \qquad \delta_{kj}\,\widetilde{\boldsymbol{\varphi}}^{k-1} = \begin{cases} \boldsymbol{0} & \text{für } k \neq j \\ \widetilde{\boldsymbol{\varphi}}^{j-1} & \text{für } k = j \end{cases}$$

Man erkennt: Aufsummieren von (8.21) über $k = 1,\ldots,m$ liefert unter Berücksichtigung von (8.22) das Verfahren (8.19), wenn alle $\boldsymbol{y}^j_{(k)}$ berechnet sind.
Berechnung der $\boldsymbol{y}^j_{(k)}$: $k \in \{1,\ldots,m\}$ fest, $j = 1,2,\ldots$. Aus (8.21), (8.22) folgt

$$\begin{aligned}
&(i) \quad j < k: \quad \boldsymbol{y}^j_{(k)} = 0 \quad \text{(da Anfangswerte} = 0, \text{ rechte Seite} = 0), \\
&(ii) \;\; j = k: \quad \frac{1}{\tau}\widetilde{\boldsymbol{B}}\,\boldsymbol{y}^k_{(k)} = \widetilde{\boldsymbol{\varphi}}^{k-1} \quad \Longrightarrow \quad \boldsymbol{y}^k_{(k)} = \tau\,\widetilde{\boldsymbol{B}}^{-1}\,\widetilde{\boldsymbol{\varphi}}^{k-1} \\
&(iii) \; j > k: \quad \frac{1}{\tau}\widetilde{\boldsymbol{B}}\left(\boldsymbol{y}^j_{(k)} - \boldsymbol{y}^{j-1}_{(k)}\right) + \widetilde{\boldsymbol{A}}\,\boldsymbol{y}^{j-1}_{(k)} = \boldsymbol{0}
\end{aligned}$$

Es können also für jedes k alle $\boldsymbol{y}^j_{(k)}$ berechnet werden und zwar, sofern sie $\neq \boldsymbol{0}$ sind als Lösungen von Differenzenverfahren, deren rechte Seite $= 0$ ist.
$\widetilde{\boldsymbol{A}} > 0,\ \sigma \ge \frac{1}{2}$ sichern nach den Sätzen 8.1 und 4.2 die Ungleichungen

$$\text{(8.23)} \qquad \|\boldsymbol{y}^j_{(k)}\|_{(0,h)} \le \|\boldsymbol{y}^{j-1}_{(k)}\|_{(0,h)} \le \ldots \le \|\boldsymbol{y}^k_{(k)}\|_{(0,h)} = \tau\,\left\|\widetilde{\boldsymbol{B}}^{-1}\,\widetilde{\boldsymbol{\varphi}}^{k-1}\right\|_{(0,h)}$$

eigentkich mit dem Index $(1,h)$, vgl. (4.12). Beachte jedoch Für $\widetilde{\boldsymbol{A}} = \boldsymbol{I}$ gilt

$$\|\boldsymbol{y}\|^2_{(0,h)} = (\boldsymbol{y},\boldsymbol{y})_{(0,h)}, \quad \|\boldsymbol{y}\|^2_{(1,h)} = (\widetilde{\boldsymbol{A}}\boldsymbol{y},\boldsymbol{y})_{(0,h)} = (\boldsymbol{y},\boldsymbol{y})_{(0,h)}$$

Einsetzen von (8.23) in die Darstellung von $\boldsymbol{y}^j$ liefert (vgl. (i))

$$\text{(8.24)} \quad \|\boldsymbol{y}^j\|_{(0,h)} \le \sum_{k=1}^{j} \|\boldsymbol{y}^j_{(k)}\|_{(0,h)} \le \sum_{k=1}^{j} \tau\,\left\|\widetilde{\boldsymbol{B}}^{-1}\,\widetilde{\boldsymbol{\varphi}}^{k-1}\right\|_{(0,h)} \qquad (\,\boldsymbol{y}^j_{(k)} = 0 \text{ für } k > j\,)$$

Nun ist wegen $\widetilde{\boldsymbol{B}} = \boldsymbol{A}^{-1}(\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A})$

$$\underbrace{\widetilde{\boldsymbol{B}}^{-1}\,\widetilde{\boldsymbol{\varphi}}^{k-1}}_{=:\,\boldsymbol{v}} = (\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A})^{-1}\,\boldsymbol{A}\,\underbrace{\boldsymbol{A}^{-1}\,\boldsymbol{\varphi}^{k-1}}_{\widetilde{\boldsymbol{\varphi}}^{k-1}} = (\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A})^{-1}\,\underbrace{\boldsymbol{\varphi}^{k-1}}_{=:\,\boldsymbol{w}}$$

Zur Abschätzung von $(\boldsymbol{I}+\sigma\,\tau\,\boldsymbol{A})^{-1}$ schätzen wir die Lösung $\boldsymbol{v}$ ab von

$$(\boldsymbol{I}+\sigma\,\tau\,\boldsymbol{A})\boldsymbol{v}=\boldsymbol{w}, \qquad (\boldsymbol{v}=\widetilde{\boldsymbol{B}}^{-1}\,\widetilde{\boldsymbol{\varphi}}^{k-1},\ \boldsymbol{w}=\boldsymbol{\varphi}^{k-1}) \tag{8.25}$$

Multiplizieren dieser Gleichung mit $(\cdot,\boldsymbol{v})_{(0,h)}$ liefert

$$\begin{array}{rrcl} & \|\boldsymbol{v}\|^2_{(0,h)}+\underbrace{\sigma\,\tau(\boldsymbol{A}\boldsymbol{v},\boldsymbol{v})_{(0,h)}}_{>0} & = & (\boldsymbol{w},\boldsymbol{v})_{(0,h)} \overset{\text{CSU}}{\leq} \|\boldsymbol{w}\|_{(0,h)}\,\|\boldsymbol{v}\|_{(0,h)} \\ \Longrightarrow & & & \\ & \|\boldsymbol{v}\|^2_{(0,h)} & \leq & \|\boldsymbol{w}\|_{(0,h)}\,\|\boldsymbol{v}\|_{(0,h)}, \quad \text{bzw. nach Kürzen} \\ & \left\|\widetilde{\boldsymbol{B}}^{-1}\,\widetilde{\boldsymbol{\varphi}}^{k-1}\right\|_{(0,h)}=\|\boldsymbol{v}\|_{(0,h)} & \leq & \|\boldsymbol{w}\|_{(0,h)}=\|\boldsymbol{\varphi}^{k-1}\|_{(0,h)} \end{array}$$

Damit erhält man aus (8.24)

$$\|\boldsymbol{y}^j\|_{(0,h)} \leq \sum_{k=1}^{j} \tau\,\|\boldsymbol{\varphi}^{k-1}\|_{(0,h)}$$

□

Bemerkung Die Stabilitätsvoraussetzung $\boldsymbol{A}>0,\ \sigma\geq\frac{1}{2}$ des Satzes kann durch jeden anderen Satz von Voraussetzungen ersetzt werden, welcher die Stabilität bzgl. der Anfangswerte garantiert.

§ 9 Gleichmäßige Stabilität und Konvergenz

Satz 9.1
Für eine parabolische Differentialgleichung mit Nullrandwerten erfülle das Verfahren

(9.1) $(\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A}^j)\boldsymbol{y}_t + \boldsymbol{A}^j\boldsymbol{y} = \boldsymbol{\varphi}^j,\ \ (\boldsymbol{A}^j\,\boldsymbol{y})^{(\sigma)} = \boldsymbol{A}^j\,\boldsymbol{y}^{(\sigma)},\ \ j \geq 0, \quad 0 \leq \sigma \leq 1$

folgende Voraussetzungen

(9.2) $\boldsymbol{A}^j = (a^j_{i,k})_{i,k=1,\dots,N-1}$ sei $\forall j$ eine tridiagonale M-Matrix, $a^j_{ik} \leq 0\ \forall i \neq k$

(9.3) $a^j_{ii} + a^j_{i,i-1} + a^j_{i,i+1} \geq 0,\ \ a_{1,0} = a_{N-1,N} = 0,\ \ \forall i = 1,\dots,N-1,\ \ j \geq 0$

(9.4) $(1-\sigma)\tau\, a^j_{ii} \leq 1$ (Schrittweitenbeschränkung)

Dann gelten

(9.5) $\|\hat{\boldsymbol{y}}\|_\infty \leq \|\boldsymbol{y}\|_\infty + \tau\,\|\boldsymbol{\varphi}\|_\infty$ und

(9.6) $\|\boldsymbol{y}^{j+1}\|_\infty \leq \|\boldsymbol{y}^0\|_\infty + \tau \displaystyle\sum_{k=0}^{j} \|\boldsymbol{\varphi}^k\|_\infty$

(Stabilität bzgl. Anfangswerten und rechter Seite)

Bemerkungen:

α) $\boldsymbol{A}^j$ bedeutet, dass $\boldsymbol{A}$ zeitabängig sein darf (vgl. etwa (8.14)). Aufgaben mit variablen Koffizienten sind erfaßbar, die Diskretisierung muß keine symmetrische Matrix $\boldsymbol{A}^j$ liefern und es sind auch keine Definitheitsvoraussetzungen für $\boldsymbol{A}^j$ erforderlich.

β) Für das einfache Verfahren (4.1): $\boldsymbol{y}^j_t = -\boldsymbol{A}^0_h\,\boldsymbol{y}^j$ mit $\boldsymbol{y}^0 = u_0\,,\ y^j_0 = y^j_N = 0\,,\ j \geq 0$ bedeutet (9.4) wegen $\sigma = 0\,,\ a_{ii} = \frac{2}{h^2}$, dass $\tau \leq \frac{h^2}{2}$ eine Voraussetzung ist, die sich – vgl. (4.2) – kaum abschwächen läßt.

γ) Auf Grund der Nullrandwerte hängt $\boldsymbol{\varphi}$ nur von der rechten Seite der Differentialgleichung ab, z.B. $\boldsymbol{\varphi}^j = \tilde{\boldsymbol{f}}^j$, wobei die „$\tilde{\ }$“ bedeutet, dass $\boldsymbol{f}$ auch an einer Zeitschicht zwischen $j+1$ und j betrachtet werden kann.

δ) **Nur für rein implizite Verfahren keine Schrittweitenbeschränkung!** (d.h. $\sigma = 1$)

Beweis
$\forall\,\boldsymbol{A}^j$ gilt: $\boldsymbol{B}^j = (\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A}^j)$ ist eine M-Matrix, denn (wir unterdrücken den Index j) die Vorzeichenregel ist erfüllt und mit $\boldsymbol{p} = (1,\dots,1)^T$ folgt

(9.7) $(\boldsymbol{B}\boldsymbol{p})_i = ((\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A})p)_i = 1 + \sigma\,\tau(a_{i,i-1} + a_{ii} + a_{i,i+1}) \overset{(9.3)}{\geq} 1 > 0\ \forall i$

Damit gilt nach Satz (6.4)

$$\underset{\substack{\uparrow\\ \text{Zeilensummennorm}}}{\|\boldsymbol{B}^{-1}\|_\infty} \le \frac{1}{\min\limits_i (\boldsymbol{B}\boldsymbol{p})_i} = 1 \tag{9.8}$$

Auflösen von (9.1) nach $\hat{\boldsymbol{y}}$ liefert

$$\underbrace{(\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A})}_{\boldsymbol{B}}\,\hat{\boldsymbol{y}} = (\boldsymbol{I} + \tau(\sigma-1)\boldsymbol{A})\,\boldsymbol{y} + \tau\boldsymbol{\varphi} \tag{9.9}$$

und

$$\begin{aligned}
|\{(\boldsymbol{I} + \tau(\sigma-1)\,\boldsymbol{A})\,\boldsymbol{y}\}_i| &= \big|(1 + \tau(\sigma-1)\,a_{ii})\,y_i + \tau(\sigma-1)\,a_{i,i-1}\,y_{i-1} + \tau(\sigma-1)\,a_{i,i+1}\,y_{i+1}\big| \\
&\le \big|\underbrace{(1 + \tau(\sigma-1)\,a_{ii})}_{\ge 0 \text{ nach } (9.4),\, a_{ii}\ge 0}\big|\,|y_i| + \big|\underbrace{\tau(\sigma-1)\,a_{i,i-1}}_{\ge 0, \text{ da } a_{i,i-1}\le 0}\big|\,|y_{i-1}| \\
&\quad + \big|\underbrace{\tau(\sigma-1)\,a_{i,i+1}}_{\ge 0, \text{ da } a_{i,i+1}\le 0}\big|\,|y_{i+1}| \\
&\le \big(1 + \underbrace{\tau(\sigma-1)}_{\le 0}\,\underbrace{(a_{ii} + a_{i,i-1} + a_{i,i+1})}_{\ge 0 \text{ nach } (9.3)}\big)\,\|\boldsymbol{y}\|_\infty \\
&\le \|\boldsymbol{y}\|_\infty,
\end{aligned}$$

deshalb folgt aus (9.9) mit (9.8)

$$\begin{aligned}
\|\hat{\boldsymbol{y}}\|_\infty &\le \underbrace{\|\boldsymbol{B}^{-1}\|_\infty}_{\le 1}(\|\boldsymbol{I} + \tau(\sigma-1)\,\boldsymbol{A}\,\boldsymbol{y})\|_\infty + \tau\|\boldsymbol{\varphi}\|_\infty) \\
&\le \|\boldsymbol{y}\|_\infty + \tau\,\|\boldsymbol{\varphi}\|_\infty \quad \text{(also (9.5))}
\end{aligned}$$

Wendet man diese Ungleichung von Zeitschicht zu Zeitschicht an, so folgt

$$\|\boldsymbol{y}^{j+1}\|_\infty \le \|\boldsymbol{y}^0\|_\infty + \sum_{k=0}^{j} \tau\,\|\boldsymbol{\varphi}^k\|_\infty\,.$$

□

Konvergenzbetrachtung

Unter den Voraussetzungen von Satz 9.1 gilt mit den Bezeichnungen aus Definition 5.3: Mit

$$\boldsymbol{z} = \underset{\substack{\uparrow\\ \text{Gitterfkt. der exakten Lösung}}}{\boldsymbol{u}} - \boldsymbol{y}$$

folgt

$$\begin{aligned}
(\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A}^j)\,\boldsymbol{z}_t + \boldsymbol{A}^j\,\boldsymbol{z} &= (\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A}^j)\,\boldsymbol{u}_t + \boldsymbol{A}^j\,\boldsymbol{u} - \big[\underbrace{(\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A}^j)\,\boldsymbol{y}_t + \boldsymbol{A}^j\,\boldsymbol{y}}_{=\boldsymbol{\varphi}^j}\big] \\
&= \boldsymbol{\psi}^j \quad \text{(Diskretisationsfehler, vgl. Definition (5.1))}
\end{aligned}$$

Wegen $\boldsymbol{z}^0 = 0, z_0^j = z_N^0 = 0$ liefert die Abschätzung (9.6)

$$\|\boldsymbol{z}^{i+1}\|_\infty \leq \tau \sum_{k=0}^{j} \|\boldsymbol{\psi}^k\|_\infty, \tag{9.10}$$

also gleichmäßige Konvergenz von der Ordnung des Diskretisierungsfehlers unter entsprechenden Differenzierbarkeitsvoraussetzungen an u.

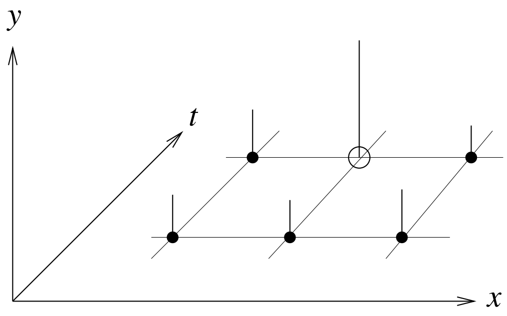
Bemerkung:

Wählt man $\sigma = \frac{1}{2}$ (wie bei Crank-Nicolson), so erhält man aus (9.10) unter der Schrittweitenbeschränkung (9.4): $\tau \leq \frac{2}{a_{ii}^j}$ $(= h^2$ im einfachsten Fall, vgl. (2.11)) eine Verfahrensordnung $O(\tau^2 + h^2)$.

Fazit:
Crank-Nicolson ohne Schrittweitenbeschränkung erst anwenden, wenn das Verfahren schon geglättet ist – sofern man für kleine Zeiten Oszilllationen vermeiden will.

Oszillationsfreiheit

Hinweise: Im Abschnitt über elliptische Differentialgleichungen werden wir ein diskretes Maximumprinzip beweisen, das im Wesentlichen für $M-$Matrizen gilt (vgl. Sätz 12.5 bis 12.5. Es hängt nur ab von der Diskretisierungsmatrix (die allerdings auch die Randwerte mit einschließen muß) für die entsprechende Aufgabe. Im parabolischen Fall muß man dazu eine "Blockmatrix" aufstellen, die über alle Zeitschichten geht. Das Maximumprinzip besagt dann, daß das Maximum nur am parabolischen Rand angenommen werden kann. Dieses Maximumprinzip gilt auch lokal.
Die folgende 2-dimensionale Abbildung (die $y-$Werte werden in den Gitterpunkten abgetragen) entspricht einer Oszillation. Gilt das Maximumprinzip (hier lokal), so muß das Maximum im parabolischen Rand angenommen werden, d.h. hier in einem der "dicken" Gitterpunkte. Die Zeichung steht also im Widerspruch zum Maximumprinzip.



§ 10 Die mehrdimensionale Wärmeleitungsgleichung

Wir beschränken uns (zunächst) auf die Aufgabe

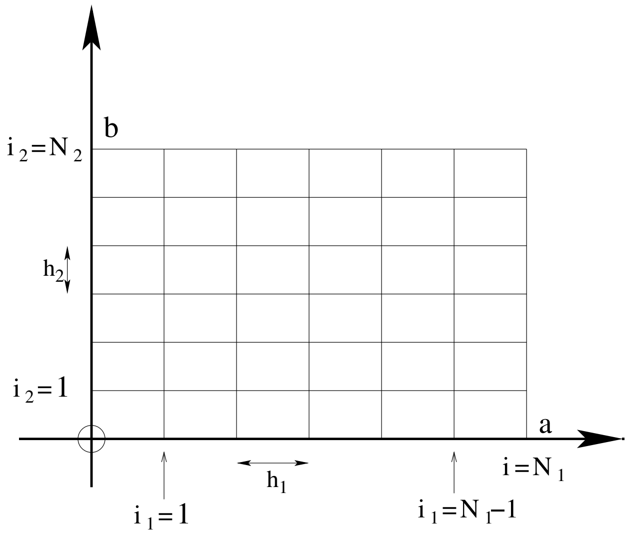
$$(10.1)\qquad \begin{aligned} &\frac{\partial u}{\partial t} = \operatorname{div}(k(x)\ \operatorname{grad} u) + f = \sum_{i=1}^{n} \frac{\partial}{\partial x_i}\left(k(x)\,\frac{\partial u}{\partial x_i}\right) + f \\ &\qquad\qquad x \in \Omega \subset \mathbb{R}^p,\ \ 0 < t < T \\ &\text{AWe: } t = 0:\ \ u(x,0) = u_0(x),\ \ x \in \overline{\Omega} \\ &\text{RWe: } \ u\Big|_{\Gamma\times[0,T]} = g,\ \ \Gamma = \partial\,\Omega \end{aligned}$$

Insbesondere betrachten wir als Ω zunächst einen Quader (analog zum 1D-Fall). Wir werden zeigen, daß diese Aufgabe im Wesentlichen auf den 1D-Fall (bzgl. des Orts) reduzierbar ist. Danach kann man wie im 1D-Fall auch $k = k(x,t)$ zulassen, sowie Konvektion und Abbaurate.
Im Weiteren beschränken wir uns zunächst (im Wesentlichen) auf den 2D-Fall, die Verallgemeinerung auf mehr Dimensionen ist dann offensichtlich.

Wir beschreiben zunächst die

Diskretisierung des div grad-Terms im Rechteck

Sei $\overline{\Omega} = [0,a] \times [0,b]$, $a = N_1 \cdot h_1$, $b = N_2 \cdot h_2$.
$\overline{\omega}$ sei die Menge aller Gitterpunkte aus $\overline{\Omega}$,
ω die Menge der inneren Gitterpunkte: $\omega \subset \Omega$,
γ die Menge der Randpunkte: $\gamma \in \partial\,\Omega$, also $\overline{\omega} = \omega + \gamma$.



Ist u eine auf Ω (bzw. $\overline{\Omega}$) definierte Funktion, so definiert sie eine Gitterfunktion $\boldsymbol{u}$, die durch die Restriktion von u auf die Gitterpunkte entsteht. Mit der Gitterfunktion $\boldsymbol{y}$ bezeichnen wir die zugehörigen Näherungswerte.

Wir bezeichnen die inneren Gitterpunkte $(\in \omega)$ auf verschiedene Weisen:

(i) Wir können die $N = (N_1 - 1) \cdot (N_2 - 1)$ inneren Gitterpunkte durchnummerieren (z.B. zeilenweise von links nach rechts und von unten nach oben oder spaltenweise von unten nach oben und von links nach rechts). Der einzelne Gitterpunkt bekommt dann seinen Zählindex

$$x_\ell, \quad \ell \in [1, \dots, N], \quad N = (N_1 - 1)\,(N_2 - 1)$$

oder

(ii) entsprechend den (i_1, i_2) Indizes, die die Stellung des Punktes im Gitter zeigen

$$x_\ell = x_{i_1, i_2} = (i_1 h_1, i_2 k_2), \quad 1 \le i_\nu \le N_\nu - 1, \quad \nu = 1, 2.$$

Damit definieren wir den Gittervektor

$$\boldsymbol{y} = (x_1, \dots, y_N)^T, \quad N = (N_1 - 1)\,(N_2 - 1).$$

Nun entspricht der Aufgabe (10.1) das Differenzenschema

$$\boldsymbol{y}_t = -\boldsymbol{A}\,\boldsymbol{y}^\sigma + \boldsymbol{f}^{j+\frac{1}{2}}, \quad \boldsymbol{A} = \sum_{i=1}^{2} \boldsymbol{A}_i \quad \left(\sum_{i=1}^{d} \text{ im } d\text{-dim. Fall} \right)$$

Dabei beschreibt $\boldsymbol{A}_i$ die Diskretisierung der Ableitungen in x_i-Richtung. Für $\boldsymbol{A}_i$ werden die Gitterpunkte in Richtung der der x_i-Achse durchnummeriert (also zeilenweise von links nach rechts und von unten nach oben für x_1, und spaltenweise von unten nach oben und links nach rechts für x_2). Daß damit die Gitterpunkte für die Ableitungen in x_1-Richtung und x_2-Richtung unterschiedlich nummeriert sind, soll uns im Augenblick nicht stören. Wir wollen zunächst nur Abschätzungen für die Normen der $\boldsymbol{A}_i$ herleiten und dafür spielt die Reihenfolge der Nummerierung keine Rolle. Dann ist

$$(\boldsymbol{A}_i\,\boldsymbol{y})_\ell = -(k\, y_{\overline{x}_\ell})_{x_\ell} \quad \text{(vgl. (7.1))}$$
$$:= \frac{1}{h_i} \left((k(x_\ell + \tfrac{h_i}{2}\,\boldsymbol{e}^i)\, \frac{y_\ell(x_\ell + h_i\,\boldsymbol{e}^i) - y_\ell(x_\ell)}{h_i} - k(x_\ell - \tfrac{h_i}{2}\,\boldsymbol{e}^i)\, \frac{y(x_\ell) - y(x_\ell - h_i\,\boldsymbol{e}^i)}{h_i} \right)$$

mit den Einheitsvektoren $\boldsymbol{e}^i$, deren Koordinaten sich natürlich auch an der Zählweise orientieren. Abgesehen von den „Randpunkten“ unter den inneren Randpunkten gilt dann

$$y_\ell \left(x_\ell \pm h_i\,\boldsymbol{e}^i \right) = y_{\ell \pm 1}.$$

$y_{\pm 1}$ bedeutet hier nur – ausgehend von x_ℓ – einen Schritt in $+x_i-$ bzw. $-x_i-$ Richtung entsprechend der jeweiligen Durchnummerierung.
Im 2D-Fall benutzen wir das Skalarprodukt $(\boldsymbol{v}, \boldsymbol{w})_{(0,H)}$, $H = h_1 h_2$:

$$(\boldsymbol{v}, \boldsymbol{w})_{(0,H)} = \sum_{\ell=1}^{N} v_\ell\, w_\ell\, h_1\, h_2 = \sum_{k=1}^{N_2-1} \sum_{i=1}^{N_1-1} v_{i,k}\, w_{i,k} \cdot h_1\, h_2 = \sum_{x \in \omega} (v(x),\, w(x)) H.$$

Man beachte, dass die Definition dieser Skalarprodukte und insbesondere ihr Wert nicht von der Nummerierung der Gitterpunkte abhängen. Wir setzen nun voraus

$$0 < c_0 \le k(x) \le c_1 \quad \forall x \in \Omega$$

und leiten Normabschätzungen her, basierend auf den bekannten Abschätzungen aus dem 1D-Fall.

Wir betrachten zunächst $\boldsymbol{A}_1$ bei zeilenweiser Durchnummerierung des Gittervektors. Die Wirkung von $\boldsymbol{A}_1$ auf den Teilgittervektor $\boldsymbol{v}^{(k)}$ der k-ten Zeile wird durch eine symmetrische, positiv definite, tridiagonale Matrix $\boldsymbol{A}_1^{(k)}$ (vgl. (7.3)) beschrieben. Die Gesamtmatrix $\boldsymbol{A}_1$ ergibt sich damit als Blockdiagonalmatrix, deren Blöcke durch die „Zeilenblöcke“ $\boldsymbol{A}_1^{(k)}$ gegeben werden. $\boldsymbol{A}_1$ ist damit eine symmetrische Tridiagonalmatrix.

Wir setzen nun voraus: $c_0 \le k \le c_1$ und erhalten unter Beachtung von $\boldsymbol{v}|_\Gamma = 0$ für den Teilgittervektor $\boldsymbol{v}^{(k)}$

$$\left(\boldsymbol{A}_1^{(k)}\boldsymbol{v}^{(k)}, \boldsymbol{v}^{(k)}\right)_{(0,H)} \overset{(7.6)}{=} \sum_{i=1}^{N_1} k^{(k)}_{i-\frac{1}{2}} \left(v^{(k)}_{\overline{x}_1,i}\right)^2 h_1\, h_2 \begin{cases} \le & c_1 \\ \ge & c_0 \end{cases} \underbrace{\sum_{i=1}^{N_1} \left(v^{(k)}_{\overline{x}_1,i}\right)^2 h_1}_{\|\boldsymbol{v}^{(k)}\|^2_{(1,h_1)}} \cdot h_2$$

$$= \left\{ \begin{matrix} c_1 \\ c_0 \end{matrix} \right. \|\boldsymbol{v}^{(k)}\|^2_{(1,h_1)} \cdot h_2 \begin{cases} \le & \dfrac{4}{h^2}\, c_1 \\ \ge & 8c_0 \end{cases} \|\boldsymbol{v}^{(k)}\|^2_{(0,h_1)} \cdot h_2$$

Für den ganzen Gittervektor $\boldsymbol{v}$ und die Matrix $\boldsymbol{A}_1$ gilt also (Summation über die Zeilen)

$$(\boldsymbol{A}_1\, \boldsymbol{v}, \boldsymbol{v})_{(0,H)} \begin{cases} \le & c_1 \\ \ge & c_0 \end{cases} \sum_{k=1}^{N_2-1} h_2 \|\, \boldsymbol{v}^{(k)}\|^2_{(1,h_1)} \begin{cases} \le & \dfrac{4c_1}{h_1^2} \\ \ge & 8c_0 \end{cases} \sum_{k=1}^{N_2-1} h_2\, \|\boldsymbol{v}^{(k)}\|^2_{(0,h_1)}$$

Wir definieren im 2D-Fall die Normen entsprechend dem Skalarprodukt durch

$$\|\boldsymbol{v}\|^2_{(1,H)} = \sum_{k=1}^{N_2-1} h_2\, \|\boldsymbol{v}^{(k)}\|^2_{(1,h_1)}, \quad \|v\|^2_{(0,H)} = \sum_{k=1}^{N_2-1} h_2\, \|\boldsymbol{v}^{(k)}\|^2_{(0,h_1)}$$

und erhalten damit

$$(\boldsymbol{A}_1\, \boldsymbol{v}, \boldsymbol{v})_{(0,H)} \left. \begin{matrix} \le & \dfrac{4c_1}{h_1^2} \\ \ge & 8c_0 \end{matrix} \right\} \|\boldsymbol{v}\|^2_{(0,H)} \tag{10.2}$$

Beachte: Ist $\boldsymbol{v}$ die Gitterfunktion einer Funktion $v(x)$, so gilt

$$\|\boldsymbol{v}\|^2_{(0,H)} \xrightarrow{h_1,h_2\to 0} \int_0^1 \int_0^1 v(x,y)^2\, dy\, dy.$$

Durch analoges Vorgehen erhalten wir für $\boldsymbol{A}_2$, indem wir nun den Gittervektor spaltenweise durchnummerieren:

$$(\boldsymbol{A}_2\,\boldsymbol{v},\boldsymbol{v})_{(0,H)} \begin{cases} \leq & \dfrac{4c_1}{h_2^2} \\ \geq & 8c_0 \end{cases} \|v\|^2_{(0,H)} \tag{10.3}$$

Bei spaltenweiser Nummerierung ist $\boldsymbol{A}_2$ eine symmetrische (, tridiagonale) Matrix, d.h.

$$(\boldsymbol{A}_2\,\boldsymbol{v},\boldsymbol{v})_{(0,H)} = (\boldsymbol{v},\boldsymbol{A}_2\,\boldsymbol{v})_{(0,H)} \tag{*}$$

Wird nun $\boldsymbol{v}$ anders nummeriert und die Indizierung der Matrixelemente entsprechend geändert, so geht die tridiagonale Form von $\boldsymbol{A}_2$ verloren, die Gleichung (*) bleibt jedoch richtig, da das Skalarprodukt unabhängig ist von der Komponentenumerierung der Vektoren, d.h. auch bei zeilenweiser Nummerierung bleibt $\boldsymbol{A}_2$ (eigentlich müßte man nach der Umnummerierung den Namen ändern) eine symmetrische Matrix. Dasselbe Argument zeigt, dass $\boldsymbol{A}_2$ positiv definit ist, denn die „Spaltenblöcke" $\boldsymbol{A}_2^{(i)}$ der Matrix $\boldsymbol{A}_2$ (ebenso wie die „Zeilenblöcke" von $\boldsymbol{A}_1$) sind positiv definit.

Aufgabe: Wie sieht $\boldsymbol{A}_2$ bei zeilenweiser Nummerierung aus?

Wir können nun wie früher zur Lösung der Aufgabe (10.1) das Verfahren

$$\boldsymbol{B}\,\boldsymbol{y}_t + \boldsymbol{A}\,\boldsymbol{y} = \boldsymbol{\varphi}, \quad \boldsymbol{B} = \boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A}, \quad \boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2 \text{ symmetr., pos. def.} \tag{10.4}$$

anwenden und erhalten dafür die bekannten Stabilitäts- und Konvergenzaussagen

$\boldsymbol{B} \geq \frac{\tau}{2}\boldsymbol{A}$ sichert die Stabilität bzgl. der Anfangswerte (vgl. Satz 8.4), die wichtigste Eigenschaft. Diese Bedingung wird mit $\boldsymbol{I} \geq \frac{\boldsymbol{A}}{\|\boldsymbol{A}\|}$ verschärft zu

$$\boldsymbol{B} \geq \left(\frac{1}{\|\boldsymbol{A}\|} + \sigma\,\tau\right)\boldsymbol{A} \overset{!}{\geq} \frac{\tau}{2}\boldsymbol{A}$$

was wegen $\boldsymbol{A} > 0$ durch

$$\sigma \geq \frac{1}{2} - \frac{1}{\tau\,\|\boldsymbol{A}\|}$$

gesichert wird. Offensichtlich ist $\underline{\sigma \geq \frac{1}{2}\text{ problemlos}}$.

Für $\sigma = 0$ (explizites Verfahren) folgt hieraus die Beschränkung

$$\tau \leq \frac{2}{\|\boldsymbol{A}\|}$$

Setzt man $h = h_1 = h_2$ so folgt aus (10.2), (10.3)

$$\|\boldsymbol{A}\|_S \leq \|\boldsymbol{A}_1\|_S + \|\boldsymbol{A}_2\|_S \leq 2\cdot\frac{4c_1}{h^2} = \underset{\substack{\uparrow\\ p\,=\text{Raumdimension}}}{p}\cdot\frac{4c_1}{h^2}$$

und damit

$$\tau \leq \frac{h^2}{2c_1\, p}$$

Dies ist um den Faktor p schlimmer als im 1D-Fall

$\Longrightarrow \quad \boldsymbol{\sigma = 0}$ **indiskutabel im $\boldsymbol{pD}$-Fall**.

Lösungsmöglichkeiten für $\boldsymbol{\sigma > 0}$ insbesondere $\boldsymbol{\sigma = \frac{1}{2}}$

Das Verfahren (10.4) läßt sich in der Form schreiben

$$(\boldsymbol{I} + \sigma\,\tau\,\boldsymbol{A})\,\boldsymbol{y}^{j+1} = (I - (1-\sigma)\tau\,\boldsymbol{A})\,\boldsymbol{y}^j + \tau\,\boldsymbol{\varphi} =: \boldsymbol{F}^j$$

Die rechte Seite der Gleichung ist bekannt. Für jeden Zeitschritt kann man dies als stationäres parabolisches, d.h. elliptisches Problem betrachten und darauf das Mehrgitterverfahren (MGV) anwenden, das wir im Kapitel für elliptische Aufgaben beschreiben. Damit läßt sich auch der Zeitaufwand für die Lösung des Gleichungssystems abschätzen. Solche Verfahren werden in der Praxis gerechnet (beachte, dass $\boldsymbol{A}$ nun keine Tridiagonalmatrix mehr ist).

Es gibt jedoch schnellere Verfahren, die auf tridiagonalen Gleichungssystemen beruhen, die Verfahren der alternativen Richtungen: (ADI-method, alternating direction implicit method) auf die wir nun eingehen.

Verfahren der Alternierenden Richtungen

(ADI-Verfahren: Alternating direction implicit method)

Das Verfahren wurde Anfang der 50er Jahre von Peachman-Rachford („Erdölleuten“) entwickelt für eine AWA mit der Differentialgleichung

$$\begin{aligned} \frac{\partial u}{\partial t} &= \sum_{i=1}^{2} L_i\, u + f \\ L_i\, u &= \frac{\partial}{\partial x_i}\left(k_i \frac{\partial u}{\partial x_i}\right) - q_i\, u \quad \text{(nicht konstante Koeffizienten möglich)} \end{aligned} \tag{10.5}$$

Die diskretisierte Aufgabe führt zum Differenzenchema (vgl. (4.7) bzw. (10.4) und setze $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2$). Wären Abbauterme $q = (q_1, q_2)$ in der Differentialgleichung enthalten, so könnten wir sie uns, wie früher, als bereits in $\boldsymbol{A}_i$ integriert vorstellen, da sie nur zur Hauptdiagonale beitragen.

$$(\boldsymbol{I} + \sigma\,\tau(\boldsymbol{A}_1 + \boldsymbol{A}_2))\,\boldsymbol{y}_t + (\boldsymbol{A}_1 + \boldsymbol{A}_2)\,\boldsymbol{y} = \boldsymbol{\varphi}, \qquad (\boldsymbol{A}_i \hat{=} \text{Diskretisierung von } L_i) \tag{10.6}$$

bzw. für $\sigma = \frac{1}{2}$ (q_i trägt nur zur Hauptdiagonalen bei)

$$\left(\boldsymbol{I} + \frac{\tau}{2}(\boldsymbol{A}_1 + \boldsymbol{A}_2)\right)\,\boldsymbol{y}_t + (\boldsymbol{A}_1 + \boldsymbol{A}_2)\,\boldsymbol{y} = \boldsymbol{\varphi}. \tag{10.7}$$

Wegen $\boldsymbol{B} = \boldsymbol{I} + \frac{\tau}{2}\,\boldsymbol{A} \geq \frac{\tau}{2}\,\boldsymbol{A} + \frac{\varepsilon}{2}\,\boldsymbol{I}$ ist die Stabilität für $\varepsilon \leq 2$ problemlos. Das $\frac{\varepsilon}{2}\,\boldsymbol{I}$ wird durch den Beitrag von q geliefert. Allerdings ist $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2$ keine Tridiagonalmatix mehr, weshalb jeder Iterationsschritt ziemlichen Aufwand erfordert.
Deshalb schlugen Peachman-Rachford statt (10.6) folgendes 2-stufige Verfahren vor

$$\begin{aligned} &\frac{\boldsymbol{y}^{j+\frac{1}{2}} - \boldsymbol{y}^j}{\tau/2} + \boldsymbol{A}_1\,\boldsymbol{y}^{j+\frac{1}{2}} + \boldsymbol{A}_2\,\boldsymbol{y}^j = \boldsymbol{\varphi}^{j+\frac{1}{2}} \\ &\frac{\boldsymbol{y}^{j+1} - \boldsymbol{y}^{j+\frac{1}{2}}}{\tau/2} + \boldsymbol{A}_1\,\boldsymbol{y}^{j+\frac{1}{2}} + \boldsymbol{A}_2\,\boldsymbol{y}^{j+1} = \boldsymbol{\varphi}^{j+\frac{1}{2}}. \end{aligned} \tag{10.8}$$

Numerisch geht man bei der Rechnung wie folgt vor:
Die erste Gleichung schreibt man als

$$(\frac{2}{\tau}\boldsymbol{I} + \boldsymbol{A}_1)\boldsymbol{y}^{j+\frac{1}{2}} = \boldsymbol{F} \text{ mit } \boldsymbol{F} = (\frac{2}{\tau} - \boldsymbol{A}_2)\,\boldsymbol{y}^j + \boldsymbol{\varphi}^{j+\frac{1}{2}}. \tag{$*$}$$

Dazu werden die $\boldsymbol{A}_i$ und entsprechend $\boldsymbol{y}^j$ zeilenweise von links nach rechts und unten nach oben geordnet, $\boldsymbol{A}_1$ wird dadurch zu einer Tridiagonalmatrix. Man berechnet $\boldsymbol{F}$ und berechnet dann $\boldsymbol{y}^{j+\frac{1}{2}}$ relativ billig mit dem Tridiagonalverfahren.
Die Matrizen $\boldsymbol{A}_i$ werden nun spaltenweise von links nach rechts und von unten nach oben umgeordnet. Ebenso wird $\boldsymbol{y}^{j+\frac{1}{2}}$ entsprechend umgeordnet. Die umgeordneten Größen werden (nicht ganz korrekt) wieder mit $\boldsymbol{A}_i$ bzw. $\boldsymbol{y}^{j+\frac{1}{2}}$ bezeichnet.
Die zweite Gleichung aus (10.8) ist äquivalent zu

$$(\frac{2}{\tau} + \boldsymbol{A}_2)\,y^{j+1} = \bar{\boldsymbol{F}} \text{ mit } \bar{\boldsymbol{F}} = (\frac{2}{\tau} - \boldsymbol{A}_1)\,\boldsymbol{y}^{j+\frac{1}{2}} + \boldsymbol{\varphi}^{j+\frac{1}{2}}. \tag{$**$}$$

Man berechnet $\bar{\boldsymbol{F}}$ bzgl. der neuen Nummerierung der Komponenten von $\boldsymbol{y}^{j+\frac{1}{2}}$ und danach wieder mit dem Tridiagonalalgorithmus $\boldsymbol{y}^{j+1}$. Dieses Spiel wiederholt sich von Zeitschritt zu Zeitschritt.

Bemerkungen:

1. Der Schritt von $t_j \to t_{j+1}$ wird in 2 Teilschritte zerlegt, der erste in x_1-Richtung, der zweite in x_2-Richtung (**alternierende Richtungen**), wobei in jedem Schritt ein Gleichungssystem mit einer **Tridiagonalmatrix** zu lösen ist, was mit geringem Aufwand (vgl. den Abschnitt: Tridiagonalverfahren) möglich ist. Physikalisch bedeutet dieses Vorgehen:
   1. Schritt: Wärmeausbreitung (Diffusion) in x_1- Richtung,
   2. Schritt: Wärmeausbreitung (Diffusion) in x_2- Richtung.

2. Beide Gleichungen (10.8) sind Approximationen an die Ausgangsgleichung, jedoch nur von 1. Ordnung. Wir werden jedoch sehen, dass das Gesamtverfahren von 2. Ordnung ist und stabil.

Wir zeigen die Stabilität gleich für die etwas allgemeinere Verfahrensklasse der **Splitting-Verfahren**, wobei wir in (10.8) verschiedene rechte Seiten zulassen, also

$$
(10.9)\qquad
\begin{aligned}
(\alpha)\quad & \frac{\boldsymbol{y}^{j+\frac{1}{2}} - \boldsymbol{y}^j}{\tau/2} + \boldsymbol{A}_1\, \boldsymbol{y}^{j+\frac{1}{2}} + \boldsymbol{A}_2\, \boldsymbol{y}^j = \boldsymbol{\varphi}_1 \\
(\beta)\quad & \frac{\boldsymbol{y}^{j+1} - \boldsymbol{y}^{j+\frac{1}{2}}}{\tau/2} + \boldsymbol{A}_1\, \boldsymbol{y}^{j+\frac{1}{2}} + \boldsymbol{A}_2\, \boldsymbol{y}^{j+1} = \boldsymbol{\varphi}_2.
\end{aligned}
$$

Zur Stabilitätsberechnung wird wird dieses 2-stufige Verfahren auf ein 1-stufiges zurückgeführt durch Elimination von $\boldsymbol{y}^{j+\frac{1}{2}}$. Danach lassen sich die bekannten Stabilitätssätze anwenden.

Wir eliminieren $\boldsymbol{y}^{j+\frac{1}{2}}$. Subtraktion von $(\beta) - (\alpha)$ liefert

$$
\begin{aligned}
\frac{\boldsymbol{y}^{j+1} - 2\boldsymbol{y}^{j+\frac{1}{2}} + \boldsymbol{y}^j}{\tau/2} &= -\boldsymbol{A}_2 \left(\boldsymbol{y}^{j+1} - \boldsymbol{y}^j\right) + \boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1 \\
\boldsymbol{y}^{j+\frac{1}{2}} &= \frac{\tau}{4} \boldsymbol{A}_2 \left(\boldsymbol{y}^{j+1} - \boldsymbol{y}^j\right) - \frac{\tau}{4}(\boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1) + \frac{1}{2}\left(\boldsymbol{y}^{j+1} + \boldsymbol{y}^j\right)
\end{aligned}
$$

Einsetzen in (β) :

$$
\frac{\boldsymbol{y}^{j+1} - \frac{1}{2}\left(\boldsymbol{y}^{j+1} + \boldsymbol{y}^j\right)}{\tau/2} + \boldsymbol{A}_1\, \boldsymbol{y}^{j+\frac{1}{2}} + \boldsymbol{A}_2 \left(\boldsymbol{y}^{j+1} - \frac{1}{2}\left(\boldsymbol{y}^{j+1} - \boldsymbol{y}^j\right)\right) + \frac{1}{2}(\boldsymbol{\varphi}_2 - \boldsymbol{\varphi}_1) = \boldsymbol{\varphi}_2
$$

Auflösen nach $\boldsymbol{y}_t^j$:

$$\begin{aligned}\frac{\frac{1}{2}\,(\boldsymbol{y}^{j+1}-\boldsymbol{y}^j)}{\tau/2} \quad &+\boldsymbol{A}_1\,\left(\tfrac{\tau}{4}\,\boldsymbol{A}_2\,\boldsymbol{y}^{j+1}-\tfrac{\tau}{4}\,\boldsymbol{A}_2\,\boldsymbol{y}^j-\tfrac{\tau}{4}(\boldsymbol{\varphi}_2-\boldsymbol{\varphi}_1)+\tfrac{1}{2}\,(\boldsymbol{y}^{j+1}+\boldsymbol{y}^j)\right)\\ &+\boldsymbol{A}_2\,\left(\tfrac{1}{2}\,(\boldsymbol{y}^{j+1}+\boldsymbol{y}^j)\right)=\tfrac{1}{2}(\boldsymbol{\varphi}_1+\boldsymbol{\varphi}_2)\end{aligned}$$

$$\begin{aligned}\frac{\boldsymbol{y}^{j+1}-\boldsymbol{y}^j}{\tau} \quad &+\tfrac{1}{2}\,\boldsymbol{A}_1\,(\boldsymbol{y}^{j+1}+\boldsymbol{y}^j)+\tfrac{1}{2}\,\boldsymbol{A}_2\,(\boldsymbol{y}^{j+1}+\boldsymbol{y}^j)+\tfrac{\tau}{4}\,\boldsymbol{A}_1\,\boldsymbol{A}_2\,(\boldsymbol{y}^{j+1}-\boldsymbol{y}^j)\\ &=\tfrac{1}{2}(\boldsymbol{\varphi}_1+\boldsymbol{\varphi}_2)+\tfrac{\tau}{4}\,\boldsymbol{A}_1(\boldsymbol{\varphi}_1-\boldsymbol{\varphi}_2)\end{aligned}$$

$$\begin{aligned}\frac{\boldsymbol{y}^{j+1}-\boldsymbol{y}^j}{\tau} \quad &+\tfrac{1}{2}(\boldsymbol{A}_1+\boldsymbol{A}_2)\,(\boldsymbol{y}^{j+1}-\boldsymbol{y}^j)+\tfrac{\tau}{4}\,\boldsymbol{A}_1\,\boldsymbol{A}_2\,(\boldsymbol{y}^{j+1}-\boldsymbol{y}^j)+(\boldsymbol{A}_1+\boldsymbol{A}_2)\,\boldsymbol{y}^j\\ &=\tfrac{\tau}{4}\,\boldsymbol{A}_1(\boldsymbol{\varphi}_1-\boldsymbol{\varphi}_2)+\tfrac{1}{2}(\boldsymbol{\varphi}_1+\boldsymbol{\varphi}_2)\end{aligned}$$

Wir erhalten mit $\boldsymbol{y}^{j+1}-\boldsymbol{y}^j=\tau\boldsymbol{y}_t^j$

(10.10)
$$\underbrace{\left(\boldsymbol{I}+\frac{\tau}{2}(\boldsymbol{A}_1+\boldsymbol{A}_2)+\frac{\tau^2}{4}\,\boldsymbol{A}_1\,\boldsymbol{A}_2\right)}_{\widetilde{\boldsymbol{B}}}\boldsymbol{y}_t^j+\underbrace{(\boldsymbol{A}_1+\boldsymbol{A}_2)}_{\boldsymbol{A}}\,\boldsymbol{y}^j=\underbrace{\frac{\tau}{4}\,\boldsymbol{A}_1(\boldsymbol{\varphi}_1-\boldsymbol{\varphi}_2)+\frac{1}{2}(\boldsymbol{\varphi}_1+\boldsymbol{\varphi}_2)}_{\boldsymbol{\varphi}}$$

bzw.

(10.11) $$\left(\boldsymbol{I}+\frac{\tau}{2}\,\boldsymbol{A}_1\right)\left(\boldsymbol{I}+\frac{\tau}{2}\,\boldsymbol{A}_2\right)\boldsymbol{y}_t^j+(\boldsymbol{A}_1+\boldsymbol{A}_2)\,\boldsymbol{y}^j=\frac{\tau}{4}\,\boldsymbol{A}_1(\boldsymbol{\varphi}_1-\boldsymbol{\varphi}_2)+\frac{1}{2}(\boldsymbol{\varphi}_1+\boldsymbol{\varphi}_2).$$

Bemerkung:

Setzt man nicht von vorneherein $\sigma=\frac{1}{2}$ wie beim Übergang von (10.6) zu (10.7), sondern betrachtet das leicht verallgemeinerte Verfahren, das man erhält, wenn man in (10.8), (10.9) $\tau/2$ durch $\tau\,\sigma$ ersetzt, so lautet die (10.10) entsprechende Form

(10.12)
$$(\boldsymbol{I}+\sigma\,\tau(\boldsymbol{A}_1+\boldsymbol{A}_2)+\sigma^2\,\tau^2\,\boldsymbol{A}_1\,\boldsymbol{A}_2)\,\boldsymbol{y}_t^j+(\boldsymbol{A}_1+\boldsymbol{A}_2)\boldsymbol{y}^j=\frac{\sigma\,\tau}{2}\,\boldsymbol{A}_1(\boldsymbol{\varphi}_1-\boldsymbol{\varphi}_2)+\frac{1}{2}(\boldsymbol{\varphi}_1+\boldsymbol{\varphi}_2).$$

Diese Form liefert bei der Abschätzung des Diskretisierungsfehlers die Möglichkeit, analog zum Vorgehen in Satz 5.2, durch entsprechende Wahlen von σ und $\boldsymbol{\varphi}$ höhere Konvergenzgeschwindigkeiten zu erreichen.

Nun ist (10.12) wieder ein Verfahren der Art

(10.13) $$\widetilde{\boldsymbol{B}}\,\boldsymbol{y}_t+\boldsymbol{A}\,\boldsymbol{y}=\boldsymbol{\varphi},\qquad \begin{aligned}\widetilde{\boldsymbol{B}}&=\boldsymbol{I}+\sigma\,\tau\,\boldsymbol{A}+\sigma^2\tau^2\,\boldsymbol{A}_1\,\boldsymbol{A}_2\\ \boldsymbol{A}&=\boldsymbol{A}_1+\boldsymbol{A}_2\\ \boldsymbol{\varphi}&=\frac{\sigma\,\tau}{2}\,\boldsymbol{A}_1(\boldsymbol{\varphi}_1-\boldsymbol{\varphi}_2)+\frac{1}{2}(\boldsymbol{\varphi}_1+\boldsymbol{\varphi}_2)\\ &(\text{wobei } \boldsymbol{\varphi}_1=\boldsymbol{\varphi}_2=\boldsymbol{f}^{j+\frac{1}{2}} \text{ üblicherweise})\end{aligned}$$

Da $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2$ symmetrisch und positiv definit ist, liegt Stabilität vor falls $\widetilde{\boldsymbol{B}} \geq \frac{\tau}{2}\,\boldsymbol{A} + \frac{\varepsilon}{2}\,\boldsymbol{I}$ ist (vgl. Satz 4.2), bzw.

$$\boldsymbol{I} + \tau\left(\sigma - \frac{1}{2}\right)\boldsymbol{A} + \sigma^2\,\tau^2\,\boldsymbol{A}_1\,\boldsymbol{A}_2 \geq \frac{\varepsilon}{2}\boldsymbol{I}. \tag{10.14}$$

Diese Bedingung ist für $\sigma \geq \frac{1}{2}$ sicher erfüllt für $\varepsilon \leq 2$ falls $\boldsymbol{A}_1\,\boldsymbol{A}_2 \geq 0$ ist.

Letztere Eigenschaft werden wir gleich zeigen. Sie bringt jedoch einschneidende Voraussetzungen bzgl. der Wahl von Ω mit sich (vgl. Satz 10.1), weshalb wir darauf hinweisen, dass in der Wahl von $\varepsilon = 2$ noch „Luft“ liegt. Wir zeigen nun

Satz 10.1
In einem Rechteck $\overline{\Omega} = [0, a_1] \times [0, a_2] \subset \mathbb{R}^2$ seien $\boldsymbol{A}_i$ die Diskretisierungsmatrizen von $L_i\,u = \dfrac{\partial}{\partial x_i}\left(k_i\,\dfrac{\partial}{\partial x_i}\,u\right) - q_i\,u\,,\ i = 1, 2\,.$
Es sei $k_i(x) = k_i(x_i)\,,\ q_i(x) = q_i(x_i)$ und $\boldsymbol{A}_i = \boldsymbol{A}_i^T > 0\,,\ i = 1, 2\,.$

Dann gilt:

a) $\boldsymbol{A}_1\,\boldsymbol{A}_2 = \boldsymbol{A}_2\,\boldsymbol{A}_1$ und

b) $(\boldsymbol{A}_1\,\boldsymbol{A}_2\,\boldsymbol{y}, \boldsymbol{y}) \geq 0 \quad \forall \boldsymbol{y} \in \mathbb{R}^n\,.$

Bemerkungen:

1. Dass $k_i(x) = k_i(x_i)$, d.h. die Diffusionskonstante $k_1(x)$ nur von x_1 - nicht von x_2 - abhängt, entsprechend für k_2, ist realistisch wenn z.B. $\overline{\Omega}$ ein Schnitt in einer geschichteten Fläche darstellt.

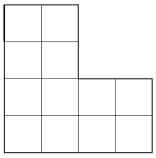
2. Entsprechende Diskretisierungsmatrizen wurden in § 7 , (7.3) hergeleitet. Dazu beachte man, dass gemäß § 8 q_i nur einen Beitrag zur Hauptdiagonalen von $\boldsymbol{A}_i$ liefert und somit weder Symmetrie noch Definitheit stört.

3. Die Voraussetzung eines Rechteckgebiets ist wesentlich, wie folgende Aufgabe zeigt.

Aufgabe:

1. Man beweise a) aus Satz 10.1

2. Für $L_1\,u = u_{x_1 x_1}, L_2\,u = u_{x_2 x_2}$ und das $L-$Gebiet, samt eingezeichnetem Gitter zeige man, dass a) für die entsprechenden Matrizen $\boldsymbol{A}_i$ nicht gilt.



3. Für $\boldsymbol{A} \in \mathbb{R}^{n\times n}\,,\ \boldsymbol{A}^T = \boldsymbol{A} > 0$ gibt es genau eine Wurzel $\boldsymbol{A}^{1/2}$ (d.h. $\boldsymbol{A}^{1/2}A^{1/2} = \boldsymbol{A}$) mit $\boldsymbol{A}^{1/2} = (\boldsymbol{A}^{1/2})^T > 0\,.$

4. Das Verfahren (Newton-Verfahren) für $\boldsymbol{A} = \boldsymbol{A}^T > 0$

$$\boldsymbol{X}_{n+1} = \frac{1}{2}\,(\boldsymbol{X}_n + \boldsymbol{X}_n^{-1}\,\boldsymbol{A}), \quad \boldsymbol{X}_1 = \boldsymbol{I} \tag{10.15}$$

ist durchführbar und konvergiert gegen das symmetrische $\boldsymbol{A}^{1/2} > 0$. Dazu gehört insbesondere der Nachweis der Existenz der $\boldsymbol{X}_n^{-1}$.

5. $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$, $\det \boldsymbol{A} \neq 0 \Longrightarrow \boldsymbol{A}\boldsymbol{B}$ und $\boldsymbol{B}\boldsymbol{A}$ haben dieselben Eigenwerte.

Beweis Satz 10.1

a) laut Aufgabe

b) wird aus a) gefolgert.

Wir zeigen zunächst für beliebige Matrizen $\boldsymbol{A} = \boldsymbol{A}^T > 0$, $\boldsymbol{B} = \boldsymbol{B}^T > 0$:

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A} \implies \sqrt{\boldsymbol{A}}\boldsymbol{B} = \boldsymbol{B}\sqrt{\boldsymbol{A}} \implies \sqrt{\boldsymbol{A}}\,\sqrt{\boldsymbol{B}} = \sqrt{\boldsymbol{B}}\,\sqrt{\boldsymbol{A}} \tag{10.16}$$

Für das Verfahren (10.15)

$$\begin{aligned} &\boldsymbol{X}_{n+1} = \tfrac{1}{2}\,(\boldsymbol{X}_n + \boldsymbol{X}_n^{-1}\,\boldsymbol{A}), \quad \boldsymbol{X}_1 = \boldsymbol{I} \\ &\text{gilt} \\ &\boldsymbol{X}_n\,\boldsymbol{B} = \boldsymbol{B}\,\boldsymbol{X}_n \quad \forall n. \end{aligned} \tag{10.17}$$

Die Behauptung ist richtig für $n = 1$, da $\boldsymbol{X}_1 = \boldsymbol{I}$.

Sie sei richtig für n, dann gilt

$$\boldsymbol{X}_{n+1}\,\boldsymbol{B} = \frac{1}{2}\,(\boldsymbol{X}_n\,\boldsymbol{B} + \boldsymbol{X}_n^{-1}\,\boldsymbol{A}\,\boldsymbol{B}) \overset{\text{a)}}{=} \frac{1}{2}\,(\boldsymbol{X}_n\,\boldsymbol{B} + \boldsymbol{X}_n^{-1}\,\boldsymbol{B}\,\boldsymbol{A})$$

und mit der Induktionsvoraussetzung folgt wegen

$$\boldsymbol{B}\,\boldsymbol{X}_n = \boldsymbol{X}_n\,\boldsymbol{B} \quad \iff \quad \boldsymbol{B} = \boldsymbol{X}_n\,\boldsymbol{B}\,\boldsymbol{X}_n^{-1} \quad \iff \quad \boldsymbol{X}_n^{-1}\,\boldsymbol{B} = \boldsymbol{B}\,\boldsymbol{X}_n^{-1}$$

also

$$\boldsymbol{X}_{n+1}\,\boldsymbol{B} = \frac{1}{2}\,(\boldsymbol{B}\,\boldsymbol{X}_n + \boldsymbol{B}\,\boldsymbol{X}_n^{-1}\,\boldsymbol{A}) = \boldsymbol{B}\,\frac{1}{2}\,(\boldsymbol{X}_n + \boldsymbol{X}_n^{-1}\,\boldsymbol{A}) = \boldsymbol{B}\,\boldsymbol{X}_{n+1}.$$

Wegen $\boldsymbol{X}_n \to \sqrt{\boldsymbol{A}}$ folgt aus (10.5) die erste Behauptung (10.16): $\sqrt{\boldsymbol{A}}\,\boldsymbol{B} = \boldsymbol{B}\,\sqrt{\boldsymbol{A}}$. Unter Benutzung dieser Behauptung zeigt man auf dieselbe Weise, in dem man das Verfahren (10.15) zur Berechnung von $\sqrt{\boldsymbol{B}}$ aufstellt, zuerst $\boldsymbol{X}_{n+1}\sqrt{\boldsymbol{A}} = \sqrt{\boldsymbol{A}}\,\boldsymbol{X}_{n+1}$ und durch Grenzübergang schließlich $\sqrt{\boldsymbol{A}}\,\sqrt{\boldsymbol{B}} = \sqrt{\boldsymbol{B}}\,\sqrt{\boldsymbol{A}}$, also Behauptung (10.16).

Wie wenden (10.16) an auf $\boldsymbol{A}_1, \boldsymbol{A}_2$ und erhalten

$$\begin{aligned} (\boldsymbol{A}_1\,\boldsymbol{A}_2\,\boldsymbol{y}, \boldsymbol{y}) &= \Big(\boldsymbol{A}_1^{1/2}\,\boldsymbol{A}_1^{1/2}\,\boldsymbol{A}_2^{1/2}\,\boldsymbol{A}_2^{1/2}\,\boldsymbol{y}, \boldsymbol{y}\Big) = \Big(\boldsymbol{A}_1^{1/2}\,\boldsymbol{A}_2^{1/2}\,\boldsymbol{A}_1^{1/2}\,\boldsymbol{A}_2^{1/2}\,\boldsymbol{y}, \boldsymbol{y}\Big) = \\ &= \Big(\boldsymbol{A}_1^{1/2}\,\boldsymbol{A}_2^{1/2}\,\boldsymbol{y}, \boldsymbol{A}_2^{1/2}\,\boldsymbol{A}_1^{1/2}\boldsymbol{y}\Big) = \Big(\boldsymbol{A}_1^{1/2}\,\boldsymbol{A}_2^{1/2}\,\boldsymbol{y}, \boldsymbol{A}_1^{1/2}\,\boldsymbol{A}_2^{1/2}\,\boldsymbol{y}\Big) = \\ &= \Big\|\boldsymbol{A}_1^{1/2}\,\boldsymbol{A}_2^{1/2}\,\boldsymbol{y}\Big\| > 0 \quad \forall \boldsymbol{y} \neq 0\,. \end{aligned}$$

$\square$

Zusammenfassend gilt

Satz 10.2 Stabilität des ADI-Verfahrens (10.9)
$\boldsymbol{A}_i = \boldsymbol{A}_i^T > 0$ seien die Diskretisierungsmatrizen von $L_i\, u = \frac{\partial}{\partial x_i}\left(k_i \frac{\partial u}{\partial x_i}\right) + q_i$, $i = 1,2$ mit $k_i(\boldsymbol{x}) = k_i(x_i)$, $q_i(\boldsymbol{x}) = q_i(x_i)$ **in einem Rechteck**.

Dann ist das ADI-Verfahren (10.9) bzw. (10.10) stabil bzgl. Anfangswerten und rechter Seite (Satz 4.2und $\varepsilon \le 2$)

$$\|\boldsymbol{y}^j\|_{(1,h)} \le \|\boldsymbol{y}^0\|_{(1,h)} + \frac{1}{\sqrt{\varepsilon}}\left(\sum_{k=0}^{j} \tau \|\boldsymbol{\varphi}^k\|^2_{(0,h)}\right)^{1/2} \quad \text{mit } \boldsymbol{\varphi} = \frac{\tau}{4}\boldsymbol{A}_1(\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2) + \frac{1}{2}(\boldsymbol{\varphi}_1 + \boldsymbol{\varphi}_2)$$

Konvergenz der mehrdimensionalen Verfahren

a) das Verfahren (10.1)

Sind die $\boldsymbol{A}_i$ die Diskretisierungsmatrizen 2. Ordnung der L_i $(i = 1,2)$, dann lautet die Diskretisierung der Aufgabe (10.1) (vgl. (4.6), (4.7) und (10.4), (10.6)) für ein geeignetes φ:

$$\underbrace{(\boldsymbol{I} + \sigma\,\tau(\boldsymbol{A}_1 + \boldsymbol{A}_2))}_{\boldsymbol{B}}\, \boldsymbol{y}_t + \underbrace{(\boldsymbol{A}_1 + \boldsymbol{A}_2)}_{\boldsymbol{A}}\, \boldsymbol{y} = \boldsymbol{I}\, \boldsymbol{y}_t + \boldsymbol{A}\, \boldsymbol{y}^{(\sigma)} = \boldsymbol{\varphi} \tag{10.18}$$

Für den Verfahrensfehler $\boldsymbol{z} = \boldsymbol{u} - \boldsymbol{y}$ dieser Diskretisierung gilt dann analog zum Beweis Satz 5.2

$$\begin{aligned} \boldsymbol{z}_t + \boldsymbol{A}\boldsymbol{z}^{(\sigma)} &= \boldsymbol{u}_t + \boldsymbol{A}\,\boldsymbol{u}^{(\sigma)} - \underbrace{\left(\boldsymbol{y}_t + \boldsymbol{A}\,\boldsymbol{y}^{(\sigma)}\right)}_{=\boldsymbol{\varphi}} = \boldsymbol{u}_t + \boldsymbol{A}\,\boldsymbol{u}^{(\sigma)} - \boldsymbol{\varphi} \\ &= \underbrace{\left(\frac{\partial u}{\partial t} - L\,u - f\right)^{j+\frac{1}{2}}}_{=0} + \mathrm{O}\left(\tau\left(\sigma - \frac{1}{2}\right) + \tau^2 + h_1^2 + h_2^2\right) \quad \text{für } \boldsymbol{\varphi} = \boldsymbol{f}^{j+\frac{1}{2}}. \end{aligned}$$

vgl. die Abschätzungen zu Satz 5.2 wobei $\boldsymbol{A}_i$ geeignete Diskretisierungsmatrizen 2. Ordnung sind. Hieraus folgt die Konvergenz von der Ordnung des Diskretisierungsfehlers (vgl. Satz 5.4.)
Wählt man $h = h_1 = h_2$, $\sigma = \frac{1-h^2}{12}$ und $\boldsymbol{\varphi} = (\boldsymbol{I} + \sum\limits_{i=1}^{2} \frac{h^2}{12} L_i)\boldsymbol{f}$, so erhält man analog zu Satz 5.4 eine Konvergenz der Ordnung $O(\tau^2 + h^4)$.

b) das ADI-Verfahren

Für den Verfahrensfehler des Verfahrens (10.12) folgt mit

$$\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2, \quad \widetilde{\boldsymbol{B}} = \boldsymbol{I} + \sigma, \quad \tau\boldsymbol{A} + \tau^2\sigma^2\boldsymbol{A}_1\boldsymbol{A}_2, \quad \boldsymbol{B} = \boldsymbol{I} + \sigma\,\tau\boldsymbol{A}$$

unter Verwendung der Abschätzung aus Teil a):

$$\begin{aligned}\widetilde{\boldsymbol{B}}\boldsymbol{z} + \boldsymbol{A}\boldsymbol{z} &= \boldsymbol{B}\boldsymbol{z} + \boldsymbol{A}\boldsymbol{z} + \tau^2\sigma^2\,\boldsymbol{A}_1\,\boldsymbol{A}_2\boldsymbol{z}\\ &= \boldsymbol{B}\,\boldsymbol{u}_t + \boldsymbol{A}\,\boldsymbol{u} + \tau^2\sigma^2\,\boldsymbol{A}_1\,\boldsymbol{A}_2\,\boldsymbol{u} - \underbrace{\left(\boldsymbol{B}\,\boldsymbol{y}_t + \tau^2\sigma^2\,\boldsymbol{A}_1\,\boldsymbol{A}_2\,\boldsymbol{y}_t + \boldsymbol{A}\,\boldsymbol{y}\right)}_{\boldsymbol{\varphi}}\\ &= \underbrace{\left(\frac{\partial u}{\partial t} - L\,u - f\right)^{j+\frac{1}{2}}}_{=0} + \tau^2\sigma^2\,\boldsymbol{A}_1\,\boldsymbol{A}_2\,\boldsymbol{u}_t + O\left(\tau(\sigma - \frac{1}{2}) + \tau^2 + h_1^2 + h_2^2\right)\end{aligned}$$

Beachte: Das $\boldsymbol{\varphi}$ ist dasselbe wie in (10.8).
Ist $A_1\,\boldsymbol{A}_2\,\boldsymbol{u}_t$ beschränkt (das kann man zeigern, wenn $u \in C^6$, beachte: $\boldsymbol{A}_1\,\boldsymbol{A}_2\,\boldsymbol{u}_t$ ist Diskretisierung einer Ableitung 5. Ordnung), so erhält man über den Stabilitätssatz dieselbe Stabilitäts- und Konvergenzordnung wie für (10.8).
Diese Überlegungen müssen (und können) für $h_1 \neq h_2$ etwas modifiziert werden. Insgesamt liefert das Vorgehen von Satz 5.2 dann

Satz 10.3 Konvergenz der ADI-Methode (Peachman/Rachford)
Für die Aufgabe

$$\begin{aligned}\frac{\partial u}{\partial t} &= \sum_{i=1}^{2} L_i\,u + f \quad \text{in} \quad \Omega = (0,a_1)\times(0,a_2) \subset \mathbb{R}^2 \qquad \text{mit}\\ L_i\,u &= \frac{\partial}{\partial x_i}\left(k_i\,\frac{\partial}{\partial x_i}\,u\right) + q_i\,u, \quad k_i(\boldsymbol{x}) = k_i(x_i),\ \ q_i(\boldsymbol{x}) = q_i(x_i)\end{aligned}$$

seien $\boldsymbol{A}_i$ geeignete symmetrische, positiv definite Diskretisierungsmatrizen 2. Ordnung für L_i, $i = 1,2$ (vgl. dazu z.B. (7.2) und § 8, 1. Abschnitt). Dann gilt:

Das ADI-Verfahren

$$\begin{aligned}\frac{\boldsymbol{y}^{j+\frac{1}{2}} - \boldsymbol{y}^j}{\tau\,\sigma} + \boldsymbol{A}_1\,\boldsymbol{y}^{j+\frac{1}{2}} + \boldsymbol{A}_2\,\boldsymbol{y}^j &= \boldsymbol{\varphi}\\ \frac{\boldsymbol{y}^{j+1} - \boldsymbol{y}^{j+\frac{1}{2}}}{\tau\,\sigma} + \boldsymbol{A}_1\,\boldsymbol{y}^{j+\frac{1}{2}} + \boldsymbol{A}_2\,\boldsymbol{y}^{j+\frac{1}{2}} &= \boldsymbol{\varphi}\end{aligned}$$

konvergiert von der Ordnung

$O\left(\tau\left(\sigma - \frac{1}{2}\right) + \tau^2 + h_1^2 + h_2^2\right)$ falls $\boldsymbol{\varphi} = \boldsymbol{f}^{j+\frac{1}{2}}, u \in C^{3,4}(\Omega)$ (3 bzgl. Zeit, 4 bzgl. Ort)

$O(\tau^2 + h^4)$ für $h = h_1 = h_2, \sigma = \frac{1}{12} - \frac{h^2}{12\tau}$, $\boldsymbol{\varphi} = \left(\boldsymbol{I} + \sum_{i=1}^{2} \frac{h^2}{12}\,L_i\right)\boldsymbol{f}$ und $u \in C^{3,6}(\Omega)$

$O(\tau^2 + h^6)$ für noch kompliziertere Wahlen von σ und $\boldsymbol{\varphi}$ und hinreichend hohe Differenzierbarkeitsordnung.

Bemerkungen

1. Damit man im 2. Schritt von (10.8) das Tridiagonalverfahren anwenden kann, muß zuerst das Gitter (und damit auch $\boldsymbol{A}_1$) spaltenweise umnummeriert werden.

2. Die Verallgemeinerung von (10.11) auf den nD-Fall lautet:

$$\prod_{i=1}^{n}\left(\boldsymbol{I}-\frac{\tau}{2}\boldsymbol{A}_i\right)\boldsymbol{y}_t+\sum_{i=1}^{n}\boldsymbol{A}_i\,\boldsymbol{y}=\boldsymbol{\varphi}.$$

   Stabilität und Konvergenz müssen neu untersucht werden (Arbeiten von Janenko, Fairwather, Mitchel, Dyakonov - Mitte der 60er Jahre).

3. Unter mehreren Varianten des Verfahrens erwähnen wir eine von Janenko:

$$(\boldsymbol{I}+\sigma\,\tau\,\boldsymbol{A}_i)\,\boldsymbol{y}^{j+\frac{i}{n}}=(\boldsymbol{I}-(1-\sigma)\,\tau\,\boldsymbol{A}_i)\,\boldsymbol{y}^{j+\frac{i-1}{n}}+\underset{\substack{\uparrow\\ \text{Kronecker}}}{\delta_{i1}}\;\boldsymbol{\varphi},\quad i=1,\ldots,n,$$

   die auch mit tridiagonalen Matrizen arbeitet (wie auch weitere Varianten von Stoyan und Dyakonov). Bemerkenswert ist, dass die Zwischenschritte dieser Verfahren die ursprüngliche Aufgabe nicht mehr approximieren. Die Approximiation kommt erst nach Elimination der Zwischenschritte zustande.

4. Man kann die Stabilität und Konvergenz der ADI-Verfahren auch für nicht vertauschbare Matrizen und andere als Rechtecksgebiete zeigen, muß dann aber Einbußen in der Ordnung in Kauf nehmen.

5. Der Stabilitätssatz 10.2 läßt sich auch mit Hilfe einer diskreten Fourier-Analyse beweisen (vgl. hierzu die einfache Demonstration des Verfahrens zum Beweis von (4.2)).

6. ADI-Verfahren lassen sich auch zur Behandlung elliptischer Probleme anwenden.

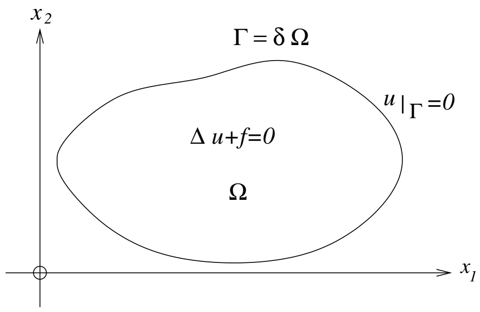
7. Literatur und einige Varianten der ADI-Verfahren (für Rechtecksgebiete) findet man in Morton/Mayers: Num. Sol. of PDEs, Cambridge Univ. Press, Abschnitt 3.2 f.

Kapitel II

Elliptische Gleichungen

§ 11 Die Poissongleichung - Einleitung

Wir stellen 2 einfache Motivationsbeispiele für die Poissonaufgabe vor:



Ein einfaches Beispiel liefert eine stationäre Flüssigkeitsströmung.

Ist $\boldsymbol{u} = (u, v)^T$ der Strömungsvektor, so bedeuten

$$-\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = 0 \quad \text{die Wirbelfreiheit}$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \quad \text{das Nichtvorhandensein von Massenquellen (Massenerhaltung)}$$

Beide Gleichungen zusammen bilden das Cauchy-Riemann'sche System (vgl. Ableitungsbedingungen für Real- und Imaginärteil komplexwertiger Funktionen).

Wird die 1. Gleichung nach x, die zweite nach y differenziert und danach die Differenz gebildet, so erhält man

$$\Delta\, v = 0$$

Vertauschung von x und y liefert $\Delta\, u = 0$.

Kompliziertere Anwendungen folgen aus den Navier-Stokes-Gleichungen (wir beschränken uns auf den 2-dimensionalen Fall). Anwendungen z.B. im Umweltschutz, z.B. Chemieunfall.

Seien $\boldsymbol{u} = (u, v)^T$ der Strömungsvektor, p der Druck, ν die kinematische Viskosität der Flüssigkeit (sie ist bei Gasen klein und kann bei Flüssigkeiten sehr groß sein) und f_i einwirkende Kräfte, so stellen die ersten beiden der folgenden Gleichungen die Impulserhaltung für die 1. bzw. 2. Komponente des Geschwindigkeitsvektors dar, die 3. Gleichung beschreibt die Massenerhaltung.

$$\begin{aligned}
\frac{\partial u}{\partial t} + u\,\frac{\partial u}{\partial x} \quad &+ \quad v\,\frac{\partial u}{\partial y} + \frac{\partial p}{\partial x} &&= \quad \nu\,\Delta\,u + f_1 \\
\frac{\partial v}{\partial t} + u\,\frac{\partial v}{\partial x} \quad &+ \quad v\,\frac{\partial v}{\partial y} + \frac{\partial p}{\partial y} &&= \quad \nu\,\Delta\,v + f_2 \\
\frac{\partial u}{\partial x} \quad &+ \quad \frac{\partial v}{\partial y} &&= \quad 0\,.
\end{aligned}$$

Vereinfachungsmöglichkeiten:

1. Bei kleinen Geschwindigkeitsänderungen $\left(\frac{\partial u}{\partial t}, \frac{\partial v}{\partial t} = 0\right)$ und bei kleinen Geschwindigkeiten ($\Rightarrow$ Vernachlässigung der nicht linearen Terme) erhält man eine Poissongleichng für die Geschwindigkeitskomponenten.

2. Daß auch die einfache Poissongleichung für die Anwendung interessant sein kann, zeigt folgende mathematische Umrechnung der Navier-Stokes-Gleichungen: Differenziert man (bei konstantem ν) die 1. Gleichung partiell nach x, die zweite nach y, vertauscht die Differentiationsreihenfolge und addiert die ersten beiden Gleichungen, so erhält man

$$\begin{array}{cccccccc}
\underbrace{\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)}_{\substack{=0 \\ \text{3.Gleichung}}} & + \left(\frac{\partial u}{\partial x}\right)^2 & \boxed{\begin{array}{c} + u\,\dfrac{\partial^2 u}{\partial x^2} \\[2ex] + u\,\dfrac{\partial^2 v}{\partial x\,\partial y} \end{array}} & + \frac{\partial v}{\partial x}\,\frac{\partial u}{\partial y} & \boxed{\begin{array}{c} + v\,\dfrac{\partial^2 u}{\partial x\,\partial y} \\[2ex] + v\,\dfrac{\partial^2 v}{\partial y^2} \end{array}} & & & \\
& + \frac{\partial u}{\partial y}\,\frac{\partial v}{\partial x} & & + \left(\frac{\partial v}{\partial y}\right)^2 & & + \Delta p & = & \nu\,\Delta \overbrace{\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)}^{=0} \\
& & = & & = & & & + \frac{\partial}{\partial x} f_1 + \frac{\partial}{\partial y} f_2 \\
& & u\,\frac{\partial}{\partial x}\underbrace{\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)}_{=0} & & v\,\frac{\partial}{\partial y}\underbrace{\left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}\right)}_{=0} & & &
\end{array}$$

also folgt

$$\left(\frac{\partial u}{\partial x}\right)^2 + \frac{\partial v}{\partial x}\,\frac{\partial u}{\partial y} + \frac{\partial u}{\partial y}\,\frac{\partial v}{\partial x} + \left(\frac{\partial v}{\partial y}\right)^2 + \Delta\,p = \frac{\partial f_1}{\partial x} + \frac{\partial f_2}{\partial y}$$

d.h. bei bekannten Geschwindigkeiten erhält man eine Poissongleichung für den Druck.

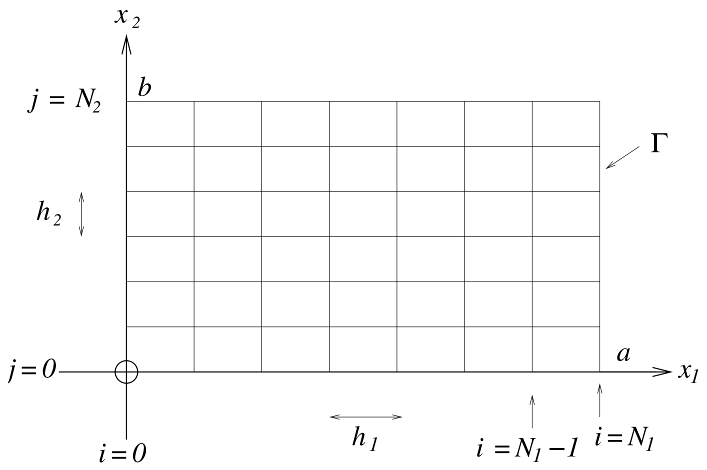
§ 12 Die erste RWA für die Poissongleichung im Rechteck

Für die Aufgabe

$$(12.1)\qquad \begin{array}{ll} \Delta u + f = 0 \ \text{ in } \ \Omega = (0,a)\times(0,b) & \text{(Poissongleichung)} \\ u = g \ \text{ auf } \Gamma, \quad \Gamma = \delta\Omega & \text{(Dirichletwerte)} \end{array}$$

haben wir die einfachste Diskretisierung schon im § 10 beschrieben.

Wir wiederholen die Bezeichnungen



$$\begin{array}{lcl} h_1 & = & \frac{a}{N_1}, \quad h_2 = \frac{b}{N_2}, \quad h = (h_1, h_2) \\ \omega_h & = & \text{Menge der inneren Gitterpunkte} \\ \gamma_h & = & \text{Menge der Randpunkte} \\ \overline{\omega_h} = \omega_h \cup \gamma_h & = & \text{Menge aller Gitterpunkte} \end{array}$$

Wir benutzen folgende **Abkürzungen**:

$\|\boldsymbol{f}\|_{C(\omega_h)} = \max\limits_{x\in\omega_h} |\boldsymbol{f}(x)|$ Maximumnorm der von f erzeugten Gitterfunktion $\boldsymbol{f}$ auf den inneren Gitterpunkten des Gitters ω_h mit den Maschenweiten $h = (h_1, h_2)$.

$\|\boldsymbol{f}\|_{C(\overline{\omega}_h)}$, $\|\boldsymbol{g}\|_{C(\gamma_h)}$ entsprechend.

y_{ij}, $(ij) \in \omega_h$ Doppelindizierung der Punkte zu ihrer Festlegung im Gitter und „ $(ij) \in \omega_h$ “ (z.B.) als Zugehörigkeit zu den inneren Gitterpunkten.

w_{ij} als Komponente einer durch eine Funkion $\boldsymbol{w}$ erzeugte Gitterfunktion.

Zur Aufstellung der Diskretisierungsmatrizen werden die Gitterpunkte zeilenweise durchnummeriert, beginnend mit $j = 0$, $i = 0, \dots, N_1$, $j = 1$, $i = 0, \dots, N_1$ usw. Für die Näherungswerte der exakten Lösung u ist als Bezeichnung üblich

$$u(x_i, x_j) \approx y_{ij}, \quad \text{wobei } x_i = i\, h_1, \;\; x_j = j\, h_2 \,.$$

Die Diskretisierung von (12.1) in inneren Gitterpunkten lautet

$$\begin{array}{lcccccc} y_{x_1 x_1, ij} & + & y_{\overline{x}_2 x_2, ij} & + & f_{ij} & = & 0 \\ \frac{1}{h_1^2}\left(y_{i-1,j} - 2y_{ij} + y_{i+1,j}\right) & + & \frac{1}{h_2^2}\left(y_{i,j-1} - 2y_{ij} + y_{i,+1j}\right) & + & f_{ij} & = & 0 \end{array} \tag{12.2}$$

wobei für die Ableitungen gemäß (2.5) gilt (bzgl. jeder Variablen in jedem Punkt)

$$u_{\overline{x}x} = \begin{cases} u'' + \frac{h^2}{24}\left(u_+^{(4')} + u_-^{(4')}\right) & \text{falls } u \in C^4 \\ u'' + \frac{h^2}{12}\, u^{(4')} + \frac{h^4}{720}\left(u_+^{(6')} + u_-^{(6')}\right) & \text{falls } u \in C^6 \end{cases} \tag{12.3}$$

Die Restglieder kann man wie folgt abschätzen: Ist $u \in C^4$, also stetig, so folgt aus dem Zwischensatz: $\exists\, \xi \in [\tilde{x}_1, \tilde{x}_2]$: $u^{(4')}(\tilde{x}_1) + u^{(4')}(\tilde{x}_2) = 2u^{(4)}(\xi)$.

Damit erhält man

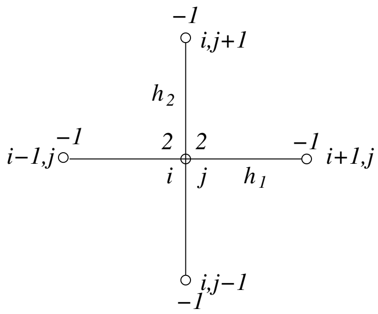
$$\left| \frac{h^2}{24}\left(u_+^{(4')} + u_-^{(4')}\right) \right| \le \frac{h^2}{12} M_4, \quad M_4 = \max_x \left| u^{(4')}(x) \right| . \tag{12.4}$$

Analog kann man im Fall $u \in C^6$ verfahren.

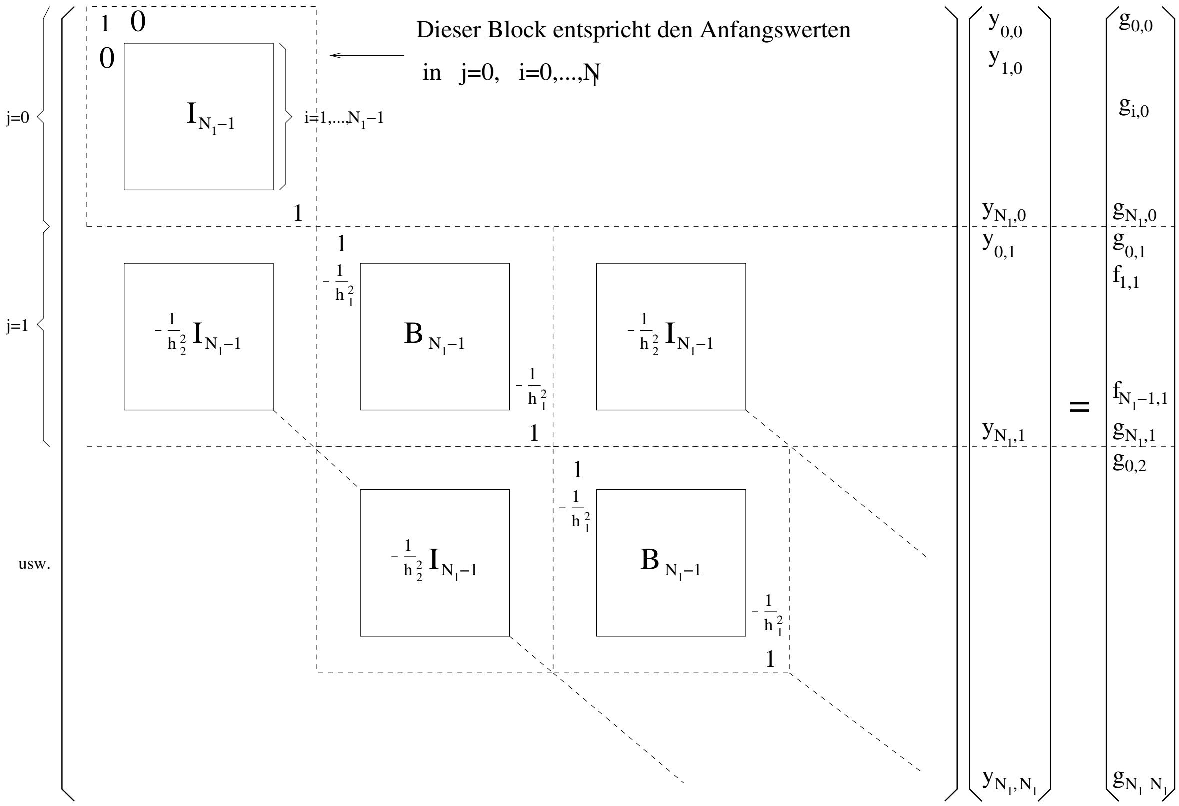
Wir wollen die Differenzengleichungen für <u>alle</u> Gitterpunkte in Matrixform aufschreiben unter Benutzung des 5-Punkte-Sterns (vgl.(12.2)):

$$\boldsymbol{A}_h\, \boldsymbol{y} = \boldsymbol{\varphi} \tag{12.5}$$

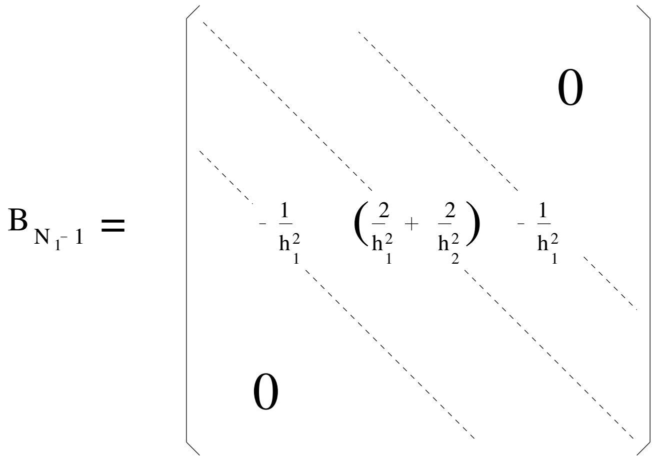
Beachte die Vorzeichen: In der Differentialgleichung stehen Δu und f auf derselben Seite. In der Matrixdarstellung stehen sie auf verschiedenen Seiten.



Die Matrix A_h und das Gleichungssystem (12.5) haben folgende Gestalt:
(Für eine geeignete Implementierung vgl. Hackbusch S. 90 f.)

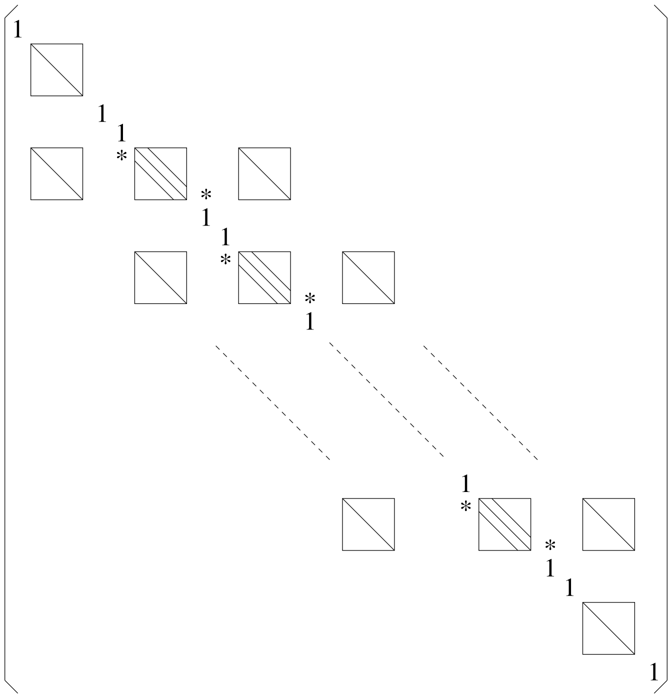


$g_{i,j}$ sind Randwerte $f_{i,j}$ Werte der rechten Seite der Differentialgleichung



Insgesamt ist A_h eine tridiagonale Blockmatrix folgender Struktur:

(12.5') $A_h =$



„ $*$ “ bedeuten die Übergangselemente am linken oder rechten Rand des Gebiets. Alle Blöcke haben die Dimension $(N_1 - 1)^2 \times (N_1 - 1)^2$ und die einzelnen Blöcke sind Diagonal- oder Tridiagonalmatrizen.

In der Gesamtmatrix stehen in der Hauptdiagonalen nur positive Einträge, alle anderen Elemente sind ≤ 0. Matrizen mit dieser Einteilung heißen **L-Matrizen**.

Allerdings ist die Matrix nicht mehr streng diagonal dominant. In den Zeilen, die einen der Blöcke $\boldsymbol{B}_{N_1-1}$ betreffen, ist die Zeilensumme =0.

Wir beweisen die Konvergenz des Verfahrens (12.5) zur Lösung der Aufgabe (12.1). Dazu zeigen wir, dass $\boldsymbol{A}_h$ eine M-Matrix ist und benutzen Satz 6.3 zur Abschätzung.

Konvergenz und Apriori Schranke für die Lösung von $\boldsymbol{A}_h\,\boldsymbol{y} = \boldsymbol{\varphi}$.

Satz 12.1
Für die Lösung u von (12.1)

$$\Delta u + f = 0 \text{ in } \Omega = (0,a)\times(a,b)\subset\mathbb{R}^2, \quad u\big|_{\partial\Omega} = g,$$

betrachten wir das Verfahren $\boldsymbol{A}_h\boldsymbol{y} = \boldsymbol{\varphi}$, (12.5), mit

$$y_{\overline{x}_1\,x_{1,ij}} + y_{\overline{x}_1\,x_{2,ij}} + f_{ij} = 0, \;\; (i,j)\in w_h$$

Dann gilt:

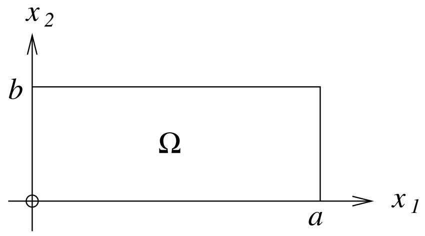
a) Die Diskretisierungsmatrix $\boldsymbol{A}_h$ aus (12.5) ist eine M-Matrix.

b) Das Verfahren konvergiert quadratisch ($O(h_1^2 + h_2^2)$)

c) Es gilt die Apriori Abschätzung

$$\text{(12.6)} \qquad \|\boldsymbol{y}\|_{C(\overline{w}_h)} \le \left(1 + \frac{a^2+b^2}{16}\right)\max\left(\|\boldsymbol{g}\|_{C(\gamma_h)}, \|\boldsymbol{f}\|_{C(\overline{w}_h)}\right)$$

Beweis a)
Die Vorzeichenverteilung einer M-Matrix wurde für $\boldsymbol{A}_h$ schon gezeigt. Zur Konstruktion eines Vektors $\boldsymbol{w} > \boldsymbol{0}$ mit $\boldsymbol{A}_h\,\boldsymbol{w} > \boldsymbol{0}$ (elementweise vgl. Definition 6.2) machen wir den Ansatz



$$w(x_1,x_2) = 4 + x_1(a-x_1) + x_2(b-x_2) \text{ auf } \overline{\Omega}\,.$$

Offensichtlich gilt: $w \ge 4$ in $[0,a]\times[0,b]$.

Nun ist

$$\text{(12.7)} \qquad (\boldsymbol{A}_h\,\boldsymbol{w})_{ij} = \begin{cases} w_{ij}, \;\; (i,j)\in\gamma_h & \text{(Randwerte)} \\ -\left(w_{\overline{x}_1\,x_1} + w_{\overline{x}_2\,x_2}\right)_{ij}, \;\; (i,j)\in w_h & \text{(innere Punkte)} \end{cases}$$

Gemäß (12.3): $u_{\overline{x}_1\,x_1} = \dfrac{\partial^2}{\partial x_1^2}u + \dfrac{h^2}{12}\dfrac{\partial^4}{\partial x_1^4}u\,(x_1 + \theta\,h_1, x_2)$, entsprechend für x_2, gilt für die quadratische Funktion w:

$$w_{\overline{x}_1\,x_1} = \frac{\partial^2}{\partial x_1^2}w, \;\; w_{\overline{x}_2\,x_2} = \frac{\partial^2}{\partial x_2^2}w, \qquad \text{da } \frac{\partial^4}{\partial x_1^4}w = \frac{\partial^4}{\partial x_4^2}w = 0.$$

Weiter ist $w_{\bar{x}_1 x_1} + w_{\overline{x}_2 x_2} = -4$, somit folgt aus (12.7)

$$(\boldsymbol{A}_h \, \boldsymbol{w})_{ij} \begin{cases} \geq 4 & (i,j) \in \gamma_h, \\ = 4 & (i,j) \in \omega_h \end{cases} \tag{12.8}$$

Bezeichnet $\boldsymbol{w}$ die durch w erzeugte Gitterfunktion, so gilt $\boldsymbol{w} \geq 4\boldsymbol{e}$, $\boldsymbol{e} = (1, 1, \ldots, 1)^T$, und nach (12.8) $(\boldsymbol{A}_h \, \boldsymbol{w})_{ij} \geq 4 > 0$, bzw. $\boldsymbol{A}_h \, \boldsymbol{w} > \boldsymbol{0}$ (elementweise).

$\Longrightarrow$ $\boldsymbol{A}_h$ ist eine M-Matrix

Beweis b)

Es ist
$$\begin{aligned} \|\boldsymbol{w}\|_{C(\overline{\omega})} &= 4 + \frac{a^2 + b^2}{4} && \text{(Ableiten, Maximum berechnen)} \\ \min_i \, (\boldsymbol{A} \, \boldsymbol{w})_i &= 4 && \text{laut (12.8)} \end{aligned}$$

Deshalb liefert Satz 6.3

$$\|\boldsymbol{A}_h^{-1}\|_\infty \leq \frac{\|\boldsymbol{w}\|_\infty}{\min\limits_i \, (\boldsymbol{A}\boldsymbol{w})_i} \leq \frac{4 + \frac{a^2+b^2}{4}}{4} = 1 + \frac{a^2 + b^2}{16} \, .$$

Der Verfahrensfehler $\boldsymbol{z} = \boldsymbol{u} - \boldsymbol{y}$ hat Nullrandwerte. Man kann das System (12.5) also beschränken auf die Matrix $\boldsymbol{A}_h^0$, die man aus $\boldsymbol{A}_h$ durch Streichen der Randterme erhält. $\boldsymbol{A}_h^0$ wirkt dann nur auf den Gittervektor $\mathring{\boldsymbol{y}}$ der inneren Punkte. Der Verfahrensfehler genügt dann dem Differenzenschema (mit $\boldsymbol{f}^0 = \boldsymbol{f}\big|_{\omega_h}$)

$$\boldsymbol{A}_h^0 \, \boldsymbol{z} = \boldsymbol{A}_h^0 \, \boldsymbol{u} - \underbrace{\boldsymbol{A}_h^0 \, \boldsymbol{y}}_{=\boldsymbol{f}^0} = \boldsymbol{u}_{\overline{x}_1 x_1} + \boldsymbol{u}_{\overline{x}_2 x_2} - \boldsymbol{f}^0 = (\underbrace{\Delta u - f}_{=0}) + \mathcal{O}(h_1^2 + h_2^2) =: \boldsymbol{\psi}$$

Aus $\boldsymbol{A}_h \, \boldsymbol{z} = \boldsymbol{\psi}$ ($\boldsymbol{\psi}, \boldsymbol{z}$ sind hier um die Nullkomponenten der Randterme angereichert) folgt somit (siehe oben)

$$\|\boldsymbol{z}\|_\infty \leq \left(1 + \frac{a^2 + b^2}{16}\right) O(h_1^2 + h_2^2).$$

Beweis c)
Aus $\boldsymbol{A}_h \, \boldsymbol{y} = \boldsymbol{\varphi}$, $\boldsymbol{\varphi}$ gemäß (12.5), erhält man ebenso

$$\begin{aligned} \|\boldsymbol{y}\|_{C(\overline{\omega}_h)} &\leq \|\boldsymbol{A}_h^{-1}\|_\infty \max(\|\boldsymbol{g}\|_{C(\gamma_h)}, \|\boldsymbol{f}\|_{C(\overline{\omega}_h)}) \\ &\leq \left(1 + \frac{a^2 + b^2}{16}\right) \max(\|\boldsymbol{g}\|_{C(\gamma_h)}, \|\boldsymbol{f}\|_{C(\overline{\omega}_h)}). \end{aligned}$$ ■

Bemerkung
Diese Abschätzung bedeutet auch die Stabilität für das Verfahren (12.5), denn die Konstante auf der rechten Seite, also die Abschätzung von $\|\boldsymbol{A}_h^{-1}\|_\infty$, ist unabhängig von der Diskretisierung.

Schärfere Abschätzungen kann man mit Hilfe des diskreten Maximumprinzips erhalten, das wir nun beweisen.

Das diskrete Maximumprinzip

Definition 12.2 Maximumprinzip für Matrizen
Für eine Matrix $\boldsymbol{A}$ und einen Vektor $\boldsymbol{y} \in \mathbb{R}^n$ definieren wir

$$N^0(\boldsymbol{Ay}) = \{i;\ (\boldsymbol{Ay})_i = 0,\ i \in \{1, ..., n\}\}$$
$$N^{\neq}(\boldsymbol{Ay}) = \{i;\ (\boldsymbol{Ay})_i \neq 0,\ i \in \{1, ..., n\}\}$$

$\boldsymbol{A}$ genügt dem (strengen) Maximumprinzip, falls gilt

a) $N^{\neq}(\boldsymbol{Ay}) = \emptyset \quad \Longrightarrow \quad \boldsymbol{y} = \boldsymbol{0}\ \forall\, \boldsymbol{y} \in \mathbb{R}^n$ (d.h. $\boldsymbol{A}$ ist regulär)

b) $\max\limits_{i \in N^0(\boldsymbol{Ay})} |y_i| <$ (bzw. $\leq$) $\max\limits_{j \in N^{\neq}(\boldsymbol{Ay})} |y_j|\ \ \forall\, \boldsymbol{y} \in \mathbb{R}^n,\ \boldsymbol{y} \neq 0$

d.h. in Worten: Dort wo die rechte Seite $\neq 0$ ist, wird das Maximum angenommen. Das Maximum der restlichen Komponenten ist kleiner oder (im nichtstrengen Fall) höchsten gleich groß.

Satz 12.3
Das strenge Maximumprinzip gilt genau dann, wenn die Hauptdiagonale in $\boldsymbol{A}$ streng überwiegt (streng diagonaldominant), d.h.

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| \ \ \forall\, i$$

Beweis „$\Longleftarrow$"

a) Wir zeigen zuerst $\exists\, \boldsymbol{A}^{-1}$.
Aus der Diagonaldomonaz folgt $a_{ii} \neq 0; \forall\, i$

$\Longrightarrow$ $\boldsymbol{D} = \operatorname{diag}(a_{11}, ..., a_{nn})$ ist invertierbar.

$\Longrightarrow$ $\boldsymbol{D}^{-1}\boldsymbol{A}$ hat in der Hauptdiagonalen nur Einsen.

$\Longrightarrow$ $\boldsymbol{B} := (b_{ij}) = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A}$ hat in der Hauptdiagonalen nur Nullen und

$$b_{ij} = -\frac{a_{ij}}{a_{ii}} \ \ \forall\, j \neq i.$$

Für die Zeilensummennorm $\|\boldsymbol{B}\|_\infty$ folgt aus der Diagonaldominanz

$$\|\boldsymbol{B}\|_\infty < 1.$$

Nach Satz 3.1 (Neumansche Reihe) $\exists\, (\boldsymbol{I} - \boldsymbol{B})^{-1} = (\boldsymbol{D}^{-1}\boldsymbol{A})^{-1}$

$\Longrightarrow$ $\boldsymbol{A}$ ist invertierbar, also a).

Beweisvariante:
In keinem Gershgorinkreis von $\boldsymbol{A}$ ist die Null enthalten auf Grund der starken Diagonaldominanz, also sind alle Eigenwerte $\lambda(\boldsymbol{A}) \neq 0 \implies \exists\, \boldsymbol{A}^{-1}$.

b) Sei $\boldsymbol{y}$ der Lösungsvektor von $\boldsymbol{Ax} = \boldsymbol{b} \neq \boldsymbol{0}$, $\implies \exists\, i: \; y_i \neq 0$ und $|y_i| = \|\boldsymbol{y}\|_\infty$. (der Lösungsvektor ist eindeutig bestimmt, da $\boldsymbol{A}$ regulär ist nach a).

Wir zeigen indirekt: $i \in N^{\neq}(\boldsymbol{Ay})$.

Annahme: $i \in N^0(\boldsymbol{Ay})$, d.h. $(\boldsymbol{Ay})_i = 0$ d.h.

$$a_{ii}y_i + \sum_{j=1,\, j\neq i}^{n} a_{ij}y_j = 0 \quad \implies$$

$$\begin{aligned}
0 =& |a_{ii}\, y_i + \sum_{j=1,\, j\neq i}^{n} a_{ij}y_j| \\
\geq& |a_{ii}y_i| - \sum_{j=1,\, j\neq i}^{n} |a_{ij}|\,|y_j| \quad \text{und wegen} \quad |y_i| = \|\boldsymbol{y}\|_\infty \\
\geq& \underbrace{|y_i|}_{>0} \underbrace{(|a_{ii}| - \sum_{j=1\, j\neq i}^{n} |a_{i\,j}|)}_{>0} \quad \implies \quad \text{W!}
\end{aligned}$$

$|y_i|$ nimmt also sein Maximum nicht für $i \in N^{\neq}(\boldsymbol{Ay})$ an, d.h. b).

„$\implies$"

Wir zeigen zuerst $a_{ii} \neq 0 \;\forall\, i$.

Wende das Maximumprinzip an auf $\boldsymbol{y} = \boldsymbol{e}^i$ ($i-$ ter Einheitsvektor).

$\boldsymbol{e}^i \neq \boldsymbol{0}$ und da nach a) $\exists\, \boldsymbol{A}^{-1} \implies N^{\neq}(\boldsymbol{Ae}^i) \neq \emptyset$, d.h. $\boldsymbol{Ae}^i = \boldsymbol{a}^i \neq \boldsymbol{0}$ ($\boldsymbol{a}^i = i-$te Spalte von $\boldsymbol{A}$).

Wegen $e^i_i = 1,\; e^i_j = 0$ für $i \neq j$, also $\max\limits_j |e^i_j| = e^i_i = 1$, muß nach b) gelten:

$$i \in N^{\neq}(\boldsymbol{Ae}^i)\,, \text{ d.h. } (\boldsymbol{Ae}^i)_i = a_{ii} \neq 0.$$

Wir zeigen die strenge Diagonaldominanz.

Für beliebiges, aber festes i definiere

$$y_i := -\sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|a_{ij}|}{a_{ii}}, \qquad y_j := \overline{\operatorname{sgn}}\, a_{ij} \text{ für } i \neq j, \text{ wo } \overline{\operatorname{sgn}}\,(t) = \begin{cases} 1 \text{ für } t \geq 0 \\ -1 \text{ für } t < 0 \end{cases}$$

$$\implies \quad \boldsymbol{y} \neq 0,$$

denn selbst wenn alle $a_{ij} = 0$ wären, folgt das aus der Definition von $\overline{\operatorname{sgn}}$.

Nun ist

$$(\boldsymbol{Ay})_i = \sum_{j=1}^{n} a_{ij} y_j = \sum_{\substack{j=1\\j\neq i}}^{n} |a_{ij}| + a_{ii}\left(-\sum_{\substack{j=1\\j\neq i}}^{n} \frac{|a_{ij}|}{a_{ii}}\right) = 0,$$

also $i \in N^0(\boldsymbol{Ay})$, d.h. $|y_j|$ nimmt sein Maximum nicht für $j = i$ an. (strenges Maximumprinzip!).
Damit folgt

$$|y_i| = \sum_{\substack{j=1\\j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} < \max_j |y_j| = 1 \quad \text{laut Definition von } y.$$

Das ist die starke Diagonaldominanz. ■

Satz 12.4
$\boldsymbol{A}$ sei invertierbar und schwach diagonaldominant ($|a_{ii}| \geq \sum\limits_{\substack{j=1\\j\neq i}}^{n} |a_{ij}| \ \ \forall\, i$)

Dann gilt das schwache Maximumprinzip.

Beweis:

Eigenschaft a) von Definition 12.2 ist erfüllt, da $\boldsymbol{A}$ invertierbar.

Da $\boldsymbol{A}$ schwach diagonal dominant und invertierbar ist, sind alle $a_{ii} \neq 0$, denn $\boldsymbol{A}$ kann keine Nullzeile haben.
Sei $\boldsymbol{A} = (a_{ij})$, so gilt für ein beliebiges $\varepsilon > 0$ mit $\tilde{\boldsymbol{\varepsilon}} = \varepsilon(\operatorname{sgn}(a_{11}), ..., \operatorname{sgn}(a_{nn}))^T$:

$\boldsymbol{A}_{\tilde{\varepsilon}} := \boldsymbol{A} + \tilde{\boldsymbol{\varepsilon}}\boldsymbol{I}$ ist $\forall\, \varepsilon > 0$ streng diagonaldominant, erfüllt also das strenge Maximumprinzip und ist invertierbar.

Für ein beliebiges $\boldsymbol{y} \neq \boldsymbol{0}$ sei $\boldsymbol{\varphi} := \boldsymbol{Ay}$.
Dadurch sind $N^0(\boldsymbol{Ay})$ und $N^{\neq}(\boldsymbol{Ay})$ festgelegt. Sei $\boldsymbol{y}^{(\tilde{\varepsilon})}$ Lösung von

$$(*) \qquad \boldsymbol{A}_{\tilde{\boldsymbol{\varepsilon}}}\boldsymbol{y}^{(\tilde{\varepsilon})} = \boldsymbol{\varphi}$$

$$\Longrightarrow \quad N^0(\boldsymbol{A}_{\tilde{\varepsilon}}\boldsymbol{y}^{(\tilde{\varepsilon})}) = N^0(\boldsymbol{Ay}), \quad N^{\neq}(\boldsymbol{A}_{\tilde{\varepsilon}}\boldsymbol{y}^{(\tilde{\varepsilon})}) = N^{\neq}(\boldsymbol{Ay}).$$

dann gilt nach dem starken Maximumprinzip für $\boldsymbol{A}_{\varepsilon}$.

$$(**) \qquad \max_{i\in N^0(\boldsymbol{Ay})} |y_i^{(\tilde{\varepsilon})}| < \max_{j\in N^{\neq}(\boldsymbol{Ay})} |y_j^{(\tilde{\varepsilon})}|$$

In $(*)$ kann der Grenzübergang $\varepsilon \to 0$ ausgeführt werden, da $\boldsymbol{A}_{\tilde{\varepsilon}}^{-1}$ existiert für $\varepsilon \geq 0$.
Dann folgt aus $(**)$ die Behauptung b). ■

Bemerkung: Für schwach diagonal dominante M-Matrizen gilt das schwache Maximumprinzip, denn M-Matrizen sind invertierbar.

Satz 12.5 Diskretes Maximumprinzip für die 1. RWA der Poissongl.
Für die Aufgabe

$$\Delta u = f \text{ in } \Omega = (0,a) \times (0,b), \quad u\big|_{\delta\Omega} = g$$

sei durch (12.5) die diskrete Form

$$\boldsymbol{A}_h \boldsymbol{y} = \boldsymbol{\varphi}$$

gegeben. Dann gilt für beliebiges $h = (h_1, h_2)$

$$\text{(12.9)} \qquad \text{falls } f = 0: \quad \|\boldsymbol{y}\|_\infty \le \|\boldsymbol{g}\|_{C(\gamma_h)}$$

$$\text{(12.10)} \qquad \text{falls } g = 0: \quad \|\boldsymbol{y}\|_\infty \le \frac{a^2+b^2}{16}\|\boldsymbol{f}\|_{C(\bar\omega_h)}$$

$$\text{(12.11)} \qquad \text{allgemein}: \quad \|\boldsymbol{y}\|_\infty \le \|\boldsymbol{g}\|_{C(\gamma)_h)} + \frac{a^2+b^2}{16}\|\boldsymbol{f}\|_{C(\bar\omega_h)}$$

Bemerkung: Die Aussage des Satzes kann direkt für parabolische Aufgaben übernommen werden. Wesentlich ist nur, daß $\boldsymbol{A}_h \boldsymbol{y} = \boldsymbol{\varphi}$ eine Diskretisierung der parabolischen Aufgabe ist, welche die Randwerte mit einschließt und über alle Zeitschichten geht..

Beweis des Satzes: Wir zerlegen die allgemeine Aufgabe $\boldsymbol{A}_h \boldsymbol{y} = \boldsymbol{\varphi}$ in zwei Teilaufgaben

1. $\boldsymbol{A}_h \boldsymbol{y}^{(1)} = \boldsymbol{\varphi}_{\gamma_h}, \quad \boldsymbol{\varphi}_{\gamma_h} = \begin{cases} 0 & \text{für } (i,j) \in \omega_h \\ g_{i,j} & \text{für } (i,j) \in \gamma_h \end{cases}$ (Fall (12.9))

2. $\boldsymbol{A}_h \boldsymbol{y}^{(2)} = \boldsymbol{\varphi}_{\omega_h}, \quad \boldsymbol{\varphi}_{\omega_h} = \begin{cases} f_{i,j} & \text{für } (i,j) \in \omega_h \\ 0 & \text{für } (i,j) \in \gamma_h \end{cases}$ (Fall (12.11))

Die Linearität der Aufgabe bewirkt (Superposition)

$$\boldsymbol{A}\boldsymbol{y} = \boldsymbol{A}\boldsymbol{y}^{(1)} + \boldsymbol{A}\boldsymbol{y}^{(2)} = \boldsymbol{\varphi}_{\gamma_h} + \boldsymbol{\varphi}_{\omega_h} = \boldsymbol{\varphi}.$$

Dies liefert (12.11), wenn (12.9) und (12.10) bewiesen sind.

Fall (12.9): $f = 0$
$\boldsymbol{A}_h$ ist eine schwach diagonal dominante Matrix gemäß (12.5) und nach Satz 12.1 eine $M-$ Matrix, also gilt das schwache Maximumprinzip (Satz 12.4).
Die Nullenverteilung von $\boldsymbol{\varphi}_{\gamma_h}$ liefert (interpretiere γ_h und ω_h als Indexmengen:)

$$N^{\neq}(\boldsymbol{A}_h y^{(1)}) \subset \gamma_h, \quad N^0(\boldsymbol{A}_h y^{(1)}) \supset \omega_h.$$

Damit folgt

$$\max_{i\in\omega_h}|y_i^{(1)}| \le \max_{i\in N^0(\boldsymbol{A}_h\boldsymbol{y}^{(1)})}|y_i^{(1)}| \le \max_{j\in N^{\neq}(\boldsymbol{A}_h\boldsymbol{y}^{(1)})}|y_j^{(1)}| \le \max_{j\in\gamma_h}|y_j^{(1)}|,$$

oder kurz

$$\|\boldsymbol{y}^{(1)}\|_{C(\omega_h)} \le \|\boldsymbol{y}^{(1)}\|_{C(\gamma_h)} = \|\boldsymbol{g}\|_{C(\gamma_h)}, \quad \text{also (12.9).}$$

Fall(12.10): $g = 0$
Nach Satz 6.3 c) gilt für $\boldsymbol{A}_h \boldsymbol{y}^{(2)} = \boldsymbol{\varphi}_{\omega_h}$

$$\|\boldsymbol{y}^{(2)}\|_{C(\bar{\omega}_h)} \leq \frac{\|\boldsymbol{p}\|_{C(\bar{\omega}_h)}}{\min\limits_{i:\,(\varphi_{\omega_h})_i \neq 0} (\boldsymbol{A}_h \boldsymbol{p})_i} \|\boldsymbol{\varphi}_{\omega_h}\|_{C(\bar{\omega}_h)} \quad \forall\, \boldsymbol{p} > 0 \text{ und } \boldsymbol{A}\boldsymbol{p} > 0 \text{ (komponentenweise)}$$

In diesem Fall genügt für p der Ansatz

$$w = \varepsilon + x_1(a - x_1) + x_2(b - x_2), \quad (x_1, x_2) \in \Omega, \quad \boldsymbol{p} = \boldsymbol{w}.$$

Dann ist

$$(\boldsymbol{A}_h \boldsymbol{p})_i \begin{cases} \geq \varepsilon > 0 & \text{auf } \gamma_h \\ = 4 & \text{auf } \omega_h \end{cases}$$

$$\|\boldsymbol{p}\|_{C(\bar{\omega}_h)} = \varepsilon + \frac{a^2 + b^2}{4},$$

$$\min_{i:\,(\varphi_{\omega_h})_i \neq 0} (\boldsymbol{A}_h \boldsymbol{p})_i = 4,$$

also

$$\|\boldsymbol{y}^{(2)}\|_{C(\bar{\omega}_h)} \leq \left(\frac{\varepsilon}{4} + \frac{a^2 + b^2}{16} \right) \|\boldsymbol{f}\|_{C(\bar{\omega}_h)} \quad \forall\, \varepsilon > 0,$$

und da diese Ungleichung von ε unabhängig ist, gilt sie auch für $\varepsilon = 0$. Daraus folgt (12.10).

Fall (12.11) durch Superposition. ■.

Bemerkungen

1. Die Abschäzungen (12.9) - (12.11) sind besser als die jeweiligen Abschätzungen aus Satz 12.1 c), die nur die $M-$ Matrixeigenschaft verwendeten.

2. Für die Aufgabe 2) am Beweisanfang (Nullrandwerte) liefert das schwache Maximunprinzip (Satz 12.4)

$$\|\boldsymbol{y}^{(2)}\|_{C_{(\gamma_h)}} \leq \|\boldsymbol{y}^{(2)}\|_{C_{(\omega_h)}},$$

   das Maximum wird also im Innern angenommen, was nicht verwunderlich ist, da die Randwerte $= 0$ sind.

3. In der praktischen Anwendung werden in $\boldsymbol{A}_h \boldsymbol{y} = \varphi$ die Randwerte immer auf die rechte Seite gebracht. Man erhält dann eine Matrix $\boldsymbol{A}_h^0$ (vgl. (12.5)), die symmetrisch ist (die Übergangselemente $\frac{1}{h_1^2}$ entfallen) und daher bessere numerische Eigenschaften hat. Man löst dann ein System $\boldsymbol{A}_h^0 \boldsymbol{y}^0 = \boldsymbol{\varphi}^0$, in dem $\boldsymbol{y}^0$ nur innere Punkte enthält. Man beachte jedoch, daß nun in den Komponenten von $\boldsymbol{\varphi}^0$ additiv zu den Komponenten von f_{ij} auch Komponenten von $\boldsymbol{g}$ auftreten. Für den Punkt $i = 1,\ j = 1$ (zum Beispiel) lautet die rechte Seite $f_{11} + \frac{1}{h_1^2} g_{10}$ (vgl. dazu auch (2.12)).

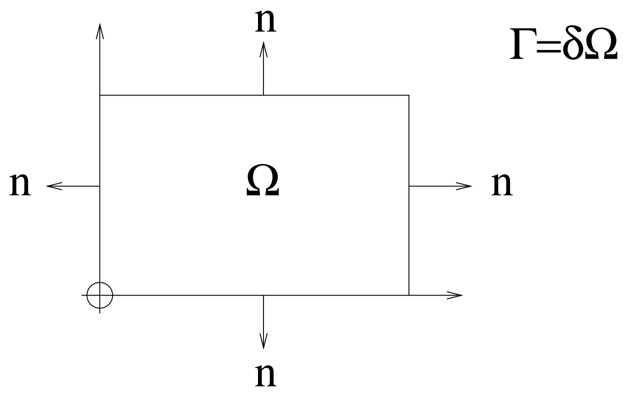
§ 13 Die 3. RWA für die Poissongleichung

Physikalische Herkunft der 3. RWA:

$\Delta u + f = 0, \quad x \in \Omega$

$\dfrac{\partial u}{\partial n} + \sigma(u - u_0) = 0$ auf $\delta\Omega$

$\sigma \geq \sigma_0 > 0,\; u_0 =$ Außentemperatur.

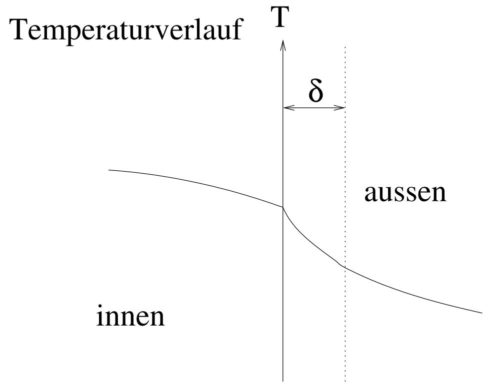


Diese Randbedingung ist bei einem stationären Wärmeleitungsprozess (z.B. in Ω wird geheizt, außerhalb ist es kälter) die richtige physikalische Formulierung dafür, daß der Wärmestrom $k\dfrac{\partial u}{\partial n}$ (wobei u = Körpertemperatur, u_0 =Außentemperatur, n = äußere Normale) stetig durch die Trennfläche geht. Sie wird bestimmt durch das

Fick'sche Gesetz: $\exists\, \alpha$:

$$\underbrace{\left(k\frac{\partial u}{\partial n}\right)\Big|_{\Gamma-\varepsilon}}_{\text{innere}} = \underbrace{\left(\alpha\frac{\partial u}{\partial n}\right)\Big|_{\Gamma+\varepsilon}}_{\text{äußere}}$$

Randgrenzwerte



Probleme in der Praxis: Der Diffusionskoeffizient k kann sich beim Durchgang durch Γ (unstetig) ändern. Dies bewirkt einen Knick in der Temperaturableitung. In der Grenzschicht (Dicke $= \delta,\; \delta = ?$) herrschen komplizierte Verhältnisse. Ist $u(\Gamma)$ die Körpertemperatur in Ω (genauer: der Grenzwerte von innen zum Rand Γ) und u_0 eine Schätzung für die Außentemperatur (, die nicht notwendig konstant sein muß,) so wird eine grobe Schätzung der rechten Seite des Fick'schen gesetzes gegeben durch

$$\left(\alpha\frac{\partial u}{\partial n}\right)\Big|_{\Gamma+\varepsilon} = \alpha\,\frac{u_0 - u(\Gamma)}{\delta} \qquad \left(= \left(k\frac{\partial u}{\partial n}\right)\Big|_{\Gamma-\varepsilon}\right)$$

woraus in der Randbedingung, die durch das Fick'sche Gesetz beschrieben wird, sich der Faktor σ ergibt zu $\sigma = \dfrac{\alpha}{\delta k}$; α, δ und k sind in der Anwendung oft schwierig zu erhalten.

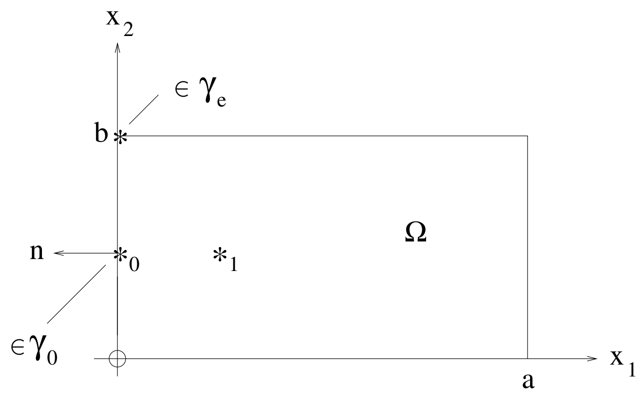
Grenzfall: Ist der Außenraum ein Isolator, so fließt keine Wärme von innen nach außen und man erhält als Randbedingung: $\dfrac{\partial u}{\partial n} = 0.$

Es sind nun folgende Probleme zu untersuchen:

1. Herstellung einer geeigneten Differenzenapproximation und Abschätzung des Diskretisationsfehlers (Taylorabgleich).
2. Nachweis, daß die Diskretisierungsmatrix eine $M-$ Matrix ist ($\Rightarrow \ \exists\, \boldsymbol{A}^{-1}$).
3. Konvergenzabschätzung.

1) Differenzenapproximation für die 3. Randbedingung: $\dfrac{\partial u}{\partial n} + \sigma(u-u_0) = 0$

Beachte: Die Diskretisierung muß für jeden Gitterpunkt, also auch jeden Randpunkt, eine Gleichung liefern weil keine Dirichletrandwerte gegeben sind.



In Ω wird wie bekannt approximiert

$$(\boldsymbol{A}_h y)_i = -(y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2})_i, \quad i \in \omega_h.$$

Diese Approximation ist von 2. Ordnung. Wir suchen deshalb für die Randpunkte ebenfalls eine Approximation 2. Ordnung. Dazu muß der Rand γ in zwei Teile zerlegt werden.

$$\gamma = \gamma_0 + \gamma_e, \quad \gamma_0 = \textit{innere}\text{ Randpunkte}, \quad \gamma_e = \text{Eckpunkte}.$$

Wir konstruieren zunächst die

Approximation am linken Rand

In obiger Abbildung bezeichnen wir mit Index 0 einen inneren Randpunkt $\in \gamma_0$, mit Index 1 seinen rechten Gitternachbarn. Die Taylorentwicklung im Randpunkt liefert

$$u_1 = u_0 + h_1 \frac{\partial u}{\partial x_1}\Big|_\Gamma + \frac{h_1^2}{2}\frac{\partial^2 u}{\partial x_1^2}\Big|_\Gamma + \frac{h_1^3}{6}\frac{\partial^3 u_z}{\partial x_1^3}$$

(13.1) $$u_{x_1,0} = \frac{u_1 - u_0}{h_1} = \frac{\partial u}{\partial x_1}\Big|_\Gamma + \frac{h_1}{2}\frac{\partial^2 u}{\partial x_1^2}\Big|_\Gamma + \frac{h_1^2}{6}\frac{\partial^3 u_z}{\partial x_1^3} \quad \Longrightarrow$$

(13.2) $$\frac{\partial u}{\partial n}\Big|_\Gamma = -\frac{\partial u}{\partial x_1}\Big|_\Gamma = -u_{x_1,0} + \frac{h_1}{2}\frac{\partial^2 u}{\partial x_1^2} + O(h_1^2)$$

Nun kann $\dfrac{\partial^2 u}{\partial x_1^2}$ im Randpunkt nicht (oder nur schlecht) approximiert werden. Zwar wäre eine Approximation von $\dfrac{\partial^2 u}{\partial x_1^2}$ von entsprechender Ordnung durch Taylorabgleich unter Hinzunahme entsprechend vieler x_1- Werte auf dem Level x_2 möglich. Dadurch wird jedoch die Struktur der Matrix beeinflußt und ein Nachweis, daß sie eine $M-$ Matrix ist, erschwert, wenn nicht unm öglich gemacht. Deshalb wird $\dfrac{\partial^2 u}{\partial x_1^2}$ mit Hilfe der Differentialgleichung ersetzt. Laut Differentialgleichung gilt

$$\frac{\partial^2 u}{\partial x_1^2} = -f - \frac{\partial^2 u}{\partial x_2^2} \tag{13.3}$$

und analog zu (12.3), falls $u \in C^3$ (Taylorreihe, vgl. z.B. (12.2)-(12.3))

$$\frac{\partial^2 u}{\partial x_2^2}\Big|_{\Gamma} = u_{\bar{x}_2 x_2,0} - \underbrace{\left(\frac{h_2}{6}\frac{\partial^3 u_+}{\partial x_2^3} + \frac{h_2}{6}\frac{\partial^3 u_-}{\partial x_2^3}\right)}_{|\ |\le \frac{h_2}{3} M_3} \tag{13.4}$$

Bemerkung: In (13.2) ist $\dfrac{\partial^2 u}{\partial x_1^2}$ schon mit einem Faktor h_1 versehen, deshalb genügt hier eine in h_2 lineare Abschätzung um gemäß $h_1 h_2 = \dfrac{1}{2}(h_1^2 + h_2^2)$ eine quadratische Abschätzung zu erhalten, insgesamt also: $\dfrac{h_1}{2}\dfrac{h_2}{3} \le M_3 \dfrac{h_1^2 + h_2^2}{6}$

Wir ersetzen in (13.2) die 2.te Ableitung gemäß (13.3), (13.4) und erhalten damit insgesamt die Approximation

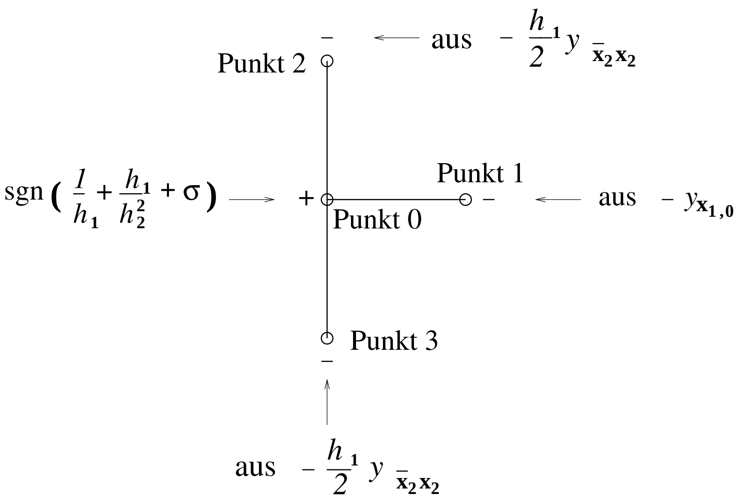
$$\begin{aligned} &\frac{\partial u}{\partial n}\Big|_{\Gamma} + \sigma(u - u_0) = -y_{x_1,0} + \frac{h_1}{2}(-f - y_{\bar{x}_2 x_2,0}) + \sigma(y_0 - u_0) + R = 0 \\ &\text{mit } |R| \le \left(\underbrace{\frac{h_2^1}{6}}_{\text{aus (13.1)}} + \underbrace{\frac{h_1^2 + h_2^2}{6}}_{\text{vgl. oben}}\right) M_3 = M_3\left(\frac{2h_1^2 + h_2^2}{6}\right) \le \frac{M_3}{3}(h_1^2 + h_2^2). \end{aligned} \tag{13.5}$$

Die Diskretisierung am <u>linken Rand</u> lautet also (bekannte Daten auf die rechte Seite und Verwendung der Punktnummern (vgl. Abb.) als Indizes)

$$(\boldsymbol{A}_h \boldsymbol{y})_0 := -y_{x_1,0} + \sigma y_0 - \frac{h_1}{2} y_{\bar{x}_2 x_2,0} = \sigma u_0 + \frac{h_1}{2} f\,, \quad \text{bzw.} \tag{L}$$

$$-\frac{y_1 - y_0}{h_1} + \sigma y_0 - \frac{h_1}{2}\frac{y_2 - 2y_0 + y_3}{h_2^2} = \sigma u_0 + \frac{h_1}{2} f$$

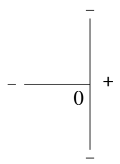
mit folgender Vorzeichenverteilung in $\boldsymbol{A}_h$



Beachte: Die Vorzeichenverteilung ist die richtige für eine $M-$ Matrix, denn die $y-$ Komponente des Punkt 0 steht in der Hauptdiagonalen, die Komponenten der $y-$ Werte der Punkte 1 und 2 stehen rechts neben der Hauptdiagonalen, der Koeffizient zur $y-$ Komponente von Punkt 3 steht links der Hauptdiagonalen.

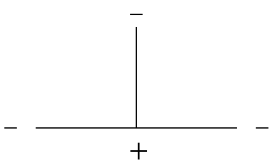
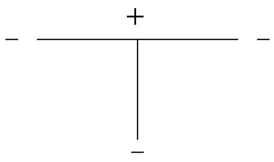
Entsprechend findet man am <u>rechten Rand</u> (beachte: $\frac{\partial}{\partial n} = +\frac{\partial}{\partial x_1}$ und $y_{\bar{x}_1, N_1}$)

(R) $\quad (\boldsymbol{A}_h y)_0 := +y_{\bar{x}_1,\, N_1} + \sigma y_{N_1} - \dfrac{h_1}{2} y_{\bar{x}_2 x_2,\, 0} = \sigma u_0 + \dfrac{h_1}{2} f.$



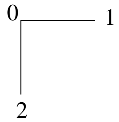
$\uparrow$
beachte $\bar{x}_1$: rückwärts genommener Differenzenquotient

Die Randbedingungen am <u>unteren</u> bzw. <u>oberen Rand</u> ergeben sich aus (L) bzw. (R) durch Vertauschung von $x_1 \leftrightarrow x_2$ und $h_1 \leftrightarrow h_2$ mit den Vorzeichenverteilungen

(U)   (O)

<u>Behandlung der 4 Eckpunkte</u>

Wir betrachten zunächst den Fall



(die Ziffern indizieren die Punkte)

Ziel:
Man konstruiert als Approximation für $\dfrac{\partial}{\partial n}$ eine Konvexkombination von $\dfrac{\partial}{\partial x_1}$ und

$\dfrac{\partial}{\partial x_2}$ derart, daß man eine $M-$ Matrix erhält. Hierbei kommen wir in den Randbedingungen ohne die Approximation der Differentialgleichung aus. Wir können analog zu (13.2) vorgehen. Die Randbedingungen lauten im Punkt 0:

$$\frac{\partial u}{\partial x_1}\Big|_{\Gamma} = -\frac{\partial u}{\partial n}\Big|_{\Gamma_l} = +\sigma(u-u_0), \qquad \Gamma_l = \text{linker Rand}$$

$$\frac{\partial u}{\partial x_2}\Big|_{\Gamma} = +\frac{\partial u}{\partial n}\Big|_{\Gamma_o} = -\sigma(u-u_0), \qquad \Gamma_o = \text{oberer Rand}$$

$$\overset{(13.2)}{\Longrightarrow} \quad u_{x_1,0} = +\sigma(u-u_0) + \frac{h_1}{2}\frac{\partial^2 u}{x_1^2} + O(h_1^2)$$

$$u_{\bar{x}_2,0} = -\sigma(u-u_0) - \frac{h_2}{2}\frac{\partial^2 u}{x_2^2} + O(h_2^2)$$

Um die zweiten Ableitungen zu beseitigen, multipliziere die 1. Gleichung mit $\frac{2}{h_1}$, die zweite mit $(-\frac{2}{h_2})$ und addiere

$$\frac{2}{h_1}u_{x_1,0} - \frac{2}{h_2}u_{\bar{x}_2,0} = \underbrace{\left(\frac{2}{h_1}+\frac{2}{h_2}\right)}_{=:\frac{2}{H}}\sigma(u-u_0) + \underbrace{\Delta u}_{=-f} + O(h_1+h_2)$$

$$\Longrightarrow \quad H = \frac{h_1 h_2}{h_1+h_2}.$$

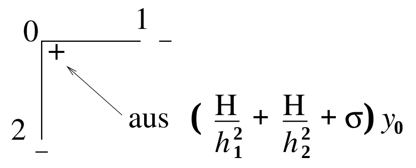
Multiplikation mit $\frac{H}{2}$ liefert die Konvexkombination

$$(13.6) \qquad \frac{H}{h_1}u_{x_1,0} - \frac{H}{h_1}u_{\bar{x}_2,0} = \sigma(u-u_0) - \frac{H}{2}f + \underbrace{\frac{1}{2}\frac{h_1h_2}{h_1+h_2}O(h_1+h_2)}_{O(h_1^2+h_2^2)}$$

Damit erhalten wir als Approximation für die Gleichung im Eckpunkt

$$(E) \qquad (\boldsymbol{A}_h\boldsymbol{y})_0 := -\frac{H}{h_1}y_{x_1,0} + \frac{H}{h_2}y_{\bar{x}_2,0} + \sigma y_0 = \sigma u_0 + \frac{H}{2}f$$

mit der für eine $M-$ Matrix richtigen Vorzeichenverteilung.



2) $\boldsymbol{A}_h$ ist eine $M-$ Matrix

Wir konstruieren mit Hilfe der Ansatzfunktion

$$w = c + x_1(a-x_1) + x_2(b-x_2), \quad c>0 \text{ geeignet}, \quad (x_1,x_2)\in[0,a]\times[0,b]$$

einen Vektor $\boldsymbol{p} > \mathbf{0}$ mit $\boldsymbol{Ap} > \mathbf{0}$ mit $\boldsymbol{p} = \boldsymbol{w}$ (Gittervektor).

Für die inneren Punkte ($\in \omega_h$) gilt (vgl. (12.7)- (12.8)): $(\boldsymbol{Ap})_l = 4,\ l \in \omega_h$.

Für die <u>linken Randpunkte</u> ($\in \gamma_0$) folgt aus (L) durch Taylorabgleich:
(beachte: Im linken Randpunkt hat der rechte Nachbar den Wert $x_1 = h_1$)

$$w_1 = w_0 + h_1 \frac{\partial w|_0}{\partial x_1} + \frac{h_1^2}{2}\frac{\partial^2 w|_0}{\partial x_1^2} + \frac{h_1^3}{6}\underbrace{\frac{\partial^3 w|_0}{\partial x_1^3}}_{=0}$$

$$\frac{w_1 - w_0}{h_1} = a - 2\cdot 0 - h_1 = a - h_1 \quad \text{und damit nach (L)}$$

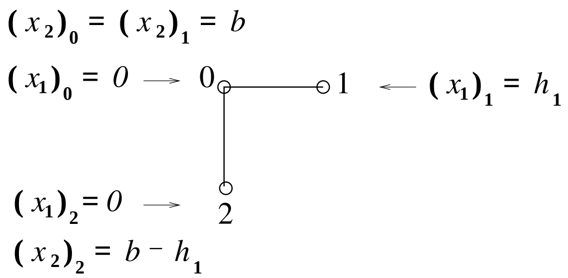
$$\begin{aligned}(\boldsymbol{A}_h\boldsymbol{w})_0 :&= -w_{x_1,0} + \sigma w_0 - \frac{h_1}{2} w_{\bar{x}_2 x_2,0} \\ &\geq -(a-h_1) + \sigma c - \frac{h_1}{2}(-2), \text{ da } x_2(b-x_2) \geq 0 \text{ und } \frac{\partial^2}{\partial x_2^2}[x_2(b-x_2)] = -2 \\ &-a + \sigma c + 2h_1 \overset{!}{\geq} 4 \quad \Longleftrightarrow \quad c \overset{!}{\geq} \frac{4 + a - 2h_1}{\sigma}.\end{aligned}$$

Die letzte Abschätzung hätte man gerne im Hinblick auf die Abschätzung von $\|\boldsymbol{A}_h^{-1}\|$: $\min\limits_l(\boldsymbol{Ap})_l$ soll nicht kleiner werden als 4, was durch die Abschätzung für die inneren Punkte schon vorgegeben ist. Mit $\sigma \geq \sigma_0 > 0$ ist sie erfüllt, falls verschärft gilt

$$c \geq \frac{4+a}{\sigma_0}.$$

Für die rechten Randpunkte folgt analog: $c \geq \dfrac{4+b}{\sigma_0}$, und entsprechend für die oberen bzw. unteren Randpunkte $c \geq \dfrac{4+b}{\sigma_0}$ bzw. $c \geq \dfrac{4+a}{\sigma_0}$.

<u>Eckpunkte (zunächst links oben)</u>



Laut (E) gilt

$$(\boldsymbol{A}_h\boldsymbol{y})_0 := -\frac{H}{h_1^2}(y_1 - y_0) + \frac{H}{h_2^2}(y_0 - y_2) + \sigma y_0.$$

Aus $w = c + x_1(a - x_1) - x_2(b - x_2)$ folgt

$$\left.\begin{array}{l} \left.\begin{array}{ll} w_0 & = c + 0 \\ w_1 & = c + h_1(a-h_1) \end{array}\right\} \ w_1 - w_0 = h_1(a-h_1) \\ w_2 \ = c + (b-h_2)(b-(b-h_2)) = c + h_2(b-h_2) \end{array}\right\} \quad w_0 - w_2 = -h_2(b-h_2)$$

Damit erhält man

$$\begin{aligned}
(\boldsymbol{A}_h\boldsymbol{w})_0 :&= -\frac{H}{h_1^2}(h_1(a-h_1)) + \frac{H}{h_2^2}(-h_2(b-h_2)) + \sigma c \\
&= -\frac{H}{h_1}(a-h_1) - \frac{H}{h_2}(b-h_2) + \sigma c \\
&= -\frac{h_2}{h_1+h_2}(a-h_1) - \frac{h_1}{h_1+h_2}(b-h_2) + \sigma c \\
&= -\underbrace{\frac{h_2}{h_1+h_2}}_{\leq 1} a - \underbrace{\frac{h_1}{h_1+h_2}}_{\leq 1} b + \underbrace{\frac{2h_1h_2}{h_1+h_2}}_{\geq 0} + \sigma c \\
&\geq -a-b+\sigma c \overset{!}{\geq} 4 \quad \text{(möchte man wieder!)}
\end{aligned}$$

Dies führt zu der hinreichenden Bedingung

$$c \geq \frac{4+a+b}{\sigma_0}, \qquad \text{(analog für die anderen Eckpunkte)} \tag{13.7}$$

Diese Forderung impliziert die vorigen, ist also insgesamt ausreichend.
Also ist $\boldsymbol{A}_h$ eine $M-$ Matrix.

3) Stabilität und Konvergenz

Nach obiger Wahl von $w = c + x_1(a-x_1) + x_2(b-x_2)$, $\boldsymbol{p} = \boldsymbol{w}$ (Gitterfunktion) erhält man (vgl. Satz 6.3 und Beweis von Satz 12.1)

$$\|\boldsymbol{A}_h^{-1}\|_{C(\omega_h)} \leq \frac{\|\boldsymbol{p}\|_\infty}{\min\limits_l (\boldsymbol{A}\boldsymbol{p})_l} \leq \frac{c + \frac{a^2+b^2}{4}}{4}, \qquad \text{also mit (13.7)}$$

$$\|\boldsymbol{A}_h^{-1}\|_{C(\omega_h)} \leq \frac{4+a+b}{4\sigma_0} + \frac{a^2+b^2}{16}. \tag{13.8}$$

Diese Abschätzung ist unabhängig von der Diskretisierung, d.h. auch die Stabilität ist gesichert.

Der Verfahrensfehler $\boldsymbol{z} = \boldsymbol{y} - \boldsymbol{u}$, ($\boldsymbol{u} =$ Gitterfunktion der exakten Lösung,) genügt $\boldsymbol{A}_h\boldsymbol{z} = \boldsymbol{\psi}$, $\quad \boldsymbol{\psi} =$ Diskretisierungsfehler.

$\boldsymbol{\psi} =$ enthält den Fehler der Differentialgleichung: $|\cdot| \leq \frac{h_1^2+h_2^2}{12}M_4$ (vgl. (12.3),(12.4)) und den Fehler des Randes $|\cdot| \leq \frac{h_1^2+h_2^2}{3}M_3$ (vgl. (13.5). Insgesamt also

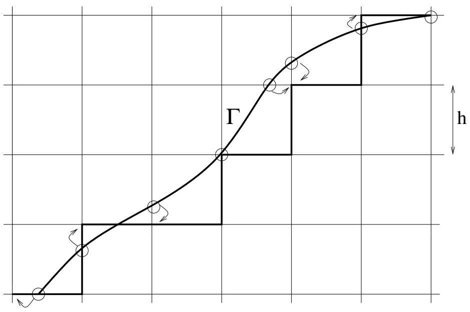
$$\|\boldsymbol{z}\|_\infty \leq \underbrace{\left(\frac{4+a+b}{4\sigma_0} + \frac{a^2+b^2}{16}\right)}_{=:C_1} \max\left(\frac{h_1^2+h_2^2}{12}M_4,\ \frac{h_1^2+h_2^2}{3}M_3\right)$$

$$\|\boldsymbol{z}\|_\infty \leq C_1 \max(4M_3,\ M_4)\frac{h_1^2+h_2^2}{12} \qquad \text{falls } u \in C^4,$$

also quadratische Konvergenz.

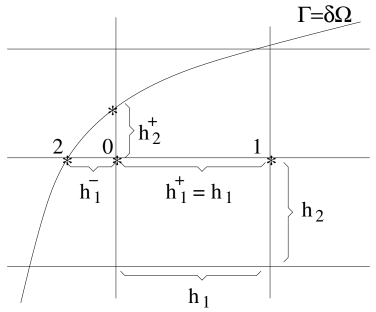
§ 14 Die 1. RWA der Poissongl. in allgemeineren Gebieten

Was macht man z.B. mit dem Rand der Nordsee? Er ist nicht eindeutig.
Ein einfaches, stabiles Verfahren erhält man, indem man die echten Randpunkte auf die benachbarten Gitterpunkte verschiebt.



Dies liefert einen Fehler 1. Ordnung. Die Ergebnisse sind jedoch, in einem gewissen Abstand vom Rand brauchbar. Wir beschreiben im Folgenden eine Diskretisierung in den randnahen inneren Punkten, die zunächst von 1. Ordnung ist, jedoch zu einem quadratisch konvergenten Verfahren für den Gesamtbereich führt.

Shortley-Wellers: Approximation und Verfahren



$h^{\pm}_{1,2}$ bezeichnen die Abstände des Bezugpunktes (hier der mit 0 bezeichnete Punkt) zu den benachbarten Gitterpunkten. Dies können innere Gitterpunkte sein (wie hier der Punkt 1), als auch Randpunkte, die durch den Schnitt des Randes von Ω mit den Gitterlinien entsteht (hier z.B. der Punkt 2).

Randnahe Gitterpunkte sind solche, für welche mindestens einer der Werte $h^{\pm}_{1,2}$ von der Maschenweite verschieden ist. Wir untersuchen, zunächst nur dür die x_1- Richtung, die Approximation von $\frac{\partial^2}{\partial x_1^2}$ im Punkt 0.

Für die Differenzenquotienten gilt (Taylorabgleich)

$$\frac{y_1 - y_0}{h_1^+} = y_0' + \frac{h_1^+}{2} y_0'' + \frac{(h_1^+)^2}{6} y_0^{(3')} + \frac{(h_1^+)^3}{24} y_{0(+)}^{(4')}$$

$$\frac{y_0 - y_2}{h_1^-} = y_0' - \frac{h_1^-}{2} y_0'' + \frac{(h_1^-)^2}{6} y_0^{(3')} - \frac{(h_1^-)^3}{24} y_{0(-)}^{(4')}$$

Beide Gleichungen weisen einen Fehler 1. Ordnung auf. Durch Subtraktion der Gleichungen und Multiplikation mit $\dfrac{2}{h_1^+ + h_1^-}$ folgt

$$\frac{2}{h_1^+ + h_1^-}\left(\frac{y_1 - y_0}{h_1^+} - \frac{y_0 - y_2}{h_1^-}\right) = y_0'' + \underbrace{\frac{2}{h_1^+ + h_1^-}\left(\frac{(h_1^+)^2}{6} - \frac{(h_1^-)^2}{6}\right)}_{\frac{h_1^+ - h_1^-}{3}} y_0^{(3')} + O((h_1^+)^2 + (h_1^-)^2)$$

$$y_0'' = \frac{2}{h_1^+ + h_1^-}\left(\frac{y_1 - y_0}{h_1^+} - \frac{y_0 - y_2}{h_1^-}\right) - \frac{h_1^+ - h_1^-}{3} y_0^{(3')} + O((h_1^+)^2 + (h_1^-)^2) \tag{14.1}$$

Das Fehlerglied folgt mit Hilfe Abschätzung

$$\frac{(h_1^+)^3}{h_1^+ + h_1^-} + \frac{(h_1^-)^3}{h_1^+ + h_1^-} \le \frac{(h_1^+)^3}{h_1^+} + \frac{(h_1^-)^3}{h_1^-} \le (h_1^+)^2 + (h_1^-)^2.$$

Beachtet man weiter

$$|h_1^+ - h_1^-| \le h_1 \quad \text{und} \quad (h_1^+)^2 + (h_1^-)^2 \le 2h_1^2$$

so läßt sich (14.1) auch schreiben als

$$y_0'' = \frac{2}{h_1^+ + h_1^-}\left(\frac{y_1 - y_0}{h_1^+} - \frac{y_0 - y_2}{h_1^-}\right) + O(h_1) + O(h_1^2) \tag{14.2}$$

$$\text{mit } |\,O(h_1)| \le \frac{h_1}{3} M_3, \quad |\,O(h_1^2)| \le \frac{h_1^2}{12} M_4, \quad M_i = \max_{\bar\Omega} |y^{(4')}|.$$

Dies ist eine Approximation 1. Ordnung, die auch bei nichtäquidistanten Gittern verwendet werden kann.

Mögliche Abhilfe: Ist h_{max} die maximale Maschenweite, so kann man durch die Forderung: $\quad |h_1^+ - h_1^-| \le M h_{max}^2$, falls $M \ge \dfrac{1}{h_{max}}$

erreichen, daß die Approximation (14.1) von 2. Ordnung ist.
Man kann Bedingungen angeben, wie ein gegebenes, nicht äquidistantes Gitter verfeinert werden kann, damit diese Forderung erhalten bleibt.

Obwohl (14.2) nur eine Abschätzung 1. Ordnung ist, werden wir zeigen, daß sie zu einem Verfahren 2. Ordnung führt.

Wir legen nun über Ω ein nicht notwendig äquidistantes Gitter mit den Maschenweiten $h = (h_1, h_2)$.
Die Menge der Gitterpunkte, die wir bei der Diskretisierung benutzen, setzt sich wie folgt zusamen aus

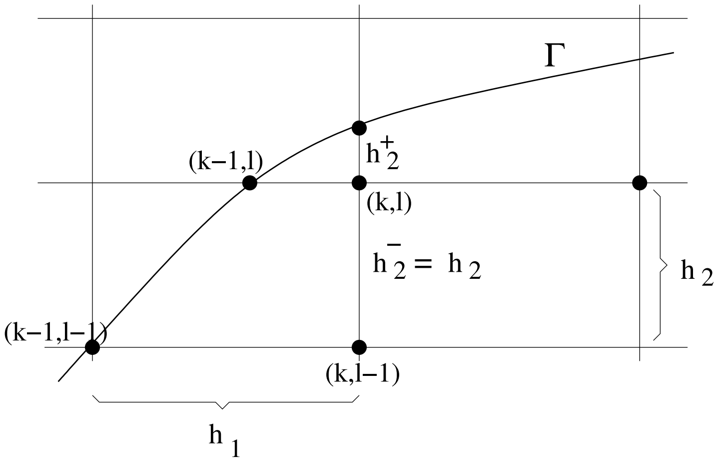
$$\bar\omega_h = \gamma_h \cup \omega_h^* \cup \omega_h^0. \tag{14.3}$$

Dabei bedeuten (vgl. Abbildung)

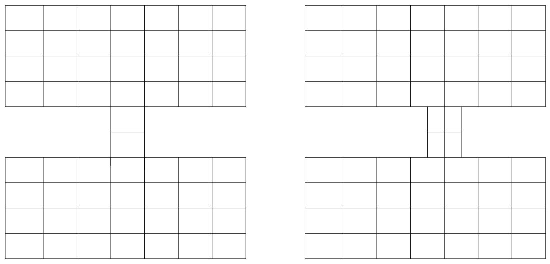
$\gamma_h =$ Randpunkte: Schnittpunkte von $\Gamma = \delta\Omega$ mit den Gitterlinien
(z.B. $(k-1,l),(k-1,l-1)$)

$\omega_h^* =$ irreguläre Punkte: Innere Gitterpunkte, die mindestens einen Randpunkt als Nachbar haben (z.B. $(k,l),(k,l-1)$).

$\omega_h^0 =$ reguläre Punkte: Innere Gitterpunkte,deren Nachbarn alle innere Punkte sind.



Das Gitter soll so fein sein (zusammenhängend), daß sich zwei beliebige innere Gitterpunkte durch einen Polygonzug aus inneren Gitterlinien verbinden lassen. Sonst zerfällt das Gleichungssystem der diskretisierten Gleichungen in separate Teilsysteme und die Kopplung zwischen den Teilsystemen geht verloren.



nicht zusammenhaengend zusammenhaengend

Bei nichtzusammenhängenden Gebieten kann man durch Verschiebung des Gitters oder Gitterverdichtung den Zusammenhang herstellen.

Wir bezeichnen nun mit $\tilde{\boldsymbol{y}}$ der Vektor, der alle Punkte aus $\bar{\omega}_h$ enthält. $\tilde{\boldsymbol{A}}_h$ sei die zugehörige Diskretisierungsmatrix.

Diskretisierung:
in ω_h^0 durch den bekannten 5-Punktestern von 2. Ordnung,

in ω_h^* gemäß (14.1) durch die Shortley-Wellers-Approximation

$$(14.4)\qquad \begin{aligned}(\tilde{A}_h\tilde{y})_{(k,l)} = \;& -\frac{2}{h_1^+ + h_1^-}\left(\frac{y_{k+1,l}-y_{k,l}}{h_1^+} - \frac{y_{k,l}-y_{k-1,l}}{h_1^-}\right) \\ & -\frac{2}{h_2^+ + h_2^-}\left(\frac{y_{k,l+1}-y_{k,l}}{h_2^+} - \frac{y_{k,l}-y_{k,l-1}}{h_2^-}\right)\end{aligned} \quad \text{mit} \quad \begin{aligned} h_1 &\geq h_1^\pm, \\ h_2 &\geq h_2^\pm \end{aligned}$$

Beachte: Für $h_1^+ = h_1^- = h_1,\ h_2^+ = h_2^- = h_2,$ wie das in inneren Punkten der Fall ist, ist das der übliche 5-Punktestern, d.h. in den entsprechenden Matrixkomponenten ergibt das die gewohnte Vorzeichenverteilung.

Dann gilt gemäß (14.2)

$$(14.5)\qquad (\tilde{A}_h\tilde{u})_{(k,l)} = -(\Delta u)_{(k,l)} + \underbrace{O(h_1+h_2)}_{|\ |\leq \frac{h_1+h_2}{3}M_3 \text{ vgl. (14.3)}} + \underbrace{O(h_1^2+h_2^2)}_{|\ |\leq \frac{h_1^2+h_2^2}{12}M_4, \text{ vgl. (12.3)}}$$

$$\text{mit } M_i = \max_{\bar{\Omega}}\left(\left|\frac{\partial^i u}{\partial x_1^i}\right|, \left|\frac{\partial^i u}{\partial x_2^i}\right|\right)$$

Dies ist offenbar eine Approximation 1. Ordnung. Trotzdem werden wir zeigen, daß sie ein konvergentes Verfahren 2. Ordnung liefert. Wir zeigen zuerst

Satz 14.1
$\tilde{A}_h$ gemäß (14.4) ist eine $M-$ Matrix und es gilt

$$\|\tilde{A}_h^{-1}\|_{C(\bar{\omega}_h)} \leq 1 + \frac{1}{4}(\text{diam}(\Omega))^2, \quad (\text{diam}(\Omega) := \max_{\boldsymbol{x},\boldsymbol{y}\in\bar{\Omega}} \|\boldsymbol{x}-\boldsymbol{y}\|_2).$$

Beweis: Im Gleichungssystem

$$\tilde{\boldsymbol{A}}_h\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{\varphi}} = \underset{\substack{\uparrow \\ \text{bedeutet keine Reihenfolge der Anordnung}}}{\begin{pmatrix}\boldsymbol{g}\\ \boldsymbol{f}\end{pmatrix}}, \qquad \begin{aligned}\boldsymbol{g} &= \text{Werte auf } \gamma_h \\ \boldsymbol{f} &= \text{rechte Seite}\end{aligned}$$

ist die Vorzeichenverteilung in den irregulären Punkten die gleiche wie in den regulären Punkten. Die echten Randpunkte liefern nur eine Eins in der Diagonalen. Die Vorzeichenverteilung in $\tilde{\boldsymbol{A}}_h$ genügt also der einer $M-$ Matrix.

Zu konstruieren ist also ein Vektor $\tilde{\boldsymbol{p}} > \boldsymbol{0}$ mit $\tilde{\boldsymbol{A}}_h\tilde{\boldsymbol{p}} > \boldsymbol{0}$.
Dazu machen wir einen quadratischen Ansatz

$$(14.6)\qquad \begin{aligned}& w(x_1,x_2) = 4 + c - ((x_1-x_{10})^2 + (x_2-x_{20})^2),\ (x_1,x_2)\in\Omega \\ & (x_{10},x_{20}) \text{ innerhalb des Umkreises um } \Omega,\ \ c = (\text{diam}(\Omega))^2.\end{aligned}$$

Begründung und Bemerkungen

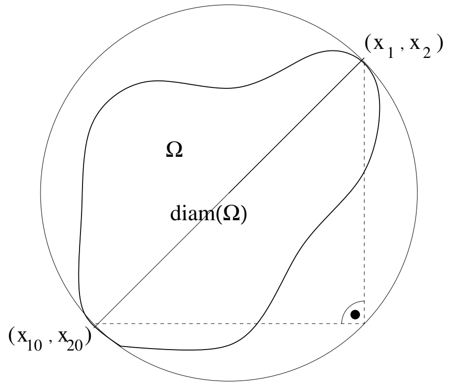
1. Wir setzen $\tilde{\boldsymbol{p}} = \boldsymbol{w}$. Die Vorzeichenverteilung der quadratischen Glieder in w ist die gleiche wie in der Ansatzfunktion (12.7). Das benötigt man für eine $M-$ Matrix.

2. $w \geq 4$ wird gebraucht im Hinblick auf die Abschätzung aus Satz 6.3, damit der Nenner nicht kleiner wird. Dies erfordert einen Korrekturterm c, der so beschaffen sein muß, daß "$c-$ quadratische Glieder" ≥ 0 ausfällt. c geht in $\|\tilde{\boldsymbol{p}}\|_\infty$ ein und sollte deshalb möglichst klein sein. Die quadratischen Glieder werden deshalb so angesetzt, daß sie der Vorzeichenverteilung genügen, aber nicht zu groß ausfallen.

3. Liegt der Punkt (x_{10}, x_{20}) im Umkreis um Ω, so ist $(x_1 - x_{10})^2 + (x_2 - x_{20})^2 \leq (\mathrm{diam}(\Omega))^2)$ (Phytagoras.) Das "$\leq$" kann zum "$=$" werden, wenn (x_{10}, x_{20}) auf dem Umkreis liegt <u>und</u> $(x_{10}, x_{20}) \in \delta\Omega$ (vgl. Zeichung).

4. Die Lage von (x_{10}, x_{20}) und die Wahl von c bewirken deshalb, daß
$$c - (x_1 - x_{10})^2 - (x_2 - x_{20})^2 \geq 0 \quad \text{und} \quad w(x_1, x_2) \geq 4 \text{ in } \bar{\Omega}.$$

Extreme Lage von (x_0, x_1) und (x_{10}, x_{20})



Mit $\tilde{\boldsymbol{p}} = \boldsymbol{w}$ gilt nach (14.4)

$$(\tilde{\boldsymbol{A}}_h \tilde{\boldsymbol{p}})_{k,l} = \begin{cases} \geq 4 & \text{auf } \gamma_h \quad \text{Randpunkte, vgl. Bemerkung d)} \\ 4 & \text{auf } \omega_h^0 \quad \text{reguläre innere Punkte, vgl. (12.9),} \\ 4 & \text{auf } \omega_h^* \quad \text{weil } w \text{ quadratisch ist und } \frac{\partial^3 w}{\partial x_i^3} = 0. \end{cases} \tag{14.7}$$

Begründung zum Fall $(k, l) \in \omega_h^*$:
$(\tilde{\boldsymbol{A}}_h \tilde{\boldsymbol{p}})_{k,l}$ wird gemäß (14.4) dargestellt und (14.1) zeigt, daß der Fehler der Darstellung erst in der 3. Ableitung nicht verschwindet. Das tut nicht weh, da w quadratisch ist.

Mit $\tilde{\boldsymbol{p}} = \boldsymbol{w}$ gilt nach (14.4)

$$(14.8) \qquad (\tilde{\boldsymbol{A}}_h \tilde{\boldsymbol{p}})_{k,l} = \begin{cases} \geq 4 & \text{auf } \gamma_h \quad \text{Randpunkte, vgl. Bemerkung 4)} \\ 4 & \text{auf } \omega_h^0 \quad \text{reguläre innere Punkte, vgl. (12.9),} \\ 4 & \text{auf } \omega_h^* \quad \text{weil } w \text{ quadratisch ist und } \frac{\partial^3 w}{\partial x_i^3} = 0. \end{cases}$$

Also ist $\tilde{A}_h$ eine $M-$ Matrix und wir erhalten nach Satz 6.3 b) die Abschätzung

$$\|\tilde{\boldsymbol{A}}_h^{-1}\|_{C(\bar{\omega}_h)} \leq \frac{\|\tilde{\boldsymbol{p}}\|_\infty}{\min\limits_i (\tilde{\boldsymbol{A}}_h \tilde{\boldsymbol{p}})_i} \leq \frac{4+c}{4} = 1 + \frac{c}{4} = 1 + \frac{1}{4}(\operatorname{diam}(\Omega))^2 \quad \text{für } c = (\operatorname{diam}(\Omega))^2.$$

$\blacksquare$.

Bemerkung:
Für den Verfahrensfehler $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{y}} - \tilde{\boldsymbol{u}}$, ($\tilde{\boldsymbol{u}} =$ Gitterfunktion der exakten Lösung u in allen Gitterpunkten inclusive Randpunkten), liefert diese Abschätzung wegen $\tilde{\boldsymbol{A}}_h \tilde{\boldsymbol{z}} = \tilde{\boldsymbol{\psi}} = O(h_1 + h_2)$ (vgl. (14.5))

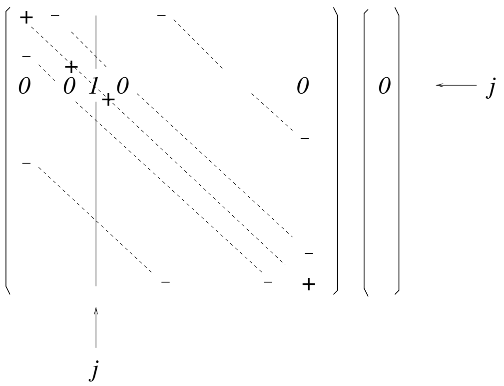
$$(14.9) \qquad \|\tilde{\boldsymbol{z}}\|_\infty \leq \|\tilde{\boldsymbol{A}}_h^{-1}\|_{C(\bar{\omega}_h)} \|\tilde{\boldsymbol{\psi}}\|_\infty = O(h_1 + h_2).$$

Dieses Ergebnis können wir jedoch verbessern mit Hilfe der schärferen Abschätzung Satz 6.3 c)

Satz 14.2 Shortly-Wellers für die 1. RWA der Poissongleichung
Falls $u \in C^4(\bar{\Omega})$ konvergiert das Verfahren gemäß (14.4) quadratisch.

Beweis: Idee:
Es ist $\bar{\omega}_h = \gamma_h \cup \omega_h^* \cup \omega_h^0$. Die Konvergenzabschätzung erhält man aus dem Gleichungssystem $\tilde{\boldsymbol{A}}_h \tilde{\boldsymbol{z}} = \tilde{\boldsymbol{\psi}}$. Nun sind bei der 1. RWA die Komponenten von $\tilde{\boldsymbol{z}}$ und $\tilde{\boldsymbol{\psi}}$, die zu den Komponenten $\in \gamma_h$ gehören, =0. Daher kann man die Randwerte aus dem Gleichungssystem eliminieren, sodaß nach der Elimination keine Randwerte mehr auftauchen.

Elimination: (vgl. Zeichnung)
Zu jedem Randpunkt gehört in $\tilde{\boldsymbol{A}}_h$ eine Zeile (z.B. j), in der die 1 in der Diagonale das einzige Element $\neq 0$ ist. Die Zeile j im Gleichungssystem $\tilde{\boldsymbol{A}}_h \tilde{\boldsymbol{z}} = \tilde{\boldsymbol{\psi}}$ und die Spalte j in $\tilde{\boldsymbol{A}}_h$ werden gestrichen, letztere weil sie mit $\tilde{z}_j = 0$ multipliziert wird. Auf diese Weise entsteht aus $\tilde{\boldsymbol{A}}_h$ die Matrix $\boldsymbol{A}_h$. Durch das Streichen der $j-$ ten Spalte werden nur Nebendiagonalelemente von $\tilde{\boldsymbol{A}}_h$ und $\boldsymbol{A}_h$ entfernt, was in den entsprechenden Zeilen die Hauptdiagonale stärkt. Das bedeutet: Geht $\boldsymbol{p}$ aus $\tilde{\boldsymbol{p}}$ hervor durch Streichen der Komponenten, die zu den Randwerten gehören, ist $(\boldsymbol{A}_h)_i$ eine Zeile, in der ein Nebendiagonalelement gestrichen wurde, $\tilde{\boldsymbol{A}}_{h\tilde{i}}$ die entsprechende Zeile von $\tilde{\boldsymbol{A}}_h$, so gilt $(\boldsymbol{A}_h)_i \boldsymbol{p} > (\tilde{\boldsymbol{A}}_h)_{\tilde{i}} \tilde{\boldsymbol{p}}$, denn die Nebendiagonalelemente sind ≤ 0. Da $\tilde{\boldsymbol{A}}_h$ eine $M-$ Matrix war, gilt dies auch für $\boldsymbol{A}_h$, denn die Vorzeichenverteilung bleibt dieselbe.



Für $\boldsymbol{A}_h$ untersuchen wir nun für die Fehlerfunktion $\boldsymbol{z}$ das System

$$\boldsymbol{A}_h \boldsymbol{z} = \boldsymbol{\psi},$$

das wir in zwei Teile zerlegen

$$\text{(14.10)} \qquad \boldsymbol{A}_h \boldsymbol{z}^0 = \boldsymbol{\psi}^0, \quad \boldsymbol{\psi}^0 = \begin{cases} \psi_{kl}, & (k,l) \in \omega_h^0 \\ 0, & (k,l) \in \omega_h^* \end{cases} \quad \Longrightarrow \quad \|\boldsymbol{\psi}^0\|_{C(\omega_h^0)} = O(h_1^2 + h_2^2)$$

$$\text{(14.11)} \qquad \boldsymbol{A}_h \boldsymbol{z}^* = \boldsymbol{\psi}^*, \quad \boldsymbol{\psi}^* = \begin{cases} \psi_{kl}, & (k,l) \in \omega_h^* \\ 0, & (k,l) \in \omega_h^0 \end{cases} \quad \overset{zeige}{\Longrightarrow} \quad \|\boldsymbol{\psi}^*\|_{C(\omega_h^*)} = O(h_1^2 + h_2^2)$$

Dann gilt

$$\text{(14.12)} \qquad \boldsymbol{A}_h \boldsymbol{z} = \boldsymbol{A}_h(\boldsymbol{z}^0 + \boldsymbol{z}^*) = \boldsymbol{\psi}^0 + \boldsymbol{\psi}^*.$$

Wir betrachten beide Aufgaben getrennt.

Zu Aufgabe (14.10) benutzen wir die verschärfte Abschätzung Satz 6.3 c)

$$\text{(14.13)} \qquad ||\boldsymbol{z}||_\infty \le \frac{||\boldsymbol{p}||_\infty}{\min\limits_{(i,k)\in\omega_h^0} (\boldsymbol{A}\boldsymbol{p})_{ik}} ||\boldsymbol{\psi}^0||_\infty \quad \forall\, \boldsymbol{p} > 0 \text{ mit } (\boldsymbol{A}\boldsymbol{p})_{ik} > \boldsymbol{0} \ \ \forall\, (i,k) \in \omega_h^0.$$

Es genügt hier $w(x_1, x_2) = c - (x_1 - x_{10})^2 - (x_2 - x_{20})^2)$, $\boldsymbol{p} = \boldsymbol{w}$ zu wählen mit $c \ge (\mathrm{diam}(\Omega))^2$, denn $\min w$ wird nur auf γ_h angenommen (vgl. Bemerkung c) nach (14.6)), also ist $\boldsymbol{w} > \boldsymbol{0}$ in $\omega_h = \omega_h^* \cup \omega_h^0$ und $\boldsymbol{A}_h \boldsymbol{w} = 4$ in ω_h^0 (vgl. (14.7)). Man erhält also

$$\text{(14.14)} \qquad \|\boldsymbol{z}^0\|_{C(\omega_h)} \le \frac{c}{4} \underbrace{\|\boldsymbol{\psi}^0\|_\infty}_{O(h_1^2+h_2^2)} \le \frac{c}{4} M_4(h_1^2 + h_2^2) \quad \text{nach(12.3)).}$$

Das ist die gewöhnliche Abschätzung für Δu in den inneren Punkten.

Zu Aufgabe (14.11) benutzen wir das schwache Maximumprinzip (vgl. dazu die Sätze 12.3 - 12.4) $\boldsymbol{A}_h$ ist eine invertierbare, zumindest schwach diagonaldominante $M-$ Matrix

(vgl. (14.16)). Wir wenden es auf $\boldsymbol{A}_h\boldsymbol{z}^* = \boldsymbol{\psi}^*$ und erhalten wegen
$\omega_h^0 \subset N^0(\boldsymbol{A}_h\boldsymbol{z}^*)$, $N^{\neq}(\boldsymbol{A}_h\boldsymbol{z}^*) \subset \omega_h^*$ (identifiziere die Gitterpunkte mit ihren Indizes)

$$\max_{i\in\omega_h^0}|z_i^*| \le \max_{i\in N^0(\boldsymbol{A}_h\boldsymbol{z}^*)}|z_i^*| \le \max_{j\in N^{\neq}(\boldsymbol{A}_h\boldsymbol{z}^*)}|z_j^*| \le \max_{j\in\omega_h^*}|z_j^*|.$$

Die maximale $z-$ Komponente gehört also zu einem Gitterpunkt $\in \omega_h^*$.
Sei $\boldsymbol{A}_h = (a_{ij}) \in \mathbb{R}^{n\times n}$, dann gilt für $i \in \omega_h^*$ mit $|z_i^*| = \max\limits_j |z_j^*|$ wegen $\boldsymbol{A}_h\boldsymbol{z}^* = \boldsymbol{\psi}^*$

$$a_{ii}|z_i^*| = |a_{ii}z_i^*| = |\psi_i^* - \sum_{\substack{j=1\\j\neq i}}^n a_{ij}z_j^*| \le |\psi_i^*| + \sum_{\substack{j=1\\j\neq i}}^n |a_{ij}|\,|z_j^*| \le |\psi_i^*| + \sum_{\substack{j=1\\j\neq i}}^n |a_{ij}|\,|z_i^*|$$

$$\Longrightarrow \quad |z_i^*|\,|a_{ii} - \sum_{\substack{j=1\\j\neq i}}^n |a_{ij}|| \le |\psi_i^*|$$

Wir zeigen, daß die Zeilensumme $\neq 0$ ist. Dann folgt

$$|z_i^*| \le \frac{1}{|a_{ii} - \sum\limits_{\substack{j=1\\j\neq i}}^n |a_{ij}||}|\psi_i^*|. \tag{14.15}$$

Eine Abschäzung der Zeilensumme nach unten wird für $|z_i^*|$ die gewünschte Abschätzung liefern.
Dazu betrachten wir zunächst für die volle Matrix $\tilde{\boldsymbol{A}}_h$ und den dazugehörigen längeren Vektor $\tilde{\boldsymbol{e}} = (1, ..., 1)^T$ die zu einem Punkt $(k, l) \in \omega_h^*$ gehörige Komponente $(\tilde{A}_h\tilde{e})_{k,l}$ in der Diskretisierung (14.4) (für zeilenweise Nummerierung)

(14.16)

$$\begin{aligned}(\tilde{\boldsymbol{A}}_h\tilde{\boldsymbol{e}})_{k,l} = \; & -\left(\frac{2}{h_2^+ + h_2^-}\frac{1}{h_2^-}\right)\tilde{e}_{k,\,l-1} - \left(\frac{2}{h_1^+ + h_1^-}\frac{1}{h_1^-}\right)\tilde{e}_{k-1,\,l} \\ & + \left(\frac{2}{h_1^+ + h_1^-}\frac{1}{h_1^-} + \frac{2}{h_1^+ + h_1^-}\frac{1}{h_1^+} + \frac{2}{h_2^+ + h_2^-}\frac{1}{h_2^+} + \frac{2}{h_2^+ + h_2^-}\frac{1}{h_2^-}\right)\tilde{e}_{k,\,l} \\ & - \left(\frac{2}{h_1^+ + h_1^-}\frac{1}{h_1^+}\right)\tilde{e}_{k+1,\,l} - \left(\frac{2}{h_2^+ + h_2^-}\frac{1}{h_2^+}\right)\tilde{e}_{l,\,k+1}\end{aligned}$$

Hierdurch wird die Zeilensumme der Zeile $(k,\,l)$ der Matrix dargestellt.
In der 1. Zeile stehen die Elemente links der Hauptdiagonalen, in der 2. Zeile die Hauptdiagonalelemente und in der 3. Zeile die Elemente rechts der Hauptdiagonalen.
Diese Darstellung gilt in ω_h^* und in ω_h^0 (wobei im letzteren Fall alle $h_{1,\,2}^{\pm} = h_{1,\,2}$).
Beachte: $\tilde{\boldsymbol{A}}_h$ ist schwach diagonaldominant. Die Zeilensummen, die zu Elementen von $\omega_h^0 \cup \omega_h^*$ gehören, sind =1.
Beim Übergang zur Matrix $\boldsymbol{A}_h$ werden die Randpunkte $\in \gamma_h$ gestrichen, d.h. in $\boldsymbol{A}_h$ fehlt in den Punkten $\in \omega_h^*$ mindestens ein Summand in (14.16), *mindestens*, weil ein irregulärer Punkt als Nachbarn mehr als einen Randpunkt haben kann.

In $\boldsymbol{A}_h$ fehlen mindestens

1. bei einem unteren Randpunkt $\tilde{e}_{k,l-1}$ eine frühere Komponente von $\boldsymbol{A}_h$
2. bei einem linken Randpunkt $\boldsymbol{e}_{k-1,l}$ eine frühere Komponente von $\boldsymbol{A}_h$
3. bei einem rechten Randpunkt $\tilde{e}_{k+1,l}$ eine spätere Komponente von $\boldsymbol{A}_h$
4. bei einem oberen Randpunkt $\tilde{e}_{l,k+1}$ eine spätere Komponente von $\boldsymbol{A}_h$

Um den jeweils gestrichenen Term überwiegt der entsprechende Term in der Hauptdiagonalen. Also ist $(\boldsymbol{A}_h\boldsymbol{e})_{k,l}$, $(k,\,l) \in \omega_h^*$ größer um den jeweils fehlenden Term, den wir abschätzen.

$$\frac{2}{h_2^+ + h_2^-}\,\frac{1}{h_2^-} \geq \frac{2}{2h_2}\,\frac{1}{h_2} = \frac{1}{h_2^2}\,, \qquad \frac{2}{h_1^+ + h_1^-}\,\frac{1}{h_1^-} \geq \frac{2}{2h_1}\,\frac{1}{h_1} = \frac{1}{h_1^2}$$

$$\text{entsprechend die anderen beiden Terme} \quad \Longrightarrow$$

$$(\tilde{\boldsymbol{A}}_h\tilde{\boldsymbol{e}})_{kl} \geq \min\left(\frac{1}{h_1^2},\frac{1}{h_2^2}\right) \quad \Longrightarrow \quad \frac{1}{(\tilde{\boldsymbol{A}}_h\tilde{\boldsymbol{e}})_{kl}} \leq \frac{1}{\min\left(\dfrac{1}{h_1^2},\dfrac{1}{h_2^2}\right)} \leq h_1^2 + h_2^2 \;\; \forall\,(k,\,l) \in \omega_h^*$$

Wir tragen diese Abschätzung in (14.14) ein und erhalten

$$|z_i^*| = \|\boldsymbol{z}^*\|_\infty \leq (h_1^2 + h_2^2)|\psi_i^*| = O(h_1^2 + h_2^2)\cdot O(h_1 + h_2) \tag{14.17}$$

■.

Bemerkungen zum Verfahren

1. Dieses Verfahren arbeitet gut bei Randwertaufgaben.
2. Bei Eigenwertaufgaben ist es problematisch. Die kontinuierliche Aufgabe ist oft selbstadjungiert, zumindest immer formal selbstadjungiert. Die Matrix $\tilde{\boldsymbol{A}}_h$ ist aber nicht symmetrisch, hat also ggf. komplexe Eigenwerte. Die Asymmetrie von $\tilde{\boldsymbol{A}}_h$ stört.

   Aufgabe:

   (a) Zeige, daß $\boldsymbol{A}_h$ (Randwerte auf die rechte Seite gebracht) symmetrisch ist bei beliebigen, einfach zusammenhängenden Gebieten.

   (b) Zeige, daß $\tilde{\boldsymbol{A}}_h$ nicht symmetrisch ist.
3. Die Abschätzung (14.17) läßt für die Diskretisierung in den irregulären Randpunkten Spielraum ohne die quadratische Konvergenz zu gefährden. Man kann also "unfaire" Approximationen für Δu benutzen, oder "gar keine", d.h. einen Fehler von $O(1)$, man verliert dann den Faktor $O(h_1+h_2)$, bzw. muß ihn durch eine Konstante ersetzen. Das tut der quadratischen Konvergenz nichts. Solche "unfairen" Approximationen werden benutzt.

   <u>Insbesondere:</u> Ersetzt man in (14.4) alle $h_{1,2}^\pm$ (, die auch noch vom irregulären

Punkt abhängen) durch $h_{1,2}$, so geht die Ordnung in (14.1) verloren und somit auch der Faktor $O(h_1+h_2)$ in (14.15). Die Abschätzungen von (14.15) zu (14.17) bleiben jedoch richtig. Der Faktor $O(h_1^2+h_2^2)$ in (14.17) bleibt erhalten, und somit auch die quadratische Konvergenz.

Daß der Einfluß der Diskretisierung in den irregulären Randpunkten auf die Konvergenzgeschwindigkeit gering ist, kann mit Hilfe der Greenschen Funktion erklärt werden, die in den randnahen Punkten klein ausfällt.

4. Statt Δu in irregulären Punkten zu diskretisieren, kann man für die Funktionswerte in inneren Punkten auch einen Wert durch lineare Interpolation der Funktionswerte in den benachbarten Punkten erhalten.

§ 15 Jacobi und Gauss-Seidel

Direkte Verfahren zur Lösung großer linearer Gleichungssysteme sind im allgemeinen zu aufwendig und zu ungenau. Deshalb werden iterative Verfahren vorgezogen.
Jacobi-Verfahren (Gesamtschrittverfahren) zur Lösung von

$$\begin{aligned} \boldsymbol{Ax} = \boldsymbol{b},\ \boldsymbol{A} \in \mathbb{R}^{n\times n},\ \boldsymbol{A} &= (\boldsymbol{L}+\boldsymbol{D}+\boldsymbol{R}) \\ \boldsymbol{L} &= \text{untere Dreiecksmatrix von } \boldsymbol{A} \\ \boldsymbol{D} &= \text{Diagonalmatrix von } \boldsymbol{A} \\ \boldsymbol{L} &= \text{obere Dreiecksmatrix von } \boldsymbol{A} \end{aligned}$$

$$\begin{aligned} \boldsymbol{x}^{m+1} &= -\boldsymbol{D}^{-1}(\boldsymbol{L}+\boldsymbol{R})\boldsymbol{x}^m + \boldsymbol{D}^{-1}\boldsymbol{b} = -\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{D})\boldsymbol{x}^m + \boldsymbol{D}^{-1}\boldsymbol{b} \\ &= (\boldsymbol{I}-\boldsymbol{D}^{-1}\boldsymbol{A})\boldsymbol{x}^m + \boldsymbol{D}^{-1}\boldsymbol{b} = \boldsymbol{x}^m + \boldsymbol{D}^{-1}(\boldsymbol{b}-\boldsymbol{Ax}^m) \end{aligned}$$

Entsprechend lautet das “gedämpfte” Jacobi-Verfahren (der Defekt wird gedämpft)

(15.1) $$\boldsymbol{x}^{m+1} = (\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A})\boldsymbol{x}^m + \omega\boldsymbol{D}^{-1}\boldsymbol{b} = \boldsymbol{x}^m + \omega\boldsymbol{D}^{-1}(\boldsymbol{b}-\boldsymbol{Ax}^m),\ \ \omega > 0.$$

Aus (15.1) liest man ab, daß alle Fixpunkte Lösungen von $\boldsymbol{Ax} = \boldsymbol{b}$ sind. Subtrahiert man die Fixpunktgleichung

$$\boldsymbol{x} = (\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A})\boldsymbol{x} + \omega\boldsymbol{D}^{-1}\boldsymbol{b}$$

so erhält man die Fehlerdarstellung

(15.2) $$\boldsymbol{e}^{m+1} := \boldsymbol{x}^{m+1} - \boldsymbol{x} = (\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A})\boldsymbol{e}^m = ... = (\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A})^{m+1}\boldsymbol{e}^0$$

$$\|\boldsymbol{e}^{m+1}\| \le \|\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A}\|^{m+1}\,\|\boldsymbol{e}^0\|.$$

Die Iterationsmatrix $\|\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A}\|$ heißt auch Konvergenzrate von (15.1) (normabhängig). Konvergenz liegt vor, wenn die Konvergenzrate < 1 ausfällt (hinreichend), bzw. wenn der Spektralradius der Iterationsmatrix $\rho(\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A}) < 1$ ist (notwendig und hinreichend). Letzteres wird bewiesen durch den

Satz
$\forall\varepsilon > 0 \wedge \boldsymbol{A} \in \mathbb{R}^{n\times n}\ \exists$ Vektornorm $\|\cdot\|_V$ und eine zugeordnete Matrixnorm $\|\cdot\|_M$ mit $\|\boldsymbol{A}\|_M \le \rho(\boldsymbol{A}) + \varepsilon$.

Bemerkungen:

1. Ist $\boldsymbol{A} = \boldsymbol{A}^T > 0$ und $\boldsymbol{D}^{-1}(\boldsymbol{A}) = c\boldsymbol{I}$ (dies ist gegeben, wenn $\boldsymbol{A}$ die Diskretisierungsmatrix von Δu ist), so ist $\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A}$ symmetrisch und es gilt
(15.3)
$$\rho(\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A}) = \max|\lambda(\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A})| = \|\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A}\|_S = \|\boldsymbol{I}-\omega c\boldsymbol{A}\|_S.$$

2. Übung: Man zeige, daß $\lambda \in \sigma(\boldsymbol{A}) \iff (1-\lambda) \in \sigma(\boldsymbol{I}-\boldsymbol{A})$.
   $\sigma(\boldsymbol{A})$ bezechnet das Spektrum von A.

Satz 15.1
Für das gedämpfte Jacobi-Verfahren zur Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ mit $\boldsymbol{A} = \boldsymbol{A}^T > 0$

$$\boldsymbol{x}^{m+1} = (\boldsymbol{I} - \omega \boldsymbol{D}^{-1}\boldsymbol{A})\boldsymbol{x}^m + \omega \boldsymbol{D}^{-1}\boldsymbol{b}, \quad \boldsymbol{D} = \operatorname{diag}(\boldsymbol{A}), \ \ \omega > 0, \ \ \boldsymbol{x}^0 \text{ gegeben}$$

gilt

Alle Eigenwerte von $\boldsymbol{D}^{-1}\boldsymbol{A}$ sind positiv.

Sind λ_{min} bzw. λ_{max} der minimale bzw. maximale Eigenwert von $\boldsymbol{D}^{-1}\boldsymbol{A}$, so wird für den Dämpfungsfaktor

$$\omega_{opt} = \frac{2}{\lambda_{min} + \lambda_{max}} \tag{15.4}$$

der Spektralradius der Iterationsmatrix $\rho(\omega) := \rho(\boldsymbol{I} - \omega \boldsymbol{D}^{-1}\boldsymbol{A})$ minimal und

$$\rho(\omega_{opt}) = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{min} + \lambda_{max}} < 1. \tag{15.5}$$

Das mit ω_{opt} gedämpfte Verfahren konvergiert für jeden Anfangswert gegen eine Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$.

Bemerkungen:

1. Für das ungedämpfte Verfahren ($\omega = 1$) ist der Satz falsch (vgl. Übungen).
2. Ist $\boldsymbol{D} = c\boldsymbol{I}$, so können in (15.5) für λ_{min} und λ_{max} die entsprechenden Eigenwerte von $\boldsymbol{A}$ eingetragen werden ($\sigma(c\boldsymbol{A}) = c\,\sigma(\boldsymbol{A})$).
3. Zur praktischen Anwendung benötigt man (zumindest Schätzungen für) λ_{min} und λ_{max}.

Beweis des Satzes:

$$\begin{aligned}
\lambda \in \sigma(\boldsymbol{D}^{-1}\boldsymbol{A}) \quad &\iff \quad \exists \boldsymbol{x} \neq 0 : \ \boldsymbol{D}^{-1}\boldsymbol{Ax} = \lambda \boldsymbol{x} \\
&\iff \quad \boldsymbol{Ax} = \lambda \boldsymbol{Dx} \\
&\implies \quad \boldsymbol{x}^*\boldsymbol{Ax} = \lambda \boldsymbol{x}^*\boldsymbol{Dx} \\
&\iff \quad \lambda = \frac{\boldsymbol{x}^*\boldsymbol{Ax}}{\boldsymbol{x}^*\boldsymbol{Dx}}
\end{aligned}$$

Da $\boldsymbol{A}$ und $\boldsymbol{D}$ positiv definit sind, folgt $\lambda > 0 \ \ \forall \lambda \in \sigma(\boldsymbol{D}^{-1}\boldsymbol{A})$.

$$0 < \lambda_{min} \leq \lambda(\boldsymbol{D}^{-1}\boldsymbol{A}) \leq \lambda_{max} \quad \stackrel{\omega>0}{\Longrightarrow} \quad 1-\omega\lambda_{max} \leq 1-\omega\lambda_{min}$$

$$\rho(\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A}) \text{ wird in Abhängigkeit von } \omega \text{ minimal, falls}$$

$$|1-\omega\lambda_{max}| = |1-\omega\lambda_{min}| \quad \text{bzw.} \quad \omega\lambda_{max}-1 = 1-\omega\lambda_{min} \quad \text{woraus folgt}$$

$$\omega_{opt} = \frac{2}{\lambda_{min}+\lambda_{max}} \quad \text{und} \quad 1-\omega\lambda_{min} = \frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}+\lambda_{min}} < 1. \blacksquare.$$

Für nichtsymmetrische Matrizen gilt

Satz 15.2
Ist $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ irreduzibel und diagonaldominant, d.h.

$$\sum_{j=1,\, j\neq i}^{n} |a_{ij}| \leq |a_{ii}| \;\forall\, i, \quad < \text{ für mindestens ein } i\,,$$

so konvergiert das Jacobi-Verfahren für jeden Startwert gegen die Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$.

Definition 15.3
$\boldsymbol{A} \in \mathbb{C}^{n\times n}$ heißt *reduzibel*, wenn gilt

1. $\exists\, N_1, N_2 \subset N = \{1,2,...,n\},\ N_1, N_2 \neq 0$
2. $N_1 \cup N_2 = N,\quad N_1 \cap N_2 = \emptyset$
3. $a_{ij} = 0 \;\;\forall\,(i,j) \in N_1 \times N_2$.

$\boldsymbol{A}$ heißt *irreduzibel* $\iff$ $\boldsymbol{A}$ ist nicht reduzibel.

Beweis von Satz (15.2): Übung.

Bemerkung zur Wahl der Startwerte:

Eine gute Wahl der Startwerte ist im Allgemeinen schwierig.
Oft wird empfohlen: $\boldsymbol{x}^0 = \boldsymbol{D}^{-1}\boldsymbol{b}$. Dies ist nichts anderes als der Iterationswert $\boldsymbol{x}^1$ für $\boldsymbol{x}^0 = 0$ und $\omega = 1$ mit dem Jacobi-Verfahren.
Beim (noch zu besprechenden) Mehrgitter-Verfahren wird eine Näherung mitgeliefert.

Abbruchkriterien

In der Praxis wird oft empfohlen: $\|\boldsymbol{x}^{m+1}-\boldsymbol{x}^m\| \leq$ eps (Maschienengenauigkeit). Dies ist problematisch, insbesondere bei schlechter Kondition, weil dieses Kriterium nichts über die Genauigkeit aussagen muß.

Empfehlung: Defektabschätzung über die Kondition.
Es ist $\boldsymbol{A}(\underbrace{\boldsymbol{x}^m - \boldsymbol{x}}_{\boldsymbol{e}^m}) = \boldsymbol{A}\boldsymbol{x}^m - \boldsymbol{b} =: \boldsymbol{d}^m$ Defektvektor, also

$$\|\boldsymbol{e}^m\| \leq \|\boldsymbol{A}^{-1}\| \, \|\boldsymbol{d}^m\|.$$

Aus $\|\boldsymbol{b}\| = \|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{x}\|$ folgt $\|\boldsymbol{x}\| \geq \frac{\|\boldsymbol{b}\|}{\|\boldsymbol{A}\|}$ also gilt für den relativen Fehler

$$\frac{\|\boldsymbol{e}^m\|}{\|\boldsymbol{x}\|} \leq \|\boldsymbol{A}\| \, \|\boldsymbol{A}^{-1}\| \frac{\|\boldsymbol{d}^m\|}{\|\boldsymbol{b}\|} = \operatorname{cond}(\boldsymbol{A}) \, \frac{\|d^m\|}{\|\boldsymbol{b}\|}.$$

Ein **vernünftiges Abbruchkriterium**

$$\frac{\|\boldsymbol{e}^m\|}{\|\boldsymbol{x}\|} \leq \operatorname{cond}(\boldsymbol{A}) \, \frac{\|\boldsymbol{d}^m\|}{\|\boldsymbol{b}\|} \overset{!}{\leq} \varepsilon \tag{15.6}$$

ist nur möglich, wenn man Werte oder Schätzungen für cond $(\boldsymbol{A})$ hat. Gute Formelpakete liefern das.

Schätzung einer Iterationszahl zur Erreichung einer vorgegebenen Genauigkeit

Satz 15.4
Für das gedämpfte Jacobi-Verfahren zur Lösung von $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ gelte $\operatorname{diag}(\boldsymbol{A}) = c\boldsymbol{I}$ und die Voraussetzungen von Satz (15.1) seien erfüllt (insbesondere gilt dann $\rho := \rho(\omega_{opt}) < 1$).
Dann gilt für den absoluten Fehler nach m Iterationen

$$\|\boldsymbol{e}^m\|_2 \leq \varepsilon,$$

falls

$$m = m(\varepsilon) \geq \frac{\log \|\boldsymbol{e}^0\|_2 + \log(1/\varepsilon)}{\log(1/\rho)} \quad \text{und}$$

$$\frac{\log \|\boldsymbol{e}^0\|_2 + \log(1/\varepsilon)}{\log(1/\rho)} \leq (\log \|\boldsymbol{e}^0\|_2 + \log(1/\varepsilon)) \, \frac{1}{2} \operatorname{cond}(\boldsymbol{A}),$$

wobei $\boldsymbol{e}^0$ = Fehler der Ausgangsnäherung.

Bemerkungen:

1. $\operatorname{diag}(\boldsymbol{A}) = c\boldsymbol{I}$ ist nur technisch. Es ist immer $\operatorname{diag}(\boldsymbol{D}^{-1}\boldsymbol{A}) = \boldsymbol{I}$.

2. Die notwendige Iterationszahl m wächst (in etwa) linear mit cond $(\boldsymbol{A})$.

3. Für $\|\boldsymbol{e}^0\|_2 \approx 1$ ist $\log \|\boldsymbol{e}^0\|_2 \approx 0$, deshalb lautet eine Schätzung (für relativ gute Anfangsnäherungen)

$$m(\varepsilon) \approx \frac{1}{2}(\log(1/\varepsilon)) \operatorname{cond}(\boldsymbol{A}) \tag{15.7}$$

Beweis:
Wir haben gezeigt (vgl. (15.2))

$$\|\boldsymbol{e}^{m+1}\|_2 \leq \|\boldsymbol{I} - \omega \boldsymbol{D}^{-1}\boldsymbol{A}\|_2^m \|\boldsymbol{e}^0\|_2 = \rho(\omega_{opt})^m \|\boldsymbol{e}^0\|.$$

Aus der Forderung $\rho^m \|\boldsymbol{e}^0\| \leq \varepsilon$ folgt (beachte: $\log \rho < 0$)

$$m \geq \frac{\log(\varepsilon) - \log(\|\boldsymbol{e}^0\|)}{\log(\rho)} = \frac{\log(\|\boldsymbol{e}^0\| + \log(1/\varepsilon))}{\log(1/\rho)} \tag{15.8}$$

Wir zeigen $\dfrac{1}{\log(1/\rho)} \leq \frac{1}{2} \operatorname{cond}_2(\boldsymbol{A})$.

Bekannt ist für die Eigenwerte von $\boldsymbol{D}^{-1}\boldsymbol{A}$ (vgl. (15.5))

$$\rho(\omega_{opt}) = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} = \frac{\frac{\lambda_{max}}{\lambda_{min}} - 1}{\frac{\lambda_{max}}{\lambda_{min}} + 1} = \frac{\kappa - 1}{\kappa + 1},$$

dabei ist wegen $\boldsymbol{D} = \operatorname{diag}\boldsymbol{A} = c\boldsymbol{I}$

$$\kappa = \operatorname{cond}_2(\boldsymbol{D}^{-1}\boldsymbol{A}) = \operatorname{cond}_2(c^{-1}\boldsymbol{A}) = \|c^{-1}\boldsymbol{A}\|_2 \, \|c\boldsymbol{A}^{-1}\|_2 = \operatorname{cond}_2(\boldsymbol{A}) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}.$$

In (15.8) schätzen wir $\log(1/\rho)$ nach unten ab. Nun ist

$$\log(1/\rho) = \log(\frac{\kappa + 1}{\kappa - 1}) = \log(1 + \frac{2}{\kappa - 1})$$

Mit der Abschätzung $\log(1 + x) \geq \frac{2x}{x+2}$ (Beweis: richtig für $x = 0$, und für $x \geq 0$ wächst die linke Seite stärker als die rechte) folgt

$$\log(1/\rho) = \log(1 + \underbrace{\frac{2}{\kappa - 1}}_{=x}) \geq \frac{2x}{x + 2} = \frac{2\,\frac{2}{\kappa-1}}{\frac{2}{\kappa-1} + 2} = \frac{4}{2(\kappa - 1) + 2} = \frac{2}{\kappa}.$$

Trägt man dies in (15.8) ein, so folgt die Behauptung. ■

Anwendung auf die Diskretisierungsmatrix $\boldsymbol{A}_h^0 = \frac{1}{h^2}\operatorname{tridiag}(-1, 2, -1)$.

Für diese Matrix (Diskretisierung von u_{xx} in $[0, 1]$) wurde gezeigt (vgl. (2.11), (3.16),(3.17))

$$8 \leq \lambda_{min} \leq \lambda_{max} \leq \frac{4}{h^2}.$$

Mit den Schätzungen

$$\|\boldsymbol{A}^{-1}\|_S = \frac{1}{\lambda_{min}(\boldsymbol{A})} \approx \frac{1}{8}, \quad \|\boldsymbol{A}\|_S = \lambda_{max}(\boldsymbol{A}) \approx \frac{4}{h^2} \quad \text{ist} \quad \mathrm{cond}_2(\boldsymbol{A}) \approx \frac{1}{2h^2}.$$

Die Schätzung (15.7) für die Iterationszahl $m(\varepsilon)$ liefert also

$$m \approx \frac{1}{2}(\log(\frac{1}{\varepsilon}))\mathrm{cond}\boldsymbol{A} = \frac{1}{4h^2}\log(\frac{1}{\varepsilon}) \overset{h=1/N}{=} N^2\frac{\log(\frac{1}{\varepsilon})}{4}.$$

Da eine Jacobiiteration mindestens N^2 Rechenoperationen braucht (N =Zahl der Unbekannten), folgt

$$m \;\approx\; N^4\,\frac{\log(\frac{1}{\varepsilon})}{4}\text{Iterarionsschritte.}$$

Zu einem vorgegebenen absoluten Fehler ε, der sich natürlich an einem sinnvollen relativen Fehler ausrichten wird, wächst also auch die Zahl der notwendigen Rechenoperationen in etwa proportional zu N^4.
Fazit: Die Lösung einer solchen Gleichung durch reine Iteration ist zu aufwendig.

Das Gauß-Seidel-Verfahren

(Einzelschrittverfahren (ESV))

Mit der Matrixzerlegung

$$\boldsymbol{A} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{R} \tag{15.9}$$

erhält man zur Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ wegen $(\boldsymbol{L}+\boldsymbol{D})\boldsymbol{x} = -\boldsymbol{Rx} + \boldsymbol{b}$ falls $\boldsymbol{D}$ regulär ist, das

$$ESV: \qquad \boldsymbol{x}^{m+1} = -(\boldsymbol{L}+\boldsymbol{D})^{-1}\boldsymbol{R}\boldsymbol{x}^m + (\boldsymbol{L}+\boldsymbol{D})^{-1}\boldsymbol{b} \tag{15.10}$$

mit der Iterationsmatrix

$$\boldsymbol{S} = -(\boldsymbol{L}+\boldsymbol{D})^{-1}\boldsymbol{R} = -(\boldsymbol{L}+\boldsymbol{D})^{-1}(\boldsymbol{A} - (\boldsymbol{L}+\boldsymbol{D})) = \boldsymbol{I} - (\boldsymbol{L}+\boldsymbol{D})^{-1}\boldsymbol{A}. \tag{15.11}$$

Beachte: Das Verfahren ist abhängig von der Anordnung der Gleichungen.

Satz 15.5
Ist $\boldsymbol{A} = \boldsymbol{A}^T > 0$ (symmetrisch, positiv definit), so konvergiert das ESV für jede Ausgangsnäherung gegen die Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$.

Beweis: Wir zeigen für eine geignete Matrixnorm $\|\boldsymbol{S}\| < 1$, genauer:
Durch $\|\boldsymbol{y}\|_A^2 := (\boldsymbol{Ay}, \boldsymbol{y}) = (\boldsymbol{y}, \boldsymbol{Ay})$ wird eine Vektornorm definiert, die zugeordnete Matrixnorm (Energienorm) ist

$$\|\boldsymbol{S}\|_A = \sup_{\boldsymbol{y}\neq 0}\frac{\|\boldsymbol{Sy}\|_A}{\|\boldsymbol{y}\|_A} = \max_{\|\boldsymbol{x}\|_A=1}\|\boldsymbol{Sx}\|_A. \tag{15.12}$$

Das Maximum wird angenommen, da $\|\boldsymbol{x}\|_A = 1$ eine kompakte Menge ist. Wir zeigen

$$\|\boldsymbol{S}\boldsymbol{y}\|_A < \|\boldsymbol{y}\|_A \quad \forall\, \boldsymbol{y} \neq 0. \tag{15.13}$$

Gemäß der Definition (15.12) gibt es dann ein $\tilde{\boldsymbol{y}}$ mit $\|\boldsymbol{S}\tilde{\boldsymbol{y}}\|_A = \|\boldsymbol{S}\|_A\, \|\tilde{\boldsymbol{y}}\|_A$. Deshalb folgt aus (15.13): $\|\boldsymbol{S}\|_A < 1$.

Beweis (15.13): (der Einfachheit halber schreiben wir $\|\cdot\| = \|\cdot\|_A$).
Mit $\boldsymbol{z} = \boldsymbol{S}\boldsymbol{y} = (\boldsymbol{I} - (\boldsymbol{L}+\boldsymbol{D})^{-1}\boldsymbol{A})\boldsymbol{y}$, also $\|\boldsymbol{z}\|^2 = \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}$ folgt

$$\begin{aligned}
\|\boldsymbol{y}\|^2 - \|\boldsymbol{S}\boldsymbol{y}\|^2 =& \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} - \boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}\\
=& \boldsymbol{y}^T\boldsymbol{A}\boldsymbol{y} - \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{A}^T(\boldsymbol{L}+\boldsymbol{D})^{-1})^{-T})\boldsymbol{A}(\boldsymbol{I} - (\boldsymbol{L}+\boldsymbol{D})^{-1}\boldsymbol{A})\boldsymbol{y}\\
&\text{ mit der Abkürzung } \boldsymbol{B} = (\boldsymbol{L}+\boldsymbol{D})\\
=& \boldsymbol{y}^T\{\boldsymbol{A} - (\boldsymbol{I} - \boldsymbol{A}^T\boldsymbol{B}^{-T})\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A}))\}\boldsymbol{y}\\
=& \boldsymbol{y}^T\{\boldsymbol{A} - (\boldsymbol{A} - \boldsymbol{A}^T\boldsymbol{B}^{-T}\boldsymbol{A})(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\}\boldsymbol{y} \quad \overset{\boldsymbol{A}=\boldsymbol{A}^T}{\Longrightarrow}\\
=& \boldsymbol{y}^T\boldsymbol{A}^T\{\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{B}^{-T}\boldsymbol{A})(\boldsymbol{I} - \boldsymbol{B}^{-1}\boldsymbol{A})\}\boldsymbol{y}\\
=& \boldsymbol{y}^T\boldsymbol{A}^T\{\boldsymbol{I} - (\boldsymbol{I} - \boldsymbol{B}^{-T}\boldsymbol{A} - \boldsymbol{B}^{-1}\boldsymbol{A} + \boldsymbol{B}^{-T}\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A})\}\boldsymbol{y}\\
=& \boldsymbol{y}^T\boldsymbol{A}^T\{\boldsymbol{B}^{-T} + \boldsymbol{B}^{-1} - \boldsymbol{B}^{-T}\boldsymbol{A}\boldsymbol{B}^{-1}\}\boldsymbol{A}\boldsymbol{y}.
\end{aligned}$$

Wegen $\boldsymbol{A} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{L}^T = \boldsymbol{B} + (\boldsymbol{L}+\boldsymbol{D})^T - \boldsymbol{D} = \boldsymbol{B} + \boldsymbol{B}^T - \boldsymbol{D}$ folgt

$$\begin{aligned}
\boldsymbol{B}^{-T}\boldsymbol{A}\boldsymbol{B}^{-1} &= \boldsymbol{B}^{-T}(\boldsymbol{B} + \boldsymbol{B}^T - \boldsymbol{D})\boldsymbol{B}^{-1}\\
&= \boldsymbol{B}^{-T}(\boldsymbol{I} + \boldsymbol{B}^T\boldsymbol{B}^{-1}) - \boldsymbol{D}\boldsymbol{B}^{-1})\\
&= \boldsymbol{B}^{-T} + \boldsymbol{B}^{-1} - \boldsymbol{B}^{-T}\boldsymbol{D}\boldsymbol{B}^{-1}.
\end{aligned}$$

Setzt man dies ein, so folgt

$$\|\boldsymbol{y}\|^2 - \|\boldsymbol{S}\boldsymbol{y}\|^2 = \boldsymbol{y}^T\boldsymbol{A}^T\boldsymbol{B}^{-T}\boldsymbol{D}\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{y} = (\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{y})^T\boldsymbol{D}(\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{y}) > 0,$$

denn $\boldsymbol{D}$ hat wegen $\boldsymbol{A} = \boldsymbol{A}^T > 0$ und $(\boldsymbol{e}^i, \boldsymbol{A}\boldsymbol{e}^i) = a_{ii} > 0$ nur positive Diagonalelemente, ist also positiv definit.

Bemerkungen

1. Die Diskretisierungsmatrizen für Δu und $\boldsymbol{A}_h$ erfüllen die Voraussetzungen von Satz (15.5).

2. Der Iterationsmatrix (15.11) der Gauß-Seidel-Iteration fehlen Symmetrieeigenschaften. Dies wird sich im Zusammenhang mit dem Mehrgitterverfahren als Nachteil erweisen (vgl. § 17 f). Wir behandeln deshalb das symmetrische Gauß-Seidel-Verfahren.

Das symmetrische Gauß-Seidel-Verfahren

Man kann Einzelschrittverfahren durch verschiedene Zerlegungen der Matrix $\boldsymbol{A}$ erhalten.

$$\boldsymbol{A} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{R}, \quad \boldsymbol{B}_1 = \boldsymbol{L} + \boldsymbol{D}, \quad \boldsymbol{B}_2 = \boldsymbol{D} + \boldsymbol{R}.$$

Das gewöhnliche Gauß-Seidel-Verfahren benutzt $\boldsymbol{B}_1$ und führt zur Iteration (vgl. (15.9),(15.10))

$$\boldsymbol{x}^{m+1} = (\boldsymbol{I} - \boldsymbol{B}_1^{-1}\boldsymbol{A})\boldsymbol{x}^m + \boldsymbol{B}_1^{-1}\boldsymbol{b}. \tag{15.14}$$

das rückwärtsgenommene ESV benutzt $\boldsymbol{B}_2$ und man erhält analog

$$\boldsymbol{x}^{m+1} = (\boldsymbol{I} - \boldsymbol{B}_2^{-1}\boldsymbol{A})\boldsymbol{x}^m + \boldsymbol{B}_2^{-1}\boldsymbol{b}. \tag{15.15}$$

Natürlich gilt Satz 15.5 auch für (15.15).
Das symmetrische ESV faßt je einen Schritt von (15.13) und (15.4) zu einem Iterationsschritt zusammen.

$$\begin{aligned} \boldsymbol{x}^{m+\frac{1}{2}} &= (\boldsymbol{I} - \boldsymbol{B}_1^{-1}\boldsymbol{A})\boldsymbol{x}^m + \boldsymbol{B}_1^{-1}\boldsymbol{b} \\ \boldsymbol{x}^{m+1} &= (\boldsymbol{I} - \boldsymbol{B}_2^{-1}\boldsymbol{A})\boldsymbol{x}^{m+\frac{1}{2}} + \boldsymbol{B}_2^{-1}\boldsymbol{b}. \end{aligned} \tag{15.16}$$

Für das symmetrische ESV erhält man deshalb

$$\boldsymbol{x}^{m+1} = \underbrace{(\boldsymbol{I} - \boldsymbol{B}_2^{-1}\boldsymbol{A})(\boldsymbol{I} - \boldsymbol{B}_1^{-1}\boldsymbol{A})}_{\boldsymbol{S}}\boldsymbol{x}^m + \underbrace{(\boldsymbol{B}_1^{-1} - \boldsymbol{B}_2^{-1}\boldsymbol{A}\boldsymbol{B}_1^{-1} + \boldsymbol{B}_2^{-1})}_{\boldsymbol{Q}}\boldsymbol{b} \tag{15.17}$$

Die Iterationsmatrix hat die Darstellung

$$\boldsymbol{S} = \boldsymbol{I} - \underbrace{(\boldsymbol{B}_2^{-1} + \boldsymbol{B}_1^{-1} - \boldsymbol{B}_2^{-1}\boldsymbol{A}\boldsymbol{B}_1^{-1})}_{=:\boldsymbol{W}^{-1}}\boldsymbol{A} \tag{15.18}$$

Satz 15.6 Eigenschaften von W
Sei $\boldsymbol{A} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{R}, \quad \boldsymbol{B}_1 = \boldsymbol{L} + \boldsymbol{D}, \quad \boldsymbol{B}_2 = \boldsymbol{D} + \boldsymbol{R}, \quad \boldsymbol{D}$ regulär.

a) Für die Iterationsmatrix des symmetrischen Gauß-Seidel-Verfahrens gilt

(15.19) $\boldsymbol{S} = \boldsymbol{I} - (\boldsymbol{B}_2^{-1} + \boldsymbol{B}_1^{-1} - \boldsymbol{B}_2^{-1}\boldsymbol{A}\boldsymbol{B}_1^{-1})\boldsymbol{A} =: \boldsymbol{I} - \boldsymbol{W}^{-1}\boldsymbol{A}$ mit $\boldsymbol{W} = \boldsymbol{A} + \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{R}$.

b) Aus $\boldsymbol{A} = \boldsymbol{A}^T > 0$ folgt $0 < \boldsymbol{A} \leq \boldsymbol{W} = \boldsymbol{W}^T$ und $\|\boldsymbol{S}\|_S < 1$ (Spektralnorm)

Bemerkungen:

1. $\boldsymbol{W}$ ist symmetrisch, falls $\boldsymbol{A}$ symmetrisch ist.
   Beachte auch $\boldsymbol{B}_1^T = \boldsymbol{B}_2$, $\boldsymbol{B}_2^T = \boldsymbol{B}_1$.

2. Die Bezeichnung
$$\boldsymbol{W}^{-1} := (\boldsymbol{B}_2^{-1} + \boldsymbol{B}_1^{-1} - \boldsymbol{B}_2^{-1}\boldsymbol{A}\boldsymbol{B}_1^{-1})$$
rührt daher, daß (15.17) durch Multiplikation mit W auch dargestellt werden kann als
$$\boldsymbol{W}(\boldsymbol{x}^{m+1} - \boldsymbol{x}^m) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^m \quad (= \boldsymbol{d}^m) \tag{15.20}$$
$\boldsymbol{W}$ ist dann eine Präkonditionsmatrix.
Hinweis: Eine Möglichkeit ein Iterationsverfahren zu beschleunigen wird untersucht durch Einführen sog. Präkonditionsmatrizen. Ist z.B. die Matrixdiagonale von A schon auf I normiert, so lautet das Gesamtschritt-Verfahren (vgl. (15.1) mit $\omega = 1$)
$$\boldsymbol{x}^{m+1} - \boldsymbol{x}^m = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^m.$$
Es wird versucht durch Multiplikation der linken Seite mit einer geeigneten (Präkonditions-)Matrix $\tilde{\boldsymbol{W}}$ ein schnelleres Verfahren der Art
$$\tilde{\boldsymbol{W}}(\boldsymbol{x}^{m+1} - \boldsymbol{x}^m) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}^m$$
zu erhalten. Für die Wahl $\tilde{\boldsymbol{W}} = \omega^{-1}\boldsymbol{I}$ mit einer geeigneten Konstanten ω erhält man z.B. das gedämpfte Gesamtschritt-Verfahren. Wählt man $\tilde{\boldsymbol{W}} = \boldsymbol{W}$, so erhält man gerade das symmetrische Gauß-Seidel-Verfahren.
3. Man kann das symmetrische Gauß-Seidel-Verfahren so programmieren, daß der Aufwand derselbe ist wie für das Gesamtschritt-Verfahren. (vgl. Hanke-Bourgeois: Grundlagen der Numerischen Mathematik und des Wissenschaftrlichen Rechnens.Teubner 2002, §II,8, S. 83).

Beweis a): Wir berechnen (vgl. (15.18))
$$\begin{aligned}\boldsymbol{B}_2^{-1}\boldsymbol{A}\boldsymbol{B}_1^{-1} &= \boldsymbol{B}_2^{-1}(\boldsymbol{B}_1 + \boldsymbol{R})\boldsymbol{B}_1^{-1} = \boldsymbol{B}_2^{-1}(\boldsymbol{B}_1 + \boldsymbol{B}_2 - \boldsymbol{D})\boldsymbol{B}_1^{-1}\\ &= (\boldsymbol{B}_2^{-1}\boldsymbol{B}_1 + \boldsymbol{I} - \boldsymbol{B}_2^{-1}\boldsymbol{D})\boldsymbol{B}_1^{-1} = \boldsymbol{B}_2^{-1} + \boldsymbol{B}_1^{-1} - \boldsymbol{B}_2^{-1}\boldsymbol{D}\boldsymbol{B}_1^{-1}.\end{aligned}$$
Eingesetzt in $\boldsymbol{W}^{-1}$ folgt (vgl. (15.19))
$$\begin{aligned}\boldsymbol{W}^{-1} &= \boldsymbol{B}_2^{-1}\boldsymbol{D}\boldsymbol{B}_1^{-1}\\ \boldsymbol{W} &= (\boldsymbol{L} + \boldsymbol{D})\boldsymbol{D}^{-1}(\boldsymbol{D} + \boldsymbol{R}) = (\boldsymbol{L} + \boldsymbol{D})(\boldsymbol{I} + \boldsymbol{D}^{-1}\boldsymbol{R})\\ &= \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{R} + \boldsymbol{R} = \boldsymbol{A} + \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{R}.\end{aligned}$$
Beweis b): Aus $\quad \boldsymbol{A} = \boldsymbol{A}^T,\ \boldsymbol{L}^T = \boldsymbol{R},\ \boldsymbol{R}^T = \boldsymbol{L} \quad$ folgt
$$\boldsymbol{W}^T = \boldsymbol{A}^T + \boldsymbol{R}^T\boldsymbol{D}^{-1}\boldsymbol{L}^T = \boldsymbol{A} + \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{R} = \boldsymbol{W}.$$
Aus $(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}) + (\boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{R}\boldsymbol{x}, \boldsymbol{x}) = (\boldsymbol{A}\boldsymbol{x}, \boldsymbol{x}) + (\boldsymbol{D}^{-1}\boldsymbol{R}\boldsymbol{x}, \boldsymbol{R}\boldsymbol{x}))$
folgt wegen $\boldsymbol{D}^{-1} > 0$ sofort
$$\boldsymbol{W} \geq \boldsymbol{A}.$$
Beweis $\|\boldsymbol{S}\|_2 < 1$ als Übung (Hinweis: Satz (15.5) verwenden).

Kapitel III

Das Mehrgitterverfahren (MGV)

§ 16 Motivation und Grobstruktur

Ausgangssituation: Löse

(16.1) $$\boldsymbol{A}_h \boldsymbol{y}_h = \boldsymbol{b}_h$$

ein lineares Gleichungssystem, das durch Diskretisierung einer Aufgabe (h =Diskretisierungsparameter, Maschenweite) entstanden ist, die ihre Struktur jeder Diskretisierung überträgt. Die Gleichungssysteme können aus elliptischen, hyperbolischen, parabolischen oder auch Integralgleichungen stammen. Auf alle diese Typen läß sich das MGV anwenden.
Um die Dimension der zu lösenden Probleme zu veranschaulichen, hier ein bescheidenes Beispiel: Poissongleichung im Einheitsquadrat mit 40×40 inneren Punkten, also 1600 Unbekannten, $1600^2 = 2.56 \cdot 10^6$ Matrixelementen. Pro Zeile sind nur 5 Elemente $\neq 0$, also 8000 Matrixelemente $\neq 0$.
Man kann sich überlegen, daß zur Lösung (z.B. für Bandmatrizen) ein Rechenaufwand von Q arithmetischen Operationen mit $Q \sim n^2$ ($\sim$ proportional) nötig ist, mit $n = (N-1)^2$ (hier $n = 1600.$)

Probleme und Fakten:

1. der Aufwand für exaktes Lösen (sowohl an Operationen als auch an Speicherplatz) steigt ungeheuerlich.
2. Eine exakte Lösung ist gar nicht sinnvoll. Es genügt eine Approximationsgenauigkeit der Lösung in der Größenordnung des Diskretisierungsfehlers.
3. Lösung der Gleichungen durch reine Iteration erfordert ebenfalls viel Aufwand, aber
4. Die "Effektivität" der Iterationsverfahren (z.B. Jacobi und Gauß-Seidel) ist in den ersten Iterationsschritten viel größer, als die lineare Konvergenzgeschwindigkeit erwarten läßt (vgl. Übungen.)

Dies führte nacheinander zu folgenden Entwicklungen.

1: Grundidee: Nachiteration (schon länger bekannt.)
Löse das Gleichungssystem $\boldsymbol{A}_h \boldsymbol{y}_h = \boldsymbol{b}_h$ "billig" (als erste Möglichkeit etwa LR-Zerlegung und einfache Genauigkeit, später werden wir noch billigere Möglichkeiten kennen lernen.) Man erhält dann eine Näherung $\boldsymbol{y}_h^0$.
Berechne den Defekt $\boldsymbol{d}_h$ exakt (z.B. mit doppelter Genauigkeit)

(16.2)
$$\boldsymbol{d}_h = \boldsymbol{b}_h - \boldsymbol{A}_h \boldsymbol{y}_h^0.$$

Löse die Korrekturgleichung exakt (was immer das numerisch bedeutet)

(16.3)
$$\boldsymbol{A}_h \boldsymbol{v}_h = \boldsymbol{d}_h.$$

Gewinne eine neue Näherung durch

(16.4)
$$\boldsymbol{y}_h^1 = \boldsymbol{y}_h^0 + \boldsymbol{v}_h.$$

Dies bewirkt:

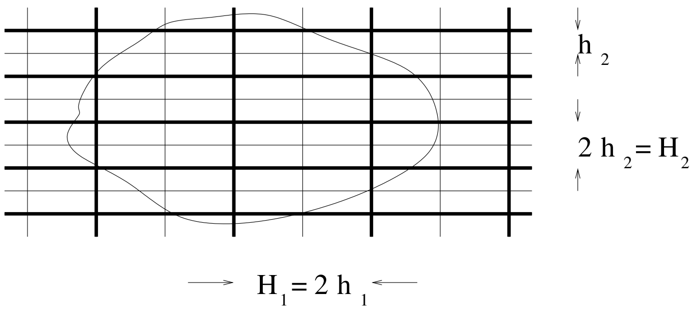
$$\boldsymbol{A}_h \boldsymbol{y}_h^1 = \boldsymbol{A}_h \boldsymbol{y}_h^0 + \boldsymbol{A}_h \boldsymbol{v}_h = \boldsymbol{A}_h \boldsymbol{y}_h^0 + \boldsymbol{d}_h \overset{(16.2)}{=} \boldsymbol{A}_h \boldsymbol{y}_h^0 + \boldsymbol{b}_h - \boldsymbol{A}_h \boldsymbol{y}_h^0 = \boldsymbol{b}_h,$$

d.h. **die neue Näherung ist die exakte Lösung.**

Ist die Lösung von (16.3) nicht exakt, so ist mindestens zu erwarten, daß die Näherung $\boldsymbol{y}_h^1$ besser ist, als $\boldsymbol{y}_h^0$. Naheliegend ist daher der Gedanke (16.3) "billiger" zu berechnen und diese Korrektur iterativ zu wiederholen. Daraus ergibt sich die

2. Grundidee: Löse $\boldsymbol{A}_h \boldsymbol{v}_h = \boldsymbol{d}_h$ "billig" (und dafür öfter) durch Ausnutzen der Aufgabenstruktur,
d.h. wir wissen, daß $\boldsymbol{A}_h \boldsymbol{y}_h = \boldsymbol{b}_h$ aus einer Diskretisierung mit der Schrittweite h stammt.
Idee: Löse die Korrekturgleichung (16.3) nur auf einem gröberen Gitter mit der Schrittweite H. (Standart: $H = 2h$). Löse also

(16.5)
$$\boldsymbol{A}_H \boldsymbol{v}_H = \boldsymbol{d}_H$$

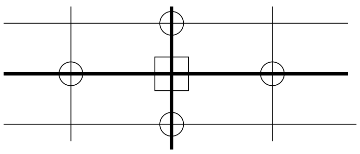


Ausgangsvoraussetzung ist also: Man hat eine Ausgangsnäherung $\boldsymbol{y}_h^0$ für (16.1). Also ist der Defekt $\boldsymbol{d}_h = \boldsymbol{b}_h - \boldsymbol{A}_h \boldsymbol{y}_h^0$ bekannt.

Problem 1: Wie macht man aus dem "großen" Vektor $\boldsymbol{d}_h$ (mit den vielen Komponenten) den "kleinen" Vektor $\boldsymbol{d}_H$?

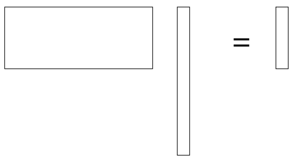
1. Möglichkeit: Man könnte die Defektkomponenten den einzelnen Gitterpunkten zuordnen und für die gemeinsamen Punkte beider Gitter (also für die Punkte des groben Gitters) $\boldsymbol{d}_h = \boldsymbol{d}_H$ setzen.
Diese Variante steigt gelegentlich aus (zur Erklärung vgl. Problem 3).

2. Möglichkeit: Man gewinnt $\boldsymbol{d}_H$ durch konvexe Mittelbildung der "umgebenden $\boldsymbol{d}_h$".



Dies kann man in Matrixschreibweise angeben mit dem Restriktionsoperator $\boldsymbol{R}_h^H$

(16.6) $$\boldsymbol{R}_h^H \quad \boldsymbol{d}_h = \boldsymbol{d}_H.$$



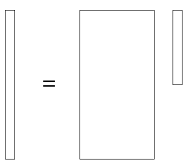
Die Matrix $\boldsymbol{R}_h^H$ wird nicht gespeichert, sondern im Programm berechnet. Dann ist

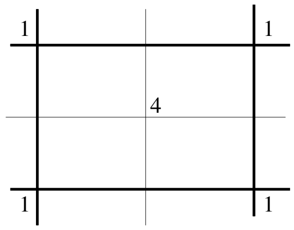
(16.7) $$\boldsymbol{A}_H \boldsymbol{v}_H = \boldsymbol{R}_h^H \boldsymbol{d}_h \quad \text{lösbar} \quad \Longrightarrow \quad \boldsymbol{v}_h \text{ berechenbar}$$

Problem 2: Die Lösung $\boldsymbol{v}_H$ ist zu "klein" (zu wenige Komponenten), also in $\boldsymbol{y}_h^1 = \boldsymbol{y}_h^0 + \boldsymbol{v}_h$ nicht einsetzbar. Man braucht eine

Prolongation $\boldsymbol{P}_H^h$:

(16.8) $$\boldsymbol{v}_h = \boldsymbol{P}_H^h \quad \boldsymbol{v}_H$$





Beim Übergang vom groben auf's feine Gitter wird linear interpoliert. Die 4 in der Mitte bedeutet nur daß der Funktionswert in der Mitte aus den 4 umgebenden, mit "1" bezeichneten Punkten gewonnen wird. (vgl. § 17)

Bemerkungen:

1. Die Indizierungen bei R und P zeigen die Richtung an, in welcher die Umformung vorgenommen wird (jeweils von unten nach oben), also vom feineren zum gröberen Gitter oder umgekehrt. Wenn man das weiß, kann man die Indizes auch weglassen.

2. Wenn das Verfahren so liefe, wie beschrieben beim Vorgehen zur Lösung von Problem 1 oder 2, wäre das vom Rechenaufwand her (für die Lösung der Korrekturgleichung) schon lohnend, denn liegt der Rechenaufwand (arithmetische Operationen) beim feinen Gitter bei Q$\approx n^2$ (größenordnungsmäßig bei n Unbekannten), so liegt er beim gröberen Gitter nur bei $Q_H \approx (\frac{1}{4}n)^2 = \frac{1}{16}Q_h$ (im 2D-Fall bei $H = 2h$, asymptotisch), d.h. man sparte mehr als eine Größenordnung.

3. Natürlich wäre mit diesem Verfahren ein Genauigkeitsverlust verbunden. Wie ernst das wäre in Anbetracht des Diskretisierungsfehlers, wäre zu überlegen.

Problem 3
Leider läuft das Verfahrenso, wie beschrieben nicht, denn $\boldsymbol{R}$ hat – notwendigerweise – einen nichttrivialen Nullraum $N(\boldsymbol{R})$.
Gilt $\boldsymbol{d}_h \in N(\boldsymbol{R}) \quad \overset{(16.7)}{\Longrightarrow} \quad \boldsymbol{v}_h = \boldsymbol{0} \quad \overset{(16.4)}{\Longrightarrow} \quad \boldsymbol{y}_h^1 = \boldsymbol{y}_h^0.$
Die Iteration steht.

Abhilfeüberlegungen

1. Man muß an Hand einer Fehleranalyse überlegen, welche Art von Fehlern mit einer "Grobgitterkorrektur" überhaupt korrigiert werden können und nicht ein $\boldsymbol{d} \in N(\boldsymbol{R})$ liefern.

2. Dann wäre es wünschenswert, eine Ausgangsnäherung $\boldsymbol{y}_h^0$ so zu wählen, oder so abzuändern, daß sie nur – oder zumindest überwiegend – Fehler enthält, die man mit der Grobgitterkorrektur abbauen kann. Dazu untersuchen wir das Fehlerverhalten bei der Lösung eines Gleichungssysterms $\boldsymbol{Ax} = \boldsymbol{b}$ mit Hilfe der Entwicklung des Fehlers nach Eigenvektoren. Dies hängt natürlich wesentlich davon ab, daß man zur Lösung ein Verfahren benutzt, das eine solche Analyse ermöglicht. Zur Erklärung des Sachverhalts benutzen wir hier das gedämpfte Jacobi-Verfahren. In der Praxis nimmt man lieber Gauß-Seidel, aber das Prinzip läßt sich hier einfacher erklären.

Fehleranalyse:

Zur Lösung von $\boldsymbol{Ax} = \boldsymbol{b}$ setzen wir voraus $\boldsymbol{A} = \boldsymbol{A}^T > \boldsymbol{0}$. Dann lautet das gedämpfte Jacobi-Verfahren (vgl. (15.1))

$$\boldsymbol{x}^{m+1} = (\boldsymbol{I} - \omega \boldsymbol{D}^{-1}\boldsymbol{A}) + \omega \boldsymbol{D}^{-1}\boldsymbol{b}, \quad \boldsymbol{D} = \mathrm{diag}(\boldsymbol{A}),$$

und für den Fehler gilt die Darstellung (15.2)

$$\boldsymbol{e}^{m+1} = \boldsymbol{x}^{m+1} - \boldsymbol{x} = (\boldsymbol{I} - \omega \boldsymbol{D}^{-1}\boldsymbol{A})\boldsymbol{e}^m = ... = (\boldsymbol{I} - \omega \boldsymbol{D}^{-1}\boldsymbol{A})^{m+1}\boldsymbol{e}^0.$$

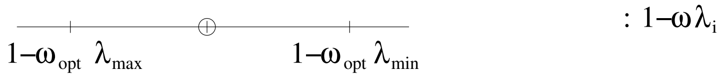
$\boldsymbol{D}^{-1}\boldsymbol{A}$ besitzt lauter positive Eigenwerte (Satz 15.1) und ein vollständiges System von Eigenvektoren $\boldsymbol{v}^{(k)}$ zu den Eigenwerten $\lambda_k,\ k = 1, ...n$. Falls $\boldsymbol{D} = \operatorname{diag}(\boldsymbol{A}) = \operatorname{diag}(c),\ c = const$, ist dies klar, weil $\boldsymbol{A}$ auf Diagonalgestalt transformierbar ist. Für eine allgemeine Diagonalmatrix findet man den Beweis in Parlettt: The symmetric Eigenvalue Problem, Chap 15.3. Wir können also $\boldsymbol{e}^0$ durch eine Eigenvektorentwicklung darstellen.

$$\begin{aligned} \boldsymbol{e}^0 &= \sum_{k=1}^{n} \alpha_k \boldsymbol{v}^{(k)} \quad \Longrightarrow \\ \boldsymbol{e}^1 &= \sum_{k=1}^{n} \alpha_k (\boldsymbol{I} - \omega \boldsymbol{D}^{-1}\boldsymbol{A})\boldsymbol{v}^{(k)} = \sum_{k=1}^{k} \alpha_k (1 - \omega\lambda_k)\boldsymbol{v}^{(k)} \\ &\vdots \\ \boldsymbol{e}^m &= \sum_{k=1}^{n} \alpha_k (1 - \omega\lambda_k)^m \boldsymbol{v}^{(k)}. \end{aligned} \tag{16.9}$$

Der Fehler konvergiert $\xrightarrow{m\to\infty} 0$, wenn $|1 - \omega\lambda_k| < 1\,\forall k$ gilt, und man erkennt, **daß die Fehlerkomponenten in Richtung der Eigenvektoren verschieden schnell gegen Null gehen**, abhängig von der Größe von $|1 - \omega\lambda_k|$, also abhängig von den jeweiligen Eigenwerten.
Beim gewöhnlichen gedämpften Jacobi-Verfahren sucht man für die Eigenwerte $\lambda_k \in \sigma(\boldsymbol{D}^{-1}\boldsymbol{A})$ die Faktoren $q_k(\omega) = 1 - \omega\lambda_k$, betragsmäßig zu minimieren durch geeignete Wahl von $\omega > 0$. Die Eigenwerte waren wie folgt verteilt

bei $\boldsymbol{D}^{-1}\boldsymbol{A}$:

bei $\boldsymbol{I} - \omega\boldsymbol{D}^{-1}\boldsymbol{A}$:



Den optimalen Wert von ω erreichte man durch die Forderung

$$|1 - \omega\lambda_{max}| = |1 - \omega\lambda_{min}|.$$

Dadurch wurde der Spektralradius von $\boldsymbol{I} - \omega\boldsymbol{D}^{-1}\boldsymbol{A}$ minimiert. Man berechnete

$$\omega_{opt} = \frac{2}{\lambda_{min} + \lambda_{max}} \Longrightarrow \quad q_{opt} = 1 - \omega_{opt}\lambda_{min} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}} < 1.$$

Betrachtet man bei dieser Wahl von ω_{opt} die Fehlerdarstellung von (16.9), so erkennt man, daß

1. die Fehleranteile der "Randeigenvektoren" (d.h. der Eigenvektoren mit kleinem und großen k, wobei $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$) in etwa gleich schnell $\xrightarrow{m\to\infty} 0$ gehen und

2. daß sie langsamer $\longrightarrow 0$ gehen, als die Anteile mit mittleren k-Werten, für die $|q_k(\omega)|$ kleiner ausfällt.

Problem 4

Das gröbere Gitter produziert Matrizen $\boldsymbol{A}_H$ mit sehr viel weniger Eigenwerten (und damit weniger Eigenvektoren) als das feinere Gitter. Nur die kleinen Eigenwerte des groben Gitters (und damit auch die "kleinen" Eigenvektoren) kann man als Näherungen für die Eigenwerte und Eigenvektoren des feineren Gitters betrachten.

Die großen Eigenwerte und Eigenvektoren des feineren Gitters kann man aus dem groben Gitter gar nicht approximieren

Folgerungen

1. Die Fehleranteile der "großen" Eigenvektoren (das sind die Eigenvektoren, die zu großen λ_k) gehören,) kann man durch Korrekturen , die aus dem groben Gitter kommen (d.h. iterative Verbesserung auf dem groben Gitter), gar nicht korrigieren.
2. Man müßte $\boldsymbol{y}_k^0$ dahingehend korrigieren, daß die Fehleranteile in $\boldsymbol{y}_k^0$, die von "großen" Eigenvektoren herrühren, möglichst klein sind und dann versuchen, durch Korrekturen, die aus dem groben Gitter errechnet werden, die Fehleranteile der kleinen Eigenvektoren zu reduzieren.
3. Man wird also nach einem (Iterations-) Verfahren zur Verbesserung und Konstruktion von $\boldsymbol{y}_k^0$ suchen, das die Fehleranteile der "großen" Eigenvektoren klein macht.

Glättung

Im 1D-Fall werden die Eigenwerte und Eigenvektoren von $\boldsymbol{D}^{-1}\boldsymbol{A}_h^0,\ \boldsymbol{D} = \operatorname{diag}(\boldsymbol{A}_h^0)$ gegeben durch (vgl. Lemma 3.7 und beachte: $\boldsymbol{D} = \operatorname{diag}(\frac{2}{h^2})$)

$$\lambda_k^h = 2\sin^2(\frac{k\pi h}{2}),\quad \boldsymbol{y}_i^k = \sin(k\pi x_i),\ i = 0, ..., N,\quad \boldsymbol{y}_0^k = \boldsymbol{y}_N^k = 0.$$

Für große k sind die $\boldsymbol{y}^k$ hochfrequente Schwingungen. Werden die Koeffizienten dieser Eigenvektoren in der Fehlerdarstellung (16.9) betragsmäßig verkleinert, so entspricht dies optisch einer Glättung des Fehlers, daher der Name für diesen Prozess.

Im 2D-Fall, $\Omega = (1,0)^2, h_1, = h_2 = h$ hat man für $(\operatorname{diag}\boldsymbol{A}_h^0)^{-1}\boldsymbol{A}_h^0$ die Eigenwerte

$$\lambda_k^h = \lambda_{k_1k_2}^h = \lambda_{k_1}^{(1)} + \lambda_{k_2}^{(2)} = \sin^2(\frac{k_1\pi h}{2}) + \sin^2(\frac{k_2\pi h}{2})$$

mit den Eigenvektoren

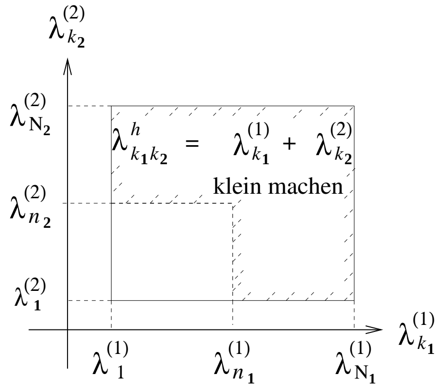
$$\boldsymbol{v}_{k_1k_2} = 2\sin(k_1\pi x_1)\sin(k_2\pi x_2),\quad \text{(komponentenweise definiert für } (x_1, x_2) \in \omega_h^0).$$

Will man wieder das gedämpfte Jacobi-Verfahren benutzen, für den Glättungsprozess, (in der Praxis wird oft Gauß-Seidel vorgezogen, darauf kommen wir später zurück), so entspricht die vorige Wahl von ω_{opt} dieser Absicht nicht.

Ziel: $|1-\omega\lambda_k|$ soll klein sein für die großen Eigenwerte, d.h. im

1D-Fall: $\frac{N}{2} \le k \le N,$

2D-Fall: $\frac{N}{4} \le k \le N,$ (genauer $k_1 \ge \frac{N_1}{2} =: n_1,\ k_2 \ge \frac{N_2}{2} =: n_2$.)



Man kann zeigen (später): Es ist möglich, die Werte $|1-\omega\lambda_k|$, die zu $\frac{N}{4} \le k \le N$ gehören (2D-Fall), unter eine von der Schrittweite unabhängige Konstante zu drücken, entsprechend im 1D-Fall.

Wir demonstrieren die Glättung an einem einfachen (1D-) Beispiel mit der Diskretisierungsmatrix, die schon im Zusammenhang mit den parabolischen Differentialgleichungen untersucht wurde. (Wegen 1D wird $\frac{N}{4}$ durch $\frac{N}{2}$ ersetzt, vgl. dazu die Eigenvektoren des 2D-Falles.)

Einfaches (1D-) Demonstrationsbeispiel zur Glättung
(unter Verwendung des gedämpften Jacobi-Verfahrens)

$$y_{\bar{x}x} + f = 0, \quad y(0) = y(1) = 1, \quad h = \frac{1}{N}.$$

Die Eigenwerte von $\boldsymbol{D}^{-1}\boldsymbol{A}$, ($\boldsymbol{A} =$ Diskretisierungsmatrix) und die zugehörigen Eigenvektoren sind (beachte: $\boldsymbol{D} = \mathrm{diag}(\frac{2}{h^2})$)

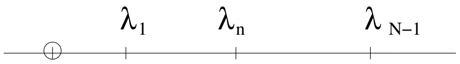
$$\lambda_k = 2\,\sin^2(\frac{k\pi h}{2}), \quad \boldsymbol{v}^{(k)} = \begin{pmatrix} \sin(k\pi x_1) \\ \vdots \\ \sin(k\pi x_{N-1}) \end{pmatrix}, \quad \text{(vgl. Lemma 3.7)}$$

wobei $N = 2n$, $\quad n =$ Matrixdimension fürs grobe Gitter,

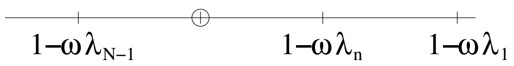
$N =$ Matrixdimension fürs feine Gitter,

$h = \dfrac{1}{2n}$ Gitterweite des feinen Gitters.

Ziel: Für $n \leq k \leq N-1$ sollen die Eigenwerte der Iterationsmatrix, also die Werte von $|1-\omega\lambda_k|$, durch geeignete Wahl von $\omega > 0$ möglichst klein ausfallen (vgl. § 15). Da alle $\lambda_k > 0$ sind, wird dies erreicht durch die Forderung



$$|1-\omega\lambda_n| = |1-\omega\lambda_{N-1}|$$



Hieraus folgt

$$\omega = \frac{2}{\lambda_n + \lambda_{N-1}} \quad \text{und} \quad q = 1 - \omega\lambda_n = \frac{\lambda_{N-1}-\lambda_n}{\lambda_{N-1}+\lambda_n}.$$

Nun gilt für die Eigenwerte des feinen Gitters

$$\lambda_n = 2\sin^2\left(\frac{n\pi h}{2}\right) \overset{nh=1/2}{=} 2\sin^2\left(\frac{\pi}{4}\right) = 1, \quad \left(\sin\left(\frac{\pi}{4}\right) = \frac{\sqrt{2}}{2}\right)$$

$$\lambda_{N-1} = 2\sin^2\left(\frac{(N-1)\pi h}{2}\right) = 2\sin^2\left(\frac{\pi h}{2} - \frac{Nh\pi}{2}\right) \overset{Nh=1}{=} 2 \underbrace{\cos^2\left(\frac{\pi h}{2}\right)}_{\approx 1 \text{ für kleine } h} \leq 2,$$

$$\omega \approx \frac{2}{3},$$

also folgt für das Betragsmaximum der Eigenwerte von $(I - \omega \boldsymbol{D}^{-1}\boldsymbol{A})$ für die Eigenwerte λ_k, $\frac{N}{2} \leq k \leq N$ von $\boldsymbol{D}^{-1}\boldsymbol{A}$

$$q = 1 - \omega\lambda_n = \frac{\lambda_{N-1}-\lambda_n}{\lambda_{N-1}+\lambda_n} = \frac{2\cos^2(\frac{\pi h}{2}) - 1}{2\cos^2(\frac{\pi h}{2}) + 1} \leq \frac{1}{3}.$$

Diese, bzw. eine entsprechende Wahl im höherdimensionalen Fall, wird dem Mehrgitterverfahren zur Konvergenz verhelfen.

Folge: Man kann also eine Ausgangsnäherung $\boldsymbol{y}_h^0$ des Gleichungssystems $\boldsymbol{A}_h\boldsymbol{y}_h = \boldsymbol{b}_h$ durch einige Schritte des beschriebenen Glättungsverfahrens so abändern, daß man Näherungen erhält, deren Fehler weniger hochfrequente Anteile erhalten.

Bemerkung: Eine Glättungsiteration muß im allgemeinen Fall nicht mehr konvergieren. Es sollen ja nur die Fehleranteile der "großen" Vektoren vermindert werden. In diesem Fall konvergiert sie "gerade noch".

$$\rho(\boldsymbol{I} - \omega\boldsymbol{D}^{-1}\boldsymbol{A}) = 1 - \omega\lambda_1 = 1 - \frac{2}{3}\, 2\sin^2\left(\frac{\pi h}{2}\right) \approx 1 - \frac{4}{3}\,\frac{\pi^2 h^2}{4} \approx 1 - \pi h^2.$$

Beachte jedoch: Die Konvergenz wird schlechter je kleiner h wird.

Zwei-Gitter-Verfahren

Mit der Bezeichnung

$$\mathcal{S} = \mathcal{S}(\boldsymbol{A}_h, \boldsymbol{b}_h) \quad \text{für den } \textit{Glättungsoperator} \quad (\mathcal{S} \mathrel{\hat{=}} \text{smoothing})$$

kann man also ein 2-Gitter-Verfahren wie folgt beschreiben:

Verlauf eines 2-Gitter-Verfahrens zur Lösung von $\boldsymbol{A}_h \boldsymbol{y}_h = \boldsymbol{b}_h$

1. Billige Beschaffung einer Ausgangsnäherung $\boldsymbol{y}_h^0$
2. Glättung der Ausgangsnäherung (durch i Schritte eines einfachen Verfahrens: Gauß-Seidel ist beliebter als Jacobi)
   $\boldsymbol{y}^{(i)} = \mathcal{S}^i(\boldsymbol{A}_h, \boldsymbol{b}_h)\boldsymbol{y}_h^0$ es verbleiben hauptsächlich niederfrequente Fehleranteile
3. $\boldsymbol{d}_h = \boldsymbol{b}_h - \boldsymbol{A}_h \boldsymbol{y}_h^{(i)}$ Defektberechnung
4. $\boldsymbol{R}_h^H \boldsymbol{d}_h = \boldsymbol{d}_H$ Restriktion auf “kleineren” Vektor
5. $\boldsymbol{A}_H \boldsymbol{v}_H = \boldsymbol{d}_H$ Korrekturgleichung lösen
6. $\boldsymbol{y}_h^{(1)} = \boldsymbol{y}_h^{(i)} + \boldsymbol{P}_H^h \boldsymbol{v}_H$ Prolongation und Verbesserung
7. $\longrightarrow$ 2)

Dieses Verfahren konvergiert schon.

Beschaffung einer Ausgangnäherung: Löse $\boldsymbol{A}_h \boldsymbol{y}_h = \boldsymbol{b}_h$ auf einem ganz groben Gitter, “fahre es hoch” durch Prolongation bis aufs Gitter h, und benutze diese Approximation als Ausgangsnäherung.

Bemerkungen:

1. Dieser Zyklus wird drei bis vier mal durchlaufen. (Es gibt Varianten mit nur 2 Zyklen.)
   Beachte: Es muß ja nur eine Lösungsgenauigkeit erreicht werden, die mit der Genauigkeit des Diskretisierungsoperators vergleichbar ist.
2. Der Schritt 2) heißt *Vorglättung*. Man könnte als Schritt 6a) nochmals eine Glättung einfügen (*Nachglättung*). Es gibt Varianten ohne Vorglättung aber mit Nachglättung.
3. Die Glättungsstrategie bewirkt, daß bei der Restriktion der Defekt $\boldsymbol{d}_H \neq 0$ ausfällt, in 5) also tatsächlich eine Verbesserung erreicht wird (genaueres muß der Konvergenzbeweis zeigen).
4. Das Verfahren klappt auch bei Finite-Element-Rechnungen, ist aber schneller beim Differenzenverfahren.

5. Selbst für nichtlineare Probleme wird es mit Erfolg angewendet (vgl. § 20). Man benötigt dafür sehr gute Lösungsverfahren auf dem untersten Gitter.

6. Überschlägt man den Aufwand an Rechenoperationen für einen Zyklus, so erkennt man leicht, daß er in allen Schritten, außer 5), linear mit der Anzahl der Unbekannten wächst.

Zur weiteren Beschreibung ändern wir die Indizierung. Da wir mehr als nur zwei Gitter einsetzen wollen, reichen zu ihrer Bezeichnung die Indizes h und H nicht mehr aus. Wir bezeichnen das Gitter, auf dem wir die Gleichung $\boldsymbol{A}_h \boldsymbol{y}_h = \boldsymbol{b}_h$ lösen wollen, mit der Gitternummer l, das gröbere Gitter (üblich ist Schrittweitenverdopplung bei der Gitterverfeinerung) mit $l-1$; $l=0$ bezeichnet das gröbste Gitter. Entsprechend werden die Gleichungsgrößen indiziert, also z.B. $\boldsymbol{A}_l \boldsymbol{y}_l = \boldsymbol{b}_l$.

Algorithmisch können wir das Zweigitter-Verfahren (ZGV, englisch TGM $\hat{=}$ <u>t</u>wo-<u>g</u>rid-<u>m</u>ethod) in einer programmierähnlichen Schreibweise wie folgt beschreiben: (vgl. Hackbusch: Multigrid Methods and Applications)

Mit den Abkürzungen

$l \;\hat{=}$ Gitternummer

$\boldsymbol{u} \;\hat{=}$ als Eingabevektor die Anfangsnäherung $\boldsymbol{y}_h^0$,
als Ausgabevektor für die durch Glättung verbesserte Approximation

$\boldsymbol{b} \;\hat{=}$ rechte Seite von $\boldsymbol{Ax} = \boldsymbol{b}$

$\boldsymbol{d} \;\hat{=}$ Defekt

$\boldsymbol{v} \;\hat{=}$ Lösung der Korrekturgleichung

lautet die Programmieranweisung

procedure ZGV $(l, \boldsymbol{u}, \boldsymbol{b})$; integer l; array $\boldsymbol{u}, \boldsymbol{b}$;

if $l = 0$ then $\boldsymbol{u} := \boldsymbol{A}_0^{-1} * \boldsymbol{b}_0$ Lösung auf dem gröbsten Gitter

else

begin array $\boldsymbol{v}, \boldsymbol{d}$; Korrekturvektor und Defekt

$\boldsymbol{u} := \mathcal{S}_l^{\nu}(\boldsymbol{u}, \boldsymbol{b})$; ν mal glätten auf Gitter l (Vorglättung)

$\boldsymbol{d} := \boldsymbol{R} * (\boldsymbol{b} - \boldsymbol{A}_l * \boldsymbol{u})$; Restriktion und Defektberechnung

$\boldsymbol{v} =: \boldsymbol{A}_{l-1}^{-1} \boldsymbol{d}$; Lösung der Korrekturgleichung auf dem gröberen Gitter

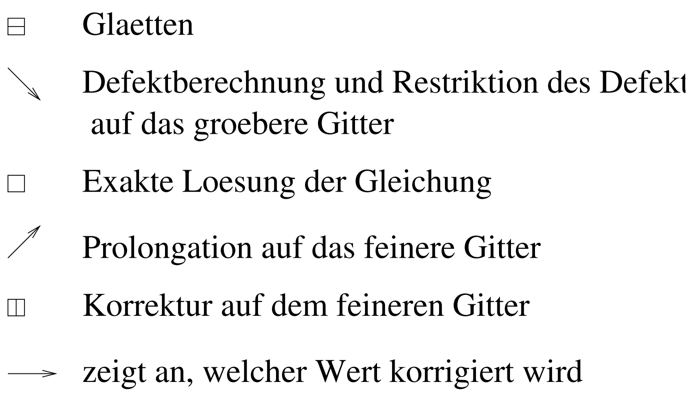
$\boldsymbol{u} =: \boldsymbol{u} + \boldsymbol{P} * \boldsymbol{v}$; Prolongation und Korrektur

$\boldsymbol{u} =: \mathcal{S}^{\nu_2}(\boldsymbol{u}, \boldsymbol{b})$; ggf. Nachglättung ($\nu_2$ Iterationen)

end ZGV

Der Aufruf ZGV $(l, \boldsymbol{u}, \boldsymbol{b})$; mit $l > 0$ beschreibt einen Schritt des ZGV.

Graphisch kann man das mit Symbolen wie folgt veranschaulichen:



So lassen sich z.B. zwei Zyklen des ZGV veranschaulichen als

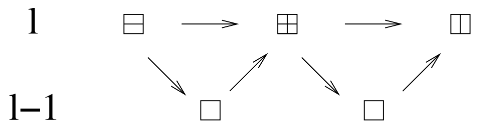


Abb. 1

Die waagrechten Pfeile werden oft weggelassen. Korrigiert wird dann immer die Lösung, die links daneben auf derselben Höhe steht.

Mehrgitter-Verfahren

In Anbetracht dessen, daß exakte Lösung der Defektgleichung auf dem gröberen Gitter + Prolongation auf das feinere Gitter + Korrektur auf dem feineren Gitter nur eine neue Näherung $\boldsymbol{y}_l^0$ liefert, kann man sich überlegen, ob die exakte Lösung auf dem gröberen Gitter nicht durch eine billigere, approximative Lösung ersetzt werden kann. Beachtet man, daß $\boldsymbol{A}_{l-1}$ ja dieselbe Struktur wie $\boldsymbol{A}_l$ besitzt, so liegt es nahe die Korrekturgleichung auf dem Gitter $l-1$ zu lösen durch Anwendung eines neuerlichen ZGV unter Benutzung eines nochmals vergröberten Gitters $l-2$. Als Ausgangsnäherung wird $\boldsymbol{v} = \boldsymbol{0}$ benutzt (graphisch durch o angedeutet).

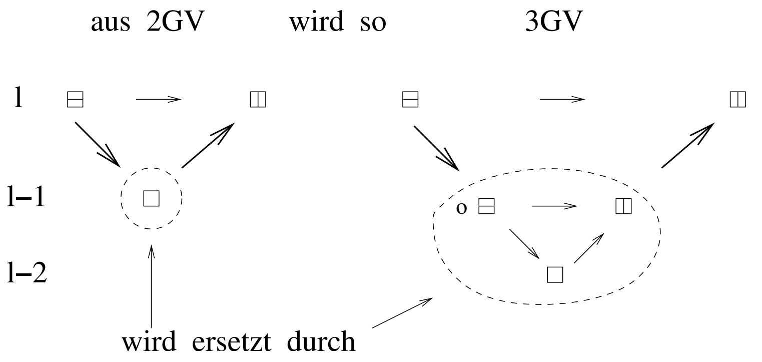


Abb. 2

Algorithmisch bedeutet das, daß innerhalb des ZGV nochmals ein ZGV aufgerufen wird (rekursiver Aufruf des ZGV) zur approximativen Lösung der Korrekturgleichung.

Algorithmisch kann man das wie folgt als Mehrgitterverfahren fassen durch mehrmaligen rekursiven Aufruf des ZGV.

procedure MGV $(l, \boldsymbol{u}, \boldsymbol{b})$; integer l; array $\boldsymbol{u}, \boldsymbol{b}$;

if $l = 0$ then $\boldsymbol{u} := \boldsymbol{A}_0^{-1} * \boldsymbol{b}_0$

else

begin array $\boldsymbol{v}, \boldsymbol{d}$;

$\boldsymbol{u} := \mathcal{S}_l^{\nu}(\boldsymbol{u}, \boldsymbol{b})$;

$\boldsymbol{d} := \boldsymbol{R} * (\boldsymbol{b} - \boldsymbol{A}_l * \boldsymbol{u})$; $\downarrow$ Änderung gegenüber ZGV

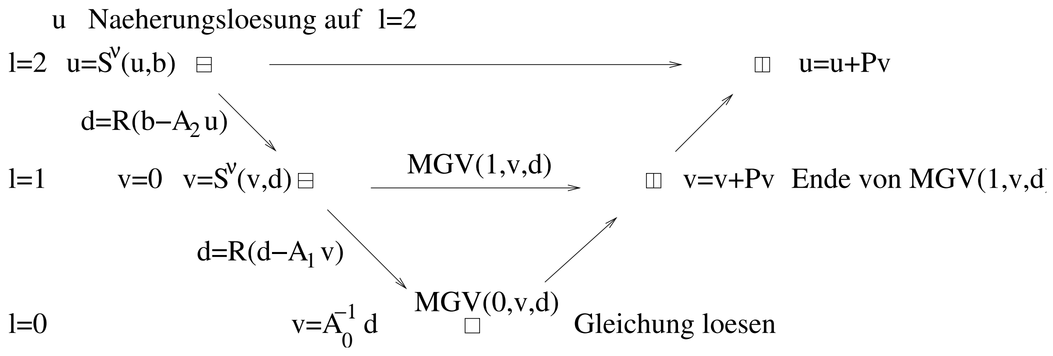
$\boxed{\boldsymbol{v} := \boldsymbol{0};\ \text{MGV}\,(l-1, \boldsymbol{v}, \boldsymbol{d});}$ Approximative Lösung der Grobgitterkorrektur

$\boldsymbol{u} = \boldsymbol{u} + \boldsymbol{P} * \boldsymbol{v}$;

$\boldsymbol{u} := \mathcal{S}^{\nu_2}(\boldsymbol{u}, \boldsymbol{b})$; als Variante gelegentlich noch eine Nachglättung

end;

Wir verfolgen den Programmverlauf an einem Beispiel für l=2.



3GV sind unüblich, weil auch die Lösung auf Gitter $l-3$ zu aufwendig ist. Man löst also die Korrekturgleichung auf $l-3$ approximativ durch ein weiteres ZGV, also MGV $(3, \boldsymbol{u}, \boldsymbol{b})$ vgl. nächstes Beispiel.

Praktisch verwendet werden $l \geq 4$ Gitter.
Graphisch sehen zwei Zyklen eines solchen 4GV wie folgt aus

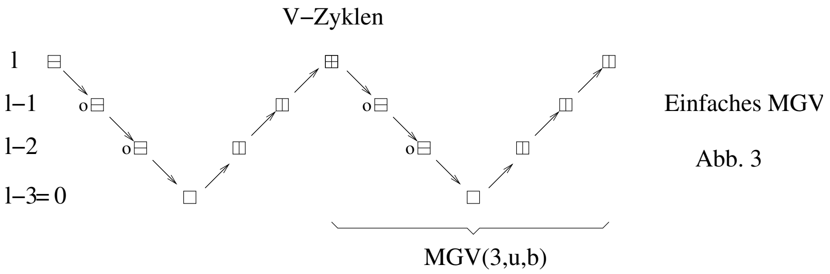


Abb. 3

In diesem Beispiel wird zur approximativen Lösung der Korrekturgleichung auf Gitter $l-1$ ein 3GV benutzt.

Der erste und zweite V-Zyklus aus Abb 3 wird beschrieben durch ein MGV $(3, \boldsymbol{u}, \boldsymbol{b})$ (beachte: das ist ein 4GV), d.h. die 1. Grobgitterkorrekturgleichung wird durch ein 3GV gelöst. Erst auf dem gröbsten Gitter $(l = 0)$ wird die dortige Gleichung exakt gelöst. Für gutartige Probleme läuft dieses Verfahren schon ganz gut.
Verfahrensvariante: In den aufsteigenden Linien wird nicht nur eine Korrektur durchgeführt, sondern auch geglättet, d.h. ersetze $\boxed{\,|\,}$ durch $\boxed{+}$

Ausbau:

Bei empfindlichen Problemen, (z.B. schlechte Kondition, zu viele Gitterverfeinerungen) kann es vorkommen, daß die Verbesserung, die man durch die approximative Lösung der Korrekturgleichung erhält, auf Grund der vielen Restriktionen und Prolongationen, nicht gut genug ist. Dem kann man begegnen, indem man den im MGV eingerahmten Prozess zur approximativen Lösung der Grobgittergleichung mehrfach ausführt, (d.h. die Lösung der Grobgitterkorrektur wird genauer), z.B. γ mal ($\gamma = 2$ wird oft benutzt), bevor man nach Prolongation die Korrektur auf dem obersten Gitter ausführt. Algorithmisch geschieht das, indem man für den eingerahmten Teil des MGV eine Wiederholungsschleife einführt. Dies bewirkt u.a., daß auf dem untersten Gitter auch das Gleichungslösen mehrfach (γ-fach im folgenden Programm) ausgeführt wird.
Wir bezeichnen diese Prozedur mit MGV γ, um anzudeuten, daß γ mal eine Nachiteration auf dem gröberen Gitter stattfindet. Mit dieser Bezeichnung ist MGV=MGV1.

procedure MGV $\gamma(l, \boldsymbol{u}, \boldsymbol{b})$; integer l; array $\boldsymbol{u}, \boldsymbol{b}$;

if $l = 0$ then $\boldsymbol{u} := \boldsymbol{A}_0^{-1} * \boldsymbol{b}_0$

else

begin $\boxed{\text{integer } j;}$ array $\boldsymbol{v}, \boldsymbol{d}$;

$\boldsymbol{u} := \mathcal{S}_l^{\nu}(\boldsymbol{u}, \boldsymbol{b})$;

$\boldsymbol{d} := \boldsymbol{R} * (\boldsymbol{b} - \boldsymbol{A}_l * \boldsymbol{u})$;

$\boldsymbol{v} := 0$

$\boxed{\text{for } j = 1 \text{ step } 1 \text{ until } \gamma \text{ do}}$ MGV$(l-1, \boldsymbol{v}, \boldsymbol{d})$;

$\boldsymbol{u} :=: \boldsymbol{u} + \boldsymbol{P} * \boldsymbol{v}$;

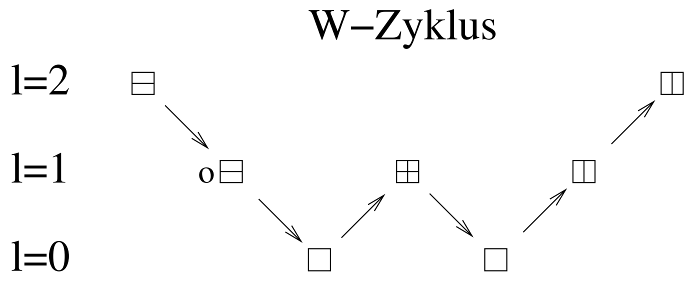
end;

Die eingerahmten Teile sind neu im Vergleich zum einfachen Verfahren MGV, das für $\gamma = 1$ in MGV1 enthalten ist.

Wir erläutern dieses Programm durch eine Reihe von Beispielen graphisch, jeweils für

γ =2 und verschiedene l.

Beispiel: Für l=2 wird ein 2-Gitter-Verfahren γ mal zur Konstruktion der Grobgitterkorrektur benutzt.

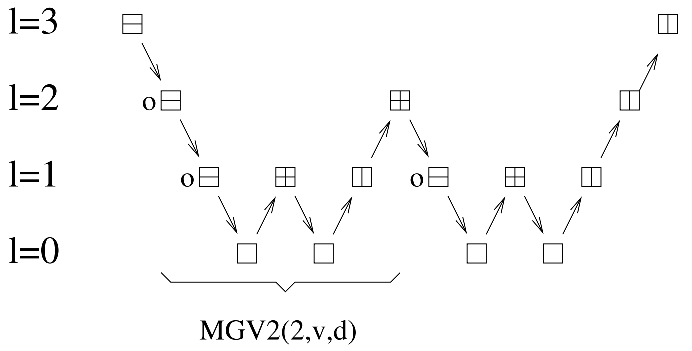


MGV2(2,u,b)

Abb. 4

Es wurden $\gamma = 2$ Schritte zur Konstruktion der Approximation für die Grobgitterlösung benutzt.

Beispiel: l=3: Die Lösung der Korrekturgleichung auf $l = 2$ wird durch $\gamma = 2$ Schritte des obigen 3-Gitter-Verfahrens approximiert.



MGV2(3,u,b)

Abb. 5

Beispiel: l=4: Die Korrekturgleichung auf $l = 3$ wird approximativ gelöst durch $\gamma = 2$ Schritte des vorigen 4-Gitter-Verfahrens.

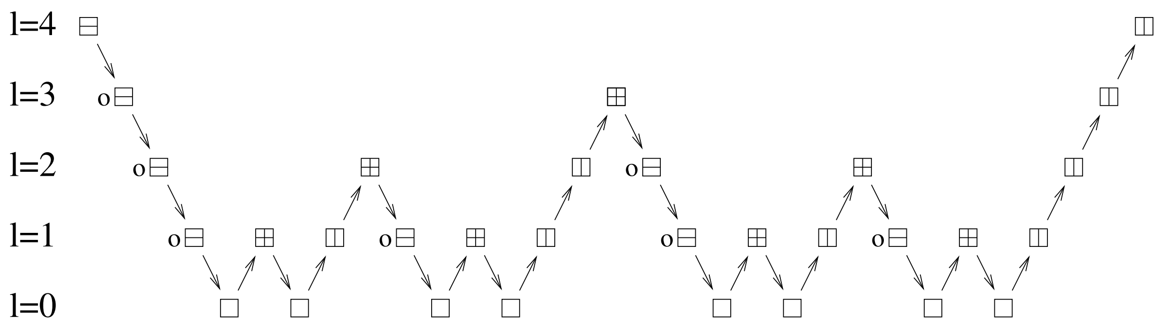


Abb. 6

Volles Mehrgitter-Verfahren

(auch *nested iteration*)

Bisher starteten wir mit einer exakten Lösung auf dem gröbsten Gitter, $\boldsymbol{A}_0\boldsymbol{u}_0 = \boldsymbol{b}_0$, die wir durch mehrere Prolongationen auf das feinste Gitter hochhoben. Die so erhaltene Ausgangsnäherung für das MGV ist im Allgemeinen nicht sehr gut, weshalb zu viele Zyklen des eigentlichen MGV durchgeführt werden müssen. Der Aufwand wird reduziert, wenn man sich eine bessere Ausgangsnäherung beschafft, wieder mit Hilfe des MGV. Natürlich starten wir mit einer Lösung von $\boldsymbol{A}_0\boldsymbol{u}_0 = \boldsymbol{b}_0$, die wir auf das nächst feinere Gitter prolongieren. Statt jedoch weiter zu prolongieren wird sie durch μ_1 Schritte eines MGV verbessert, erst dann wird prolongiert und dann wieder verbessert, usw.

Der Prozess läßt sich algorithmisch wie folgt beschreiben:

$\widetilde{\boldsymbol{u}}_0 := \boldsymbol{A}_0^{-1} * \boldsymbol{b}_0$ Lösung auf dem gröbsten Gitter
for $l = 1$ step 1 until l_{max} do
begin $\widetilde{\boldsymbol{u}}_l := \widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{u}}_{l-1}$; Prolongation: $\widetilde{\boldsymbol{P}} = \boldsymbol{P}$ möglich, $\widetilde{\boldsymbol{P}}$ "besser" als $\boldsymbol{P}$ ist manchmal wünschenswert
for $j = l$ step 1 until μ_1 do MGV $\gamma(l, \widetilde{\boldsymbol{u}}_l, \boldsymbol{b}_l)$;
end;

Auf jedem Level $k = 1, ..., l_{max}$ werden also μ_1 Iterationen durchgeführt, d.h. die jeweils aktuelle Hauptgleichung wird μ_1 mal verbessert. Der Startpunkt kommt vom gröbsten Gitter und wird auf jedem Gitter verbessert.Der dicke Pfeil (in Abb. 5) bezeichnet den Einsatz einer möglicherweise besseren Prolongation. Wir beschreiben den Verlauf symbolisch für das

Beispiel: $l_{max} = 3,\ \mu_1 = 2,\ \gamma = 2;$

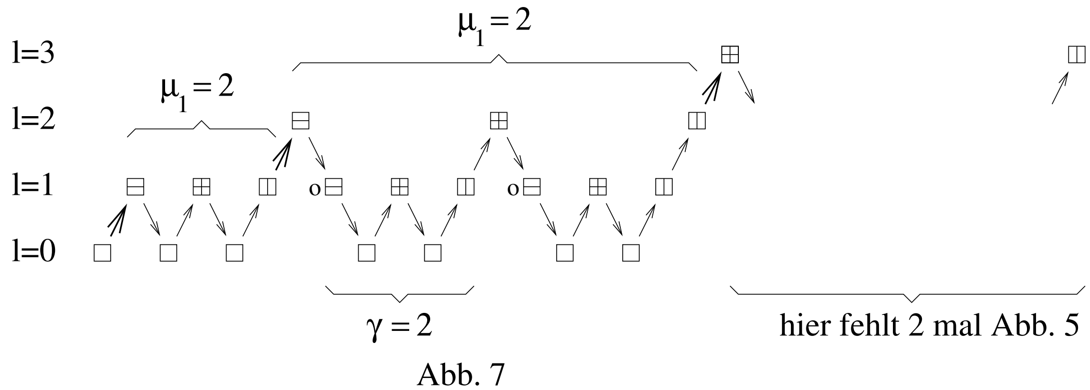


Abb. 7

Als volles Mehrgitterverfahren (nested iteration) bezeichnet man die Kombination der obigen Anlauf-Iteration mit dem MGV. Betrachtet man obige Anlaufiteration und anschließend μ_2 Zyklen des eigentlichen MGV, so kann man das wie folgt beschreiben:

Bekannt seien $(\boldsymbol{A}_l)_{l=0}^{l_{max}}$, $(\boldsymbol{b}_l)_{l=0}^{l_{max}}$, gesucht sei die Lösung von $\boldsymbol{A}_{l_{max}}\boldsymbol{u} = \boldsymbol{b}_{l_{max}}$.

$\boldsymbol{u}_0 := \boldsymbol{A}_0^{-1}\boldsymbol{b}_0$

for $l = 1$ step l until l_{max} do

begin $\boldsymbol{u}_l := \widetilde{\boldsymbol{P}}\boldsymbol{u}_{l-1}$;

$\mu :=$ if $l = l_{max}$ then μ_2 else μ_1;

for $j = 1$ step 1 until μ do MGV $\gamma(l, \widetilde{\boldsymbol{u}}_l, \boldsymbol{b}_l)$;

end;

Man hangelt sich also mit jeweils μ_1 Schritten MGV hoch bis l_{max} und rechnet dann μ_2 Zyklen von MGV. Ist l die aktuelle Stufe, so wird γ mal die Korrekturgleichung (also auf Stufe $l-1$) durch ein MGV1 gelöst.

Bemerkungen zu den Iterationszahlen

Üblich sind $\nu = \nu_1 + \nu_2 \leq 3$, wenn man die Anzahl ν der Glättungen in Vor- und Nachglättung aufteilt. Vernünftig sind 1-2 Vorglättungen und eine Nachglättung zum Glätten höherer Frequenzen, die von der Prolongation kommen könnten.
$\mu_1 \leq 2 \leq \mu_2 \leq 4$, die beste Wahl ist problemabhängig, $\gamma = 2$ ist normal.
Die Nachglättung wird in MGV1 eingebaut als letzter Schritt.

Geschichtlicher Überblick

(genaueres bei Hackbusch 2.6.5)

1961 Federenko: 1) Glättung, 2) Grobgitterkorrektur

1964 Federenko: Konvergenzbeweis für Laplace im Einheitsquadrat mit Hilfe der Fourier-Analyse (Entwicklung nach Eigenvektoren)

1966 Bachvalov: kompletterKonvergenzbeweis für das volle MGV für allgemeine elliptische Operatoren inclusive Nachweis der Optimalität bzgl. des Rechenaufwandes

1974-1975 Achi Brand: Erfuhr auf Umweg über Amerika von Schülern der Moskauer Schule die Idee des Verfahrens, hat es weiterentwickelt und auf nichttriviale Proleme der amerikanischen Navy angewandt, ohne auf Konvergenzbetrachtungen einzugehen.

1976 Hackbusch: Unabhängige Neuerfindung des MGV. Von ihm stammt das erste Buch über MGV.

1989 erstes russisches Buch über MGV.

Zum Paragraphenabschluß zitieren wir aus dem Buch von A. Iserless: A first course in the numerical Analysis for Differential Equations (Cambridge University Press 1996): The number of different strategies and implementations of multigrid is a source of major preoccupation to professionels, allthough it might be at times baffling to other numerical analysts and to users of computational algorithms.

§ 17 Glättung, Restriktion, Prolongation

Wir befassen uns erst mit dem Problem, den Glättungseffekt zu "messen". In der Literatur werden dazu zwei Möglichkeiten betrachtet.

1) Die Glättungsrate (eingeführt von A.Brand 1977)
Unsere Glättungsüberlegungen beruhen auf der in § 16 geschilderten Grundidee.
Sei n_l die Anzahl der Gitterpunkte (=Anzahl der Eigenwerte) des l-ten Gitters und n_{l-1} die des nächst gröberen Gitters.
Bei der Verdoppelung der Maschenweite gilt (vgl. S. 125)

im 1D-Fall: $n_{l-1} \approx \frac{n_l}{2}$,
im 2D-Fall: $n_{l-1} \approx \frac{n_l}{4}$.

Erinnerung: Die Eigenvektorentwicklung der Fehlerdarstellung (vgl. (15.2),(16.9)) war eine Entwicklung nach den Eigenvektoren der Iterationsmatrix $\boldsymbol{S}$. Die Iterationsmatrizen haben die Gestalt $\boldsymbol{S} = \boldsymbol{I} - \omega \boldsymbol{MA}$, wo $\boldsymbol{A}$ die Diskretisietungsmatrix ist und z.B. $M = \boldsymbol{D}^{-1}$. Sind λ_i die Eigenwerte von $\boldsymbol{MA}$ zu den Eigenvektoren $\boldsymbol{v}^{(i)}$, so hat die Iterationsmatrix die Eigenwerte $(1 - \omega\lambda_i)$ zu den Eigenvektoren $\boldsymbol{v}^{(i)}$.
Die Eigenvektoren $\boldsymbol{v}_l^{(\mu)}$ der Iterationsmatrixsmatrix wurden eingeteilt in

niedere Frequenzen: $1 \leq \mu \leq n_{l-1}$ und
hohe Frequenzen $n_{l-1} + 1 \leq \mu \leq n_l$.

Sei σ_μ der μ-te Eigenwert der Iterationsmatrix $\boldsymbol{S}_l$ des Dämpfungsverfahrens, also $\boldsymbol{S}_l \boldsymbol{v}_l^{(\mu)} = \sigma_\mu \boldsymbol{v}_l^{(\mu)}$, so definiert man für ν-fache Glättung (wenn l=1 das gröbste Gitter bezeichnet) die

$$\text{(17.1)} \qquad \textit{Glättungsrate} \quad \rho_B = \sup_{l \geq 1} \max\{|\sigma_\mu|^\nu; n_{l-1} + 1 \leq \mu \leq n_l\}$$

Beachte: Es ist (sollte sein): $|\sigma_\mu| < 1$
Im praktischen Fall ist ihre Bestimmung naturgemäß mit einigem Aufwand verbunden
Einen etwas anderen Zugang liefert die

2) Die Glättungsnummer (Hackbusch)
Die Glättung einer Näherungslösung $\boldsymbol{u}_l^j$ für $\boldsymbol{A}_l \boldsymbol{u}_l = \boldsymbol{b}_l$ wird durch eine Glättungsiteration beschrieben (z.B. gedämpfter Jacobi oder Gauß-Seidel)

$$\bar{\boldsymbol{u}}_l = \boldsymbol{S}\boldsymbol{u}_l^j + \boldsymbol{T}\boldsymbol{b}_l, \qquad \boldsymbol{S} = \text{ Iterationsmatrix}$$

$\boldsymbol{T}\boldsymbol{b}_l$ bezeichnet in der Iteration den Restsummanden, der nicht von der Iterationsmatrix beeinflußt wird.
(Beispiel: gedämpfter Jacobi, falls Diagonale zu 1 normiert: $\boldsymbol{S} = (\boldsymbol{I} - \omega \boldsymbol{A})$.)

Die exakte Lösung ist immer ein Fixpunkt der Iteration

$$\boldsymbol{u}_l = \boldsymbol{S}\boldsymbol{u}_l^j + \boldsymbol{T}\boldsymbol{b}_l.$$

Nur Verfahren mit dieser Eigenschaft werden zur Glättung benutzt.

Geglättet wird nicht die Näherungslösung $\boldsymbol{u}^j$, sondern ihr Fehler $\boldsymbol{e}^j$ (vgl. (16.9))

$$\bar{\boldsymbol{e}} := \bar{\boldsymbol{u}}_l - \boldsymbol{u}_l = \boldsymbol{S}(\underbrace{\boldsymbol{u}_l^j - \boldsymbol{u}_l}_{\boldsymbol{e}^j}) =: \boldsymbol{S}\boldsymbol{e}^j$$

bzw. bei mehrfacher Glättung

$$\bar{\boldsymbol{e}}^j = \boldsymbol{S}^\nu \boldsymbol{e}^j.$$

Ein Fehlervektor $\boldsymbol{v} = (v_1, ..., v_{n_l})^T$ kann im 1D-Fall anschaulich als glatt bezeichnet werden ("glatt" im Sine "geringes Oszillieren"), wenn eine Norm der Differenz "benachbarter Steigungen": $\dfrac{v_{j+1} - v_j}{h} - \dfrac{v_j - v_{j-1}}{h}$ klein ausfällt.
Ein spezieller Differenzenoperator 2. Ordnung, der diese Differenzen beschreibt, wird durch die Diskretisierungsmatrix von Δu gegeben. Wir bezeichnen sie hier mit $\boldsymbol{L}_l$.

$$(\boldsymbol{L}_l \boldsymbol{v})_i = \frac{1}{h}\left(\frac{v_{j+1} - v_j}{h} - \frac{v_j - v_{j-1}}{h}\right) \quad \text{im 1D-Fall.}$$

Zum 2D-Fall vgl. (12.2).

Die Norm der Steigungsdifferenzen des geglätteten Fehlervektors

$$\|\boldsymbol{L}_l \bar{\boldsymbol{e}}^j\| = \|\boldsymbol{L}_l \boldsymbol{S}^\nu \boldsymbol{e}^j\| \leq \|\boldsymbol{L}_l \boldsymbol{S}^\nu\| \, \|\boldsymbol{e}^j\|$$

sollte klein ausfallen im Vergleich zur Norm der Steigungsdifferenzen des ungeglätteten Fehlers

$$\|\boldsymbol{L}_l \boldsymbol{e}^j\| \leq \|\boldsymbol{L}_l\| \, \|\boldsymbol{e}^j\|$$

Wir motivieren so (wieder sei l=1 das gröbste Gitter) die

(17.2) *Glättungsnummer* $\displaystyle \rho_L(\nu) = \sup_{l \geq 1} \frac{\|\boldsymbol{L}_l \boldsymbol{S}^\nu\|}{\|\boldsymbol{L}_l\|},$

die möglichst klein ausfallen sollte. (Beispiele werde wir in den Übungen kennen lernen.) Die Glättungsnummer ist normabhängig. Das wird im Konvergenzbeweis eine Rolle spielen.

Glättungsiterationen

Als Glättungsverfahren werden benutzt

1. Gauß-Seidel (Standard),

2. gedämpftes Jacobi-Vefahren: Iterationsmatrix: $\boldsymbol{I} - \omega \boldsymbol{D}^{-1} \boldsymbol{A}$,

3. Tschebyscheff-Verfahren, Iterationsmatrix: $\boldsymbol{I} - \omega_i \boldsymbol{A}$ (in den einzelnen Iterationsschritten werden verschiedene ω_i benutzt),

4. SOR klappt nicht, da nur auf Konvergenz ausgerichtet, Dämpfung nicht möglich

5. Konjugierte Gradienten-Verfahren: Da der Aufwand relativ hoch ist, ist es nur sinnvoll als äußere Iteration mit MGV als Präkonditionierung. Ansonsten ist es gut (insbesondere bei dünn besetzten Matrizen, wenn man keine Glättungsiteration finden kann (z.B. bei Navier-Stokes-Gleichungen)).

Wir kennen bereits das **gedämpfte Jacobi-Verfahren**.

Beachte:

1. Dieses Verfahren ist unabhängig von der Nummerierung der Unbekannten.

2. Auf Grund seiner Struktur läßt es sich leicht parallelisieren (z.B. pro Zeile ein Prozessor).

Dieses Verfahren läßt sich schreiben als

$$\boldsymbol{u}_l^{j+1} = \boldsymbol{u}_l^j - \omega \boldsymbol{D}^{-1}(\boldsymbol{A}_l \boldsymbol{u}_l^j - \boldsymbol{b}_l) \tag{17.3}$$

Die Bestimmung eines optimalen Dämpfungsparameters ist im allgemeinen Fall schwierig. In vielen Anwendungen sind die Diagonaleinträge von $\boldsymbol{A}$ von der Art $\frac{k}{h^{2m}}$, $k = const$, wo $2m$ die Ordnung des diskreten Differentialoperators ist. Man untersucht leichter (im Hinblick auf die Glättungseigenschaften) statt (17.3) ein Verfahren der Art

$$\boldsymbol{u}_l^{j+1} = \boldsymbol{u}_l^j - \omega \frac{h_l^{2m}}{k}(\boldsymbol{A}_l \boldsymbol{u}_l^j - \boldsymbol{b}_l)$$

Ist $\operatorname{diag} \boldsymbol{A}_l = \operatorname{diag}(\frac{k}{h^{2m}})$, so ist ein beliebtes, weil bequem zu berechenbares ω, das gute Dämpfungseigenschaften liefert,

$$\omega = \omega_l = \frac{k}{\|h_l^{2m} \boldsymbol{A}_l\|_S} \qquad \text{(Spektralnorm)} \tag{17.4}$$

Wir zeigen dies im nächsten Paragraphen (vgl. Satz 18.2 und (18.29))

Das am meisten benutzte Verfahren ist

die Gauß-Seidel-Iteration (Einzelschritt-Verfahren)

Man kann zeigen, daß sie für eine große Klasse von Matrizen (nicht für alle) schneller konvergiert als die Jacobi-Iteration (vgl. Satz 18.4,Satz von Stein-Rosenberg (zu finden in Varga: Matrix Iterative Analysis, Prentice Hall 1962, vgl. dazu auch die Konvergenzbeispiele in den Übungen)
Zudem sind die Glättungseigenschaften besser als die des gedämpften Jacobi (vgl. Übungen).

Ein Iterationsschritt des Verfahrens zur Lösung der Gleichung $\boldsymbol{Ax} = \boldsymbol{b}$, $\boldsymbol{A} = (a_{ij})$, kann algorithmisch beschrieben werden durch

(17.5) $$for\ i = 1\ step\ 1\ until\ n_l\ do\ x_i := \frac{-1}{a_{ii}} \left(\sum_{j=1, j\neq i}^{n_l} a_{ij} x_j - b_j \right)$$

Man beachte, daß in jedem Schritt die "alten" Komponenten des Vektors durch die neuen überschrieben werden (Speicherersparnis gegenüber Jacobi).

Beachte:

1. Dieses Verfahren ist abhängig von der Reihenfolge, in der die Unbekannten iterativ verbessert werden (Anordnung der Gitterpunkte).
2. Ein Iterationsschritt durchläuft "sequentiell" die Zeilen des Systems und kann deshalb nicht in dieser Form parallelisiert werden,
   jedoch, eine geschickte Anordnung der Zeilen des Systems, d.h. eine vorgeschriebene Ordnung, in der die Gitterpunkte abgearbeitet werden, kann die Glättungseigenschaften gewährleisten, sogar verbessern, und abhängig von der Diskretisierungsmatrix, eine gewisse Parallelisierung ermöglichen.

Parallelisierung

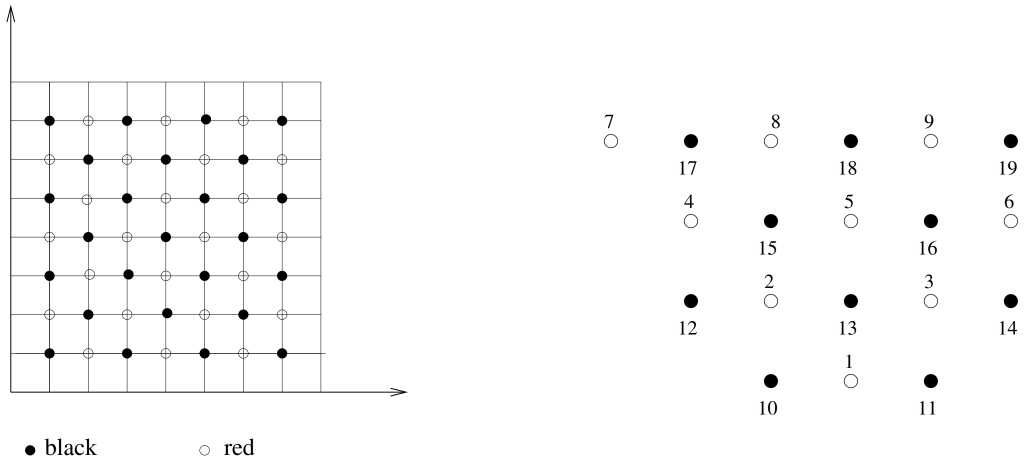
Wir besprechen einige der gebräuchlichsten und erfolgreichsten Anordnungen der Gitterpunkte, die in Verbindung mit dem Gauß-Seidel-Verfahren für die Diskretisierung von Δu gute Dämpfungseigenschaften liefern und eine Parallelisierung ermöglichen. Wir beschränken uns auf den 2D-Fall.

1. Schachbrettanordnung: schwarz-weiß oder weiß-schwarz (Chequer bord ordering, red-black ordering),
2. 4-Farben-Ordnung (Four-colour-ordering),
3. zeilen- oder spaltenweise Anordnung (lexicographical ordering oder rotated lexicographical ordering),

Für weitere Möglichkeiten vgl. etwa Hackbusch: Multigrid Methods, § 3.3 und 6.2; oder
Großmann/Roos: Numerik partieller Differentialgleichungen, § 5.2, 5.6

Diese Anordnungen betreffen jeweils eine Einteilung der Gitterpunkte in Klassen, die im Iterationsverfahren einzeln abgearbeitet werden. Die Klasseneinteilung ist abhängig von der Diskretisierungsmatrix, wie wir sehen werden.

Schachbrettanordnung (red-black): Gut geeignet für den 5-Punktestern für Δu. Wir illustrieren die Klasseneinteilung für das Einheitsquadrat und die Nummerierung der Punkte am Beispiel eines beliebigen Gebiets.



Zuerst werden die roten, dann die schwarzen Punkte zeilenmäßig von unten nach oben durchgezählt. Anfangspunkt ist der erste rote Punkt (red-black).

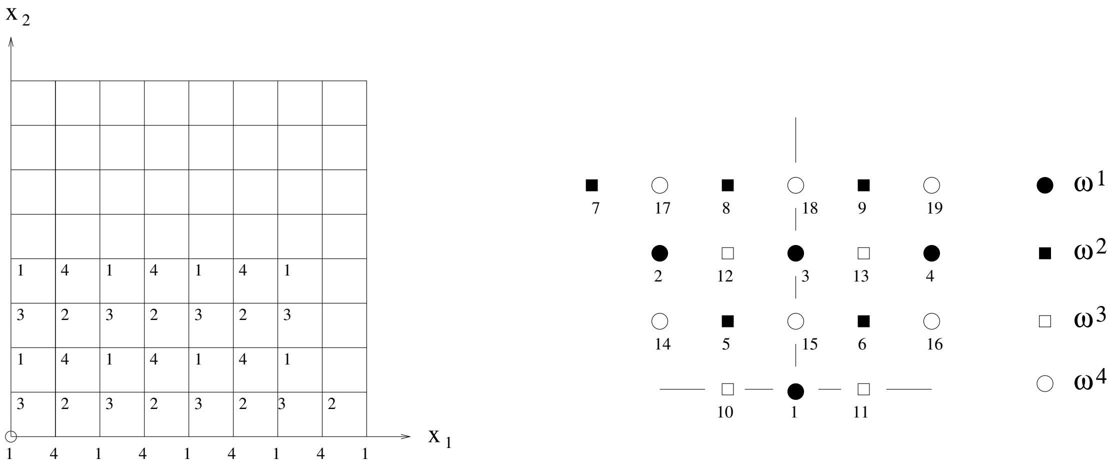
Parallelisierung

Wir führen das GS-Verfahren (17.5) zunächst für die roten Punkte durch. Die iterative Verbesserung durch das GS für einen "roten"Punkt (rote Unbekannte) benötigt nur schwarze Nachbarpunkte (vgl. obige Abbildungen und den 5-Punktestern (12.5)), d.h. innerhalb der roten Punkteklasse ist die Reihenfolge der Punkte beim GS-Verfahren beliebig.
Dies kann zur Parallelisierung benutzt werden: z.B. einen extra Prozessor für jede Zeile zur Verbesserung der Punkte der roten Klasse, die in dieser Zeile enthalten sind.

Nachdem alle roten Komponenten verbessert wurden, arbeitet man die schwarzen Komponenten ab, deren Verbesserung nur von den (schon verbesserten) roten Punkten abhängt. Für die Parallelisierung gilt das oben Gesagte.
Welche der Anordnungen bessere Glättungseigenschaften zeigt (red-black oder black-red), ist problemabhängig.
Beachte: Wir haben hier die Parallelisierung für Δu besprochen. Nicht alle Diskretisierungsmatrizen müssen eine solche Einteilung erlauben. Muß z.B. eine Ableitung höherer Ordnung diskretisiert werden, so benötigt dies üblicherweise mehr als nur die beiden unmittelbaren Gitternachbarn. Dann ist die beschriebene Parallelisierung nicht mehr möglich.

Sehr beliebt (auch für den 5-Punkte-Stern) ist auch die

Vierfarben-Ordnung: Gut geeignet für den 9-Punkte-Stern für Δu.



Die Punkte werden in 4 Klassen $\omega^1, ..., \omega^4,$ eingeteilt. Im Einheitsquadrat bezeichnen wir die Klasseneinteilung der Punkte durch Zahlen 1 bis 4.
Beim allgemeinen Gebiet führen wir zur Beschreibung ein zweites Koordinatenkreuz ein, dessen Ursprung bei zeilenweiser Nummerierung von unten nach oben im 2.ten Gitterpunkt liegt. Die einzelnen Klassen sind durch verschiedene Symbole gekennzeichnet. Die Nummerierung gibt die Reihenfolge an, in der die Punkte verbessert werden.
Wir beschreiben die Klassen wie folgt:

- $\omega^1 = \{(\nu h, \mu h);\ \ \nu, \mu$ gerade$\}$

■ $\omega^2 = \{(\nu h, \mu h);\ \ \nu, \mu$ ungerade$\}$

□ $\omega^3 = \{(\nu h, \mu h);\ \ \ \nu$ gerade, μ ungerade$\}$

○ $\omega^4 = \{(\nu h, \mu h);\ \ \ \nu$ ungerade, μ gerade$\}$

Parallelisierung:

Innerhalb ω^1 werden nur die zu ω^1 gehörigen Komponenten verbessert. Die Verbesserung der Punkte $\in \omega^1$ mit dem 9-Punktestern benötigt nur Punkte aus $\omega^2, \omega^3, \omega^4$. Die Reihenfolge der Nummerierung innerhalb ω^1 ist dann bedeutungslos. Dasselbe gilt für die anderen Klassen.
Beispiel: Zuerst werden alle ω^1-Punkte verbessert (z.B. ein Prozessor für jede Zeile, in der Punkte $\in \omega$ vorkommen), dann (wieder parallel) alle ω^2 -Punkte mit den verbesserten Werten aus ω^1. Die Verbesserung der Punkte $\in \omega^2$ benötigt nur Punkte aus $\omega^1, \omega^3, \omega^4,$ usw.
Die Vierfarben-Ordnung wird auch in Verbindung mit einem 5-Punkte-Stern angewandt (vgl. Hackbusch: Abschnitt 3.3.3 und Exercise 3.9.2)

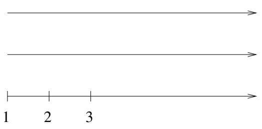
Wir erwähnen weiter die

Lexikographische Ordnung: In jeder Zeile von links nach rechts und zeilenweise von unten nach oben oder
rotierte Lexikographische Ordnung: In jeder Spalte von unten nach oben und spaltenweise von links nach rechts.

Beispiel: stationäre Strömungsaufgaben. Bei der Gleichung

$$\nu\Delta u - v\,\mathrm{grad}\,u + f = 0, \quad v > 0.$$

läuft die Stömung von links nach rechts.



Zeilenweise Numerierung: gegen die Strömung zu rechnen oder senkrecht zur Strömung ist hier nicht sinnvoll. In diesem Beispiel verteilt sich die Information von links nach rechts.

Die Anordnung der Punkte in der Zeichnung ist sinnvoll, da der Zustand in “2” nur vom Zustand in “1” beeinflusst wird.
Ist der Diffusionskoeffizient $\nu = 0$, so liefert das Verfahren die exakte Lösung, da es auf die Transportgleichung reduziert ist.

Man erkennt die Abhängigkeit der Wahl von der Anwendung.

Ein Spezialfall der lexikographischen Anordnung im Hinblick auf Parallelisierung beim obigen Beispiel wäre etwa die
zebra-line-ordering: 1.te Klasse: die Zeilen 1,3,5,...;
2.te Klasse: die Zeilen 2,4,6,...;
Auch red-black mit dem 5-Punktestern und zentralen Differenzen für die $\frac{\partial u}{\partial x_i}$ (und dann wieder zeilenweise) wäre möglich.

Für weitere Anordnungen siehe Hackbusch: Multigrid Methods.

Prolongation und Restriktion

Vernünftigerweise gibt es genau ein Vorgehen für die Prolongation, nämlich die Interpolation. Dabei hat sich gezeigt, daß für Gleichungen vom Grad $2m$ eine Interpolation der Ordnung m genügt. Wir beschränken uns hier auf den 2D-Fall und Δu, d.h. also, lineare Interpolation. Man beachte, daß auf Grund des GSV immer Defekte restringiert und prolongiert werden. Defekte haben (bei vorgegebenen Dirichlet-Randwerten, auf die wir uns hier beschränken,) immer Nullrandwerte. Dies muß bei der Prolongation berücksichtigt werden. Wir beschränken uns hier, der einfacheren Schreibweise wegen, auf den Fall gleicher Maschenweiten $h_1 = h_2 = h$. Das Vorgehen für $h_1 \neq h_2$ ist völlig analog.

Im Einheitsquadrat bezeichnen wir
Grobgitterpunkte durch $\odot$,
Feingitterpunkte nur durch einen Punkt,
$2h\hat{=}$Grobgittermaschenweite,
$h\hat{=}$ Feingittermaschenweite.
Es seien
$(0,0),(0,2h),(2h,2h),(2h,0) \in \omega_l \cap \omega_{l-1}$

$$
\begin{array}{lllllllll}
x_2 \\
\uparrow \\
\odot & \cdot & \odot & \cdot & \odot & \cdot & \odot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\odot & \cdot & \odot & \cdot & \odot & \cdot & \odot \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
\odot & \cdot & \odot & \cdot & \odot & \cdot & \odot & \rightarrow x_1
\end{array}
$$

Dann kann man die Prolongation einer auf ω_{l-1} (=Gitterpunkte auf Gitter $l-1$) definierten Funktion v (Lösung der Korrekturgleichung) nach ω_l wie folgt beschreiben:

9-Punkte-Prolongation
Auf dem Grobgitter bleiben die Werte unverändert.

$$
\begin{array}{lllll}
v_l(0,0) &=& v_{l-1}(0,0), & v_l(0,2h) &=& v_{l-1}(0,2h); \\
v_l(2h,0) &=& v_{l-1}(2h,0), & v_l(2h,2h) &=& v_{l-1}(2h,2h);
\end{array} \tag{17.6}
$$

Für die Feingitterpunkte, die zwischen 2 Grobgitterpunkten liegen, liefert die lineare Interpolation

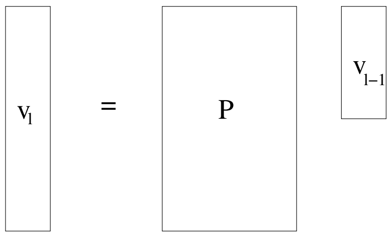
$$
\begin{array}{lll}
v_l(0,h) &=& \frac{1}{2}v_{l-1}(0,0) + \frac{1}{2}v_{l-1}(0,2h) \\
v_l(h,0) &=& \frac{1}{2}v_{l-1}(0,0) + \frac{1}{2}v_{l-1}(2h,0) \\
v_l(2h,h) &=& \frac{1}{2}v_{l-1}(2h,0) + \frac{1}{2}v_{l-1}(2h,2h) \\
v_l(h,2h) &=& \frac{1}{2}v_{l-1}(0,2h) + \frac{1}{2}v_{l-1}(2h,2h)
\end{array} \tag{17.7}
$$

und schließlich

$$
\begin{aligned}
v_l(h,h) &= \frac{1}{2}v_{l-1}(0,h) + \frac{1}{2}v_{l-1}(2h,h) \\
&= \frac{1}{4}v_{l-1}(0,0) + \frac{1}{4}v_{l-1}(0,2h) + \frac{1}{4}v_{l-1}(2h,0) + \frac{1}{4}v_{l-1}(2h,2h)
\end{aligned} \tag{17.8}
$$

Die Formeln (17.6) – (17.8) heißen 9-Punkte-Prolongation.
Natürlich kann man diese Formeln in Matrixnotation aufschreiben.

$$v_l = P v_{l-1}$$



Billiger und Speicherplatz sparender ist jedoch, die prolongierten Werte direkt zu berechnen.
Wir demonstrieren dies für den Fall des Einheitsquadrats $\bar{\Omega}$. Die Randpunkte müssen für die Interpolation mitgeführt werden. Dann haben wir das

$$\text{Grobgitter:}\qquad \omega_{l-1} = \{(\nu h_{l-1}, \mu h_{l-1}) \in \bar{\Omega};\ 0 \le \nu, \mu \le \tfrac{1}{h_{l-1}}\},$$

$$\text{Feingitter:}\qquad \omega_l = \{(\nu h_l, \mu h_l) \in \bar{\Omega}; 0 \le \nu, \mu \le \tfrac{1}{h_l}\}, \quad h_{l-1} = 2h_l.$$

Beachte: Bei Dirichlet-Randwerten sind die Defekte in den Randwerten $=0$, deshalb werden die Randpunkte mit $v = 0$ vorbesetzt.
Dann wird, unter der Voraussetzung, daß v_{l-1} bekannt ist, die Prolongation beschrieben durch

for $x := 0$ step $2h_l$ until 1 do for $y = h_l$ step $2h_l$ until $1 - h_l$ do

$$v(x,y) := [v(x, y-h_l) + v(x, y+h_l)]/2;$$

for $y := 0$ step h_l until 1 do for $x = h_l$ step $2h_l$ until $1 - h_l$ do

$$v(x,y) := [v(x-h_l, y) + v(x+h_l, y)]/2;$$

d.h. zuerst werden die Feingitterpunkte unter- und oberhalb der Grobgitterpunkte (also in vertikaler Richtung) besetzt, danach die restlichen.

Natürlich sollen Interpolation und Restriktion sich "irgendwie entsprechen". Der Konvergenzbeweis zeigt, daß dies der Fall ist, wenn sie zueinander adjungiert sind. Natürlich kann auch funktionalanalytisch begründet werden, warum das sinnvoll ist.

Definition 17.1 Adjungierte Operatoren
Seien H_H und H_h Hilberträume mit den inneren Produkten $(\ ,\ )_H$ und $(\ ,\ )_h$, und

$$\boldsymbol{R}:\ H_h \xrightarrow{auf} H_H$$

eine lineare Abbildung.
Die Abbildung

$$\boldsymbol{P}:\ H_H \xrightarrow{linear} H_h$$

heißt adjungiert zu $\boldsymbol{R}$, falls gilt

$$(\boldsymbol{P}\boldsymbol{u}^H, \boldsymbol{v}^h)_h = (\boldsymbol{u}^H, \boldsymbol{R}\boldsymbol{v}^h)_H \quad \forall\, \boldsymbol{u}^H \in H,\ \ \boldsymbol{v}^h \in H_h). \tag{17.9}$$

In unserem Fall sind $H_H = \mathbb{R}^{n_{l-1}}$, $H_h = \mathbb{R}^{n_l}$, dabei sind n_{l-1} bzw. n_l die Anzahl der Grobgitter- bzw. Feingitterpunkte im Einheitsquadrat inclusive der Randpunkte.

$$(\boldsymbol{u}^h, \boldsymbol{v}^h)_h = \sum_{j=1}^{n_l} \boldsymbol{u}_j^h \boldsymbol{v}_j^h h^{(2)}, \qquad h^{(2)} = h_1 h_2, \qquad h_i = \text{Maschenweite in } x_i\text{-Richtung},$$

$$(\boldsymbol{u}^H, \boldsymbol{v}^H)_H = \sum_{i=1}^{n_{l-1}} \boldsymbol{u}_i^H \boldsymbol{v}_i^H H^{(2)} \qquad H^{(2)} = H_1 H_2, \qquad H_i = \text{Maschenweite in } x_i\text{-Richtung}.$$

Die 2 im Exponenten ist ein Hinweis auf den 2-dimensionalen Fall.

$\boldsymbol{P}$ und $\boldsymbol{R}$ sind Prolongation und Restriktion, wir setzen $H_i = 2h_i$. (Das ist Standart, aber kein “Muß”.) Dann lautet (17.9) in Matrixschreibweise

$$(\boldsymbol{P}\boldsymbol{u}^H)^T\boldsymbol{v}^h h^{(2)} = (\boldsymbol{u}^H)^T\boldsymbol{P}^T\boldsymbol{v}^h h^{(2)} \overset{(17.9)}{=} (\boldsymbol{u}^H)^T\boldsymbol{R}\boldsymbol{v}^h \cdot H^{(2)} = (\boldsymbol{u}^H)^T\boldsymbol{R}\boldsymbol{v}_h \cdot 4h^{(2)}$$

oder

$$(\boldsymbol{u}^H)^T\boldsymbol{P}^T\boldsymbol{v}^h = (\boldsymbol{u}^H)^T\boldsymbol{R}\boldsymbol{v}^h \cdot 4.$$

Dies liefert als Konstruktionsvorschrift für eine Restriktionsmatrix

$$\boldsymbol{R} = \frac{1}{4}\boldsymbol{P}^T. \tag{17.10}$$

Benutzt man die 9-Punkte-Prolongation, so erhält man als Restriktion im 2D-Fall die

9-Punkte-Restriktion (Standardvariante) für den Defekt, (folgt aus (17.10) mit der Prolongation (17.6)-(17.8)) (Übung)

$$\frac{1}{16}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}$$

$\bigcirc$ ist der Grobgitterpunkt, die Zahlen zeigen die Gewichte der Defekte d_h in den einzelnen Punkten. (Konvexkombination der beteiligten Punkte)

$$\begin{aligned} d^H_{ij} = &\frac{1}{16}\{d^h_{i+1,j+1} + d^h_{i-1,j+1} + d^h_{i-1,j-1} + d^h_{i+1,j-1} \\ &+ 2(d^h_{i,j+1} + d^h_{i,j-1} + d^h_{i+1,j} + d^h_{i-1,j}) \\ &+ 4d^h_{i,j}\} \end{aligned} \tag{17.11}$$

Bemerkung: Beim FEM-Verfahren (Finite Element Method) sind Restiktion und Prolongation vorgeschrieben. Beim Differenzenverfahren hat man Wahlmöglichkeiten.

Aufgabe:

1. Man beschreibe die 9-Punkte-Prolongationsmatrix und leite daraus die zugehörige 9-Punkte-Restriktion ab. (inclusive Aufstellung der Restriktionsmatix)

2. Wie sieht die Restriktionsmatrix für die 5-Punkte-Restriktion aus?

   $$\frac{1}{8}\begin{pmatrix} & 1 & \\ 1 & 4 & 1 \\ & 1 & \end{pmatrix}$$

   Wie sieht die zugehörige Prolongationsmatrix aus? Beschreibt sie eine stückweis lineare Interpolation?

3. Statt (17.8) könnte man auch wählen

   $$v_l(h,h) = \frac{1}{2}v_{l-1}(0,0) + \frac{1}{2}v_{l-1}(2h,2h)$$

   Wie sehen die zugehörige Prolongations- und Restriktionsmatrix aus?

Bemerkung: Es liegt nahe, als Restriktion die triviale Injektion zu wählen , d.h.

$$(\boldsymbol{R}\boldsymbol{d}_l)(\boldsymbol{x}) = \boldsymbol{d}_l(\boldsymbol{x}) \qquad \boldsymbol{x} \in \omega_{l-1} \subset \omega_l.$$

Man kann zeigen (vgl. Hackbusch 3.5), daß für diese Restriktion in Verbindung mit der red-black-Ordnung, dem GS-Verfahren und dem 5-Punkte-Stern für Δu die Iteration zum Stehen kommen kann.

§ 18 Konvergenz des ZGV

Idee: Zur approximativen Lösung von $\boldsymbol{A}_k \boldsymbol{y}_k = \boldsymbol{b}_k$ mit dem ZGV leiten wir (wie beim Jacobi-Verfahren) für den Fehler $\boldsymbol{e}^m$ der $m-$ ten Iteration mit der Ausgangsnäherung $\boldsymbol{y}_h^0$ eine Fehlergleichung her mit einer noch aufzustellenden Iterationsmatrix $\boldsymbol{K}_h$

$$\boldsymbol{e}_h^{m+1} = \boldsymbol{K}_h \boldsymbol{e}^m, \quad \boldsymbol{e}_h^m = \boldsymbol{y}_h^m - \boldsymbol{y}_h^0, \quad \boldsymbol{K}_h = \text{Iterationsmatrix des ZGV.}$$

Konvergenz liegt vor, wenn der Spektralradius $\rho(\boldsymbol{K}_h)$ kleiner als 1 ausfällt.

Zur Herleitung der Fehlerdarstellung betrachten wir folgendes ZGV (mit Vor- und Nachglättung

Gegeben: Ausgangsnäherung $\boldsymbol{y}_h^0$ für $\boldsymbol{A}_h \boldsymbol{y}_h = \boldsymbol{b}_h$. $\boldsymbol{S}$ sei die Iterationsmatrix des Glättungsverfahrens, welches durch $\bar{\boldsymbol{y}}_h = \boldsymbol{S}\boldsymbol{y}_h^0 + \boldsymbol{T}\boldsymbol{b}_h$ beschrieben wird. Dann ist für die $\nu-$ malige Glättung die Restmatrix $\boldsymbol{T}_\nu$ aus der Iterationsvorschrift berechenbar.

(18.1) Vorglättung: $\bar{\boldsymbol{y}}_h = \mathcal{S}(\boldsymbol{A}_h, \boldsymbol{b}_h)\boldsymbol{y}_h^0 = \boldsymbol{S}^{\nu_1}\boldsymbol{y}_h^0 + \boldsymbol{T}_{\nu_1}\boldsymbol{b}_h$

(18.2) Defektberechnung: $\boldsymbol{d}_h = \boldsymbol{b}_h - \boldsymbol{A}_h \bar{\boldsymbol{y}}_h$

(18.3) Restriktion und Grobgitterlösung: $\boldsymbol{v}_H = \boldsymbol{A}_H^{-1}(\boldsymbol{R}\boldsymbol{d}_h)$

(18.4) Prolongation und Korrektur: $\bar{\bar{\boldsymbol{y}}}_h = \bar{\boldsymbol{y}}_h + \boldsymbol{P}\boldsymbol{v}_H$

(18.5) Nachglättung: $\boldsymbol{y}_h^1 = \boldsymbol{S}^{\nu_2}\bar{\bar{\boldsymbol{y}}}_h + \boldsymbol{T}_{\nu_2}\boldsymbol{b}_h$

Wir konstruieren zunächst die

Iterationsmatrix des ZGV

Als Glättungsiteration werden nur Verfahren verwendet, welche die exakte Lösung $\boldsymbol{y}_h$ von $\boldsymbol{A}_h \boldsymbol{y}_h = \boldsymbol{b}_h$ als Fixpunkt haben (vgl. S 135), d.h. $\boldsymbol{y}_h = \boldsymbol{S}^{\nu_1}\boldsymbol{y}_h + \boldsymbol{T}_{\nu_1}\boldsymbol{b}_h$.
Abziehen dieser Gleichung von (18.1) liefert

(18.6)

$$\bar{\boldsymbol{e}}_h := \bar{\boldsymbol{y}}_h - \boldsymbol{y}_h = \boldsymbol{S}^{\nu_1}\boldsymbol{e}_h^0, \quad \text{mit} \quad \boldsymbol{e}_h^0 = \boldsymbol{y}_h^0 - \boldsymbol{y}_h.$$

Mit $\boldsymbol{d}_h \overset{(18.2)}{=} \boldsymbol{b}_h - \boldsymbol{A}_h\bar{\boldsymbol{y}}_h = \boldsymbol{A}_h(\boldsymbol{y}_h - \bar{\boldsymbol{y}}_h) = -\boldsymbol{A}_h\bar{\boldsymbol{e}}_h$ folgt aus (18.3)

(18.7)

$$\boldsymbol{v}_H \overset{(18.3)}{=} \boldsymbol{A}_H^{-1}\boldsymbol{R}\boldsymbol{d}_h \overset{(18.2)}{=} \boldsymbol{A}_H^{-1}\boldsymbol{R}(\boldsymbol{b}_h - \boldsymbol{A}_h\bar{\boldsymbol{y}}_h) = -\boldsymbol{A}_H^{-1}\boldsymbol{R}(\boldsymbol{A}_h\bar{\boldsymbol{y}}_h - \boldsymbol{A}_h\boldsymbol{y}_h) = -\boldsymbol{A}_H^{-1}\boldsymbol{R}\boldsymbol{A}_h\bar{\boldsymbol{e}}_h$$

Für den Fehler nach Prolongation und Korrektur erhält man

$$\begin{aligned}
\bar{\bar{\boldsymbol{e}}}_h := \bar{\bar{\boldsymbol{y}}}_h - \boldsymbol{y}_h &\overset{(18.4)}{=} \bar{\boldsymbol{y}}_h + \boldsymbol{P}\boldsymbol{v}_H - \boldsymbol{y}_h \\
&\overset{(18.6)}{=} \boldsymbol{S}^{\nu_1}\boldsymbol{e}_h^0 + \boldsymbol{P}\boldsymbol{v}_H \overset{(18.7)}{=} \boldsymbol{S}^{\nu_1}\boldsymbol{e}_h^0 - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\boldsymbol{A}_h\bar{\boldsymbol{e}}_h \\
&\overset{(18.6)}{=} \boldsymbol{S}^{\nu_1}\boldsymbol{e}_h^0 - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\boldsymbol{A}_h\boldsymbol{S}^{\nu_1}\boldsymbol{e}_h^0 = (\boldsymbol{I} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\boldsymbol{A}_h)\boldsymbol{S}^{\nu_1}\boldsymbol{e}_h^0
\end{aligned}$$

(18.8)
$$= (\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})\boldsymbol{A}_h\boldsymbol{S}^{\nu_1}\boldsymbol{e}_h^0$$

Die Nachglättung führt unter Verwendung der Fixpunktgleichung zu

$$\begin{aligned} \boldsymbol{e}_h^1 := \boldsymbol{y}_h^1 - \boldsymbol{y}_h &\overset{(18.5)}{=} \boldsymbol{S}^{\nu_2}\bar{\bar{\boldsymbol{y}}}_h + \boldsymbol{T}_{\nu_2}\boldsymbol{b}_h - (\boldsymbol{S}^{\nu_2}\boldsymbol{y}_h + \boldsymbol{T}_{\nu_2}\boldsymbol{b}_h) \\ &= \boldsymbol{S}^{\nu_2}(\bar{\bar{\boldsymbol{y}}}_h - \boldsymbol{y}_h) = \boldsymbol{S}^{\nu_2}\bar{\bar{\boldsymbol{e}}}_h \qquad \text{(Definition von } \bar{\bar{\boldsymbol{e}}}_h) \\ &\overset{(18.8)}{=} \boldsymbol{S}^{\nu_2}\underbrace{(\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})}_{=:\boldsymbol{C}_h}\boldsymbol{A}_h\boldsymbol{S}^{\nu_1}\boldsymbol{e}_h^0 \end{aligned}$$

Damit erhalten wir die Iterationsmatrix des ZGV

$$\boldsymbol{K}_h(\nu_1,\nu_2) = \boldsymbol{S}^{\nu_2}(\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})\boldsymbol{A}_h\boldsymbol{S}^{\nu_1} = \boldsymbol{S}^{\nu_2}\boldsymbol{C}_h\boldsymbol{S}^{\nu_2}. \tag{18.9}$$

Für die Iterationsmatrix $\boldsymbol{K}_h(\nu_1,\nu_2)$ des ZGV muß gezeigt werden

$$\|\boldsymbol{K}_h(\nu_1,\nu_2)\| < 1 \quad \text{bzw.} \quad \rho(\boldsymbol{K}_h(\nu_1,\nu_2)) < 1.$$

Führt man μ Schritte des ZGV aus, so erhält man mit der Abkürzung

$$\boldsymbol{C}_h := (\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})$$

$$\boldsymbol{K}_h^\mu = \boldsymbol{S}^{\nu_2}\boldsymbol{C}_h\boldsymbol{A}_h\boldsymbol{S}^{\nu_1}\boldsymbol{S}^{\nu_2}\boldsymbol{C}_h\boldsymbol{A}_h\boldsymbol{S}^{\nu_1}...\boldsymbol{S}^{\nu_2}\boldsymbol{C}_h\boldsymbol{A}_h\boldsymbol{S}^{\nu_1} \tag{18.10}$$

Wir wollen Vor- und Nachglättung zusammenfassen, $\nu = \nu_1 + \nu_2$, und dazu in (18.10) geeignet klammern

$$\boldsymbol{K}_h^\mu = \underbrace{\boldsymbol{S}^{\nu_2}}_{\boldsymbol{F}}\underbrace{(\boldsymbol{C}_h\boldsymbol{A}_h\boldsymbol{S}^{\nu})^{\mu-1}\boldsymbol{C}_h\boldsymbol{A}_h}_{\boldsymbol{U}}\boldsymbol{S}^{\nu_1} \tag{18.11}$$

Nun gilt für quadratische Matrizen $\boldsymbol{F},\boldsymbol{U}$ gleichen Formats: $\rho(\boldsymbol{F}\boldsymbol{U}) = \rho(\boldsymbol{U}\boldsymbol{F})$ (Übung, vgl. unten), sodaß

$$\rho(\boldsymbol{K}_h^\mu) = \rho[(\boldsymbol{C}_h\boldsymbol{A}_h\boldsymbol{S}^\nu)^\mu] = [\rho(\boldsymbol{C}_h\boldsymbol{A}_h\boldsymbol{S}^\nu)]^\mu \qquad \text{und} \tag{18.12}$$

$$\|\boldsymbol{K}_h^\mu\|_S = \rho(\boldsymbol{K}_h)^\mu \quad \forall \mu \in \mathbb{N}$$

Bemerkung: Für ein ZGV ohne Nachglättung gilt (18.12) mit $\nu = \nu_1 + 0$, d.h.

$$\boldsymbol{K}_h^\mu(\nu_1, 0) =: \boldsymbol{K}_h^\mu(\nu) = [\boldsymbol{K}_h(\nu)]^\mu = [(\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})\boldsymbol{A}_h\boldsymbol{S}^\nu]^\mu = [\boldsymbol{C}\boldsymbol{A}_h\boldsymbol{S}^\nu]^\mu$$

Konvergenz liegt vor, wenn

$$\|\boldsymbol{K}_h(\nu)\| < 1 \quad \text{bzw.} \quad \|\boldsymbol{K}_h(\nu_1,\nu_2)\| < 1. \tag{18.13}$$

Der Vergleich zeigt, daß ein Nachweis von $\rho((\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})\boldsymbol{A}_h\boldsymbol{S}^\nu) < 1$ gleichermaßen die Konvergenz von Verfahren mit und ohne Nachglättung sichert.
Nötig dazu ist, wie oben erwähnt, folgende

Aufgabe: Zeige für quadratische Matrizen $\boldsymbol{A},\boldsymbol{B}$: $\rho(\boldsymbol{A}\boldsymbol{B}) = \rho(\boldsymbol{B}\boldsymbol{A})$.
Hinweise:
(i) Zeige es zuerst unter der Voraussetzung, daß eine der Matzrizen, z.B. $\boldsymbol{B}$ regulär ist.

(ii) Verallgemeinere (i) durch ein Stetigkeitsargument, angewandt auf eine "leicht gestörte" Matrix $\boldsymbol{B}$ (Jordan-Normalform).

Konvergenzbeweise für das ZGV werden üblicherweise geführt, indem man die Iterationsmatrix

$$\boldsymbol{K}_h(\nu) = (\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})\boldsymbol{A}_h\boldsymbol{S}^\nu$$

aufteilt in

(18.14) Glättungsteil $\boldsymbol{A}_h\boldsymbol{S}^\nu$ und

(18.15) Approximationsteil $\boldsymbol{C}_h = (\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})$.

Definition 18.1 (Hackbusch)
Das ZGV besitzt die Glättungseigenschaft, falls

(18.16) $$\|\boldsymbol{A}_h\boldsymbol{S}^\nu\| \leq c_S(\nu) \quad \text{mit} \quad c_S(\nu) \to 0 \text{für } \nu \to infty$$

mit einer von ν abhängigen Konstanten $c_S = c_S(\nu)$.

Das ZGV besitzt die Approximationseigenschaft, falls

(18.17) $$\|\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R}\| \leq c_a(h) \quad (= const).$$

Bemerkungen:

1. Ziel eines Konvergenzbeweises ist die Ungleichung

$$\begin{aligned}\|\boldsymbol{K}_h(\nu)\| &= \|(\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})\boldsymbol{A}_h\boldsymbol{S}^\nu\| \\ &\overset{(*)}{\leq} \|(\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})\|\,\|\boldsymbol{A}_h\boldsymbol{S}^\nu\| \leq c_a c_S \overset{!}{<} 1 \quad \text{für kleines } \nu.\end{aligned}$$

   Die Abschätzungen (18.16),(18.17) müssen also möglichst scharf ausfallen. Sie werden üblicherweise gesondert bewiesen, müssen beide aber bzgl. Normen ausgeführt werden, für welche die Abschätzung $(*)$ gilt

2. c_a in (18.17) ist eine feste Zahl, die < 1 sein kann, aber nicht muß. Im letzteren Fall hängt die Konvergenz von $\boldsymbol{S}$ ab. Eine optimale Glättung durch $\boldsymbol{S}$ muß also nicht notwendigerweise mit einer optimalen Konvergenz äquivalent sein. (vgl. Bemerkung auf S. 126)

3. Bei der Durchführung des MGV hat man immer zwei benachbarte Aufgaben zu untersuchen (auf dem feineren und auf dem gröberen Gitter)

$$\boldsymbol{A}_h\boldsymbol{y}_h = \boldsymbol{d}_h, \quad \boldsymbol{A}_H\boldsymbol{y}_H = \boldsymbol{d}_H \qquad \text{(Erinnerung: } \boldsymbol{d}_H = \boldsymbol{R}\boldsymbol{d}_h\text{)}$$

   Die Lösungen dieser Defektgleichungen werden verglichen (Nachiteration). Es gilt

   (18.18) $$\|\boldsymbol{y}_h - \boldsymbol{P}\boldsymbol{y}_H\| = \|(\boldsymbol{A}_h^{-1} - \boldsymbol{P}\boldsymbol{A}_H^{-1}\boldsymbol{R})\boldsymbol{d}_h\| = \|\boldsymbol{C}_h\boldsymbol{d}_h\| \leq \|\boldsymbol{C}_h\|\,\|\boldsymbol{d}_h\|.$$

   Hieraus erklärt sich der Name Approximationseigenschaft. Bei vorgegebenem Gitter kann $\|C_h\|$ nicht gegen Null gehen.

4. Es ist einleuchtend, dass man Abschätzungen der Art

(18.19) $$\|\boldsymbol{y}_h - \boldsymbol{P}\boldsymbol{y}_H\| \le c_\alpha h^\alpha \quad \text{(Standart: } \alpha = 2 \text{ oder } 0\text{)}$$

zeigen kann. Solche Abschätzungen wurden u.a. von Dryja (polnischer Mathematiker), Hackbusch und Bachvalov bewiesen. Sie sind sehr aufwendig, weshalb wir hier darauf verzichten. (Wir verweisen z.B. auf das Hackbusch-Buch, Abschnitt 6.3.2)

Dies führt dazu, dass man in den Konvergenzbeweisen die Eigenschaften (18.16),(18.17) gelegentlich abändert zu

(18.16′) $$\|\boldsymbol{A}_h\boldsymbol{S}^\nu\| \le \tfrac{1}{h^\alpha} c_S(\nu), \quad c_S(\nu) \xrightarrow{\nu\to\infty} 0,$$

(18.17′) $$\|\boldsymbol{C}_h\| \le c_\alpha h^\alpha\,.$$

In unseren Beispielen zum Beweis der Glättungseigenschaft werden wir sehen, daß sich Abschätzungen der Art (18.16’) “natürlich” ergeben.
$\alpha = 0$ entspricht (18.16),(18.17), $\alpha = 0$ ist üblich für Δu.

Wir zeigen im nächsten Abschnitt die Glättungseigenschaften einiger Verfahren. Beliebt sind dabei Normen, die sich (unter gewissen Symmetrievoraussetzungen) auf die Spektralnorm stützen. Dann kann nämlich mit dem Spektralradius (und damit mit den Eigenwerten) statt mit Normen gerechnet werden.

Glättungseigenschaften des gedämpften Jacobi-Verfahrens

Die Iterationsmatrix des gedämpften Jacobi-Verfahrens lautet

$$\boldsymbol{S} = \boldsymbol{I} - \omega \boldsymbol{D}_h^{-1}\boldsymbol{A}_h, \quad \boldsymbol{D}_h = \text{Diagonalmatrix von } \boldsymbol{A}_h.$$

Die Dämpfungseigenschaft ist also neben dem Verfahren auch von A_h abhängig.

Wir beschränken unsere Untersuchungen auf Matrizen $\boldsymbol{A}_h$ mit

(18.20) $$\boldsymbol{A}_h = \boldsymbol{A}_h^T > 0, \quad \boldsymbol{D}_h = \frac{k}{h^{2m}}\boldsymbol{I}, \quad k = const.$$

$\boldsymbol{A}$ hat dann nur positive Eigenwerte. Die Iterationsmatrix $\boldsymbol{S}$ des Jacobi-Verfahrens ist symmetrisch läßt sich nun schreiben als (vgl. (18.4))

(18.21) $$\boldsymbol{S} =: \boldsymbol{I} - \widetilde{\omega}\,\boldsymbol{A}_h \quad \text{mit } \widetilde{\omega} = \frac{h^{2m}}{k}\,\omega.$$

Zum Beweis der Glättungseigenschaft ist mit einer geeigneten Norm der Ausdruck $\|\boldsymbol{A}\boldsymbol{S}^\nu\|$ abzuschätzen. Wegen (18.20) gilt mit der Spektralnorm $\|\cdot\|_S$ (Beachte: $\boldsymbol{S}^\nu$ ist ein Polynom in $\boldsymbol{S}$, also ist $\boldsymbol{A}_h\boldsymbol{S}^\nu$ symmetrisch)

(18.22) $$\|\boldsymbol{A}_h\boldsymbol{S}^\nu\|_S = \max_{\lambda\in\sigma(A_h)} |\lambda(1-\widetilde{\omega}\lambda)^\nu|,$$

Um mit Eigenwerten rechnen zu können, legen wir deshalb für die weiteren Untersuchungen die Spektralnorm zu Grunde.

Ziel: Gesucht ist eine Abschätzung dieses Ausdrucks durch eine in ν fallende Schranke, die auch Auskunft über zulässige $\widetilde{\omega}$-Werte gibt.
(Bemerkung: In Hackbusch (6.2.3) wird eine $\widetilde{\omega}$-Schranke nicht hergeleitet, vgl. jedoch Exercise 6.6.3)

Satz 18.2
Seien $\boldsymbol{A}_h = \boldsymbol{A}_h^T > 0, \quad \boldsymbol{D}_h := \operatorname{diag}(\boldsymbol{A}_h) = \frac{k}{h^{2m}}\boldsymbol{I}, \quad k = const.,$ und

$\boldsymbol{S} = \boldsymbol{I} - \widetilde{\omega}\boldsymbol{A}_h, \ \widetilde{\omega} = \frac{h^{2m}}{k}\omega$ die Iterationsmatrix des gedämpften Jacobi-Verfahrens

Dann gilt für

$$\omega \leq \frac{1.2}{\|\boldsymbol{A}_h\|_S}\,\frac{k}{h^{2m}} \quad \text{(Spektralnorm)}$$

die Glättungseigenschaft

$$\|\boldsymbol{A}_h\boldsymbol{S}^\nu\|_S \leq \frac{\|\boldsymbol{A}_h\|_S}{1.2}\,\frac{1}{2(\nu+1)}.$$

Beweis
Wir setzen $t = \widetilde{\omega}\lambda$ und erhalten für ein noch zu bestimmendes t_{max}

$$\|\boldsymbol{A}_h\boldsymbol{S}^\nu\|_S = \frac{1}{\widetilde{\omega}}\max_t |t(1-t)^\nu| \quad \text{für} \quad 0 \leq t \leq t_{max}. \tag{18.23}$$

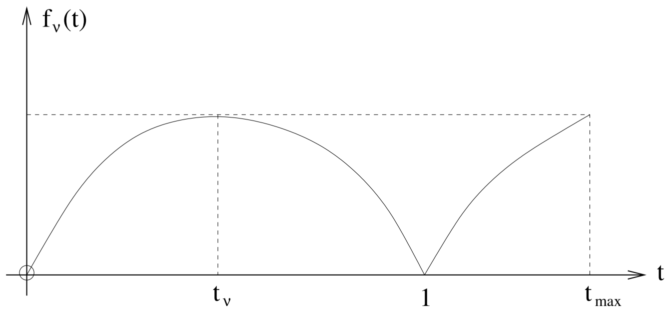
Die Untersuchung dieses Ausdrucks ist ein Kernstück vieler Glättungseigenschaftsnachweise (vgl. z.B. symmetrisches GSV und GSV mit Schachbrettanordnung.)

Vorgehen: Durch Abschätzung des Ausdrucks

$$f_\nu(t) = |t(1-t)^\nu|$$

nach oben leiten wir eine Ungleichung her, die uns in Verbindung mit $\lambda_{max}(\boldsymbol{A}_h) = \|\boldsymbol{A}_h\|_S$ eine Schranke für $\widetilde{\omega}$ liefern wird (beachte: alle Eigenwerte von $\boldsymbol{A}$ sind positiv).

$f_\nu(t) = |t(1-t)^\nu|$ hat für alle $\nu \geq 1$ qualitativ das folgende Aussehen (vgl.Abbildung), wobei für $\nu = 1$ in $t = 1$ eine Spitze auftritt.



In $0 \le t \le 1$ gilt $f_\nu(t) = t(1-t)^\nu$. Gesucht: $\max f_\nu(t)$ in $[0,1]$

$$f'_\nu(t) = (1-t)^\nu - t\nu(1-t)^{\nu-1} = (1-t)^{\nu-1}[1-t-t\nu] \overset{!}{=} 0$$

$$1 - t(1+\nu) = 0 \quad \Longrightarrow \quad t_\nu = \frac{1}{(\nu+1)} \quad \Longrightarrow$$

$$f_\nu(t_\nu) = \frac{1}{(\nu+1)}\left(1 - \frac{1}{\nu+1}\right) = \frac{1}{(\nu+1)}\left(\frac{1}{(1+\frac{1}{\nu})^\nu}\right)$$

wegen $(1+\frac{1}{\nu})^\nu \nearrow e$ für $\nu \to \infty$ gilt für $\nu = 1$

$$f_\nu(t) = \frac{1}{(\nu+1)}\left(\frac{1}{(1+\frac{1}{\nu})^\nu}\right) \le \frac{1}{(\nu+1)}\,\frac{1}{2} \quad \forall t \in [0,1].$$

Wir bestimmen, zunächst für $\nu = 1$, das maximale Intervall $(0, t_{max}]$ in dem diese Ungleichung gilt. Da $f_\nu(t) = t(t-1)$ für $t \ge 1$, ergibt sich t_{max} aus

$$t(t-1) \le \frac{1}{2}\frac{1}{(\nu+1)} \overset{\nu=1}{=} \frac{1}{4} \qquad \text{zu}$$

(18.24) $$t_{max} = \frac{1+\sqrt{2}}{2}.$$

Damit gilt in $[0, t_{max}]$ für $\nu = 1$: (beachte $f_\nu(t)$ ist monoton wachsend für $t \ge 1$)

(18.25) $$|t(1-t)^\nu| \le t_{max}(t_{max}-1) = \frac{1+\sqrt{2}}{2}\left(\frac{1+\sqrt{2}}{2} - 1\right)^\nu \le \frac{1}{2(\nu+1)}$$

Mit $\frac{1+\sqrt{2}}{2} \approx 1.2$ ist die zweite Ungleichung äquivalent zu

(18.26) $$2.4(\nu+1)\cdot 0.2^\nu \le 1, \quad \text{für } \nu = 1.$$

Die linke Seite dieser Ungleichung ist für $\nu \ge 1$ monoton fallend in ν (Ableitung ausrechnen). Deshalb gilt die Ungleichung (18.26) für alle $\nu \ge 1$, also gelten auch die Ungleichungen (18.25) für alle $\nu \ge 1$ und $0 \le t \le 1.2$.

Von Bedeutung für t sind die Werte $t_i = \widetilde{\omega}\lambda_i(\boldsymbol{A}_h)$. Nun gilt $\lambda_i(\boldsymbol{A}_h) \le \lambda_{max} = \|\boldsymbol{A}_h\|_S$. Man beachte, daß (18.25) bis auf den Faktor $\frac{1}{\widetilde{\omega}}$ eine Abschätzung für den Glättungsterm liefert (vgl. (18.23)). Um die Glättungseigenschaft zu gewährleisten (d.h. um die Ungleichung (18.25) zu erhalten), fordern wir deshalb die Ungleichung

$$\widetilde{\omega}\lambda_i \le \widetilde{\omega}\lambda_{max} = \widetilde{\omega}\|\boldsymbol{A}_h\|_S \overset{!}{\le} t_{max} = 1.2,$$

woraus sich für $\widetilde{\omega}$ folgende Schranke ergibt

(18.27) $$\widetilde{\omega} \le \frac{1.2}{\|\boldsymbol{A}_h\|_S} \quad \text{bzw. } \omega = \widetilde{\omega}\frac{k}{h^{2m}}. \quad \text{(vgl. Satz (18.2))}$$

Die Ungleichung (18.25) gilt für t_{max}. Aus (18.22) erhält man deshalb die Glättungseigenschaft

(18.28) $$\|\boldsymbol{A}_h\boldsymbol{S}^\nu\|_S = \frac{1}{\widetilde{\omega}} \max_{t\in(0,t_{max}]} |t(1-t)| \le \frac{1}{\widetilde{\omega}}\frac{1}{2(\nu+1)} = \frac{\|\boldsymbol{A}_h\|_S}{1.2}\cdot\frac{1}{2(\nu+1)}.$$

■.

Bemerkungen:

1. Satz 18.2 zeigt, daß der in (17.4) angegebene Dämpfungsparameter zulässig ist.

   (18.29) $$\omega_l = \frac{k}{h^{2m}\|A_h\|_S} \leq \widetilde{\omega}\frac{k}{h^{2m}} = \frac{1.2\,k}{h^{2m}\|A_h\|_S}.$$

2. Ist $\boldsymbol{A}_h$ die Diskretisierungsmatrix von $-\Delta u$ so gilt im Fall

   1D: $$\|\boldsymbol{A}_h\|_S \leq \frac{4}{h^2}$$ (vgl. Lemma 3.7)

   2D: $$\|\boldsymbol{A}_h\|_S \leq \frac{8}{h^2}$$

   Wird dies in (18.28) eingesetzt, erhält man eine Abschätzung vom Typ (18.16') (z.B. für 1D).

   $$\|\boldsymbol{A}_h\boldsymbol{S}^\nu\|_S =\leq \frac{\|\boldsymbol{A}_h\|_S}{1.2} \cdot \frac{1}{2(\nu+1)} \leq \frac{4}{1.2h^2} \cdot \frac{1}{2(\nu+1)}$$

3. Im 1D-Fall ist $h^{2m} = h^2,\ k = 2$ (vgl. (18.21)). Dann folgt aus (18.28):

   $$\omega \leq \frac{1.2k}{h^{2m}\|A_h\|_S} = \frac{2 \cdot 1.2}{h^2\frac{4}{h^2}} = 0.6$$

   Als optimalen Dämpfungsparameter hatten wir in § 15 im Demonstrationsbeispiel $\omega_{opt} = \frac{2}{3} = 0.6...$ ermittelt. ω_{opt} fällt also nicht in die zulässige Schranke. Nun kann man in (18.24) natürlich eine größere Schranke für $\max|t(t-1)|$ zulassen und dadurch (vgl. die Zeichnung) das Intervall $(0, t_{max}]$ so erweitern, daß ω_{opt} auch noch in den zulässigen Bereich fällt. Rechnet man dies nach, so stellt man fest, daß sich die Abschätzung für $\|\boldsymbol{A}_h\boldsymbol{S}^\nu\|_S$ insgesamt verschlechtert. Um die Konvergenzbedingung zu erfüllen, muß ein größeres ν gewählt werden.

   Man erkennt also: Optimal Glättung und optimale Konvergenz sind nicht äquivalent.

4. Beachte: Der Beweis von Satz 18.2 beruht auf der Gleichung (18.22). Diese Darstellung hat als wesentliche Voraussetzung die Symmetrie der Iterationsmatrix $\boldsymbol{S}$ des Jacobi-Verfahrens und damit des Glättungsteils $\boldsymbol{A}_h\boldsymbol{S}^\nu$. Sonst ist eine Normdarstellung (18.22) nicht möglich und damit auch kein Rechnen mit dem Spektralradius, bzw. mit der Spektralnorm.)
   Die Iterationsmatrix des GSV ist nicht symmetrisch, weshalb für die Glättungseigenschaft dieses Verfahrens bei zeilenweiser Nummerierung der Punkte (bisher) kein Beweis existiert.
   Man kann die Glättungseigenschaft für das GSV retten unter Zuhilfenahme zusätzlicher Struktureigenschaften der Iterationsmatrix in Zusammenhang mit einer geeigneten Nummerierung der Giterpunkte (ohne Beweis).
   Wir werden sehen, daß für das symmetrische GSV der Glättungsteil wieder symmetrisch ist.

Das symmetrische Gauß-Seidel-Verfahren als Glättungsiteration

Vorbemerkung: Wir zeigen, daß sich für das symmetrische GSV unter Verwendung einer geeigneten Vektor- und Matrixnorm die Glättungseigenschaft beweisen läßt, und zwar unabhängig von den Schrittweiten h_1, h_2 des Gitters.
Bezüglich einer geeigneten Matrixnorm $\|\cdot\|_M$ muß gezeigt werden, (vgl.Bemerkung 1) nach Definition 18.1)

$$\|\boldsymbol{K}_h(\nu)\|_M = \|\boldsymbol{C}_h\boldsymbol{A}_h\boldsymbol{S}^\nu\|_M \le c_a \cdot c_S(\nu) \xrightarrow{\nu\to\infty} 0.$$

Zur Definition der Vektornorm benutzen wir die in der Iterationsmatrix des symmetrischen GSV vorkommende Matrix (vgl. Satz 15.6)

$$\boldsymbol{W}_h = \boldsymbol{A}_h + \boldsymbol{L}_h\boldsymbol{D}^{-1}\boldsymbol{U}_h, \quad (\text{für } \boldsymbol{A} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{U})$$

die mit $\boldsymbol{A}_h$ symmetrisch und positiv definit ist. Also existiert genau eine Wurzel $\boldsymbol{W}_h^{\frac{1}{2}} = (\boldsymbol{W}_h^{\frac{1}{2}})^T > 0$ (vgl. Aufgabe 3) nach Satz 10.1). Mit ihrer Hilfe definieren wir

$$(18.30) \qquad \|\boldsymbol{x}\|^2_{\boldsymbol{W}_h} = (\boldsymbol{W}_h\boldsymbol{x}, \boldsymbol{x}) = (\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{x}, \boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{x}) = \|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{x}\|_2^2.$$

Hierzu gehört die induzierte Matrixnorm (Operatornorm)

$$\|\boldsymbol{K}_h\|_{\boldsymbol{W}_h} := \sup_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{K}_h\boldsymbol{x}\|_{\boldsymbol{W}_h}}{\|\boldsymbol{x}\|_{\boldsymbol{W}_h}} = \|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{K}_h\boldsymbol{W}_h^{-\frac{1}{2}}\|_S, \quad (\text{Spektralnorm})$$

denn

$$\|\boldsymbol{K}_h\boldsymbol{x}\|_{\boldsymbol{W}_h} = \|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{K}_h\boldsymbol{x}\|_S = \|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{K}_h\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{x}\|_S \le \|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{K}_h\boldsymbol{W}_h^{-\frac{1}{2}}\|_S\|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{x}\|_S,$$

und da $\|\cdot\|_2$ und $\|\cdot\|_S$ einander zugeordnet sind, existiert zu $\boldsymbol{B} := \boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{K}_h\boldsymbol{W}_h^{-\frac{1}{2}}$ ein $\boldsymbol{y}$ $(= \boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{x})$ sodaß

$$\|\boldsymbol{B}\boldsymbol{y}\|_2 = \|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{K}_h\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{x}\|_S = \|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{K}_h\boldsymbol{W}_h^{-\frac{1}{2}}\|_S\,\|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{x}\|_2 = \|\boldsymbol{B}\|_S\|\boldsymbol{y}\|_2.$$

Die Konvergenzabschätzung für das ZGV wird wie folgt geführt (vgl. und (18.12))

$$(18.31) \qquad \begin{aligned} \|\boldsymbol{K}_h(\nu)\|_{W_h} &= \|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{C}_h\boldsymbol{A}_h\boldsymbol{S}^\nu\boldsymbol{W}_h^{-\frac{1}{2}}\|_S \\ &= \|\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{C}_h\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{A}_h\boldsymbol{S}^\nu\boldsymbol{W}_h^{-\frac{1}{2}}\|_S \\ &\le \|\underbrace{\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{C}_h\boldsymbol{W}_h^{\frac{1}{2}}}_{\text{Approximationsanteil}}\|_S\,\|\underbrace{\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{A}_h\boldsymbol{S}^\nu\boldsymbol{W}_h^{-\frac{1}{2}}}_{=:\,\boldsymbol{G},\text{ Glättungsteil}}\|_S \end{aligned}$$

Wir führen nur den Nachweis der Glättungseigenschaft. Der Beweis wird die Einführung der $\boldsymbol{W}_h$-Norm begründen (vgl. dazu die Nachbemerkung).

Satz 18.3
Sei $\boldsymbol{A}_h = \boldsymbol{L}+\boldsymbol{D}+\boldsymbol{U}$, $\boldsymbol{A}_h = \boldsymbol{A}_h^T > 0$ und $\boldsymbol{S} = \boldsymbol{I}-\boldsymbol{W}^{-1}\boldsymbol{A}_h$ die Iterationsmatrix des symmetrischen Gauß-Seidel-Verfahrens mit $\boldsymbol{W} = \boldsymbol{A}_h + \boldsymbol{L}\boldsymbol{D}^{-1}\boldsymbol{U}$ (vgl. Satz 15.6).
Dann gilt (mit der Spektralnorm) folgende Glättungseigenschaft

$$\|\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{A}_h\boldsymbol{S}^\nu\boldsymbol{W}_h^{-\frac{1}{2}}\|_S \le \frac{1}{2(\nu+1)}$$

Beweis
Idee: Wir formen den Glättungsteil $\boldsymbol{G}$ so um, daß wir (wie beim Jacobi-Verfahren) eine zu (18.23) analoge Darstellung erhalten.
Für den Glättungsteil gilt

$$\begin{aligned}
\boldsymbol{G} &= \underbrace{\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{A}_h\boldsymbol{W}_h^{-\frac{1}{2}}}_{=:\boldsymbol{X}_h}\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{S}^\nu\boldsymbol{W}_h^{-\frac{1}{2}}, \qquad \boldsymbol{X}_h := \boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{A}_h\boldsymbol{W}_h^{-\frac{1}{2}} \\
&= \boldsymbol{X}_h\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{S}\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{S}\boldsymbol{W}_h^{-\frac{1}{2}}\ldots\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{S}\boldsymbol{W}_h^{-\frac{1}{2}} \\
&= \boldsymbol{X}_h(\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{S}\boldsymbol{W}_h^{-\frac{1}{2}})^\nu.
\end{aligned}$$

Nun ist gemäß (15.19) (Darstellung der Iterationsmatrix $\boldsymbol{S}$ des symmetrischen Gauß-Seidel)

$$\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{S}\boldsymbol{W}_h^{-\frac{1}{2}} = \boldsymbol{W}_h^{\frac{1}{2}}(\boldsymbol{I}-\boldsymbol{W}_h^{-1}\boldsymbol{A}_h)\boldsymbol{W}_h^{-\frac{1}{2}} = \boldsymbol{I}-\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{A}_h\boldsymbol{W}_h^{-\frac{1}{2}} = \boldsymbol{I}-\boldsymbol{X}_h$$

also (vgl. die 3. Formelzeile des Beweises)

(18.32) $\qquad \boldsymbol{G} = \boldsymbol{X}_h(\boldsymbol{I}-\boldsymbol{X}_h)^\nu, \qquad$ Darstellung des Glättungsteils.

Da $\boldsymbol{W}$ und $\boldsymbol{A}_h$ symmetrisch sind, gilt dies auch für $\boldsymbol{W}_h^{-\frac{1}{2}}$ und $\boldsymbol{X}_h$.
Die Matrix des Glättungsteils ist also symmetrisch. (beachte: $(\boldsymbol{I}-\boldsymbol{X}_h)^\nu$ ist ein Polynom in $\boldsymbol{X}_h$.).
Wir zeigen

(18.33) $\qquad 0 < \boldsymbol{X}_h \le \boldsymbol{I}, \qquad$ positive Definitheit.

Beweis von (18.33):

$$(\boldsymbol{X}_h\boldsymbol{x},\boldsymbol{x}) = (\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{A}_h\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{x},\boldsymbol{x}) = (\boldsymbol{A}_h\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{x},\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{x}) > 0 \quad \Longrightarrow \quad \boldsymbol{X}_h > 0, \text{ da } \boldsymbol{A}_h > 0.$$

Wegen $\qquad (\boldsymbol{A}_h\boldsymbol{y},\boldsymbol{y}) \le (\boldsymbol{W}_h\boldsymbol{y},\boldsymbol{y}) \ \forall\, \boldsymbol{y} \qquad$ (vgl. Satz 15.6 b)) ist

$$(\boldsymbol{A}_h\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{x},\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{x}) \le (\boldsymbol{W}_h\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{x},\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{x}) = (\boldsymbol{W}_h^{\frac{1}{2}}\boldsymbol{x},\boldsymbol{W}_h^{-\frac{1}{2}}\boldsymbol{x}) = (\boldsymbol{I}\boldsymbol{x},\boldsymbol{x}) \quad \text{daraus folgt}$$

$$\boldsymbol{X}_h \le \boldsymbol{I}.$$

Wegen (18.33) gilt für die Eigenwerte $\lambda \in \sigma(\boldsymbol{X}_h)$: $\ 0 < \lambda \le 1$, denn

$$0 \le \frac{\boldsymbol{x}^T\boldsymbol{X}_h\boldsymbol{x}}{\|\boldsymbol{x}\|_2^2} \le \frac{(\boldsymbol{x},\boldsymbol{x})}{\|\boldsymbol{x}\|_2^2} \le 1 \qquad \text{(Rayleighquotient)}$$

Deshalb folgt für die Abschätzung des Glättungsteils aus (18.32)- (18.33) (Spektralnorm)

$$(18.34)\quad \|\boldsymbol{G}\|_S \leq \max_{\lambda\in[0,1]} |\lambda(1-\lambda)^\nu| \leq \frac{1}{2(1+\nu)} \qquad \text{(vgl. (18.23) – (18.25) mit } \widetilde{\omega}=1\text{)}$$

Dies ist die gewünschte Glättungseigenschaft. Eine Abschätzung des Approximationsteils (vgl. (18.31)) durch eine Konstante c_α (das ist schwächer als durch $c_\alpha h^\alpha$) beendet den Konvergenzbeweis für hinreichend großes ν. ■

Nachbemerkungen

1. Angenehm ist, daß die Abschätzung

   a) unabhängig ist von h (vgl. dazu die Nachbemerkung 3)) und

   b) daß man sich nicht um Dämpfungsparameter kümmern muß.

2. Wesentliche Grundvoraussetzung für beide Glättungsbeweise war jeweils die Symmetrie von $\boldsymbol{D}^{-1}\boldsymbol{A}$ und $\boldsymbol{W}^{-1}$. Sie führte beim gedämpften Jacobi-Verfahren zur Gleichung (18.22), beim GSV zu (18.32) und (18.34). Da diese Symmetrie beim gewöhnlichen GSV nicht vorliegt, konnte eine Glättungseigenschaft für dieses Verfahren bisher nicht gezeigt werden. Trotzdem wird der Gauß-Seidel mit Erfolg benutzt.

3. Beim Jacobi-Verfahren war die Glättungseigenschaft in natürlicher Weise von h abhängig über die Diskretisierungsmatrix (vgl. Bemerkung 2 nach dem Beweis von Satz 18.2). Symmetrisches GSV und Jacobi-Verfahren lassen sich in einer gemeinsamen Form schreiben:

   $$\begin{aligned} \boldsymbol{x}^{m+1} &= (\boldsymbol{I}-\boldsymbol{W}^{-1}\boldsymbol{A}_h)\boldsymbol{x}^m + \boldsymbol{W}^{-1}\boldsymbol{b}, \quad &&\text{(symmetrisches GSV, vgl. (15.17),(15.18))} \\ \boldsymbol{x}^{m+1} &= (\boldsymbol{I}-\omega\boldsymbol{D}^{-1}\boldsymbol{A}_h)\boldsymbol{x}^m + \omega\boldsymbol{D}^{-1}\boldsymbol{b}, \quad &&\text{(Jacobi-Verfahren, vgl. (15.1))} \end{aligned}$$

   Dadurch erklären sich auch die Ähnlichkeiten in den Glättungsbeweisen. Der Glättungsbeweis des symmetrischen GSV läßt sich damit auch auf das Jacobiverfahren übertragen. Daß die Glättungsabschätzung für das symmetrische GSV unabhängig von h ausfällt, liegt an der verwendeten Norm, welche die h-Abhängigkeit auf den Approximationsteil überträgt.

4. Glättungseigenschaftern für nichtsymmetrische Verfahren stecken noch in den Kinderschuhen.

5. Die Konvergenzeigenschaften bei der Verwendung des symmetrischen GSV werden durch den Beweis sicherlich unterschätzt. Sie fallen für Jacobi-Verfahren und symmetrisches GSV etwa gleich aus, was nicht ganz einsichtig ist, wenn man sich überlegt, daß ein Schritt symmetrisches GSV zwei Schritten GSV entspricht und daß das GSV üblicherweise schneller konvergiert als der Jacobi.

6. Zum Vergleich der Konvergenzgeschwindigkeit der Verfahren ziteren wir, ohne Beweis (vgl. Varga: Matrix Iterative Analysis, Kap. 3.3)

Satz 18.4 Stein-Rosenberg
Sei $\boldsymbol{A} \in C^{n \times n}$, $\boldsymbol{A} = \boldsymbol{L} + \boldsymbol{D} + \boldsymbol{R}$,

$\boldsymbol{S} = -\boldsymbol{D}^{-1}(\boldsymbol{L} + \boldsymbol{R})$ die Iterationsmatrix des Jacobi-Verfahrens mit
$s_{ij} \leq 0 \;\forall\, i \neq j, \;\; s_{ii} = 0 \;\forall i,$

$\boldsymbol{G} = -(\boldsymbol{L} + \boldsymbol{D})^{-1}\boldsymbol{R}$ die Iterationsmatrix des Gauß-Seidel-Verfahrens.

Dann gilt genau eine der folgenden Aussagen

(a) $\rho(\boldsymbol{S}) = \rho(\boldsymbol{G}) = 0$

(b) $0 < \rho(\boldsymbol{G}) < \rho(\boldsymbol{S}) < 1$

(c) $1 = \rho(\boldsymbol{G}) = \rho(\boldsymbol{S})$

(d) $1 < \rho(\boldsymbol{S}) < \rho(\boldsymbol{G})$

Setzt man $\boldsymbol{L} = \boldsymbol{D}\widetilde{\boldsymbol{L}}$ und $\boldsymbol{R} = \boldsymbol{D}\boldsymbol{U}$, so erhält man die Darstellung aus Varga.

Beachte: Für unsere Diskretisierungen sind alle diese Voraussetzungen erfüllt, denn alle Nebendiagonalelemente von $\boldsymbol{A}$ sind ≤ 0, alle Diagonalelemente von $\boldsymbol{A}$ sind positiv. Es liegt Fall b) vor, denn die Konvergenzeigenschaft wird durch Satz 15.5 gesichert.
Insbesondere: Jacobi ist langsamer als Gauß-Seidel und Gauß-Seidel braucht weniger Speicherplatz als Jacobi.
Letzteres folgt aus der Darstellung (17.5). Die alten Komponenten werden beim Gauß-Seidel sofort durch die neuen überschrieben. Beim Jacobi müssen sie aufbewahrt werden, bis alle Komponenten berechnet sind.

§ 19 Konvergenz des Mehrgitterverfahrens

Der Konvergenzbeweis beruht natürlich auf einer Untersuchung der Iterationsmatrix des MGV. Zu ihrer Beschreibung benötigen wir einige

Vorbetrachtungen über Iterationssverfahren zur Lösung der Korrekturgleichungen

Jedes lineare Iterationsverfahren zur Lösung von

$$\boldsymbol{Av} = \boldsymbol{d}, \quad \boldsymbol{A} \in \mathbb{R}^{n\times n} \tag{19.1}$$

kann in der Form

$$\boldsymbol{v}^{j+1} = \boldsymbol{\varphi}(\boldsymbol{v}^j, \boldsymbol{d}) := \boldsymbol{Mv}^j + \boldsymbol{Nd}, \quad \boldsymbol{M}, \boldsymbol{N} \in \mathbb{R}^{n\times n} \tag{19.2}$$

dargestellt werden. Als Minimalvoraussetzung wird vom Verfahren verlangt, daß die Lösung von $\boldsymbol{Av} = \boldsymbol{d}$ ein Fixpunkt des Verfahrens ist. Diese Forderung erlaubt es $\boldsymbol{N}$ durch $\boldsymbol{M}$ auszudrücken, d.h.

$$\begin{gathered}\boldsymbol{v} = \boldsymbol{\varphi}(\boldsymbol{v}, \boldsymbol{d}) := \boldsymbol{Mv} + \boldsymbol{Nd} = \boldsymbol{Mv} + \boldsymbol{NAv} \quad \text{und damit} \\ \boldsymbol{I} - \boldsymbol{M} = \boldsymbol{NA}.\end{gathered}$$

Ist $\boldsymbol{A}$ invertierbar, was wir immer voraussetzen, so kann man $\boldsymbol{N}$ ausdrücken durch

$$\boldsymbol{N} = (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{A}^{-1}. \tag{19.3}$$

Das Iterationsverfahren ist dann durch seine Iterationsmatrix schon eindeutig bestimmt und hat dann die Gestalt

$$\boldsymbol{v}^{j+1} = \boldsymbol{Mv}^j + (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{A}^{-1}\boldsymbol{d}. \tag{19.4}$$

Ist $\boldsymbol{v}^0 = \boldsymbol{0}$ Ausgangsnäherung des Verfahrens, was typisch ist als Anfangsnäherung für die Lösung der Korrekturgleichung, so folgt

$$\begin{aligned}\boldsymbol{v}^1 &= \boldsymbol{Mv}^0 + \boldsymbol{Nd} = \boldsymbol{Nd} \\ \boldsymbol{v}^2 &= \boldsymbol{Mv}^1 + \boldsymbol{Nd} = \boldsymbol{MNd} + \boldsymbol{Nd} \\ \boldsymbol{v}^3 &= \boldsymbol{Mv}^2 + \boldsymbol{Nd} = (\boldsymbol{M}^2\boldsymbol{N} + \boldsymbol{MN} + \boldsymbol{N})\boldsymbol{d}\end{aligned}$$

und durch vollständige Induktion

$$\boldsymbol{v}^\gamma = \sum_{j=0}^{\gamma-1} \boldsymbol{M}^j\boldsymbol{Nd} \overset{(19.3)}{=} \sum_{j=0}^{\gamma-1} \boldsymbol{M}^j(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{A}^{-1}\boldsymbol{d} = \sum_{j=0}^{\gamma-1}(\boldsymbol{M}^j - \boldsymbol{M}^{j+1})\boldsymbol{A}^{-1}\boldsymbol{d}$$

$$\boldsymbol{v}^\gamma = (\boldsymbol{I} - \boldsymbol{M}^\gamma)\boldsymbol{A}^{-1}\boldsymbol{d}, \tag{19.5}$$

und falls $\quad \|\boldsymbol{M}\| < 1 \quad \Longrightarrow \quad \boldsymbol{v}^\gamma \xrightarrow{\gamma\to\infty} \boldsymbol{A}^{-1}\boldsymbol{d},$

d.h. wenn das Verfahren konvergiert, dann gegen die Lösung von (19.1), denn

$$\|\boldsymbol{M}\| < 1 \Longrightarrow \|\boldsymbol{M}^\gamma\| \le \|\boldsymbol{M}\|^\gamma < 1 \xrightarrow{\gamma\to\infty} 0 \Longrightarrow \boldsymbol{M} \to \boldsymbol{0}.$$

Beim MGV werden nur Glättungsiterationen benutzt $\tilde{\boldsymbol{u}} = \mathcal{S}(\boldsymbol{u}, \boldsymbol{d})$, welche die Lösung von $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{d}$, bzw. $\boldsymbol{A}\boldsymbol{y} = \boldsymbol{b}$ als Fixpunkt haben (z.B. Jakobi, GSV, symmetrisches GSV). Hieraus folgt

(19.6) Auf jedem Gitterlevel ist die Lösung $\boldsymbol{v}_l$ von $\boldsymbol{A}_l\boldsymbol{v}_l = \boldsymbol{d}_l$ bzw. $\boldsymbol{A}_l\boldsymbol{y} = \boldsymbol{b}_l$ ein Fixpunkt der Multigriditeration,

denn:

$\boldsymbol{v}_l$ ist Fixpunkt der Vorglättung	$\Longrightarrow$	
Für den Defekt gilt:	$\boldsymbol{d}_l - \boldsymbol{A}_l\boldsymbol{v}_l$	$= \boldsymbol{0}$
Restriktion des Defekts		$= \boldsymbol{0}$
Grobgitterlösung der Defektgleichung		$= \boldsymbol{0}$
Prolongation der GrobgitterLösung		$= \boldsymbol{0}$
Korrektur von $\boldsymbol{v}_l$:	$\boldsymbol{v}_l$	$= \boldsymbol{v}_l + \boldsymbol{0}$
$\boldsymbol{v}_l$ ist Fixpunkt der Nachglättung		

Man kann also für die Konstruktion der Iterationsmatrix des MGV, das ein lineares Verfahren ist, die Eigenschaft (19.3) verwenden und weiß dann, daß das MGV gegen die Lösung der Aufgabe konvergiert, falls $\|\boldsymbol{M}\| < 1$ ist.

Konstruktion der Iterationsmatrix $\boldsymbol{M}_l$ des MGV auf Level l

Auf Grund der Eigenschaften des MGV ist die Konstruktion notwendigerweise rekursiv. Wir konstruieren $\boldsymbol{M}_l$ in Abhängigkeit von der Iterationsmatrix $\boldsymbol{M}_{l-1}$ des MGV auf Gitter $l-1$, indem wir von Gitter l auf Gitter $l-1$ heruntersteigen. Dort ist die Korrekturgleichung $\boldsymbol{A}_{l-1}\boldsymbol{v}_{l-1} = \boldsymbol{d}_{l-1}$ mit der Ausgangsnäherung $\boldsymbol{v}_{l-1}^0 = \boldsymbol{0}$ zu lösen. Dies geschieht wieder durch das MGV indem man pro Iterationsschritt bis zum gröbsten Gitter $l = 0$ heruntersteigt und dann wieder bis zum Level $l-1$ hochkommt. Dabei benutzen wir für $\boldsymbol{M}_{l-1}$ die Darstellung (19.5). Schließlich steigen wir wieder zum Gitter l auf.
Wir listen die einelnen Schritte auf. Hierbei ist das Aussehen des Terms der jeweiligen Iterationen, der nur von der rechten Seite der zu lösenden Gleichung abhängt, zwar konstruierbar, aber für die Konstruktion der Iterationsmatrix ohne Interesse. Wir konstruieren die Iterationsmatrix für das MGV.

$$\boldsymbol{y}_l^0 = \text{Ausgangsnäherung auf Gitter } l \text{ zur Lösung von } \boldsymbol{A}_l\boldsymbol{y}_l = \boldsymbol{b}_l$$

$$\bar{\boldsymbol{y}}_l = \boldsymbol{S}^{\nu_1}\boldsymbol{y}_l^0 + \boldsymbol{Q}_{\nu_1}(\boldsymbol{b}_l) \quad \text{Vorglättung}$$

$$\bar{\bar{\boldsymbol{y}}}_l = \bar{\boldsymbol{y}}_l + \boldsymbol{P}\boldsymbol{v}_{l-1}^\gamma$$

Dabei ist $\boldsymbol{v}_{l-1}^{\gamma}$ die Näherung für die Lösung der Korrekturgleichung, die man, ausgehend von $\boldsymbol{v}_{l-1}^{0} = \mathbf{0}$ nach γ Schritten des MGV auf dem Level $l-1$ erhält.

$$\begin{aligned}\boldsymbol{y}_l^1 &= \boldsymbol{S}^{\nu_2}\bar{\bar{\boldsymbol{y}}}_l + \boldsymbol{Q}_{\nu_2}(\boldsymbol{b}_l) \quad \text{Nachglättung} \\ &= \boldsymbol{S}^{\nu_2}(\bar{\boldsymbol{y}}_l + \boldsymbol{P}\boldsymbol{v}_{l-1}^{\gamma}) + \boldsymbol{Q}_{\nu_2}(\boldsymbol{b}_l).\end{aligned}$$

Mit $\bar{\boldsymbol{v}}_{l-1}^{\gamma} \overset{(19.5)}{=} (\boldsymbol{I} - \boldsymbol{M}_{l-1}^{\gamma})\boldsymbol{A}_{l-1}^{-1}\boldsymbol{d}_{l-1}$ und $\boldsymbol{d}_{l-1} = \boldsymbol{R}(\boldsymbol{b}_l - \boldsymbol{A}_l\bar{\boldsymbol{y}}_l)$ folgt

$$\begin{aligned}\boldsymbol{y}_l^1 &= \boldsymbol{S}^{\nu_2}\{\bar{\boldsymbol{y}}_l + \boldsymbol{P}(\boldsymbol{I} - \boldsymbol{M}_{l-1}^{\gamma})\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}(\boldsymbol{b}_l - \boldsymbol{A}_l\bar{\boldsymbol{y}}_l)\} + \boldsymbol{Q}_{\nu_2}(\boldsymbol{b}_l) \\ &= \boldsymbol{S}^{\nu_2}\{\bar{\boldsymbol{y}}_l - \boldsymbol{P}(\boldsymbol{I} - \boldsymbol{M}_{l-1}^{\gamma})\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l\bar{\boldsymbol{y}}_l\} + \boldsymbol{Q}_3(\boldsymbol{b}_l)\end{aligned}$$

$\boldsymbol{Q}_3(\boldsymbol{b}_l), \boldsymbol{Q}_4(\boldsymbol{b}_l)$ bezeichnen Restausdrücke, die nur von den jeweiligen rechten Seiten abhängig sind

Ausmultiplizieren, Ausklammern und Einsetzen von $\bar{\boldsymbol{y}}_l$ ergibt

$$\begin{aligned}\boldsymbol{y}_l^1 &= \boldsymbol{S}^{\nu_2}\{\boldsymbol{I} - \boldsymbol{P}\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l + \boldsymbol{P}\boldsymbol{M}_{l-1}^{\gamma}\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l\}\underbrace{(\boldsymbol{S}^{\nu_1}\boldsymbol{y}_l^0 + \boldsymbol{Q}_{\nu_1}(\boldsymbol{b}_l))}_{\bar{\boldsymbol{y}}_l} + \boldsymbol{Q}_3(\boldsymbol{b}_l) \\ &= \underbrace{\{\underbrace{\boldsymbol{S}^{\nu_2}[\boldsymbol{I} - \boldsymbol{P}\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l]\boldsymbol{S}^{\nu_1}}_{\boldsymbol{K}_l(\nu_1,\nu_2)} + \boldsymbol{S}^{\nu_2}\boldsymbol{P}\boldsymbol{M}_{l-1}^{\gamma}\underbrace{\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l\boldsymbol{S}^{\nu_1}}_{=:\boldsymbol{E}_{l-1}}\}}_{\textit{Iterationsmatrix des MGV}}\boldsymbol{y}_l^0 + \boldsymbol{Q}_4(\boldsymbol{b}_l)\end{aligned}$$

$\boldsymbol{K}_l(\nu_1, \nu_2)$ ist die Iterationsmatrix des ZGV (vgl. (18.9)) auf Level l, $l \geq 1$, wenn $l = 0$ das gröbste Gitter ist.
Wir haben damit

Satz 19.1 Iterationsmatrix des MGV für $l \geq 1$
Bezeichnet $l = 0$ das gröbste Gitter, so gilt für die Iterationsmatrix $\boldsymbol{M}_l$ des MGV auf Gitter l

$$\boldsymbol{M}_l = \boldsymbol{K}_l(\nu_1, \nu_2) + \boldsymbol{S}^{\nu_2}\boldsymbol{P}\boldsymbol{M}_{l-1}^{\gamma}\boldsymbol{E}_{l-1}, \quad l \geq 1, \ \boldsymbol{M}_0 = \mathbf{0}$$

mit der Iterationsmatrix des ZGV

$$\boldsymbol{K}_l(\nu_1, \nu_2) = \boldsymbol{S}^{\nu_2}[\boldsymbol{I} - \boldsymbol{P}\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l]\boldsymbol{S}^{\nu_1}$$

und

$$\boldsymbol{E}_{l-1} = \boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l\boldsymbol{S}^{\nu_1}$$

Beachte zum Anfangswert: Für $l = 1$ liegt ein ZGV vor, also gilt

$$\boldsymbol{S}^{\nu_2}\boldsymbol{P}\boldsymbol{M}_{l-1}^{\gamma}\boldsymbol{E}_{l-1} = \mathbf{0}.$$

Dies wird erfüllt durch $\boldsymbol{M}_0 = \mathbf{0}$.

Vorbemerkungen zum Konvergenzsatz

1. Man kann die Iterationsmatrix des MGV betrachten als die mit einer Störung versehene Iterationsmatrix des ZGV. Den Konvergenzbeweis kann man führen, indem man durch Voraussetzungen dafür sorgt, daß die Summe beider Terme normmäßig < 1 ausfällt.

2. Ausschlaggebend für die Konvergenz ist der Spektralradius ρ der Iterationsmatrix. Er hängt, soweit es die Glättungen betrifft, wegen $\rho(\boldsymbol{UF}) = \rho(\boldsymbol{FU})$ (vgl. die Bemerkung nach (18.12)) nur von der absoluten Zahl $\nu = \nu_1 + \nu_2$ der Glättungen ab und nicht davon, ob sie auf Vor- oder Nachglättung verteilt sind.
   Wir führen den Konvergenzbeweis für die Variante $\nu = \nu_1$ ($\nu_2 = 0$, ohne Nachglättung.

3. Sind H_1, H_2 endlichdimensionale Hilberträume mit den inneren Produkten $(\ ,\ )_{H_1}$ und $(\ ,\ )_{H_2}$ und den von ihnen erzeugten Normen und ist

$$\boldsymbol{F} :\ H_1 \xrightarrow{linear} H_2$$

   eine lineare Abbildung, so wird die Operatornorm definiert durch

$$\|\boldsymbol{F}\|_{H_1 \to H_2} := \sup_{\boldsymbol{x} \in H_1,\, \boldsymbol{x} \neq 0} \frac{\|\boldsymbol{Fx}\|_{H_2}}{\|\boldsymbol{x}\|_{H_1}}. \tag{19.7}$$

   Man beachte, daß dadurch auch für nicht quadratische Matrizen eine Norm erklärt wird. (Übung: Normeigenschaften nachweisen)
   Wir verwenden diese Normdefinition für nichtquadratische Matrizen.
   Ist $H_1 = H_2 = \mathbb{R}^\mu$, $\mu \in \mathbb{N}$, so wird durch (19.7) die Spektralnorm definiert.

   Zur Definition der Norm der Prolongationsmatrix gemäß (19.7) verwenden wir für benachbarte Gitter folgende folgende innere Produkte und die von ihnen erzeugten Normen (vgl. dazu Definition (17.1) und die anschließende Anwendung):

$$\begin{array}{rcl} (\ ,\ )_{l,2} & = & \text{euklidisches Produkt in } \omega_l \\ \left\{ \begin{array}{l} (\ ,\ )_l \\ (\ ,\ )_{l-1} \end{array} \right. & \begin{array}{c} = \\ = \end{array} & \left. \begin{array}{l} (\ ,\ )_{l,2} h_1 h_2 \\ (\ ,\ )_{l-1,2}\, 4 h_1 h_2 \end{array} \right. \quad \text{im 2D-Fall} \end{array} \tag{19.8}$$

   Ansonsten verwenden wir durchgehend, ohne Kennzeichnung, auf allen Gittern die Euklidische Vektornorm und die zugeordnete Operatornorm (Spektralnorm).

Satz 19.2 Konvergenzsatz für das MGV (ohne Nachglättung)
Es seien folgende Voraussetzungen erfüllt

(19.9) $\|\boldsymbol{P}\| \leq 1 \quad \boldsymbol{P} = \text{Prolongationsmatrix } (vgl.(19.7), (19.8))$

(19.10) $\|\boldsymbol{Px}\| \geq c_p \|\boldsymbol{x}\| \ \forall \boldsymbol{x},$ für ein $c_p > 0$ ($c_p \leq 1$ wegen (19.9))

(19.11) $\|\boldsymbol{S}\| \leq 1, \quad \boldsymbol{S} = \text{Iterationsmatrix der Glättung}$

(19.12) $\|\boldsymbol{K}_l(\nu)\| \leq \xi(\nu) < \dfrac{\gamma - 1}{\gamma} \left(\dfrac{c_p}{2\gamma}\right)^{\frac{1}{\gamma-1}}$ für

(19.13) $\gamma \geq 2, \qquad \gamma = \text{Wiederholungszahl des MGV auf den Gittern } l \leq l_{max} - 1.$

Dabei ist $\xi(\nu)$ eine Normschranke für die Iterationsmatrix $\boldsymbol{K}_l$ des ZGV auf Level l $1 \leq l \leq l_{max}$, $(\nu = \nu_1,\ \nu_2 = 0)$.

Dann gilt für die Iterationsmatrix $\boldsymbol{M}_l$ des MGV für $l \geq 1$

(19.14) $$\zeta_l := \|\boldsymbol{M}_l\| \leq \frac{\gamma}{\gamma - 1}\,\xi(\nu) < \left(\frac{c_p}{2\gamma}\right)^{\frac{1}{\gamma-1}} \qquad (< 1 \text{ gemäß } (19.10))$$

Bemerkungen zu den Voraussetzungen:

zu (19.9)

$\|\boldsymbol{P}\| \leq 1$ ist für die Interpolation als Prolongation erfüllt. Für die lineare Interpolation (9-Punkte-Stern) rechnet man das im Einheitsquadrat (etwas langwierig, aber ohne Schwierigkeiten) nach.

Hinweise: Gemäß (19.7), (19.8) ist

$$\|\boldsymbol{P}\| = \sup_{x \neq 0} \sqrt{\frac{(\boldsymbol{Px}, \boldsymbol{Px})_l}{(\boldsymbol{x}, \boldsymbol{x})_{l-1}}}$$

und damit

$$\|\boldsymbol{P}\| = \sup_{x \neq 0} \sqrt{\frac{(\boldsymbol{Px}, \boldsymbol{Px})_{l,2}\, h_1 h_2}{(\boldsymbol{x}, \boldsymbol{x})_{l-1,\,2}\, 4 h_1 h_2}} = \sup_{x \neq 0} \sqrt{\frac{1}{4}\, \frac{(\boldsymbol{Px}, \boldsymbol{Px})_{l,2}}{(\boldsymbol{x}, \boldsymbol{x})_{l-1,\,2}}}$$

Unter Benutzung der Interpolationsformeln (17.6)-(17.8) schätzt man ab

$$(\boldsymbol{Px}, \boldsymbol{Px})_{l,2} \leq 4\,(\boldsymbol{x}, \boldsymbol{x})_{l-1,\,2} \quad \Longrightarrow \quad \|\boldsymbol{P}\| \leq 1.$$

Hinweis: Verwende bei der Abschätzung die Ungleichung $2x_i x_{i+1} \leq (x_i^2 + x_{i+1}^2)$ für aufeinanderfolgende Variablenwerte im groben Gitter.

zu (19.10)

Dies ist eine Invertierbarkeitsvoraussetzung für $\boldsymbol{P}$ auf dem Bildraum von $\boldsymbol{P}$. Man sieht, (zumindest bei allen uns bekannten Prolongationen), daß unsere Prolongationsmatrizen Höchstrang haben, woraus folgt

$$\text{Ker}\boldsymbol{P} = \{0\} \quad \Longrightarrow \quad \min_{\|\boldsymbol{x}\|_{l-1,\,2}=1} \|\boldsymbol{Px}\|_{l,\,2} =: c_p > 0.$$

So ein c_p existiert also. Werte für c_p erhält man aus der Konstruktion von $\boldsymbol{P}$ im Zusammenhang mit dem Beweis der Approximationseigenschaft. (vgl. etwa Hackbusch: Multigrid Methods Lemma 6.3.13)

zu (19.11)

1. Das gedämpfte Jacobi-Verfahren:
   In der Anwendung haben wir uns auf Matrizen $\boldsymbol{A} = \boldsymbol{A}^T > 0$ beschränkt, für die die Diagonalelemente alle konstant waren. Die Iterationsmatrix $\boldsymbol{S}$ des gedämpften Jacobi-Verfahrens war

   $$\boldsymbol{S} = \boldsymbol{I} - \tilde{\omega}\boldsymbol{A} \quad \text{(vgl. (18.21))}$$

   und es galt für die Spektralnorm

   $$\|\boldsymbol{S}\|_S = \|\boldsymbol{I} - \tilde{\omega}\boldsymbol{A}\|_S = \max_{i=1,\dots,n} |1 - \tilde{\omega}\lambda_i|, \quad \lambda_i \in \sigma(\boldsymbol{A})$$

   Nun ist $\|\boldsymbol{S}\| \leq 1$ (19.11) erfüllt, falls $|1 - \tilde{\omega}\lambda_i| \leq 1 \ \forall\, i$.
   Wegen $\boldsymbol{A} = \boldsymbol{A}^T > 0$ ist $\lambda_i > 0 \ \forall\, i$, also ist die Forderung wegen $\tilde{\omega} > 0$ äquivalent zu

   $$\tilde{\omega}\lambda_i \leq 2 \ \forall\, i \quad \text{bzw.} \quad \tilde{\omega} \leq \frac{2}{\lambda_i} \ \forall\, i,$$

   was erfüllt ist, falls

   $$\tilde{\omega} \leq \frac{2}{\|\boldsymbol{A}\|_S}.$$

   Da die Glättungseigenschaft $\tilde{\omega} \leq \dfrac{1.2}{\|\boldsymbol{A}\|_S}$ verlangte, gilt (19.11) für den gedämpften Jacobi.

2. Das Gauß-Seidel-Verfahren
   erfüllt $\|S\| \leq 1$ (vgl. Satz (15.6)). Damit ist die Voraussetzung bei der Schachbrettanordnung erfüllt.

3. Das symmetrische Gauß-Seidel-Verfahren
   Sind $\boldsymbol{S}_1$, bzw. $\boldsymbol{S}_2$ die Iterationsmatrizen des vorwärts- bzw. rückwärtsgenommenen Gauß-Seidel-Verfahrens, so ist die Iterationsmatrix des symmetrischen Verfahrens $\boldsymbol{S} = \boldsymbol{S}_1\boldsymbol{S}_2$ (vgl. (15.17)).
   $\|\boldsymbol{S}_1\| < 1$ wurde in Satz (15.5) gezeigt, analog zeigt man $\|\boldsymbol{S}_2\| < 1$ und damit folgt $\|\boldsymbol{S}\| = \|\boldsymbol{S}_1\|\,\|\boldsymbol{S}_2\| < 1$.

zu (19.12)

Dies ist eine verschärfte Forderung an die Iterationsmatrix des ZGV auf Level l, die durch hinreichend großes $\nu = \nu_1 + \nu_2$ immer erfüllt werden kannn. Glättungs- und Approximationseigenschaften sind in (19.12) enthalten.

Da $c_p \leq 1$, erhält man folgende Schranken für $\xi(\nu)$

$$\frac{\gamma-1}{\gamma}\left(\frac{c_p}{2\gamma}\right)^{\frac{1}{\gamma-1}} \leq \begin{cases} \frac{1}{8} & = 0.125 c_p & \leq 0.125 & \text{für } \gamma = 2 \\ \frac{2}{3}\sqrt{\frac{1}{6}}\sqrt{c_p} & \approx 0.272\sqrt{c_p} & \leq 0.272 & \text{für } \gamma = 3 \\ \frac{4}{5}\sqrt[3]{\frac{1}{8}}\sqrt[3]{c_p} & = 0.4\sqrt[3]{c_p} & \leq 0.4 & \text{für } \gamma = 4 \end{cases}$$

zu (19.13)

$\gamma \geq 2$ ist eine beweistechnische Voraussetzung. Für $\gamma = 1$ muß ein anderer Beweis gemacht werden, der wesentlich aufwendiger ist.

Der **Beweis des Konvergenzsatzes** gliedert sich in zwei Teile

1. Gemäß der rekursiven Definition der Iterationsmatrix wird eine Rekursionformel für $\zeta_l = \|\boldsymbol{M}_l\|$ hergeleitet.
2. Beweis des Konvergenzsatzes induktiv mit Hilfe von 1: Zeige $\|\boldsymbol{M}_l\| < \frac{\gamma}{\gamma-1}\,\xi(\nu)$.

Beweis 1:
Aus der Gestalt von $\boldsymbol{M}_l$ (Satz (19.1)) folgt für $l \geq 1$ und $\nu = \nu_1,\ \nu_2 = 0$

$$\begin{aligned} \zeta_l := \|\boldsymbol{M}_l\| &\leq \xi(\nu) + \|\boldsymbol{P}\boldsymbol{M}_{l-1}^{\gamma}\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l\boldsymbol{S}^{\nu}\| \\ &\overset{\|\boldsymbol{P}\|\leq 1}{\leq} \xi(\nu) + \|\boldsymbol{M}_{l-1}^{\gamma}\|\,\|\underbrace{\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l\boldsymbol{S}^{\nu}}_{\boldsymbol{E}_{l-1}}\| \\ &\leq \xi(\nu) + \zeta_{l-1}^{\gamma}\|\boldsymbol{E}_{l-1}\|. \end{aligned} \tag{19.15}$$

Wegen $\boldsymbol{M} = \boldsymbol{0}$ ist $\zeta_0 = 0$.
$\boldsymbol{E}_{l-1}$ ist im Wesentlichen aus der Darstellung der Iterationsmatrix $\boldsymbol{K}_l(\nu)$ des ZGV bekannt (vgl. (18.9)). Dies wird zur Abschätzung benutzt.

$$\boldsymbol{K}_l(\nu) = \boldsymbol{S}^{\nu} - \boldsymbol{P}\underbrace{\boldsymbol{A}_{l-1}^{-1}\boldsymbol{R}\boldsymbol{A}_l\boldsymbol{S}^{\nu}}_{\boldsymbol{E}_{l-1}}.$$

Hieraus folgt mit der Dreiecksungleichung (rückwärts)

$$\|\boldsymbol{P}\boldsymbol{E}_{l-1}\| \leq \underbrace{\|\boldsymbol{K}_l(\nu)\|}_{<1} + \underbrace{\|\boldsymbol{S}^{\nu}\|}_{\leq 1} < 2. \qquad \text{(Brutalabschätzung)}$$

Wegen $\|\boldsymbol{P}\boldsymbol{x}\| \geq c_p\|\boldsymbol{x}\|$ (vgl. (19.10)), folgt (indirekter Beweis unter Beachtung von (19.7))

$$\begin{aligned} c_p\|\boldsymbol{E}_{l-1}\| &\leq \|\boldsymbol{P}\boldsymbol{E}_{l-1}\| < 2, \\ \|\boldsymbol{E}_{l-1}\| &< \frac{2}{c_p}, \end{aligned}$$

Aus (19.15) folgt damit für $l \geq 1$

(19.16) $$\zeta_l \leq \xi(\nu) + \frac{2}{c_p}\,\zeta_{l-1}^{\gamma}.$$

Beweis 2:
Wir beweisen durch vollständige Induktion: (Abkürzung: $\xi := \xi(\nu)$)

(19.17) $$\zeta_l \leq \xi(\nu) + \frac{2}{c_p}\,\zeta_{l-1}^{\gamma} \leq \frac{\gamma}{\gamma-1}\,\xi(\nu).$$

$l = 1:$ richtig nach (19.15), da $\zeta_0 = 0$.
Induktionsvoraussetzung: $\zeta_{l-1} \leq \dfrac{\gamma}{\gamma-1}\,\xi$.
Aus (19.16) folgt durch Eintragen der Induktionsvoraussetzung

$$\zeta_l \leq \xi + \frac{2}{c_p}\,\zeta_{l-1}^{\gamma} \leq \xi + \frac{2}{c_p}\left(\frac{\gamma}{\gamma-1}\xi\right)^{\gamma}.$$

Wir zeigen, daß in Verschärfung von(19.17) sogar gilt

$$\xi + \frac{2}{c_p}\left(\frac{\gamma}{\gamma-1}\,\xi\right)^{\gamma} \leq \frac{\gamma}{\gamma-1}\,\xi,$$

wie sich aus der folgenden, äquivalenten Umformung ergibt

$$\begin{aligned}
\xi + \frac{2}{c_p}\left(\frac{\gamma}{\gamma-1}\,\xi\right)^{\gamma} &\leq \frac{\gamma}{\gamma-1}\,\xi \\
1 + \frac{2}{c_p}\left(\frac{\gamma}{\gamma-1}\right)^{\gamma}\xi^{\gamma-1} &\leq \frac{\gamma}{\gamma-1} \\
\left(\frac{\gamma}{\gamma-1}\right)^{\gamma}\xi^{\gamma-1} &\leq \frac{c_p}{2(\gamma-1)} \\
\xi^{\gamma-1} &\leq \frac{c_p}{2}\,\frac{(\gamma-1)^{\gamma-1}}{\gamma^{\gamma}} = \frac{c_p}{2\gamma}\left(\frac{\gamma-1}{\gamma}\right)^{\gamma-1} \\
\xi &\leq \frac{\gamma-1}{\gamma}\left(\frac{c_p}{2\gamma}\right)^{\frac{1}{\gamma-1}}
\end{aligned}$$

Die letzte Ungleichung ist aber die Voraussetzung (19.12). ■

Überlegungen zur Iterationszahl des einfachen MGV

(d.h. ohne Nachglättung)

zur Erreichen der vorgegebenen Genauigkeit

Sei
$\boldsymbol{M}_l$ =die Iterationsmatrix des MGV auf dem l-ten Gitter

$\boldsymbol{y}_l$ = exakte Lösung von $\boldsymbol{A}_l\boldsymbol{y}_l = \boldsymbol{b}_l$

$\boldsymbol{y}_l^{(0)}$ = Ausgangsnäherung für $\boldsymbol{y}_l$

$\boldsymbol{y}_l^{(m)}$ = m-te Näherung für $\boldsymbol{y}_l$ nach m Schritten des MGV, ausgehend von $\boldsymbol{y}_l^{(0)}$.

Da $\boldsymbol{y}_l$ ein Fixpunkt des MGV ist, da auf allen Gitern nur Iterationen verwendet werden, welche die Lösung der jeweiligen gleichung als Fixpunkt haben (vgl. u.a. (19.6)), gilt für den Fehler

$$\boldsymbol{y}_l^{(m)} - \boldsymbol{y}_l = \boldsymbol{M}_l(\boldsymbol{y}_l^{(m)} - \boldsymbol{y}_l), \quad \text{also mit (19.14) und } \xi := \xi(\nu)$$

$$\begin{aligned} \|\boldsymbol{y}_l^{(m)} - \boldsymbol{y}_l\| &\le \left(\frac{\gamma}{\gamma-1}\,\xi\right) \|\boldsymbol{y}_l^{(m-1)} - \boldsymbol{y}_l\| \\ &\le \left(\frac{\gamma}{\gamma-1}\,\xi\right)^m \|\boldsymbol{y}_l^{(0)} - \boldsymbol{y}_l\|. \end{aligned}$$

Eine naheliegende Forderung für eine ε-Fehlergenauigkeit könnte sein

$$\left(\frac{\gamma-1}{\gamma}\,\xi\right)^m \le \varepsilon \approx \text{Approximierungsfehler}, \quad \|\boldsymbol{y}_l^{(0)} - \boldsymbol{y}_l\| \approx 1. \tag{19.18}$$

Für die notwendige Iterationszahl folgt hieraus

$$m = \left[\frac{\ln\varepsilon}{\ln\left(\frac{\gamma-1}{\gamma}\,\xi\right)}\right],$$

wobei [ ] die nächst größere, ganze Zahl bedeutet. Beachte: $\ln(\varepsilon) < 0$, es ist an $\varepsilon < 1$ gedacht, ebenso im Nenner $\ln\left(\frac{\gamma-1}{\gamma}\,\xi\right) < 0$.

Die Abschätzung ist unbefriedigend, weil der Diskretisierungsfehler bestenfalls asymptotisch bekannt ist und man somit keine Aussage über die notwendige Iterationszahl erhält, und weil auf dem Gitter l die Differenz $\|\boldsymbol{y}_l^{(0)} - \boldsymbol{y}_l\|$ auch nicht so ohne weiteres abschätzbar ist, $\boldsymbol{y}_l$ ist ja unbekannt.
Abhilfe kann hier die Untersuchung des vollen MGV (also mit Anlaufiteration, vgl.§ 16, Volles MGV) schaffen, die ja die Anfangsnäherungen mitliefert.

Gewünscht wird eine Abschätzung $\|\boldsymbol{y}_l^{(0)} - \boldsymbol{y}_l\| = O(h_l^\kappa,)$, wobei $O(h_l^\kappa)$ der Diskretisierungsfehler auf Level l ist, unter Beachtung des Ausgangsfehlers einer Mehrgitteriteration. Wir beweisen eine entsprechende Abschätzung unter den

Voraussetzungen (19.19)-(19.21):

$$h_l = 2\,h_{l-1}. \tag{19.19}$$

Dies beschreibt den Standartfall der Schrittweitenhalbierung von einem Gitter zum nächst feineren.
(19.20)
$$\|\widetilde{\boldsymbol{P}}\| \le 1, \quad \|\widetilde{\boldsymbol{P}}\boldsymbol{y}_l - \boldsymbol{y}_{l+1}\| \le c_0 h_{l+1}^{\kappa}. \quad \text{(Approx.-Eigenschaft, vgl. (18.18),(18.19))}$$

Hierbei ist $\widetilde{\boldsymbol{P}}$ die ggf. bessere Prolongation, die bei der Anlaufiteration des vollen MGV benutzt wird.
h^{κ} beschreibt die Diskretisierungsordnung, $\boldsymbol{y}_l, \boldsymbol{y}_{l+1}$ sind die exakten Lösungen der Gleichungen auf Level l, bzw. $l+1$. (vgl. dazu die Abschätzungen zum Approximationsteil)
$c_0 = c_0(u)$ sei eine Konstante, die unabhängig vom Gitter ist, aber abhängig sein darf von der exakten Lösung u der Randwertaufgabe.
Von der Qualität her besagt diese Voraussetzung, daß $\widetilde{\boldsymbol{P}}$ so gut sein soll, dass die Approximation $\widetilde{\boldsymbol{P}}\boldsymbol{y}_k$ die Approximationsgenauigkeit der Diskretisierung nicht verdirbt.

$$\|\boldsymbol{M}_l\| \le \zeta_l, \quad \zeta := \max_l \zeta_l, \tag{19.21}$$

d.h. ζ, unabhängig von l_{max}, ist eine gleichmäßige obere Schranke für die Kontraktionszahlen der Iterationsmatrizen des MGV auf allen Gittern. (vgl. (19.14)). Wenn man die Abschätzungszahlen für das ZGV kennt, kann man aus Satz 19.2 Werte für ζ erhalten, natürlich in Abhängigkeit von γ.
Wir beweisen den

Satz 19.3
Die Voraussetzungen (19.19)-(19.21) seien erfüllt. Sei $\boldsymbol{y}_0$ die exakte Lösung von $\boldsymbol{A}_0\boldsymbol{y}_0 = \boldsymbol{d}_0$ auf dem gröbsten Gitter $l = 0$.
Das volle MGV, das auf jedem Gitterniveau μ $(\mu = \mu_1 = \mu_2)$ einfache MGV-Schritte verwendet, liefert, ausgehend von $\boldsymbol{y}_0$, die Näherungslösung $\tilde{\boldsymbol{y}}_l$ für $\boldsymbol{A}_l\boldsymbol{y}_l = \boldsymbol{b}_l$.
Sei $\zeta = \max_l \zeta_l = \max_l \|\boldsymbol{M}_l\|$ und μ genüge der Forderung

$$\zeta^{\mu} \le \frac{1}{1+2^{\kappa}}, \qquad (\kappa = \text{Diskretisierungsordnung}) \tag{19.22}$$

so gilt

$$\|\widetilde{\boldsymbol{y}}_l - \boldsymbol{y}_l\| \le c_0 h^{\kappa} \qquad (c_0 \text{ laut } (19.20)) \tag{19.23}$$

Beachte: (19.22) ist eine Bedingung an die Normen $\|\boldsymbol{M}_l\|$, die durch entsprechend großes ν (Glättungsanzahl) erfüllt werden kann.

Wir verdeutlichen die Bedeutung der Konstanten μ (siehe obiger Satz) und γ (vgl. Satz 19.2 und Beweis des Satzes 19.3) durch Beispiele: **Eine** Mehrgitteriteration mit der Iterationsmatrix $\boldsymbol{M}_l$ und $\gamma = 2$ ist für $l = 2,3,4$ in den Abb. 4-6 auf S. 132

dargestellt. μ gibt an, wie oft eine Iteration mit $\boldsymbol{M}_l$ auf jedem Gitterlevel ausgeführt werden soll.

Beweis: Wir zeigen induktiv

$$\|\widetilde{\boldsymbol{y}}_l - \boldsymbol{y}_l\| \le c_0 \frac{\zeta^\mu}{1-2^\kappa\zeta^\mu} h_l^\kappa. \tag{19.24}$$

Beachte hierzu: Der Bruch in (19.24) ist in ζ monoton wachsend. Wird in (19.24) die Abschätzung (19.22) eingetragen: $\zeta^\mu \le \frac{1}{1+2^\kappa}$, also $2^\kappa\zeta^\mu \le \frac{2^\kappa}{1+2^\kappa} < 1$, so ist der Nenner in (19.24) erklärt und es folgt

$$\frac{\zeta^\mu}{1-2^\kappa\zeta^\mu} \le \frac{\frac{1}{1+2^\kappa}}{1-\frac{2^\kappa}{1+2^\kappa}} = 1, \quad \text{also die Behauptung (19.23).}$$

Für $l = 0$ ist (19.24) erfüllt, da auf $l = 0$ exakt gelöst wird.

Sei $\widetilde{\boldsymbol{y}}_l$ die Näherungslösung für $\boldsymbol{A}\boldsymbol{y}_l = \boldsymbol{b}_l$ nach μ Iterationsschritten, ausgehend von der Näherungslösung $\widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{y}}_{l-1}$, so folgt (beachte: Die exakte Lösung $\boldsymbol{y}_l$ von $\boldsymbol{A}_l\boldsymbol{y}_l = \boldsymbol{d}$, bzw. $\boldsymbol{b}$ ist Fixpunkt der Iterationsverfahren auf dem jeweiligen Gitter

$$\begin{aligned}
\widetilde{\boldsymbol{y}}_l - \boldsymbol{y}_l &= (\boldsymbol{M}_l)^\mu(\widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{y}}_{l-1} - \boldsymbol{y}_l) \\
\|\widetilde{\boldsymbol{y}}_l - \boldsymbol{y}_l\| &\le \zeta_l^\mu \|\widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{y}}_{l-1} - \boldsymbol{y}_l\| \quad \Longrightarrow \\
&\le \zeta^\mu(\|\widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{y}}_{l-1} - \widetilde{\boldsymbol{P}}\boldsymbol{y}_{l-1}\| + \|\widetilde{\boldsymbol{P}}\boldsymbol{y}_{l-1} - \boldsymbol{y}_l\|) \quad \xRightarrow{(19.20),(19.21)} \\
&\le \zeta^\mu(\|\widetilde{\boldsymbol{y}}_l - \boldsymbol{y}_l\| + c_0 h_l^\kappa) \quad \xRightarrow{\text{Induktionsvoraussetzung}} \\
&\le \zeta^\mu\Big(c_0 h_{l-1}^\kappa \frac{\zeta^\mu}{1-2^\kappa\zeta^\mu} + c_0 h_l^\kappa\Big) \quad \xRightarrow{(19.19)} \\
&\le \zeta^\mu\Big(c_0 h_l^\kappa \frac{2^\kappa\zeta^\mu}{1-2^\kappa\zeta^\mu} + c_0 h_l^\kappa\Big) \\
&\le c_0 h_l^\kappa \zeta^\mu\Big(\frac{2^\kappa\zeta^\mu}{1-2^\kappa\zeta^\mu} + 1\Big) \\
&= c_0 h_l^\kappa \zeta^\mu \frac{1}{1-2^\kappa\zeta^\mu}. \qquad \blacksquare
\end{aligned}$$

Beachte zur **Anwendung von Satz** (19.3):

Es hat sich als numerisch sinnvoll erwiesen zu wählen:

$$\begin{aligned}
&\nu \approx 3 \\
&\gamma \ge 3 \qquad \text{im 2-dimensionalen Fall} \\
&\gamma \ge 7 \qquad \text{im 3-dimensionalen Fall}
\end{aligned}$$

Wir machen eine Überschlagrechnung zur Anwendung von Satz (19.3) für $\gamma = 3$ und

$\kappa = 2$ (quadratische Dikretisierungsordnung). Wir erinnern dazu an das Herkommen der verschiedenen Konstanten.

$$\|\boldsymbol{P}x\| \geq c_p\|x\|, \quad 0 < c_p \leq 1 \quad \text{vgl. (19.10)}$$

$$\xi(\nu) < \frac{\gamma-1}{\gamma}\left(\frac{c_p}{2\gamma}\right)^{\frac{1}{\gamma-1}} \overset{c_p\leq 1}{\leq} \frac{\gamma-1}{\gamma}\left(\frac{1}{2\gamma}\right)^{\frac{1}{\gamma-1}} \qquad \text{vgl. Satz 19.2 und}$$

$$\frac{\gamma-1}{\gamma}\left(\frac{1}{2\gamma}\right)^{\frac{1}{\gamma-1}} \leq 0.272\sqrt[3]{c_p} \quad \text{für } \gamma = 3 \quad \text{(vgl. Satz 19.3)}$$

$$\zeta_l \leq \zeta \leq \frac{\gamma}{\gamma-1}\xi(\nu) \quad \text{(vgl. (19.20))} \quad \Longrightarrow$$

$$\zeta \leq \frac{3}{2}0.272\sqrt[3]{c_p} = 0.408\sqrt[3]{c_p} \quad \text{und mit (19.22)}$$

$$(0.408\sqrt[3]{c_p})^{\mu} \leq \frac{1}{1+2^2} = \frac{1}{5}$$

$$\mu \geq \frac{|\ln 5|}{|\ln(0.408) + \ln(\sqrt[3]{c_p})|} \overset{c_p\approx 1}{\approx} \frac{|\ln 5|}{|\ln(0.408)|} \approx 1.8,$$

d.h. mit $\mu = 2$ ist man gut beraten.

Bemerkung: Eine analoge Berechnung für $\gamma = 2$ zeigt, daß auch in diesem Fall $\mu = 2$ ausreichend ist. Man beachte jedoch, daß dann die Konvergenzbedingung von

Satz 19.3: $\xi(\nu) < \frac{\gamma-1}{\gamma}\left(\frac{c_p}{2\gamma}\right)^{\frac{1}{\gamma-1}}$ für $\nu = 3$ nicht erfüllbar ist.

§ 20 MGV für nichtlineare Probleme

Gelöst werden soll

$$\boldsymbol{A}_l(\boldsymbol{y}_l) = \boldsymbol{f}_l, \tag{20.1}$$

ein nichtlineares Gleichungssystem, das z.B. aus der Diskretisierung der folgenden Aufgabe kommen könnte

$$\sum_{i=1}^{3} \frac{\partial}{\partial x_i}\left(k(u)\frac{\partial u}{\partial x_i}\right) + f(u) = 0 \quad \text{in} \quad \Omega \subset \mathbb{R}^3, \;\; u|_{\delta\Omega} = g. \tag{20.2}$$

k und f können nichtlinear sein.

Das $\boldsymbol{f}_l$ aus (20.1) muß nicht aus dem f aus (20.2) stammen. Man könnte (auf dem feinsten Gitter) das $\boldsymbol{f}_l$ auch auf die linke Seite bringen und ein System der Art

$$\widetilde{\boldsymbol{A}}_{l_{max}}(\boldsymbol{y}_{l_{max}}) := \boldsymbol{A}_{l_{max}}(\boldsymbol{y}_{l_{max}}) - \boldsymbol{f}_{l_{max}} = \boldsymbol{0} \tag{20.3}$$

lösen.
Naheliegende Idee zur Lösung von (20.1) ist die Anwendung des Newton-Verfahrens. Statt (20.1) löst man eine Folge von linearisierten Problemen (Newton-Iterationsvorschrift) mit Hilfe des linearen MGV. Mehrere Anwendungsvarianten dieser Art wurden schon untersucht (vgl. Hackbusch, Kap. 9)
Man kann jedoch die **Verwendung von Ableitungen** vermeiden durch Anwenden eines nichtlinearen MGV, von dem wir nun eine Variante, die von A. Brand vorgeschlagen wurde, beschreiben.
Für einen Konvergenzbeweis verweisen wir auf die Literatur (vgl. Hackbusch: Multigrid Methods and Applications, und die dort zitierte Literatur).

Zu lösen ist also (20.1), wobei $\boldsymbol{f}_l = 0$ erlaubt ist.

Glättungsiteration

Üblich ist ein nichtlinearer Gauß-Seidel
Grobbeschreibung:

$$\text{Löse } \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}, \quad \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{f} = (f_1, ..., f_n)^T.$$

Dann lautet der i-te Teilschritt des nichtlinearen GS-Verfahrens innerhalb des k-ten Gesamtschritts

$$0 = f_i(x_1^{(i)}, ..., x_i^{(i)}, x_{i+1}^{(i-1)}, ..., x_n^{(i-1)}), \;\; i = 1, .., n,$$

d.h. im i-ten Schritt wird aus der i-ten Komponente f_i von $\boldsymbol{f}$ das $x_i^{(i)}$ berechnet (eine Gleichung mit einer Unbekannten).
Empfohlen wird: Statt der Löung der Gleichung führe man einen Iterationsschritt nach dem Newtonverfahren aus und zwar nach einem diskretisierten Newton-Verfahren (ohne Verwendung von Ableitungen.)

Beachte: Der nichtlineare GS braucht ohne Voraussetzungen (z.B. Zeilensummenkriterium und Irreduzibilität) nicht zu konvergieren.

Restriktionen und Prolongation: wie im linearen Fall

Grobgitterkorrektur

Auf Level l sei eine Ausgangsnäherung $\boldsymbol{y}_l^0$ gegeben. Gelöst werden soll auf Gitter l (vgl. die Nachiteration in (§ 7))

$$\boldsymbol{A}(\boldsymbol{y}_l^0 + \boldsymbol{v}_l) = \boldsymbol{f}_l \tag{20.4}$$

Gesucht wird also die Korrektur $\boldsymbol{v}_l$, die $\boldsymbol{y}_l^0$ zur exakten Lösung ergänzt.
Beachte: Auch wenn $\boldsymbol{f}_l = \boldsymbol{0}$ ist für $l = l_{\max}$, muß das, wie wir sehen werden, für $l < l_{\max}$ nicht gelten. So ist etwa nicht $\boldsymbol{f}_{l-1}$ die Restriktion von $\boldsymbol{f}_l$. Man erinnere sich, dass beim MGV nicht die rechten Seiten restringiert werden, sondern die Defekte. Wenn schon (20.4) nicht exakt gelöst wird, so soll doch eine Verbesserung $\boldsymbol{y}_l^1 = \boldsymbol{y}_l^0 - \boldsymbol{v}_l$ konstruiert werden, wobei $\boldsymbol{y}_l^1$ eine bessere Näherung für die Lösung von (20.4) sein soll als $\boldsymbol{y}_l^0$. Dabei will man die Verbesserung $\boldsymbol{v}_l$ als $\boldsymbol{P}\boldsymbol{v}_{l-1}$ aus einer Rechnung aus Level $l-1$ haben.

Auf $l = l_{max}$ seien gegeben seien gegeben $\boldsymbol{A}_l, \boldsymbol{y}_l^0$ und $\boldsymbol{f}_l$ (z.B. $\boldsymbol{f}_{l_{max}} = \boldsymbol{0}$). Die Aufgabe besteht darin, auf Level $l-1$ eine geeignete Gleichung zu finden, aus der $\boldsymbol{v}_{l-1}$ berechnet werden kann. Als Ansatz soll die zu konstruierende Gleichung folgende Form haben

$$\boldsymbol{A}_{l-1}(\boldsymbol{R}\boldsymbol{y}_l^0 + \boldsymbol{v}_{l-1}) = \boldsymbol{f}_{l-1}. \tag{20.5}$$

$\boldsymbol{A}_{l-1}$, $\boldsymbol{y}_l^0$, $\boldsymbol{R}$ sind bekannt, $\boldsymbol{f}_{l-1}$ muß geeignet bestimmt werden. Dann ist $\boldsymbol{v}_{l-1}$ berechenbar. Aus (20.5) erhält man durch Linearisierung (Taylor)

$$\boldsymbol{A}_{l-1}(\boldsymbol{R}\boldsymbol{y}_l^0 + \boldsymbol{v}_{l-1}) = \boldsymbol{A}_{l-1}(\boldsymbol{R}\boldsymbol{y}_l^0) + \underbrace{\boldsymbol{A}'_{l-1}(\boldsymbol{R}\boldsymbol{y}_l^0)}_{\text{Jakobimatrix}}\boldsymbol{v}_{l-1} + \ldots \tag{20.6}$$

Die rechte Seite dieser Gleichung könnte ein geeignetes $\boldsymbol{f}_{l-1}$ sein (unter Vernachlässigung von "+..."'), wäre da nicht die Ableitung.
Idee: Ersetze (20.4) durch die linearisierte Fassung

$$\boldsymbol{f}_l = \boldsymbol{A}_l(\boldsymbol{y}_l^0 + \boldsymbol{v}_l) \approx \boldsymbol{A}_l(\boldsymbol{y}_l^0) + \underbrace{\boldsymbol{A}'_l(\boldsymbol{y}_l^0)}_{\text{Jacobi-Matrix}} \boldsymbol{v}_l, \tag{20.7}$$

berechne hieraus $\boldsymbol{A}'_l(\boldsymbol{y}_l^0)\boldsymbol{v}_l \approx \boldsymbol{f}_l - \boldsymbol{A}_l(\boldsymbol{y}_l^0)$ und verwende dessen Restriktion

$$\boldsymbol{R}(\boldsymbol{A}'_l(\boldsymbol{y}_l^0)\boldsymbol{v}_l) \approx \boldsymbol{R}(\boldsymbol{f}_l - \boldsymbol{A}_l(\boldsymbol{y}_l^0))$$

als Approximation für $\boldsymbol{A}'_{l-1}(\boldsymbol{R}\boldsymbol{y}_l^0)\boldsymbol{v}_{l-1}$. Damit folgt aus (20.6) die approximierte Fassung

$$\boldsymbol{A}_{l-1}(\underbrace{\boldsymbol{R}\boldsymbol{y}_l^0 + \boldsymbol{v}_{l-1}}_{\boldsymbol{w}_{l-1}}) = \boldsymbol{A}_{l-1}(\boldsymbol{R}\boldsymbol{y}_l^0) + \boldsymbol{R}(\boldsymbol{f}_l - \boldsymbol{A}_l(\boldsymbol{y}_l^0) := \boldsymbol{f}_{l-1}. \tag{20.8}$$

Die rechte Seite betrachten wir als Definitionsgleichung für $\boldsymbol{f}_{l-1}$.
Beachte: Auch wenn $\boldsymbol{f}_l = \mathbf{0}$ wäre, müßte nicht $\boldsymbol{f}_{l-1} = \mathbf{0}$ gelten.
Wir lösen nun (näherungsweise) das System (20.8), (ein nichtlineares System mit viel weniger Variablen als in (20.4)), erhalten $\boldsymbol{w}_{l-1} = \boldsymbol{R}\boldsymbol{y}_l^0 + \boldsymbol{v}_{l-1}$, setzen

$$\boldsymbol{v}_{l-1} = \boldsymbol{w}_{l-1} - \boldsymbol{R}\boldsymbol{y}_l^0 \tag{20.9}$$

und definieren

$$\boldsymbol{y}_l^1 = \boldsymbol{y}_l^0 + \boldsymbol{P}\boldsymbol{v}_{l-1}. \tag{20.10}$$

Bemerkungen: (20.8), (20.9), (20.10) sind Schritte zu einer verbesserten Näherung $\boldsymbol{y}_l^0$.
Ein Konvergenzbeweis stammt von Hackbusch.
Statt (20.4) wird (20.8) (näherungsweise) gelöst, ein System mit vielweniger Variablen. (Bis hierher wäre das ein ZGV). Auf die beschriebene Weise steigt man ab bis zum gröbsten Gitter, auf dem dann wirklich die Gleichung $\boldsymbol{A}_0(\boldsymbol{R}\boldsymbol{y}_1^0 + \boldsymbol{v}_0) = \boldsymbol{f}_0$ (mit entsprechend konstruiertem $\boldsymbol{f}_0$) gelöst werden muß.
Dies ist a) die entscheidende Klippe, liefert
b) aber auch Chancen.

zu a): Die Lösung auf dem gröbsten Gitter muß gut sein, sonst erhält man möglicherweise Divergenz.

zu b): Das System (20.4) kann mehrere Lösungen haben. Es kann leichter sein, auf dem gröbsten Gitter die richtige Lösung rauszusuchen und dann hochzutransportieren (Beispiel bei Bifurkation), als die Auswahl erst auf dem feinen Gitter zu treffen.

Die Eindeutigkeit der Lösung kann bei konkavem $\boldsymbol{f}$ (oft) gezeigt werden. Bei nichtlineare Aufgaben kann man oft mittels Monotonie Ober- und Unterfunktionen konstruieren und dazwischen die Eindeutigkeit beweisen.

Wesentliche Punkte sind also:

1. Auf $l = 1$ wird ein gutes Verfahren für wenige Punkte benötigt. Es muß nicht die exakte Lösung sein, sollte den Defekt aber deutlich mindern.

2. Im Unterschied zum linearen Verfahren wird auch die Näherungslösung restringiert. (vgl. (20.6),(20.8))

3. Die γ Iterationen werden, wie bekannt, durchgeführt.

4. Dieses so beschriebene Verfahren ist eine direkte Verallgemeinerung des linearen Verfahrens. Ist (20.4) eine lineare Gleichung, so fällt es mit dem bekannten linearen MGV zusammen. Dies erkennt man wie folgt:
Betrachte die Korrekturgleichung (20.8).
Im linearen Fall ist $\boldsymbol{A}_{l-1}(\boldsymbol{R}\boldsymbol{y}_l^0+\boldsymbol{v}_{l-1})$ ein Produkt aus der Matrix $\boldsymbol{A}_{l-1}$ und dem Vektor $\boldsymbol{R}\boldsymbol{y}_l^0 + \boldsymbol{v}_{l-1}$. Entsprechendes gilt für $\boldsymbol{R}(\boldsymbol{f}_l - \boldsymbol{A}_l(\boldsymbol{y}_l^0))$. Man kann (20.8)

mit $\boldsymbol{A}_{l-1}^{-1}$ multiplizieren und erhält

$$\boldsymbol{R}\boldsymbol{y}_l^0 + \boldsymbol{v}_{l-1} = \boldsymbol{R}\boldsymbol{y}_l^0 + \boldsymbol{A}_{l-1}^{-1}\underbrace{\boldsymbol{R}(\underbrace{\boldsymbol{f}_l - \boldsymbol{A}_l(\boldsymbol{y}_l^0)}_{\boldsymbol{d}_l})}_{\boldsymbol{d}_{l-1}})$$

also $$\boldsymbol{v}_{l-1} = \boldsymbol{A}_{l-1}^{-1}\boldsymbol{d}_{l-1}\,.$$

Die ist die Korrekturgleichung, die aus dem linearen Fall bekannt ist. $\boldsymbol{R}\boldsymbol{y}_l^0$ fällt weg.
Beachte jedoch: Im nichtlinearen Fall kann man (in dieser Version) nicht aufsteigen. *Man kommt von oben.* Eine Näherungslösung wird restringiert.

Nachbenmerkungen: Neuere Entwicklungen zum MGV gibt es bei GMD: Chemnitz; Inria, Kiel bzw Bremen (Hackbusch); USA, Conrad Zuse Institut (Deuflhard).

Von Deuflhard stammt eine Anwendung auf Differentialgleichungen, wo nicht immer zum gröberen Gitter heruntergestiegen wird, sondern vom gröbsten aufs feinste Giter aufgestiegen wird. Wesentliche Punkte sind dazu

1. Konjugierte Gradienten-Verfahren
2. Genau dosierte Iterationszahlen auf dem jeweiligen Gitter für das Konjugierte Gradienten-Verfahren
3. $l \to l+1$ durch geeignete Interpolation

Shaidurov hat, (Anfang der 90 ger Jahre einen Beweis dazu gemacht.)

Literaturempfehlung: Shaidurov/Marshuk: Difference Methods and their Extrapolation, Springer 1993.

MGV zur Lösung von Intergralgleichungen

Ein Beispiel: Löse
$$y(x) = \int_\Omega K(x,\xi)y(\xi)\,d\xi = f.$$

Diskretisiere das Integral nach ξ. Dies liefert

$$y(x) = \sum_{i=1}^{n} a_i(x,\xi_i)y(\xi_i)h, \qquad h = \xi_{i+1} - \xi_i,$$

mit von x abhängigen Gewichten a_i.
Löse diese Gleichung an den Punkten $x = \xi_j, \;\; j = 1, ..., n.$

Die ergibt ein Gleichungssystem

$$
\begin{gathered}
(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{y} = \boldsymbol{f} \\
\text{mit } \boldsymbol{y} = (y_1(\xi), ..., y_n(\xi))^T \\
\boldsymbol{A} = (a_{i,j}), \;\; a_{i,j} = a(\xi_i, \, \xi_j).
\end{gathered}
$$

Dies kann mit einem MGV behandelt werden.
Für Integralgleichungen wurde schon vor Aufkommen des MGV gezeigt, daß es solche Verfahren gibt.

Kapitel IV

Hyperbolische Differentialgleichungen

§ 21 Die Wellengleichung

Wir beginnen unsere Untersuchungen mit einem (bekannten) Satz aus der Theorie der Partiellen Differentialgleichungen.

Satz 21.1 (d'Alembert, eigentlich Euler)
Die Anfangswertaufgabe

$$\begin{array}{lcll} u_{tt} & = & a^2 u_{xx} + f, & a \in \mathbb{R}, \quad a \neq 0, \quad u \in C^2(\mathbb{R}^2), f \in \mathbb{C}^1(\mathbb{R}^2) \\ u(x,0) & = & u_0(x), & u_0 \in C^2(\mathbb{R}), \\ u_t(x,0) & = & u_1(x), & u_1 \in C^1(\mathbb{R}). \end{array}$$

ist sachgemäß.
Ihre Lösung wird gegeben durch die

D'Alembert'sche Lösungsformel

$$u(x,t) = \frac{1}{2}(u_0(x+at) + u_0(x-at)) + \frac{1}{2a}\int\limits_{x-at}^{x+at} u_1(\xi)\, d\xi + \frac{1}{2a}\int\limits_0^t \int\limits_{x-a(t-\tau)}^{x+a(t-\tau)} f(\xi,\tau)\, d\xi d\tau$$

Abhängigkeitsverhältnisse nach der Lösungsformel (für $t \geq 0$)

Zur Bestimmung von $u(x_0, t_0)$ werden abgefragt
Anfangspositionen in $x_0 \pm at_0$,
Anfangsgeschwindigkeit im Intervall $[x_0 - at_0, x_0 + at_0]$
f im gesamten Dreieck.

Für die Randwertaufgabe gilt

Satz 21.2
Die 1. Randwertaufgabe

$$\begin{array}{lcll}
u_{tt} & = & a^2u_{xx}, & a \in \mathbb{R}, \quad a \neq 0, \quad u \in C^2([0,\ell] \times [0,\infty)), \\
u(x,0) & = & u_0(x), & u_0 \in C^2([0,\ell]), \\
u_t(x,0) & = & u_1(x), & u_1 \in C^1([0,\ell]), \\
u(0,t) & = & g_1(t), & \\
u(\ell,t) & = & g_2(t), & g_i \in C^2([0,\infty)),
\end{array}$$

besitzt genau dann eine eindeutige Lösung $\in C^2([0,\ell] \times [0,\infty))$, wenn in jedem Randpunkt α_i, $(\alpha_1 = 0,\ \alpha_2 = \ell)$, folgende Verträglichkeitsbedingungen erfüllt sind

$$(21.1) \qquad \left\{ \begin{array}{ll}
g_i(0) = u_0(\alpha_i) & \text{(Stetigkeit der Funktion u)} \\
g_i'(0) = u_1(\alpha_i) & \text{(Stetigkeit von } u_t\text{)} \\
g_i''(0) = a^2u_0''(\alpha_i) & \text{(Erfülltheit der Differentialgleichung).}
\end{array} \right.$$

Numerische Differenzenapproximation
Wir benutzen äquidistante Gitter in Raum und Zeit. Mit den inneren Gitterpunkten

$$\omega_h := \{x_i = +ih;\ i = 1, ..., N-1,\ h = \frac{1}{n}\}$$

$$\omega_\tau := \{t_j = j\tau;\ j = 1, ..., m,\ m = \frac{T}{\tau}\}$$

approximieren wir (vgl. (2.4),(2.7))

mit $\quad u(x_i,t_j) \approx y_i^j, \quad \boldsymbol{y}^{j+1} \hat{=} \hat{\boldsymbol{y}}, \quad \boldsymbol{y}^j \hat{=} \boldsymbol{y}, \quad \boldsymbol{y}^{j-1} \hat{=} \bar{\boldsymbol{y}}$

$$\frac{\partial^2 u}{\partial t^2} \approx \frac{\hat{\boldsymbol{y}} - 2\boldsymbol{y} + \bar{\boldsymbol{y}}}{h^2} =: \boldsymbol{y}_{\bar{t}t} \quad \text{für beliebige Zeitschichten, d.h.}$$

(21.2)
$$\frac{\partial^2 u(x_i,t_j)}{\partial t^2} \approx \frac{y_i^{j+1} - 2y_i^j + y_i^{j-1}}{\tau^2} =: y_{\bar{t}t,i}^j = \frac{\partial^2 u}{\partial t^2} + \frac{\tau^2}{12}\frac{\partial^4 u}{\partial t^4}(t_j + \theta_j\tau) \quad \text{falls } u \in \mathbb{C}^{0,4}$$

(21.3)
$$\frac{\partial^2 u(x_i,t_j)}{\partial x^2} \approx \frac{y_{i+1}^j - 2y_i^j + y_{i-1}^i}{h^2} =: y_{\bar{x}x,i}^j = \frac{\partial^2 u}{\partial x^2} + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(x_i + \eta_i h) \quad \text{falls } u \in \mathbb{C}^{4,0}$$

Da in der Differentialgleichung eine 2. Zeitableitung vorliegt, brauchen wir 3 Zeitschichten zur Diskretisierung:

$$\boldsymbol{y}^\sigma := \sigma\hat{\boldsymbol{y}} + (1-2\sigma)\boldsymbol{y} + \sigma\bar{\boldsymbol{y}} = (\boldsymbol{y} + \sigma\tau^2\boldsymbol{y}_{\bar{t}t}) = \boldsymbol{y} + \boldsymbol{y}_{\bar{t}t} + O(\tau^2).$$

Die Diskretisierung der Differentialgleichung liefert damit

$$\boldsymbol{y}_{\bar{t}t} = a^2(\boldsymbol{y}^{(\sigma)})_{\bar{x}x} + \boldsymbol{\varphi} \quad \text{in } \omega_h \times \omega_\tau \quad \text{und z.B. } \boldsymbol{\varphi} = \boldsymbol{f}.$$

Es kann jedoch opportun sein als $\boldsymbol{\varphi}$ das $\boldsymbol{f}$ auf einer Zwischenschicht zu wählen.
Diskretisierung der Anfangswerte

1. Funktionswerte: $y_i^0 = u_0(x_i)$ ist trivial.

2. Ableitungswerte: Die Taylorentwicklung von $u(x, t_1)$ liefert

$$u(x,t_1) = u(x,0) + \tau \dot{u}(x,0) + \frac{\tau^2}{2}\ddot{u}(x,0) + O(\tau^3).$$

Unter Benutzung der Differentialgleichung und der Anfangswerte

$$\begin{aligned} u(x,t_1) &= u_0(0) + \tau u_1(0) + \frac{\tau^2}{2}\big(a^2 u_{xx}(x,0) + f(x,0)\big) + O(\tau^3) \\ &= u_0(0) + \tau u_1(0) + \frac{\tau^2}{2}\big(a^2 y^0_{\bar{x}x} + f(x,0) + O(h^2)\big) + O(\tau^3) \\ &= u_0(0) + \tau u_1(0) + \frac{\tau^2}{2}\big(a^2 y^0_{\bar{x}x} + f(x,0)\big) + \frac{\tau^2}{2}O(h^2) + O(\tau^3) \end{aligned}$$

erhält man mit $\boldsymbol{y}^1 \approx u(x,t_1)$ als 2. Anfangsbedingung

$$\begin{aligned} \frac{\boldsymbol{y}^1 - \boldsymbol{y}^0}{\tau} &= \boldsymbol{u}_1 + \frac{\tau}{2}(a^2(\boldsymbol{u}_0)_{\bar{x}x} + \boldsymbol{f}^0) + \frac{\tau}{2}O(h^2),\ (\frac{\tau}{2}O(h^2) = O(\tau^2 + h^4)) \quad \text{bzw.} \\ \boldsymbol{y}^1 &= \boldsymbol{y}^0 + \tau \boldsymbol{u}_1 + \frac{\tau^2}{2}(a^2(\boldsymbol{u}_0)_{\bar{x}x} + \boldsymbol{f}^0). \end{aligned}$$

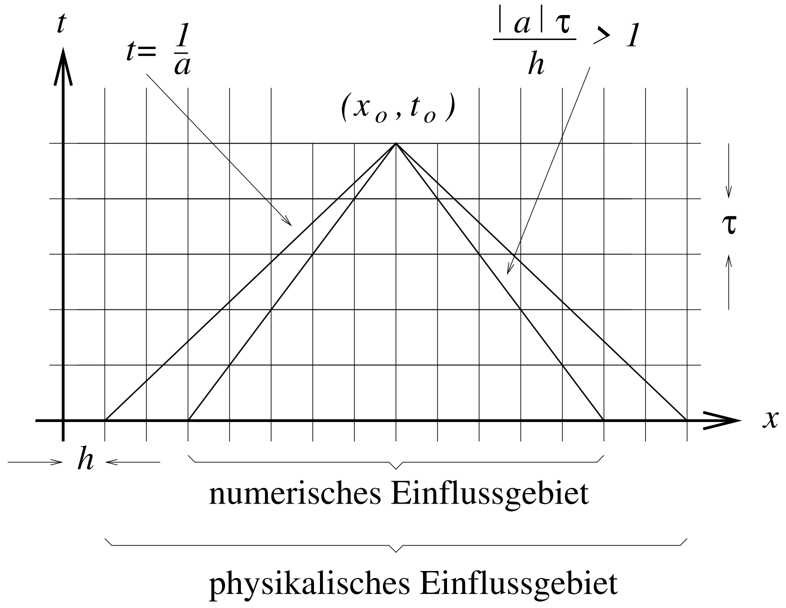
Diskretisierung der Randwerte: $y_0^{j+1} = g_0(t_{j+1}), \quad y_N^{j+1} = g_1(t_{j+1})$.

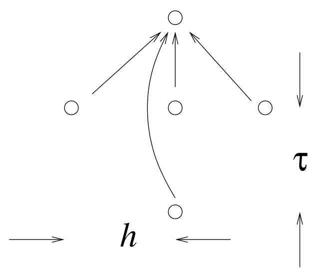
Das komplette Diferenzenschema lautet damit (nach Zeitschichten geordnet)

$$\begin{aligned} \frac{1}{\tau^2}y^{j+1} - a^2\sigma y^{j+1}_{\bar{x}x} &= a^2(1-2\sigma)y^j_{\bar{x}x} + a^2\sigma y^{j-1}_{\bar{x}x} + \frac{2}{\tau^2}y^j - \frac{1}{\tau^2}y^{j-1} + f_i^j \\ & \qquad\qquad i = 1,..,N-1,\ j = 1,...,m-1, \\ y_0^{j+1} = g_0(t_{j+1}), \quad y_N^{j+1} &= g_1(t_{j+1}) \qquad j = 0,...,m-1, \\ y_i^0 &= u_0(x_i), \qquad i = 0,...,N, \\ y_i^1 &= y_i^0 + \tau u_1(x_i) + \frac{\tau^2}{2}\big(a^2 y^0_{\bar{x}x} + f_i^0\big), \qquad i = 1,...,N-1. \end{aligned}$$

Folge: $\sigma = 0$ explizites Verfahren
$\sigma > 0$ implizites Verfahren

Wir untersuchen zunächst den expliziten Fall und vergleichen im Punkt (x_0, t_0) den numerischen und den physikalischen Einflußbereich.





Verfolge den Differenzenstern auf dem Gitter.
Gemäß dem zur Diskretisierung verwendeten Differenzenstern beschreiben Geraden durch die Gitterpunkte mit der Gitter-Steigung $\frac{\tau}{h}$ die Grenzen des numerischen Einflußbereichs. Ideal ist, wenn die Gittersteigung mit der Steigung der Charakteristiken (der physikalischen Steigung) $\frac{1}{a}$ übereinstimmt. (Beachte dazu: $t = \pm\frac{1}{a}x + k$), also

$$\frac{\tau}{h} = \frac{1}{|a|} \iff \gamma := \frac{|a|\tau}{h} = 1.$$

Ist die Gittersteigung größer als die physikalische Steigung ($\frac{\tau}{h} > \frac{1}{a} \iff \gamma := \frac{|a|\tau}{h} > 1$), so ist der numerische Einflußbereich kleiner als der physikalische Einflußbereich.

Wir bezeichnen mit	$\frac{h}{\tau}$	die Gittergeschwindigkeit
	$\|a\|$	die Wellengeschwindigkeit
	$\gamma := \frac{\|a\|\tau}{h}$	die **C**ourant-**L**evy-**F**riedrich-Zahl,CLF-Zahl.

Für die CLF-Zahl haben wir also folgende 3 Fälle:

$$\gamma = \frac{|a|\tau}{h} \quad \begin{cases} = 1 & \text{idealer Fall: numerisches = physikalisches Einflußgebiet} \\ > 1 & \text{physikalisches Einflußgebiet > numerisches Einflußgebiet} \\ < 1 & \text{physikalisches Einflußgebiet < numerisches Einflußgebiet} \end{cases}$$

Man erkennt sofort (vgl. Zeichnung, heuristische Überlegung):
Ist $\gamma > 1$, so ändert sich die physikalische (exakte) Lösung, wenn die Anfangswerte außerhalb des numerischen, aber innerhalb des physikalischen Einflußgebiets geändert werden. Die numerische Lösung bleibt jedoch unverändert.
Dieser Fall muß ausgeschlossen werden.

$\dfrac{|a|\tau}{h} \leq 1$ ist ein notwendiges, kein hinreichendes Stabilitätskriterium (vgl. Beispiel 4) (CLF-Kriterium, Courant-Levy-Friedrich).

Diese Bezeichnung muß natürlich mathematisch gerechtfertigt werden.
Wir zeigen am einfachen Beispiel der homogenen Wellengleichung, daß die Verletzung dieses Kriteriums zu Instabilitäten führt.

Die Differentialgleichung

$$u_{tt} = a^2 u_{xx}$$

hat die Diskretisierung

$$\frac{y_k^{j+1} - 2y_k^j + y_k^{j-1}}{h^2} = a^2 \frac{y_{k+1}^j - 2y_k^j + y_{k-1}^j}{h^2}.$$

Durch den (für Differenzengleichungen üblichen) Ansatz

$$y_k^j = \mu^k \lambda^j$$

bestimmen wir eine Lösung der Differenzengleichung an der man ablesen können wird, daß keine Stabilität vorliegt.
Bemerkung: Eigentlich müßten die Zeitindizes j bei den y_k^j-Werten in Klammern eingeschlossen werden, besonders bei skalaren Werten, um Verwechslungen mit Potenzen zu vermeiden. Man hat sich jedoch daran gewöhnt (vermutlich aus Bequemlichkeitsgründen) diese Klammern wegzulassen. Bei y_k^j-Werten bedeutet deshalb das j im Exponenten grundsätzlich nur die Zeitschicht, bzw. bei Folgen (in Zeitrichtung, vgl. Satz 22.1) die Laufvariable der Folge.
Setzt man den obigen Ansatz in die Differenzengleichung ein, so erhält man

$$\frac{\mu^k \lambda^j (\lambda - 1)^2}{\tau^2 \lambda} = a^2 \frac{\mu^k \lambda^j (\mu - 1)^2}{h^2 \mu}$$

woraus mit $\gamma = \dfrac{|a|\tau}{h}$ folgt

$$\frac{(\lambda - 1)^2}{\lambda} = \gamma^2 \frac{(\mu - 1)^2}{\mu}.$$

Setzen wir in dieser Gleichung $\mu = -1$ und ersetzen λ durch $-\lambda$, so erhält man

$$(\lambda + 1)^2 = 4\gamma^2 \lambda \quad \text{bzw.}$$
$$\lambda^2 - (4\gamma^2 - 2)\lambda + 1 = 0$$

mit den Wurzeln

$$\lambda_{1,2} = 2\gamma^2 - 1 \pm 2\sqrt{\gamma^2(\gamma^2 - 1)}.$$

Wir untersuchen die bekannten 3 Fälle für γ:

$\gamma < 1:\quad \lambda_{1,2}$ komplex , $|\lambda_{1,2}| = 1$, Betrag ausrechnen

$\gamma = 1:\quad \lambda_1 = \lambda_2 = 1$,

$\gamma > 1:\quad \lambda_1 = 2\gamma^2 - 1 + 2\sqrt{\gamma^2(\gamma^2 - 1)} > 1.$

Da wir eine Differenzengleichung 2. Ordnung (in der Zeit) haben, muß man für eine eindeutige Lösung Anfangswerte auf 2 Zeitschichten vorgeben. Wir geben beschränkte Anfangswerte vor

$$|y_k^0| = 1, \qquad |y_k^1| = \lambda_1.$$

Sie werden durch die Lösung $y_k^j = \mu^k \lambda^j$ angenommen. (Wir erinnern uns an $\lambda_1 > 1$ und $\mu = -1$).
In einem Punkt $(x_k, \tilde{t})$ hat die exakte Lösung den Wert $u(x_k, \tilde{t})$. Für ein gegebenes Gitter sei $\tilde{t} = t_j$. Eine Lösung der Differenzengleichung wird durch $(-1)^k \lambda_1^j$ gegeben. Verfeinert man das Gitter, indem man γ einfriert, d.h. das Verhältnis $\frac{h}{\tau}$ bleibt konstant, so ändert sich das numerische Abhängigkeitsgebiet nicht.
Wird die Gitterweite z.B. halbiert, so wird $\tilde{t} = t^{2j}$ und die numerische Lösung für $(x_k, \tilde{t})$ gegeben durch $(-1)^k \lambda^{2j}$, d.h. bei weitergehender Verfeinerung gilt wegen $\lambda_1 > 1$ für ein beliebiges, festes, z.B. geradzahliges k

1. für $k:\quad y_k^j \xrightarrow{j \to \infty} +\infty$
2. für $k+1:\quad y_{k+1}^j \xrightarrow{j \to \infty} -\infty,$

d.h. die Lösung geht mit k oszillierend gegen $\pm\infty$, ist also sicher nicht stabil im Sinne der alten Stabilitätsdefinition (vgl. parabolische Gleichungen).
Wir erhalten also :

$\gamma > 1$ ist eine hinreichende Instabilitätsbedingung.

Man beachte, daß diese Stabilitätsaussage zumindest nicht unmittelbar mit dem vorliegenden Anfangswertproblem verknüpft ist.

§ 22 Die Neumann'sche Stabilitätsanalyse

Diese Analyse untersucht die Stabilität bzl. der Anfangswerte. Wir beschreiben hier die Stabilitätsanalyse für lineare Differentialgleichungen mit konstanten Koeffizienten. Für weitere Untersuchungen, insbesondere für mehrdimensionale Differentialgleichungen und Systeme von Gleichungen, verweisen wir etwa auf das Buch von Godunov/Ryabenki: Difference Schemes, North Holland.
Diese Stabilitätsanalyse ist nicht auf hyberbolische Gleichungen beschränkt. Eine erste erste Anwendung haben wir schon zu Beginn der Vorlesung (§ 4) bei der Untersuchung der Wärmeleitungsgleichung kennengelernt. Ausgangspunkt unserer Überlegungen ist das vorige Beispiel ebenso wie das Beispiel aus § 4.
Wird eine Differentialgleichung diskretisiert, so erhält man bei festgehaltenem Ort (k bzw. x_k fest) eine Differenzengleichungen (lineare) in y^j der Ordnung N (=Anzahl der Zeitschichten - 1)

Über die Lösung solcher, homogener Differenzengleichungen gilt folgender Satz (vgl. z.B. Henrici: Elemente der Numerischen Analysis I, BI 1964, Satz 6.8)

Satz 22.1
Das charakteristische Polynom einer linearen Differenzengleichung N-ter Ordnung besitze die verschiedenen Nullstellen $\lambda_1, ..., \lambda_k,\ k \leq N,$ mit der Vielfachheit $m_i + 1,\ (\sum m_i = N - k)$ von λ_k . Dann bilden die Folgen

$$y^{(n)} = \left(\prod_{\mu=0}^{m-1} (n - \mu) \right) \lambda_i^{n-m},\ m = 0, ..., m_i,\ i = 1, .., k \tag{22.1}$$

ein System von N linear unabhängigen Lösungen.

$$y^{(n)} = n^m \lambda_i^n,\ m = 0, ..., m_i,\ i = 1, .., k \tag{22.2}$$

sind ebenfalls ein System von N linear unabhängigen Lösungen.

Man kann also in jeder Ortsschicht alle partiellen Lösungen der Differenzengleichung (alle!) in der genannten Form darstellen mit einem konstanten Faktor als Koeffizienten (- im vorigen Beispiel μ^k -) zur Beschreibung der Ortsabhängigkeit), der bei einer homogenen Differenzengleichung keine Rolle spielt. Wir bekommen auf jeder Ortschicht dann die Zeitabhängigkeit in Potenzform von λ (bei einfachen Nullstellen, bei mehrfachen Nullstellen mit einer natürlichen Zahl als Koeffizienten).

Bei Stabilitätsuntersuchungen ist man daran interessiert, Aussagen zu gewinnen, wann die Lösungen der Differenzengleichungen mit beschränkten Anfangswerten in Zeitrichtung bei Gitterverfeinerung beschränkt bleiben.

Falls ein $|\lambda_i| > 1$ ausfällt, ist dies offensichtlich nicht der Fall. Dann kann das vorgelegte Differenzenschema ausgeschieden werden. Bei mehrfachen Nullstellen, z.B. Lösungen

der Art $n\lambda^n$, muß in Abhängigkeit von der Größe von λ untersucht werden, ob der Ausdruck beschränkt bleibt. Man erhält auf diese Weise Bedingungen für Instabilität. Bedingungen für Stabilität eines Verfahrens kann man nur erhalten, wenn man **alle** Lösungen des Differenzenschemas (nicht nur bei festgehaltenem Index k) in dieser "potenzförmigen" Zeitabhängigkeit beschreiben kann, und wenn alle Lösungen beschränkt bleiben.
Dies ist insbesondere der Fall, wenn alle Lösungen λ_i des charakteristischen Polynoms $|\lambda_i| < 1$ erfüllen und einfach sind.
Der Fall mehrfacher Nullstellen muß gesondert untersucht werden. Wir könen also festlegen:

Definition 22.2
Ein Mehrschichtschema (j – Index für die Zeitschichten, k – Index für die Punkte in x-Richtung) heißt instabil, falls unter den Lösungen

$$y_k^j := \mu^k \lambda^j$$

solche sind, für die $|\lambda| > 1$ und $|\mu| = 1$ (oder umgekehrt $|\lambda| = 1$ und $|\mu| > 1$) gilt.

Es muß also untersucht werden, wann alle Lösungen des Differenzenschemas in der oben genannten Form dargestellt werden können.
Dazu normieren wir das Intervall, das wir untersuchen auf $[0,1]$. Man kann sich bei hyperbolischen Differentialgleichungen darauf beschränken auf Grund der Abhängigkeitsgebiete.
Wir benutzen ein Ortsgitter

$$x_k = k \cdot h, \; k = 0, 1, ..., n+1, \quad \text{also } (n+1)h = 1.$$

Die Feinheit des Gitters wird über $h = \frac{1}{n+1}$ bestimmt.
Bei Trennungsansätzen der Art $y_k^j := \mu^k \lambda^j$ muß $|\mu| = 1$ sein, weil sonst bei Gitterverfeinerung ($n \to \infty$) die Anfangswerte auf der Zeitschicht $j = 0$ gegen ∞ wachsen oder gegen Null fallen und somit eine Untersuchung des Verhaltens der Lösungen illusorisch wird. Sinnvoll sind für μ also Ausdrüke der Art $e^{i2\pi lx}$. (Dabei kann l als "äußerer" Index dienen), die wir als Gitterfunktion (in x) auffassen können, also $x = x_k$ einsetzen.
Wir betrachten die Vektoren

$$\boldsymbol{v}^{(\ell)} := (1, e^{2\pi i\ell h}, e^{2\pi i\ell 2h}, \ldots, e^{2\pi i\ell nh})^T \text{ für } l = 0, \ldots, n, \quad h = \frac{1}{n+1}.$$

Man beachte, daß die Funktionen $e^{2\pi i\ell x}$ in x periodisch sind ($x = kh$):

$$e^{2\pi i\ell\cdot 0} = 1 = e^{2\pi i\ell\cdot(n+1)h} = e^{2\pi i\ell\cdot 1} = \cos(2\ell\pi) + i\sin(2\pi\ell) = 1.$$

Dies überträgt sich natürlich auf die $\boldsymbol{v}^{(\ell)}$, d.h. für $\boldsymbol{v}^{(\ell)} = \boldsymbol{v}^{(\ell)}(h)$ gilt somit $\boldsymbol{v}^{(\ell)}(0) = \boldsymbol{v}^{(\ell)}((n+1)h)$.
Wir zeigen zunächst, daß die $\boldsymbol{v}^{(\ell)}$, $\ell = 0, \ldots, n$ ein Orthonormalssystem bzgl. des diskreten $L_2(\omega_h)$-Skalarprodukts

$$(\boldsymbol{u}, \boldsymbol{w})_{(0,h)} := \sum_{k=0}^{N} u_k \overline{w}_k h \tag{22.3}$$

bilden ($\overline{w}_k$ ist der konjugiert komplexe Wert), denn

$$(\boldsymbol{v}^{(\ell)}, \boldsymbol{v}^{(m)})_{(0,h)} = \sum_{k=0}^{n} e^{2\pi i(\ell-m)kh} h =: \sum_{k=0}^{n} w^k h \quad \text{mit} \quad w = e^{2\pi i(\ell-m)h} h.$$

Hieraus folgt

$$(\boldsymbol{v}^{(\ell)}, \boldsymbol{v}^{(m)})_{(0,h)} = \begin{cases} 1,\ \ell = m, & \text{weil } \sum\limits_{k=0}^{n} h = (n+1)h = 1, \\ \frac{w^{n+1}-1}{w-1} = 0,\ \ell \neq m, & \text{endliche geometrische Reihe und Periodizität.} \end{cases}$$

Daraus folgt: Ist y_k^0 periodisch in k mit der Periode $k = n+1$, so kann man jede Anfangswert(gitter)funktion $\boldsymbol{y}^0$ darstellen als

$$\boldsymbol{y}^0 = \sum_{l=0}^{n} c_l^{(0)} \boldsymbol{v}^{(\ell)}, \quad \text{bzw. komponentenweise} \quad y_k^0 = \sum_{l=0}^{n} c_l^{(0)} v_k^{(\ell)}, \tag{22.4}$$

und jede Gitterfunktion auf Level j durch

$$\boldsymbol{y}^j = \sum_{l=0}^{n} c_l^{(j)} \boldsymbol{v}^{(\ell)}, \quad \text{bzw. komponentenweise} \quad y_k^j = \sum_{k=l}^{n} c_l^{(j)} v_k^{(\ell)}. \tag{22.5}$$

Wir bezeichnen mit $\boldsymbol{y}_{(i)}^j$ eine partielle Lösung der Differenzengleichung. Dabei ist j der Laufindex der Folge und i die Anzeige, zu welcher Nullstelle der charakteristischen Gleichung die Folge gehört, i hat nichts mit der Ortsschicht zu tun.

Wir setzen nun voraus, daß alle Lösungen der charakteristischen Gleichung einfach sind.

Dann hat jede solche partielle Lösung $\boldsymbol{y}_{(i)}^j$ mit der Anfangswertfunktion $\boldsymbol{y}^0$ nach Satz 22.1 die Darstellung (wobei wir den Index i bei $\boldsymbol{y}_{(i)}^j$ zunächst unterdrücken)

$$\boldsymbol{y}^j = \sum_{l=0}^{n} c_l^{(0)} \boldsymbol{v}^{(\ell)} \lambda_i^j. \tag{22.6}$$

Der Vergleich von (22.5) und (22.6) liefert

$$c_\ell^{(j)} = c_\ell^{(0)} \lambda_i^j. \tag{22.7}$$

Insbesondere erhält man aus (22.7)

$$|c_l^{(j)}| \leq |c_l^{(j-1)}| \quad \Longleftrightarrow \quad |\lambda_i| \leq 1. \tag{22.8}$$

Da die $\boldsymbol{v}^{(\ell)}$ ein Orthonormalsystem bilden, gilt auf jeder Zeitschicht

$$\begin{aligned} (\boldsymbol{y}^j, \boldsymbol{y}^j)_{(0,h)} &= \left(\sum_{l=0}^{n} c_l^{(j)} \boldsymbol{v}^{(l)}, \sum_{m=0}^{n} c_m^{(j)} \boldsymbol{v}^{(m)} \right) = \sum_{m,l=0}^{n} c_l^{(j)} \overline{c_m^{(j)}} (\boldsymbol{v}^{(l)}, \boldsymbol{v}^{(m)}) \\ &= \sum_{l=0}^{n} |c_l^{(j)}|^2 = \|\boldsymbol{y}^j\|_{(0,h)}^2 \end{aligned}$$

Mit $|\lambda_i| \leq 1$ folgt aus (22.8) also

$$\|\boldsymbol{y}^{j+1}\|^2_{(0,h)} = \sum_{l=0}^{n} |c_l^{j+1}|^2 \leq \sum_{l=0}^{n} |c_l^j|^2 = \|\boldsymbol{y}^j\|^2_{(0,h)} \leq \dots \leq \|\boldsymbol{y}^0\|^2_{(0,h)}. \tag{22.9}$$

Mit diesem Stabilitätsbegriff arbeitet Neumann.
Die allgemeine Lösung der Differenzengleichung in jeder Ortsschicht x_k wird durch eine Linearkombination der Lösungen $\sum\limits_{i=1}^{n} a_{i,(k)}\lambda_i^j$ gegeben, die an Stelle von λ_i in (22.6)-(22.7) eingetragen werden muß. Auf Grund der Homogenität der Differenzengleichung spielt ein konstanter Faktor für die Lösung keine Rolle, weswegen jeweils einer der Koeffizienten $a_{i,(k)}$ zu 1 normiert werden kann. Bei einer Differentialgleichung n-ter Ordnung in der Zeitableitung, sowie der zugehörigen Differenzengleichung, wird die eindeutige Lösung durch $n-1$ Anfangswerte bestimmt, $(u, u_t, \dots)$ bei der Differentialgleichung, bzw. den Vorgaben auf $n-1$ Zeitschichten bei der Differenzengleichung. Die letzteren werden in jedem Punkt x_k durch die ersteren bestimmt und legen damit in jedem Punkt x_k die Koeffizienten $a_{i,(k)}$ fest.

Wir halten fest: Sind die Anfangswerte der Differentialgleichung periodisch unde alle λ einfach, so kann man auf die beschriebene Weise die allgemeine Lösung darstellen.

Der für die Stabilität schlimmste Fall tritt ein, wenn in (22.7) das betragsgrößte λ_i eingetragen wird. Sind alle $|\lambda_i| \leq 1$, so bleibt (22.9) unberührt. Andernfalls explodiert die Lösung. Wir haben also folgendes Ergebnis:

Satz 22.3

1. Sind die Anfangswerte einer Aufgabe periodisch und sind alle Lösungen λ_i der charakteristischen Gleichung einfach, so gilt:

   $\max\limits_i |\lambda_i| \leq 1 \quad \Longrightarrow$ Das Verfahren ist stabil (notwendige und hinreichende Bedingung), d.h. es gilt (vgl. (22.9))

   $$\|\boldsymbol{y}^{j+1}\|^2_{(0,h)} \leq \|\boldsymbol{y}^0\|^2_{(0,h)} \quad \forall\, j.$$

   $\exists\, i : \ |\lambda_i| > 1 \quad \Longrightarrow$ Das Verfahren ist instabil (d.h. für beschränkte Anfangswerte kann bei Gitterverfeinerung eine unbeschränkte Lösung existieren)

2. Sind die Anfangswerte nicht periodisch (man kann dann nicht alle Lösungen der Differenzengleichung darstellen), so erhält man nur hinreichende Instabilitätsbedingungen, falls ein $|\lambda_i| > 1$ auftritt. Dann existiert eine explodierende Lösung. Dazu müssen die Nullstellen der charakteristischen Gleichung nicht einfach sein.

Damit erhält man für den praktischen Verlauf der Stabiltätsuntersuchung nach Neumann

1. Man setzt den Ansatz $y_k^j = \mu^k \lambda^j = e^{i\varphi k}\lambda^j$ in die Differenzengleichung ein.
   Falls die Aufgabe periodische Anfangswerte besitzt, ist $\varphi = 2\pi\ell h$ zu setzen. Im nicht periodischen Fall ist φ frei (ggf. geeignet) wählbar: man untersucht dann ja "nur" , ob etwas schief geht.

2. Bestimme $\lambda = \lambda(\varphi)$ so, daß der Ansatz eine Lösung der Differenzengleichung liefert.
   Falls ein $|\lambda_i| > 0$ existiert $\Longrightarrow$ Instabilität
   Falls alle $|\lambda_i| \leq 1$ und die Differenzengleichung in Zeitrichtung nur einfache Wurzeln hat $\Longrightarrow$ Stabilität.
   Der Fall mehrfacher Wurzeln muß, ggf. in Abhängigkeit von φ, extra untersucht werden (vgl. Beispiel 4).

3. Falls in den Differenzengleichungen in Zeitrichtung ein τ auftritt, kann es auch zu Abschätzungen für hinreichend kleine τ kommen.

   $$|\lambda| \leq 1 + c_0\tau, c_0 = \text{konst.} \quad \text{Neumannsches Stabilitätskriterium.}$$

   Dann folgt
   $$|\lambda^j| \leq (1+c_0\tau)^j \leq e^{c_0\tau j} \leq e^{c_0 T} := M \text{ für } \tau j \leq T$$
   Beachte: $(1{+}c_0\tau)$ ist der Anfang der Taylorreihe für $e^{c_0\tau j}$ und damit (vgl. (22.8))

   $$|c_\ell^{(j)}| \leq |c_\ell^{(0)}|M, \quad \text{also} \quad \|\boldsymbol{y}^j\|_{(0,h)} \leq M\|\boldsymbol{y}^0\|_{(0,h)},$$

   d.h. man erhält Stabilität auf jeder endlichen Zeitschicht.

Schwierigkeiten bei der Neumann'schen Stabilitätsanalyse

1. mehrfache Wurzeln,
2. bei impliziten Verfahren können rationale Bedingungen für λ auftreten (Formelmanipulationssysteme!, z.B. Maple oder Mathematika),
3. oft erhält man hinreichende Bedingungen für Instabilität statt hinreichender Bedingungen für Stabilität.

Beispiele zur Stabilitätsanalyse

Beispiel 1: Wärmeleitungsgleichung aus § 4.

Beispiel 2: Wellengleichung in § 21.

Beispiel 3: Wir betrachten ein implizites Verfahren für die Wellengleichung

$$\boldsymbol{y}_{\bar{t}t} = a^2 \hat{\boldsymbol{y}}_{\bar{x}x}.$$

Einsetzen des dem Ansatzes $y_k^j = e^{i\varphi k}\lambda^j,\ \varphi = 2\pi\ell h$ in die Wellengleichung liefert

$$\begin{aligned}
(y_{\bar{x}x})_k^{j+1} &= \frac{\lambda^{j+1}}{h^2}(e^{i\varphi(k+1)} - 2e^{i\varphi k} + e^{i\varphi(k-1)}) \\
&= \lambda^{j+1} e^{i\varphi k}\left(\frac{e^{i\varphi} - 2 + e^{-i\varphi}}{h^2}\right) \\
&= \frac{\lambda y_k^j}{h^2}(-4)\left(\frac{e^{i\frac{\varphi}{2}} - e^{-i\frac{\varphi}{2}}}{2i}\right)^2 \\
&= \frac{-4y_k^j\lambda}{h^2}\sin^2(\frac{\varphi}{2}) \\
(y_{\bar{t}t})_k^j &= \frac{e^{i\varphi k}}{\tau^2}\,\frac{\lambda^j}{\lambda}(\lambda^2 - 2\lambda + 1) = y_k^j \frac{1}{\tau^2\lambda}(1-\lambda)^2.
\end{aligned}$$

Damit folgt für die Differenzengleichung

$$\frac{(1-\lambda)^2}{\tau^2\lambda} = -\frac{4}{h^2}a^2\lambda\sin^2(\frac{\varphi}{2})$$

Mit $s = \sin\frac{\varphi}{2}$ und $\gamma = \frac{\tau|a|}{h}$ erhält man

$$(1-\lambda)^2 = -\gamma^2\lambda^2\cdot 4s^2 \quad \text{bzw. } (1+4\gamma^2 s^2)\lambda^2 - 2\lambda + 1 = 0$$

$$\begin{aligned}
\lambda_{1,2} &= \frac{1}{2(1+4\gamma^2 s^2)}\left(2 \pm \sqrt{4 - 4(1+4\gamma^2 s^2)}\right) \\
&= \frac{1}{2(1+4\gamma^2 s^2)}\left(2 \pm \sqrt{-16\gamma^2 s^2}\right) \\
&= \frac{1 \pm \sqrt{-4\gamma^2 s^2}}{1+4\gamma^2 s^2} \quad \text{und daher (Betrag ausrechnen)} \\
|\lambda_{1,2}| &= \frac{1}{\sqrt{1+4\gamma^2 s^2}} < 1 \quad \text{unabhängig von } \gamma
\end{aligned}$$

also: Stabilität für periodische Anfangswerte.

Beispiel 4: Stabilität bei mehrfachen Nullstellen

Wir betrachten nochmals die Wellengleichung mit periodischen Anfangswerten

$$u_{tt} = a^2 u_{xx}$$

Bekannt ist: $\gamma > 1 \implies$ Instabilität, $\qquad \gamma \le 1$ notwendig für Stabilität.

Wir untersuchen den Fall $\gamma = 1$ für die Diskretisierung

$$\frac{y_k^{j+1} - 2y_k^j + y_k^{j-1}}{\tau^2} = a^2 \frac{y_{k+1}^j - 2y_k^j + y_{k-1}^j}{h^2}$$

nochmals genauer mit dem Ansatz

$$y_k^j = \lambda^j e^{i\varphi k}, \quad \varphi = 2\pi\ell h, \quad \ell = 1, \ldots, n, \quad h = \frac{1}{n+1}, \qquad \Longrightarrow$$

$$\begin{aligned} \frac{(\lambda-1)^2}{\tau^2\lambda} &= a^2 \frac{e^{i\varphi} - 2 + e^{-i\varphi}}{h^2} = \frac{a^2}{h^2}\left(e^{i\frac{\varphi}{2}} - e^{-i\frac{\varphi}{2}}\right)^2 \\ &= -4\sin^2(\frac{\varphi}{2}). \end{aligned}$$

Mit $\gamma = \dfrac{|a|\tau}{h}$, $\quad s = \sin(\dfrac{\varphi h}{2})$ folgt

$$\begin{aligned} \lambda^2 - 2\lambda + 1 &= -4\gamma^2 s^2 \lambda \\ \lambda_{1,2} &= 1 - 2\gamma^2 s^2 \pm 2\sqrt{\gamma^2 s^2(\gamma^2 s^2 - 1)}. \end{aligned}$$

Bei $\gamma = 1$ kann nur eine mehrfache Wurzel auftreten für $s^2 = 1$, also

$$\begin{aligned} \sin(\frac{2\pi\ell h}{2}) &= \sin(\pi\ell h) = \sin(\frac{\pi\ell}{n+1}) = \pm 1, \text{ d.h. für ganzzahliges } z \\ \frac{\pi\ell}{n+1} &= \frac{\pi}{2} + z\pi, \iff \frac{\ell}{n+1} = \frac{1}{2} + z \iff \ell = \frac{n+1}{2}(1+2z) \end{aligned}$$

Die letzte Gleichung hat nur für ungerades n eine ganzzahlige Lösung für l und n. Für geradzahliges n ist $s^2 \neq 1$, also $|\sin^2(\frac{\pi\ell}{n+1})| < 1$ und damit $\lambda_1 \neq \lambda_2$. Beide Nullstellen sind einfach und wegen $|\lambda_{1,2}| = 1$ (nachrechnen!) folgt somit:

Für n geradzahlig liegt Stabilität vor.

Für n nicht geradzahlig liefert die Lösungsformel $\lambda_1 = \lambda_2 = -1$. Nach Satz 22.1 erhält man der doppelten Nullstelle wegen die Lösungen $y_k^j = e^{i\varphi k}(-1)^j$ und $y_k^j = e^{i\varphi k} j(-1)^j$. Einsetzen in die Differenzengleichung bestätigt die Lösungseigenschaft. Damit ist $e^{i\varphi k} j(-1)^j$ eine Lösung mit polynomialen Wachstum.

Der Grenzfall $\gamma = 1$, n, gerade gehört nicht zum Stabilitätsbereich.

$\gamma = 1$ genügt also nicht für die Stabilität

Beispiel 5: Wir untersuchen eine Approximation 2. Ordnung in Orts- und Zeitrichtung für die Wärmeleitungsgleichung

$$\frac{y^{j+1} - y^{j-1}}{2\tau} = y_{\bar{x}x}$$

Mit dem Ansatz $y_k^j = e^{i\varphi k}\lambda^j$ folgt aus der Differentialgleichung

$$\begin{aligned}
\frac{\lambda - \frac{1}{\lambda}}{2\tau} y_k^j \cdot &= \frac{1}{h^2}(e^{i\varphi} - 2 + e^{-i\varphi}) y_k^j \\
\frac{\lambda - \frac{1}{\lambda}}{2\tau} &= \frac{2\cos\varphi - 2}{h^2} = -\frac{2\sin^2\frac{\varphi}{2}}{h^2} \cdot 2 = \frac{4}{h^2} \text{ mit } s = \sin^2\frac{\varphi}{2} \\
\lambda^2 - 1 &= -8\frac{\tau}{h^2}\lambda s^2
\end{aligned}$$

$\gamma := \frac{\tau}{h^2}$ ist eine Art Courant-Zahl für die Wärmeleitungsgleichung (vgl. (4.2))

$$\lambda^2 + 8\lambda\gamma s^2 - 1 = 0$$

$$\lambda_{1,2} = \frac{1}{2}(-8\gamma s^2 \pm \sqrt{64\gamma^2 s^4 + 4}).$$

Da $\gamma > 0$, folgt $|\lambda_2| > 0$, unabhängig von der Größe von γ

Instabilität, unabhängig vom Verhältnis τ zu h^2. Verfahren unbrauchbar.

Daß die Mittelpunktregel, die zur Approximation von y_t verwandt wurde, nicht immer schlechte Ergebnisse liefern muß, zeigt das

Beispiel 6: Die Schrödingergleichung

$$-\frac{\hbar^2}{2m}\Delta u + Uu = i\hbar\frac{\partial u}{\partial t}$$

hierbei sind: $\hbar$=Planck'sches Wirkungsquatum
m = Masse
U = einPotential.

Wir behandeln nur den 1D-Fall (ohne physikalische Konstanten) wie folgt

$$-i\alpha\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} - \beta u, \quad \alpha, \beta > 0$$

mit der zeitlichen Approximation durch die Mittelpunktsregel (auf der Zeitschicht j)

$$-i\alpha\frac{y^{j+1} - y^{j-1}}{2\tau} = y_{\bar{x}x} - \beta y.$$

Aus dem üblichen Ansatz (vgl. Vorseite, inclusive Abkürzungen) folgt

$$-i\alpha\frac{\lambda-\frac{1}{\lambda}}{2\tau}=-\frac{4}{h^2}s^2-\beta$$

$$-i\alpha\lambda^2-8\gamma s^2\lambda-2\tau\beta\lambda=0,\qquad \gamma=\frac{\tau}{h^2}$$

$$\lambda^2+i\underbrace{\left(\frac{8\gamma s^2}{\alpha}+\frac{2\tau\beta}{\alpha}\right)}_{=:d}\lambda-1=0$$

$$\lambda_{1,2}=\frac{1}{2}(-di\pm\sqrt{-d^2+4})\qquad\Longrightarrow$$

$$|\lambda_1\lambda_2|=1$$

und $|\lambda_i|=1$ solange $d^2\le 4$ (konjugiert komplexe Wurzeln)

$$d^2\le 4\qquad\Longleftrightarrow\qquad 8\gamma s^2+2\tau\beta\le 2\alpha$$

$s^2=1$ im schlimmsten Fall

$$8\frac{\tau}{h^2}+2\beta\le 2\alpha$$

Folge: Falls $\tau=O(h^2)$ erhält man Spielraum für eine stabile Rechnung trotz Verwendung der Mittelpunktregel.
Falls $d^2>4$ hat man Instabilität.

Diskretisierungsfehler für die 1D-Wellengleichung

$$\psi=u_{\bar{t}t}-a^2u^{\sigma}_{\bar{x}x}-\varphi,\quad \text{wobei } u=\text{exakte Lösung}$$

Wir benutzen die bekannten Entwicklungen aus § 2 für $u\in C^{4,4}$.

$$u_{\bar{t}t}=\frac{\partial^2u}{\partial t^2}+\frac{h^2}{12}\frac{\partial^4u}{\partial t^4}+O(\tau^4)$$

$$u_{\bar{x}x}=\frac{\partial^2u}{\partial x^2}+\frac{h^2}{12}\frac{\partial^4u}{\partial x^4}+O(h^4)$$

$$u^{(\sigma)}=(I+\sigma\tau^2\frac{\partial^2u}{\partial x^2})u+O(\tau^4)$$

Aus den letzten beiden Gleichungen folgt

$$\begin{aligned}u^{(\sigma)}_{\bar{x}x}&=(\boldsymbol{I}+\sigma\tau^2\frac{\partial^2u}{\partial x^2})(\frac{\partial^2u}{\partial x^2}+\frac{h^2}{24}\frac{\partial^4u}{\partial x^4})u\\&=\frac{\partial^2u}{\partial x^2}+\frac{h^2}{12}\frac{\partial^4u}{\partial x^4}+\sigma\tau^2\frac{\partial^4u}{\partial x^2\partial t^2}+O(\tau^4+h^4)\end{aligned}$$

Damit erhält man (für $\varphi = f^j$, üblicherweise)

$$\psi = \left(\frac{\partial^2 u}{\partial t^2} - a^2\frac{\partial^2 u}{\partial x^2} - f\right) + \frac{\tau^2}{12}\frac{\partial^4 u}{\partial t^4} - a^2\left(\frac{h^2}{12}\frac{\partial^4 u}{\partial x^4} + \sigma\tau^2\frac{\partial^4 u}{\partial x^2 \partial t^2}\right) + O(\tau^4 + h^4).$$

$$\psi = O(\tau^2 + h^2) \quad \text{falls } u \in C^{4,4}.$$

Eine Voraussetzung $u \in C^{6,6}$ ist jedoch nicht überflüssig, denn eine weitere Rechnung liefert

$$\psi = O(\tau^4 + h^4) \quad \text{falls } u \in C^{6,6} \text{ und}$$

$$\sigma = \frac{1 - \frac{1}{\gamma^2}}{12},\ \gamma = \frac{|a|\tau}{h},\ \varphi = f + \frac{1}{12}\left(\tau^2\frac{\partial^2 f}{\partial t^2} + h^2\frac{\partial^2 f}{\partial x^2}\right)$$

Bemerkung: Hier, wie im parabolischen und elliptischen Fall erhält man die Konvergenz aus Stabilität und Diskretisierungsordnung.

Beachte: Für $\gamma = 1$, also $\sigma = 0$ erhält man so ein explizites Verfahren 4. Ordnung (im Unterschied zum parabolischen Fall). Beachte jedoch Beispiel 4.
Noch höhere Ordnungen sind möglich, falls noch höhere Differenzierbarkeitsordnung vorausgesetzt wird.
Folge: Der Bedarf nach σ ist im hyperbolischen Fall kleiner als im parabolischen Fall. Verfahren mit $\sigma > 0$ werden wichtig bei Eigenwertaufgaben, falls die Approximationen für die höheren Eigenwerte schlecht sind und wenn sich diese höheren Eigenwerte bei Schwingungsaufgaben bemerkbar machen.
Dann benutzt man implizite Verfahren, die automatisch eine Dämpfung des Einflusses der höheren Eigenwerte bewirken (ohne Beweis).

Bemerkung: Die Stabilitätsabschätzungen, die wir in § 4 und § 8 (Stabilität bzgl. der rechten Seite auf Grund der Stabilität bzgl. der Anfangswerte) hergeleitet haben, lassen sich auf 3-Schicht-Schemata (für hyperbolische und parabolische Differentialgleichungen) verallgemeinern (natürlich unter entsprechenden Voraussetzungen). Dadurch werden die Neumann'schen Stabilitäts-Analysen auf Grund anderer Voraussetzungen und auch wegen der möglichen Instabilitätsbeweise, die bei vielen Differentialgleichungen Diskretisierungen, die "eigentlich auf der Hand liegen", ausscheiden.

§ 23 Literatur

1. Samarsky, A.A.: *Theorie der Differenzenverfahren*, Akad. Verlagsgesellschaft Geest u. Portig, Leipzig 1984

2. Morton, K.W. / Mayers, D.F.: *Num. Solution of Partial Diff. Equ.*, Cambridge University Press 1994

3. Iserles, A.:] *A first course in the Num. Analysis of Diff. Equ.*, Cambridge Texts in Appl. Math. 1996

4. Grossmann, Ch. / Roos, H.-G.: *Numerik partieller Differentialgleichungen*, Teubner Studienbücher 1993

5. Richtmyer, R.D. / Morton, K.W.: *Difference Methods for Initial-Value-Problems*, Interscience Publishers (Wiley and Sons) 1967

6. Forsythe, G.E. / Wasow, W.R.: *Finite Difference Methods for PDE*, Wiley and Sons 1960

7. Hackbusch, W.: *Theorie und Numerik elliptischer Dgln*, Teubner Studienbücher 1986

8. Hackbusch, W.: *Multi-Grid Methods and Application*, Springer 1986

9. Trottenberg, U. / Oosterlee, C. / Schüller, A.: *Multigrid*, Academic Press 2001

10. Brandt, A. *A guide to multigrid development: Multigrid Methods I*, Lecture Notes in Mathematics 960 S 230-312, 1982

11. Hackbusch, W. / Trottenberg, U.: *Multigrid II*, Lecture Notes in Mathematics 1228, 1986

Index