Über Differenzenverfahren zur numerischen Lösung stark gekoppelter Systeme nichtlinearer Diffusionsgleichungen

Karl Graf Finck von Finckenstein

Fachbereich Mathematik der Technischen Hochschule Darmstadt, Darmstadt

Z. Naturforsch. **42 a**, 1133–1140 (1987); eingegangen am 23. Juni 1987

Herrn Professor Dieter Pfirsch zum 60. Geburtstag gewidmet

Difference Methods for Strongly Coupled Systems of Nonlinear Diffusion Problems

A class of nonlinear implicit one step difference methods for quasilinear strongly coupled parabolic systems in two space variables is considered. The main part of the paper deals with proving convergence of the discrete approximations for vanishing step sizes. For this purpose, bounds for the inverse difference operators have to be derived previously. This is possible subject to a condition which can be considered as a generalization of the concept "parabolic" to systems. Finally, it is shown that the nonlinear systems arising from the discretizations have one and only one solution for all step sizes being sufficiently small.

1. Problemstellung

Viele physikalische Vorgänge werden durch Diffusionsgleichungen bzw. durch Systeme von solchen beschrieben. Ein wichtiges Beispiel aus der Plasmaphysik ist die Wärmeleitung in geheizten Plasmen, wobei man sich etwa für die zeitliche Entwicklung der Temperaturen und Dichten von Elektronen und Ionen interessiert. Man kommt dann auf Anfangs-Randwertprobleme der folgenden Art:

$$\begin{aligned} &\frac{\partial u}{\partial t} = \operatorname{div}(\varphi(u, \operatorname{grad} u) \cdot \operatorname{grad} u) \quad \text{in } G \times (0, T), \\ &u(\boldsymbol{x}, 0) = f(\boldsymbol{x}) \qquad \text{für } \boldsymbol{x} \in G, \\ &u(\boldsymbol{x}, t) = g(\boldsymbol{x}, t) \quad \text{für } (\boldsymbol{x}, t) \in \partial G \times [0, T], \end{aligned} \tag{1}$$

wobei $G \subset \mathbb{R}^d$ ($d = 1, 2$ oder 3) ein einfach zusammenhängendes beschränktes Gebiet mit stückweise glattem Rande ∂G und T die Zeitspanne ist, in der der Diffusionsvorgang betrachtet werden soll.

Die gesuchte Lösung $u(\boldsymbol{x}, t)$ sowie die Anfangs- und Randfunktionen $f(\boldsymbol{x})$, $g(\boldsymbol{x}, t)$ sind l-komponentige Vektorfunktionen, und

$$\varphi = \varphi(\xi_1^{(0)}, \ldots, \xi_l^{(0)}, \xi_1^{(1)}, \ldots, \xi_l^{(1)}, \ldots, \xi_1^{(d)}, \ldots, \xi_l^{(d)})$$

ist eine $l \times l$-Matrixfunktion, in der jede Komponente eine Funktion von $(d+1)\,l$ Variablen ist. φ ist im allgemeinen vollbesetzt; ihre Abhängigkeit nicht nur von den Lösungsfunktionen u, sondern auch von deren Ortsableitungen grad u (den Flüssen) ist von besonderem physikalischen Interesse und wird als *starke Kopplung* des nichtlinearen Systems (1) bezeichnet. Der Deutlichkeit halber schreiben wir das System (1) noch einmal ausführlich hin:

$$\frac{\partial u_i}{\partial t} = \sum_{k=1}^{d} \sum_{j=1}^{l} \frac{\partial}{\partial x_k} \left(\varphi_{ij} \cdot \frac{\partial u_j}{\partial x_k} \right), \quad 1 \leqq i \leqq l. \tag{1*}$$

Im folgenden soll stets angenommen werden, daß φ, f, g hinreichend glatte Funktionen sind, und daß (1) in $G \times (0, T)$ eine wohlbestimmte hinreichend glatte Lösung $u(\boldsymbol{x}, t)$ besitzt.

Über Differenzenverfahren zur Lösung nichtlinearer parabolischer Anfangs-Randwertprobleme existiert eine umfangreiche Literatur, so daß die am Schluß dieser Arbeit aufgeführten Zitate nur als repräsentativer Querschnitt zu verstehen sind: In [1] werden – wohl erstmalig – Stabilitätsuntersuchungen bei Differenzenverfahren für quasilineare Probleme durchgeführt. Die Arbeiten [3] bis [8] befassen sich mit der Konstruktion und Untersuchung von impliziten Differenzenverfahren für Diffusionsprobleme, die teilwise von Fragestellungen aus der Plasmaphysik herrühren. Von Interesse sind dabei unter anderem sogenannte linearisierte Verfahren, d. h. solche, die bei jedem Zeitschritt nur die Lösung eines linearen Gleichungssystems erfordern. Die Differenzenverfahren dieser Art sind in [5], [6] und auch in [10] behandelt.

Verallgemeinerungen können in zwei Richtungen erfolgen: Zum einen kann man von einzelnen Diffusionsgleichungen auf Systeme von solchen

Reprint requests to Prof. Dr. K. Graf Finck von Finckenstein, Fachbereich Mathematik, Technische Hochschule Darmstadt, Schloßgartenstraße 7, D-6100 Darmstadt.

0932-0784 / 87 / 1000-1133 $ 01.30/0. – Please order a reprint rather than making your own copy.

übergehen, wie es in [2] bzw. [9] für schwach gekoppelte und in [4] bzw. [5] für stark gekoppelte Systeme geschehen ist. Zum anderen aber können mehrere Ortsvariable in die Betrachtung einbezogen werden, wie es die Autoren der Arbeiten [2] und [10] getan haben.

Die vorliegende Arbeit befaßt sich mit stark gekoppelten Systemen in zwei Ortsvariablen. Ihr Ziel ist die Konstruktion und Untersuchung gewisser impliziter Differenzenapproximationen zur numerischen Lösung von (1). Dabei soll im Vordergrund unseres Interesses die Frage nach der Konvergenz der Näherungslösungen gegen die exakte Lösung u in den Gitterpunkten für verschwindende Schrittweiten stehen. Ist $G \subset \mathbb{R}^2$ ein Kreis (z. B. der Querschnitt eines Kreistorus), dann liegt es nahe, (1) in Polarkoordinaten umzuschreiben; ist das Problem dann außerdem noch kreissymmetrisch, so haben wir ein 1-dimensionales Problem. Für eine Klasse von Standard-Einschritt-Differenzenverfahren liegt in diesem Falle neben einer Reihe von numerischen Experimenten, die vor einiger Zeit am Max-Planck-Institut für Plasmaphysik durchgeführt wurden, ein Konvergenzbeweis in einer gewichteten L_2-Norm vor (vgl. [4]). Die Konvergenzordnung ist in dieser Norm gleich der Konsistenzordnung.

Hier nun wollen wir uns dem Fall von zwei Raumdimensionen zuwenden. Um die Verhältnisse möglichst übersichtlich beschreiben zu können, soll das Gebiet $G \subset \mathbb{R}^2$ ein zu einem kartesischen Koordinatensystem achsenparalleles Rechteck sein. Die Untersuchungen sind aber auf allgemeinere Gebiete und auch auf Zylinderkoordinaten ohne weiteres übertragbar.

Es sei ein kurzer Überblick über die vorliegende Arbeit gegeben: Der folgende Abschnitt beschreibt das allgemeine Vorgehen, um Konvergenz von impliziten Einschritt-Verfahren für nichtlineare parabolische Systeme der Gestalt (1) zu beweisen. Hierbei wird wesentlich eine Stabilitätseigenschaft benötigt, die äquivalent ist mit einer gewissen Definitheitsbedingung der zu den Differenzenoperatoren gehörigen Funktionalmatrizen. Nach Konstruktion der betreffenden Differenzenapproximation im dritten Abschnitt konzentriert sich die Untersuchung denn auch auf den Nachweis dieser Definitheitseigenschaft. Es zeigt sich, daß sie genau das diskrete Analogon zu einer der vielen in der Literatur vorhandenen Elliptizitätsdefinitionen für Systeme zweiter Ordnung ist; von J. Nečas wird sie unter der Bezeichnung „very strong ellipticity" (vgl. [11]) eingeführt. Ferner ist bemerkenswert, daß sich das naheliegende achsenparallele Rechteckgitter für die Diskretisierung von (1) keineswegs so gut eignet, wie das hier benutzte Gitter, das wir wegen seiner geometrischen Struktur „Rautengitter" nennen wollen. Im letzten Abschnitt schließlich wird unter Benutzung eines bekannten Satzes von Hadamard gezeigt, daß das durch die Diskretisierung von (1) entstehende nichtlineare Gleichungssystem eindeutig lösbar ist, falls die Schrittweiten hinreichend klein gewählt sind. Auch hier geht wesentlich die oben erwähnte Definitheitseigenschaft der Funktionalmatrix ein.

Numerische Testrechnungen mit dem betrachteten Verfahren werden zur Zeit im Rahmen einer Diplomarbeit durchgeführt.

Viele wertvolle Anregungen zur mathematischen Untersuchung von Differenzenverfahren verdanke ich meiner etwa 7jährigen Tätigkeit am Max-Planck-Institut für Plasmaphysik in Garching. Dem Leiter der dortigen theoretischen Abteilung, Herrn Professor Dr. Dietrich Pfirsch, ist daher diese Arbeit anläßlich seines 60. Geburtstages in freundschaftlicher Verbundenheit gewidmet.

2. Konvergenz von impliziten Einschritt-Verfahren

Im folgenden soll ein Vektor $z \in \mathbb{R}^m$ stets einen Spaltenvektor bezeichnen, z^T sei der entsprechende Zeilenvektor. Eine analoge Bezeichnung verwenden wir bei Matrizen $A \in \mathbb{R}^{m,n}$. Für Vektoren $z = (z_1, \ldots, z_m)^T$ führen wir die diskrete L_2-Norm ein

$$\| z \| := \left(\frac{z^T z}{m} \right)^{1/2}.$$

Bekanntlich gilt für die induzierte Matrixnorm

$$\| A \| = \sqrt{\varrho(A \cdot A^T)}, \quad A \in \mathbb{R}^{m,m},$$

wobei $\varrho(\cdot)$ den Spektralradius bezeichnet.

Nun sei das Gebiet $G \times [0, T] \subset \mathbb{R}^{d+1}$ mit einem Gitternetz versehen. Es bezeichne N die Anzahl der (irgendwie numerierten) inneren Gitterpunkte in jeder Zeitschicht $n \Delta t$ mit $n = 0, 1, \ldots, M$ und $M \Delta t = T$. Eine Ortsdiskretisierung des Differentialoperators auf der rechten Seite von (1) unter Einbeziehung der Randbedingungen definiert eine Vektorfunktion

$$F(v) = (F_1^T(v_1, \ldots, v_N), \ldots, F_N^T(v_1, \ldots, v_N))^T,$$

wobei die „Komponenten" v_ν bzw. $F_\nu(\ldots)$ ihrerseits l-komponentige Vektoren sind; es ist also

$$F(v)\colon \mathbb{R}^{Nl} \to \mathbb{R}^{Nl}.$$

Bei der Zeitdiskretisierung sollen zwei t-Schichten mit der Gewichtung $1-\alpha$, α ($\alpha \in [0,1]$) miteinander verbunden werden. Bezeichnen wir mit $U^n := (U_1^n, \ldots, U_N^n)^T \in \mathbb{R}^{Nl}$ die durch die Diskretisierung auf den Gitterpunkten definierte Näherungslösung, dann läßt sich das Differenzenverfahren folgendermaßen schreiben:

$$\begin{aligned} U^{n+1} + \alpha\, \Delta t\, F(U^{n+1}) &= U^n - (1-\alpha)\, \Delta t\, F(U^n), \\ & \qquad n = 0, 1, \ldots, M-1, \quad (2) \\ U^0 &= f. \end{aligned}$$

Hier ist also für jeden t-Schritt ein Nl-reihiges nichtlineares Gleichungssystem zu lösen.

Es sei nun $u^n := (u_1^n, \ldots, u_N^n)^T \in \mathbb{R}^{Nl}$ der Vektor der exakten Lösungswerte in den Gitterpunkten der t-Schicht Nr. n. Durch Taylor-Entwicklung erhalten wir

$$\begin{aligned} & u^{n+1} + \alpha\, \Delta t\, F(u^{n+1}) \\ & \quad = u^n - (1-\alpha)\, \Delta t\, F(u^n) + \Delta t\, \tau^n(\alpha), \qquad (3) \\ & u^0 = f, \quad n = 0, 1, \ldots, M-1, \end{aligned}$$

wobei $\tau^n(\alpha)$ der lokale Abschneidefehler des Verfahrens ist.

Um zu einer Konvergenzaussage des Differenzenverfahrens (2) zu kommen, werden die Gln. (2) von den Gln. (3) subtrahiert: wir bezeichnen mit $e^n := u^n - U^n$ den Verfahrensfehler und erhalten wegen der Beziehung

$$F(u^n) - F(U^n) = \int_0^1 F'(U^n + \sigma\, e^n)\, \mathrm{d}\sigma\, e^n =: \Gamma_n\, e^n \quad (4)$$

unmittelbar

$$\begin{aligned} & (I + \alpha\, \Delta t\, \Gamma_{n+1})\, e^{n+1} \\ & \quad = (I - (1-\alpha)\, \Delta t\, \Gamma_n)\, e^n + \Delta t\, \tau^n(\alpha), \qquad (5) \\ & e^0 = 0, \quad n = 0, 1, \ldots, M-1. \end{aligned}$$

Hier ist mit $F'(\cdot) \in \mathbb{R}^{Nl, Nl}$ die Funktionalmatrix von $F(\cdot)$ bezeichnet, und die Integration in (4) ist elementweise zu verstehen.

Unser Ziel ist eine Abschätzung für den in (5) auftretenden Verfahrensfehler e^n in der diskreten L_2-Norm durch den Abschneidefehler $\tau^n(\alpha)$. Hierzu wollen wir an die Funktionalmatrix von $F(\cdot)$ die folgende grundlegende Bedingung stellen:

Es existiert eine Konstante $K \geqq 0$ so, daß für alle Vektoren $v, w \in \mathbb{R}^{Nl}$ mit $w^T w = 1$ gilt:

$$w^T (F'(v) + F'(v)^T)\, w \geqq -K. \qquad (6)$$

Aus dieser Bedingung, die später bei der Betrachtung unseres speziellen Differenzenverfahrens nachgewiesen werden muß, ergibt sich zunächst:

Lemma: Für $\frac{1}{2} \leqq \alpha \leqq 1$, für alle hinreichend kleine Schrittweiten Δt und für $n = 0, 1, \ldots, M$ gelten unter der Voraussetzung (6) die folgenden Ungleichungen:

$$\| (I + \alpha\, \Delta t\, \Gamma_n)^{-1} \| \leqq 1 + \Delta t\, K, \qquad (7\mathrm{a})$$

$$\begin{aligned} & \| (I + \alpha\, \Delta t\, \Gamma_n)^{-1} \cdot (I - (1-\alpha)\, \Delta t\, \Gamma_n) \| \\ & \qquad \leqq 1 + \Delta t\, K, \qquad (7\mathrm{b}) \end{aligned}$$

wobei Γ_n die in (4) auftretende Matrix und K die in (6) auftretende Konstante ist.

Beweis: Aus (6) und der Definition der Matrizen Γ_n folgt unmittelbar für $w^T w = 1$:

$$w^T (\Gamma_n + \Gamma_n^T)\, w \geqq -K, \quad n = 0, 1, \ldots, M. \quad (6^*)$$

Im folgenden lassen wir der Einfachheit halber den Index n fort.

a) Sei $\mu > 0$ ein beliebiger Eigenwert der positiv definiten Matrix

$$\begin{aligned} P := [(I + \alpha\, \Delta t\, \Gamma) \cdot (I + \alpha\, \Delta t\, \Gamma^T)]^{-1}\colon\ & P w = \mu\, w, \\ & w^T w = 1. \end{aligned}$$

Wir multiplizieren diese Eigenwertgleichung nacheinander von links mit P^{-1} und w^T und erhalten auf Grund der gemachten Voraussetzungen für hinreichend kleine Δt

$$\begin{aligned} \mu &= \frac{1}{w^T [I + \alpha\, \Delta t (\Gamma + \Gamma^T) + \alpha^2\, \Delta t^2\, \Gamma \Gamma^T]\, w} \\ &\leqq \frac{1}{1 - \alpha\, \Delta t\, K} \leqq 1 + 2\, \Delta t\, K. \end{aligned}$$

Also ist $\varrho(P) = \| (I + \alpha\, \Delta t\, \Gamma)^{-1} \|^2 \leqq 1 + 2\, \Delta t\, K$ und (7a) ist bewiesen.

b) Ungleichung (7b) zeigt man ganz analog, indem man dieselben überlegungen mit der Matrix

$$\begin{aligned} \hat{P} := & [(I + \alpha\, \Delta t\, \Gamma) \cdot (I + \alpha\, \Delta t\, \Gamma^T)]^{-1} \\ & \cdot (I + (\alpha - 1)\, \Delta t\, \Gamma) \cdot (I + (\alpha - 1)\, \Delta t\, \Gamma^T) \end{aligned}$$

durchführt. q.e.d.

Wir können nun unsere gewünschte Abschätzung herleiten: Mit dem „transformierten" Verfahrensfehler

$$\tilde{e}^n := (I + \alpha\, \Delta t\, \Gamma_n)\, e^n$$

ergibt sich aus (5) unmittelbar

$$\tilde{e}^{n+1} = (I + \alpha\, \Delta t\, \Gamma_n)^{-1} \cdot (I - (1-\alpha)\, \Delta t\, \Gamma_n)\, \tilde{e}^n + \Delta t\, \tau^n(\alpha), \qquad (5^*)$$
$$\tilde{e}^0 = 0, \quad n = 0, 1, \ldots, M-1.$$

Normbildung und Benutzung von (7b) liefert

$$\| \tilde{e}^{n+1} \| \leqq (1 + \Delta t\, K) \| \tilde{e}^n \| + \Delta t\, \tau(\alpha),$$
$$\| \tilde{e}^0 \| = 0, \quad n = 0, 1, \ldots, M-1,$$

mit

$$\tau(\alpha) := \max_{0 \leqq n \leqq M} \| \tau^n(\alpha) \|.$$

Dies ergibt induktiv unter Verwendung von (7a) und wegen $n \Delta t \leqq T$ die folgende Abschätzung für $n = 0, 1, \ldots, M$:

$$\begin{aligned} \| e^n \| &\leqq (1 + \Delta t\, K) \| \tilde{e}^n \| \\ &\leqq \Delta t\, \tau(\alpha)(1 + \Delta t\, K) \sum_{v=0}^{n-1} (1 + \Delta t\, K)^v \\ &= \Delta t\, \tau(\alpha) \sum_{v=1}^{n} (1 + \Delta t\, K)^v \\ &\leqq \Delta t\, \tau(\alpha)\, n (1 + \Delta t\, K)^n \leqq T e^{KT} \tau(\alpha), \end{aligned}$$

womit gezeigt ist

Satz: Für das Differenzenverfahren (2) gelte die Bedingung (6), es sei $\alpha \in [\frac{1}{2}, 1]$ und die Schrittweite Δt sei hinreichend klein. Dann gilt für den Verfahrensfehler die Abschätzung

$$\max_n \| e^n \| \leqq T e^{KT} \max_n \| \tau^n(\alpha) \|. \qquad (8)$$

Dabei ist K die in (6) auftretende Konstante und $\tau^n(\alpha)$ der in (3) auftretende Abschneidefehler.

Anm. 1: Abgesehen von den erforderlichen Einschränkungen an die Schrittweitenverhältnisse hat es sich in der Praxis nicht bewährt, $\alpha < \frac{1}{2}$ zu wählen. Wegen der etwas höheren Konsistenzordnung wird öfter $\alpha = \frac{1}{2}$ gewählt (Crank-Nicholson-Verfahren), was andererseits Nachteile hat, auf die wir hier nicht eingehen können.

Anm. 2: Die Abschätzung (8) liefert eine Konvergenzaussage, falls $\tau(\alpha)$ für verschwindende Schrittweiten gegen Null konvergiert. Jedes vernünftige Differenzverfahren besitzt natürlich diese Eigenschaft, die man als Konsistenz bezeichnet.

3. Ein Differenzenverfahren in zwei Raumdimensionen

Wir betrachten nun den zweidimensionalen Fall und bezeichnen die Ortsvariablen mit x und y. Vor der Konstruktion des Differenzenverfahrens wollen wir die schon im 1. Abschnitt angekündigte Elliptizitätsbedingung für den Differentialoperator auf der rechten Seite von (1) formulieren. Aus (1) folgt durch Ausdifferenzieren

$$\begin{aligned} u_t &= \frac{\partial}{\partial x}(\varphi(u, u_x, u_y)\, u_x) + \frac{\partial}{\partial y}(\varphi(u, u_x, u_y)\, u_y) \\ &= a\, u_{xx} + (e + d)\, u_{xy} + c\, u_{yy} + (D_1 \varphi\, u_x)\, u_x \\ &\qquad + (D_1 \varphi\, u_y)\, u_y \end{aligned}$$

mit

$$a := \varphi + D_2 \varphi\, u_x, \quad e := D_3 \varphi\, u_x,$$
$$c := \varphi + D_3 \varphi\, u_y, \quad d := D_2 \varphi\, u_y. \qquad (9)$$

Dabei sind die $D_1 \varphi$, $D_2 \varphi$, $D_3 \varphi$ die Ableitungen der Matrixfunktion φ nach den ersten l, den zweiten l bzw. den dritten l Variablen. Die $D_v \varphi$ sind also Tensoren 3. Stufe, während $D_1 \varphi\, u_x$ usw. natürlich wieder $l \times l$-Matrizen sind.

In Anlehnung an Nečas (vgl. [11]) stellen wir jetzt die folgende Forderung an die rechte Seite von (1), die zweifellos auch von physikalischer Bedeutung sein dürfte:

Es existiert eine feste Konstante $K_1 > 0$ so, daß für alle Vektoren $\xi, \eta, \zeta, z_1, z_2 \in \mathbb{R}^l$ gilt:

$$\begin{aligned} &z_1^T a(\xi, \eta, \zeta)\, z_1 + z_2^T c(\xi, \eta, \zeta)\, z_2 + z_2^T d(\xi, \eta, \zeta)\, z_1 \\ &\qquad + z_1^T e(\xi, \eta, \zeta)\, z_2 \geqq K_1 (z_1^T z_1 + z_2^T z_2). \end{aligned} \qquad (10a)$$

Ferner benötigen wir noch die folgende Voraussetzung:

Die Tensoren $\varphi(\xi, \eta, \zeta)$, $D_v \varphi(\xi, \eta, \zeta)$ $(v = 1, 2, 3)$ sind für alle $\xi, \eta, \zeta \in \mathbb{R}^l$ elementweise betragsmäßig nach oben beschränkt. (10b)

Anm. 3: Man überlegt sich unmittelbar, daß (10a) äquivalent damit ist, daß der Rayleigh-Quotient der symmetrischen Matrizen

$$\begin{pmatrix} a + a^T & d^T + e \\ d + e^T & c + c^T \end{pmatrix} \in \mathbb{R}^{2l, 2l}$$

für alle $\xi, \eta, \zeta \in \mathbb{R}^l$ größer oder gleich $2 K_1$ ist.

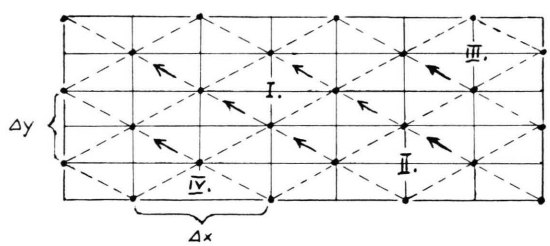
Anm. 4: Die Bedingungen (10a), (10b) scheinen auf den ersten Blick hin reichlich restriktiv zu sein. Daß dieses nicht der Fall ist, zeigt folgende Überlegung:

Die Lösung $u(x, y, t)$ sowie deren Ableitungen $u_x(x, y, t)$, $u_y(x, y, t)$ besitzen in $\bar{G} \times [0, T] \subset \mathbb{R}^3$ untere und obere Schranken, die natürlich von φ, f, g abhängen. Durch diese Schranken wird ein (kompakter) Quader $Q \subset \mathbb{R}^{3l}$ definiert. Wir denken uns nun die Komponenten der Matrixfunktion φ außerhalb Q derart (hinreichend glatt) fortgesetzt, daß sie außerhalb eines Q umfassenden Quaders $\hat{Q}$ geeignete Konstanten sind. Die Ableitungen $D_\nu \varphi$ ($\nu = 1, 2, 3$) haben dann $\hat{Q}$ als Träger, und in $\mathbb{R}^{3l} \backslash \hat{Q}$ sind die Koeffizienten in (10a) konstant. Durch geeignete Wahl der Konstanten lassen sich dann die Bedingungen (10a), (10b) in ganz $\mathbb{R}^l$ erfüllen. Dabei ist die Lösung des Anfangs-Randwertproblems (1) durch die Fortsetzung von φ nicht beeinflußt.

Es sei jetzt $G \subset \mathbb{R}^2$ ein achsenparalleles Rechteck und Δx, Δy Schrittweiten, für die stets gelten soll

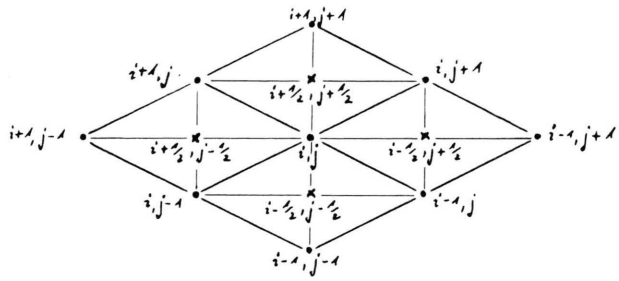
$$0 < \tilde{R} \leqq \frac{\Delta x}{\Delta y} \leqq \hat{R}, \quad \tilde{R}, \hat{R} \text{ konstant.} \tag{11}$$

G wird, wie in Fig. 1 skizziert, mit einem Punktgitter versehen, das wir *Rautengitter* nennen wollen:



Es sei N die Anzahl der inneren Punkte des Gitters, die in irgendeiner Weise numeriert seien (z.B. in der durch die Pfeile angedeuteten Weise). Zur Aufstellung des Differenzenverfahrens jedoch wollen wir die Punkte durch Doppelindizes kennzeichnen. Die Gitterpunkte P_{ij} (bzw. die Zwischengitterpunkte $P_{i-\frac{1}{2}\,j-\frac{1}{2}}$ usw.) sind in Fig. 2 der Übersicht halber nur durch $i, j, i-\frac{1}{2}, j-\frac{1}{2}$ usw. gekennzeichnet.

Im folgenden lassen wir aus Gründen der Schreibersparnis die Variable t bzw. den Index n fort und definieren für l-komponentige Vektorfunktionen



$v(x, y)$:

$$v_{ij} := v(P_{ij}),$$

$$\varphi_{i-\frac{1}{2}\,j-\frac{1}{2}} := \varphi\left(\frac{v_{i-1\,j-1} + v_{ij-1} + v_{ij} + v_{i-1\,j}}{4}, \right.$$

$$\left. \frac{v_{i-1\,j} - v_{ij-1}}{\Delta x}, \frac{v_{ij} - v_{i-1,\,j-1}}{\Delta y}\right) \quad \text{usw.}$$

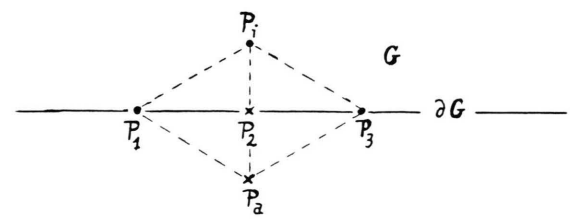
Sodann ersetzen wir den Differentialoperator auf der rechten Seite von (1) im Punkte P_{ij} durch

$$\begin{aligned} F_{ij}(v_{i-1\,j-1} \ldots v_{i+1\,j+1}) \\ := \frac{1}{\Delta x^2} [\varphi_{i+\frac{1}{2}\,j-\frac{1}{2}}(v_{ij} - v_{i+1\,j-1}) \\ - \varphi_{i-\frac{1}{2}\,j+\frac{1}{2}}(v_{i-1\,j+1} - v_{ij})] \\ + \frac{1}{\Delta y^2} [\varphi_{i-\frac{1}{2}\,j-\frac{1}{2}}(v_{ij} - v_{i-1\,j-1}) \\ - \varphi_{i+\frac{1}{2}\,j+\frac{1}{2}}(v_{i+1\,j+1} - v_{ij})]. \end{aligned} \tag{12}$$

Diese 9-Punkte-Formel wird für jeden inneren Punkt des Gitters aufgestellt. Dabei benötigt man für gewisse innere Punkte noch außerhalb von G liegende Hilfspunkte, in denen die benötigten Funktionswerte nach folgender Formel von 2. Ordnung approximiert werden (vgl. Figur 3):

$$v(P_a) = 4v(P_2) - v(P_i) - v(P_1) - v(P_3). \tag{13}$$

Faßt man für die N inneren Punkte P_{ij} die l-komponentigen Vektoren v_{ij} bzw. $F_{ij}(\cdot)$ zu Nl-komponentigen Vektoren v bzw. $F(\cdot)$ zusammen und führt



die t-Diskretisierung wie oben beschrieben durch, so erhält man aus (12) und (13) ein Differenzenverfahren der Form (2), welches konsistent von 2. Ordnung in $\Delta x, \Delta y$ ist, wie man durch Taylor-Entwicklung feststellt. Wir erhalten also für den in (3) auftretenden Abschneidefehler

$$\tau^n(\alpha) = \begin{cases} O(\Delta x^2 + \Delta y^2 + \Delta t) & \text{für} \quad \alpha \neq \frac{1}{2}, \\ O(\Delta x^2 + \Delta y^2 + \Delta t^2) & \text{für} \quad \alpha = \frac{1}{2}. \end{cases} \tag{14}$$

Zum Nachweis der Konvergenz des beschriebenen impliziten Einschritt-Verfahrens ($\alpha \geqq \frac{1}{2}$) muß auf Grund der Ausführungen von Abschnitt 2 nur noch die Bedingung (6) nachgewiesen werden. Hier wird sich die Elliptizitätsbedingung (10a) als wesentliche Voraussetzung erweisen. Wir betrachten nun die Matrix

$$A(v) := \Delta x\, \Delta y \cdot (F'(v) + F'(v)^T) \in \mathbb{R}^{Nl, Nl},$$

also den mit den Schrittweiten multiplizierten symmetrischen Anteil der zu der Vektorfunktion $F(v)$ gehörigen Funktionalmatrix. Zu zeigen ist, daß ein festes $K > 0$ existiert mit

$$w^T A(v)\, w \geqq -K\, \Delta x\, \Delta y \quad \text{für alle } v, w \in \mathbb{R}^{Nl}, w^T w = 1. \tag{15}$$

$A(v)$ ist eine Bandmatrix mit einer doppelten Blockstruktur, da wir es einerseits mit zwei Raumdimensionen, andererseits mit einem System von l Differentialgleichungen zu tun haben. Die kleinen Blöcke sind l-reihige Matrizen, die sich aus den Matrizen – bzw. deren Transponierten – der folgenden Gestalt zusammensetzen:

$$a_{i-\frac{1}{2}\,j-\frac{1}{2}} := \varphi_{i-\frac{1}{2}\,j-\frac{1}{2}} + D_2\, \varphi_{i-\frac{1}{2}\,j-\frac{1}{2}} \frac{v_{i-1j} - v_{ij-1}}{\Delta x},$$

$$c_{i-\frac{1}{2}\,j-\frac{1}{2}} := \varphi_{i-\frac{1}{2}\,j-\frac{1}{2}} + D_3\, \varphi_{i-\frac{1}{2}\,j-\frac{1}{2}} \frac{v_{ij} - v_{i-1j-1}}{\Delta y},$$

$$d_{i-\frac{1}{2}\,j-\frac{1}{2}} := D_2\, \varphi_{i-\frac{1}{2}\,j-\frac{1}{2}} \frac{v_{ij} - v_{i-1j-1}}{\Delta y},$$

$$e_{i-\frac{1}{2}\,j-\frac{1}{2}} := D_3\, \varphi_{i-\frac{1}{2}\,j-\frac{1}{2}} \frac{v_{i-1j} - v_{ij-1}}{\Delta x}, \tag{16}$$

$$b_{i-\frac{1}{2}\,j-\frac{1}{2}} := d_{i-\frac{1}{2}\,j-\frac{1}{2}} + e^T_{i-\frac{1}{2}\,j-\frac{1}{2}},$$

$$\alpha_{i-\frac{1}{2}\,j-\frac{1}{2}} := \tfrac{1}{4}\, D_1\, \varphi_{i-\frac{1}{2}\,j-\frac{1}{2}} \frac{v_{i-1j} - v_{ij-1}}{\Delta x},$$

$$\gamma_{i-\frac{1}{2}\,j-\frac{1}{2}} := \tfrac{1}{4}\, D_1\, \varphi_{i-\frac{1}{2}\,j-\frac{1}{2}} \frac{v_{ij} - v_{i-1j-1}}{\Delta y}.$$

Jetzt betrachten wir das Punktgitter (vgl. Fig. 1), das unser Gebiet G in rautenförmige Zellen zerlegt; Ihre Mittelpunkte sind durch die Doppelindizes der Form $i - \frac{1}{2}\, j - \frac{1}{2}$ usw. gekennzeichnet (vgl. Fig. 2) und entsprechen den in (16) angegebenen l-reihigen Matrizen.

Der Fig. 1 entnimmt man weiterhin, daß 4 Typen von solchen Zellen auftreten:

1. Zellen, deren Eckpunkte 4 innere Gitterpunkte sind (Typ I),
2. Zellen, deren Eckpunkte aus 3 inneren und einem Randpunkt bestehen (Typ II),
3. Zellen, deren Eckpunkte 2 innere und 2 Randpunkte sind (Typ III),
4. Zellen, deren Eckpunkte ein innerer, 2 Randpunkte und ein äußerer Hilfspunkt sind (Typ IV).

Durch die Zuordnung der Zellmittelpunkte zu den gemäß ihrer Indizierung entsprechenden l-reihigen Matrizen (16) wird eine Zerlegung der Matrix $A(v)$ definiert: Die l-reihigen Block elemente zu einer Zelle stehen genau in denjenigen Blockreihen der Matrix $A(v)$, die der Numerierung der Rautenecken entsprechen. Nimmt man diese Blöcke aus $A(v)$ heraus und schiebt sie zu einer $4l$-reihigen Matrix zusammen, dann entsteht im Falle von Typ I eine symmetrische Matrix der Gestalt

$$\left(\begin{array}{cc|cc} \frac{\Delta x}{\Delta y} \cdot \hat{c} - \Delta x\, \hat{\gamma} & b - \Delta y\, \alpha^T - \Delta x\, \gamma & -b + \Delta y\, \alpha^T - \Delta x\, \gamma & -\frac{\Delta x}{\Delta y}\, \hat{c} - \Delta x\, \check{\gamma} \\ b^T - \Delta y\, \alpha - \Delta x\, \gamma^T & \frac{\Delta y}{\Delta x}\, \hat{a} - \Delta y\, \hat{\alpha} & -\frac{\Delta y}{\Delta x}\, \hat{a} - \Delta y\, \check{\alpha} & -b^T - \Delta y\, \alpha + \Delta x\, \gamma^T \\ \hline -b^T + \Delta y\, \alpha - \Delta x\, \gamma^T & -\frac{\Delta y}{\Delta x}\, \hat{a} + \Delta y\, \check{\alpha} & \frac{\Delta y}{\Delta x}\, \hat{a} + \Delta y\, \hat{\alpha} & b^T + \Delta y\, \alpha + \Delta x\, \gamma^T \\ -\frac{\Delta x}{\Delta y}\, \hat{c} + \Delta x\, \check{\gamma} & -b - \Delta y\, \alpha^T + \Delta x\, \gamma & b + \Delta y\, \alpha^T + \Delta x\, \gamma & \frac{\Delta x}{\Delta y}\, \hat{c} + \Delta x\, \hat{\gamma} \end{array}\right) =: T^{(1)}_{i-\frac{1}{2}\,j-\frac{1}{2}} \in \mathbb{R}^{4l, 4l}, \tag{17}$$

wobei wir die Doppelindizes $i-\frac{1}{2}\,j-\frac{1}{2}$ auf der linken Seite übersichtshalber fortgelassen sowie noch folgende Abkürzungen verwendet haben:

$$\hat{a} := a + a^T, \quad \check{\alpha} := \alpha - \alpha^T \quad \text{usw.}$$

Ferner stellt man fest, daß die den Zellen vom Typ II (bzw. vom Typ III) entsprechenden Matrizen 3 l-reihige (bzw. 2 l-reihige) Hauptuntermatrizen von denen des Typs I sind. Die den Zellen des Typs IV zugeordneten Matrizen schließlich sind l-reihig und haben die Gestalt $2\dfrac{\Delta x}{\Delta y}\,\hat{c}$ oder $2\dfrac{\Delta y}{\Delta x}\,\hat{a}$. Wir wollen diese Matrizen entsprechend dem Typ der ihnen zugeordneten Zellen mit

$$T^{(k)}_{i-\frac{1}{2}\,j-\frac{1}{2}} \in \mathbb{R}^{(5-k)l,\,(5-k)l}, \quad k = 1, 2, 3, 4$$

bezeichnen.

Wir betrachten nun die quadratische Form (15) und erhalten auf Grund der durchgeführten Überlegungen

$$w^T A(v)\, w = \sum_{k=1}^{4} \sum_{\mu} w_\mu^{(k)T} \cdot T_\mu^{(k)} \cdot w_\mu^{(k)}, \tag{18}$$

wobei μ zur Abkürzung für die Doppelindizes $i-\frac{1}{2}$, $j-\frac{1}{2}$ steht und $\sum_{\mu}$ die Summation über alle Zellen vom Typ k bezeichnet. Der Abschnittsvektor $w_\mu^{(k)} \in \mathbb{R}^{(5-k)l}$ besteht aus denjenigen Komponenten von $w \in \mathbb{R}^{Nl}$, die in der Numerierung den (inneren) Eckpunkten der Zelle Nr. μ vom Typ k entsprechen.

Wir wenden uns nun der Untersuchung der Teilmatrizen (17) zu, wobei der Einfachheit halber die Doppelindizes $i-\frac{1}{2}$, $j-\frac{1}{2}$ fortgelassen werden. Wir setzen

$$g := \begin{pmatrix} \sqrt{\dfrac{\Delta y}{\Delta x}} \cdot \alpha & \sqrt{\dfrac{\Delta y}{\Delta x}} \cdot \alpha \\ \sqrt{\dfrac{\Delta x}{\Delta y}} \cdot \gamma & \sqrt{\dfrac{\Delta x}{\Delta y}} \cdot \gamma \end{pmatrix}, \quad p := \begin{pmatrix} \dfrac{\Delta y}{\Delta x} \cdot \hat{a} & b^T \\ b & \dfrac{\Delta x}{\Delta y} \cdot \hat{c} \end{pmatrix},$$

$$h := \sqrt{\Delta x\, \Delta y}, \quad \tilde{T} := \begin{pmatrix} p - h\,\hat{g} & -p - h\,\check{g} \\ -p + h\,\check{g} & p + h\,\hat{g} \end{pmatrix}.$$

Man prüft jetzt leicht nach, daß $\tilde{T} \in \mathbb{R}^{4l,\,4l}$ zu der in (17) auftretenden Matrix $T^{(1)} = T^{(1)}_{i-\frac{1}{2}\,j-\frac{1}{2}}$ ähnlich ist, da $\tilde{T}$ aus $T^{(1)}$ durch Vertauschen der beiden ersten Blockzeilen und Bockspalten hervorgeht.

Als nächstes untersuchen wir die quadratische Form der Matrizen p:

Wegen der Elliptizitätsbedingung (10a) (vgl. auch Anm. 3) und wegen der Schrittweitenbedingung (11) ergibt eine leichte Rechnung

$$\tilde{z}^T p\, \tilde{z} \geqq 2K_1 \operatorname{Min}\{\tilde{R}, \hat{R}^{-1}\}\, \tilde{z}^T \tilde{z} =: 2K_2\, \tilde{z}^T \tilde{z}$$

für alle $\tilde{z} \in \mathbb{R}^{2l}$. Wegen der Beschränktheitsbedingung (10b) folgt also für hinreichend kleine h und für alle Matrizen p und $\hat{g}$:

$$\tilde{z}^T (p + h\,\hat{g})\, \tilde{z} \geqq K_2\, \tilde{z}^T \tilde{z} \quad \text{für alle} \quad \tilde{z} \in \mathbb{R}^{2l}. \tag{19a}$$

Weiterhin existiert wegen (10b) und (11) eine Konstante $\mathcal{K}$, so daß für alle Matrizen g

$$\|g^T\| \leqq \tfrac{1}{4}\,\mathcal{K} \tag{19b}$$

erfüllt ist. Nun sei $z = \begin{pmatrix} z^1 \\ z_2 \end{pmatrix}$ mit $z_1, z_2 \in \mathbb{R}^{2l}$ und $z_1^T z_1 + z_2^T z_2 = 1$. Wir setzen zur Abkürzung $\tau := [(z_1 - z_2)^T \cdot (z_1 - z_2)]^{1/2}$ und erhalten mit (19a), (19b) und der Schwarzschen Ungleichung

$$\begin{aligned} z^T T z &= z_1^T (p - h\,\hat{g})\, z_1 + z_2^T (p + h\,\hat{g})\, z_2 \\ &\quad - z_2^T (p - h\,\check{g})\, z_1 - z_1^T (p + h\,\check{g})\, z_2 \\ &= (z_1 - z_2)^T (p + h\,\hat{g})\, (z_1 - z_2) - 4h\, z_1^T g^T (z_1 - z_2) \\ &\geqq K_2\, \tau^2 - h\,\mathcal{K}\,\tau \geqq -\frac{h^2\,\mathcal{K}^2}{4K_2} =: -K_3\, h^2. \end{aligned}$$

Da aber $\tilde{T}$ und $T^{(1)}$ ähnlich sind, haben wir für alle in (18) auftretenden Matrizen $T_\mu^{(1)}$ die Abschätzung

$$z^T T_\mu^{(1)}\, z \geqq -K_3\, \Delta x\, \Delta y, \quad z^T z = 1$$

erhalten. Da die $T_\mu^{(2)}$ und $T_\mu^{(3)}$ Hauptuntermatrizen gewisser $T_\mu^{(1)}$ sind, erhalten wir natürlich auch für diese die obige Abschätzung. Schließlich noch sind die $T_\mu^{(4)}$ wegen (10a) positiv definit, erfüllen also obige Abschätzung ebenfalls. Zusammenfassend folgt damit

$$z^T T_\mu^{(k)}\, z \geqq -K_3\, \Delta x\, \Delta y \quad \text{für alle} \quad z \in \mathbb{R}^{(5-k)l} \tag{20}$$

mit $z^T z = 1$ und für alle in (18) auftretenden Matrizen $T_\mu^{(k)}$.

Jede Komponente des Einheitsvektors $w \in \mathbb{R}^{Nl}$, der in (15) bzw. (18) auftritt, kommt in der Summe auf der rechten Seite von (18) genau viermal vor, da

jeder innere Punkt des Gitters Eckpunkt von genau vier Rauten ist. Daher erhalten wir aus (18) und (20)

$$w^T A(v)\, w \geqq -K_3\, \Delta x\, \Delta y \sum_{k=1}^{4} \sum_{\mu} w_\mu^{(k)^T} w_\mu^{(k)}$$

$$= -4 K_3\, \Delta x\, \Delta y =: -K\, \Delta x\, \Delta y,$$

womit (15) gezeigt ist. Da aber (15) zu (6) äquivalent ist, folgt die gewünschte Konvergenz unseres Differenzenverfahrens.

Anm. 5: Die Teilmatrizen $\hat{g}$ bzw. $\check{g}$ in $\tilde{T}$ sind die diskreten Analoga der Konvektionsanteile unseres Divergenzoperators in (1). Sie treten in $\tilde{T}$ als Störung 1. Ordnung (bzgl. der Schrittweite h) auf; trotzdem wird der minimale Eigenwert nur von 2. Ordnung in h gestört. Dieser Sachverhalt ist das diskrete Analogon der Tatsache, daß in einem Differentialoperator der Elliptizitätscharakter durch Terme 1. Ordnung nicht beeinflußt wird.

Anm. 6: Die durchgeführten Untersuchungen sind natürlich auch für ein achsenparalleles Rechteckgitter möglich; es zeigt sich jedoch, daß die Differenzenapproximationen komplizierter werden, wenn man die Bedingung (6) erhalten will. Im Falle der schwachen Kopplung allerdings (also φ unabhängig von grad u) ist das Rechteckgitter vorzuziehen, da dann die 5-Punkte-Formel bereits geeignet ist.

4. Existenz und Eindeutigkeit der Näherungslösung

In unseren Konvergenzbetrachtungen waren wir von der Annahme ausgegangen, daß die nichtlinearen Gleichungssysteme in (2) für jedes n eindeutig lösbar sind. Daß dieses für hinreichend kleine Schrittweiten $\Delta x, \Delta y, \Delta t$ auch in der Tat richtig ist, zeigt folgende Überlegung.

Es sei die Lösung U^n für $t_n = n\, \Delta t$ bereits bestimmt. Wir betrachten dann das Gleichungssystem

$$\Phi(X) = b \quad \text{mit} \quad b := U^n - (1-\alpha)\, \Delta t\, F(U^n) \qquad (2^*)$$

und

$$\Phi(X) := X + \alpha\, \Delta t\, F(X)\colon \mathbb{R}^{Nl} \to \mathbb{R}^{Nl}.$$

Mit der gleichen Methode wie im Beweis des Lemmas läßt sich nun unter der Voraussetzung (6) für die Funktionalmatrix

$$\Phi'(X) = I + \alpha\, \Delta t\, F'(X)$$

die folgende, (7a) entsprechende Abschätzung herleiten:

$$\| \Phi'(X)^{-1} \| \leqq 1 + \Delta t\, K \leqq \gamma < \infty \qquad (21)$$

für alle $X \in \mathbb{R}^{Nl}$. γ ist dabei eine von den Schrittweiten unabhängige Konstante.

Damit ist ein Satz von Hadamard anwendbar (vgl. z. B. [12], Theorem 5.3.10), der insbesondere besagt, daß $\Phi(X)$ bijektiv ist; somit ist (2*) eindeutig lösbar.

[1] J. Douglas, Jr., Trans. Amer. Math. Soc. **89,** 484 (1958).
[2] J. Douglas, Jr. u. J. E. Gunn, J. Assoc. Comput. Mach. **9,** 4, 450 (1962).
[3] K. Eriksson u. V. Thomée, Galerkin Methods for Singular Boundary Value Problems in One Space Dimension, S-412 96, Dept. Math., University of Göteborg 1982.
[4] K. v. Finckenstein, Numerical Methods for Partial Differential Equations **3** (1987), in press.
[5] K. v. Finckenstein, Methoden Verfahren math. Phys. **20,** 7 (1980).
[6] K. v. Finckenstein u. D. Düchs, Lecture Notes Math. **395,** 3 (1974).
[7] K. v. Finckenstein u. D. Düchs, Methoden Verfahren math. Phys. **15,** 75 (1976).
[8] K. v. Finckenstein u. K. v. Hagenow, Numer. Math. **20,** 372 (1973).
[9] R. Gorenflo, Computing **8,** 343 (1971).
[10] L. J. Hayes, SIAM J. Num. Anal. **18,** 781 (1981).
[11] J. Nečas, Introduction to the Theory of Nonlinear Elliptic Equations, B. G. Teubner, Leipzig 1983.
[12] J. M. Ortega u. W. C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, London 1970.