Numerische Simulation von Diffusionsprozessen mit nichtnegativitätserhaltenden konservativen Differenzenverfahren

Rudolf Gorenflo
Drittes Mathematisches Institut der Freien Universität Berlin

Angelika Kuban
Institut für Geologie — Mathematische Geologie — der Freien Universität Berlin

Z. Naturforsch. **37 a**, 759–768 (1982); eingegangen am 29. April 1982

To Professor Arnulf Schlüter on his 60th Birthday

Numerical Simulation of Diffusion Processes with conservative difference schemes of non-negative type

After analyzing the general linear equation of diffusion (of a substance or of energy) with source term and given influx across the boundary we describe a method for constructing explicit and implicit conservative difference schemes which also preserve nonnegativity. We call a scheme "conservative" if via a convenient sum-analogue it does exactly imitate the conservation of a substance or energy. We concretize this method for the spatially two-dimensional heat equation in a rectangle with given influx. We also present a conservative implicit scheme with alternating directions.

§ 1. Einleitung

Bei der numerischen Simulation von Evolutionsprozessen durch Diskretisierung ist es wünschenswert und nützlich, wichtige ihrer Eigenschaften möglichst gut zu imitieren. Solche Eigenschaften sind beispielsweise Integralinvarianten (zeitunabhängige Integrale über den Ortsbereich), Erhaltung der Nichtnegativität, Ausgleichsvorgänge (Einebnung lokaler Extrema bei der Diffusion, auch „Variationsverminderung" genannt), asymptotisches Verhalten. Die beste Imitation einer solchen Eigenschaft ist ihr exakter Nachvollzug. Differenzenschemata, welche dieser Absicht möglichst guter oder sogar exakter Imitation nachkommen, haben eine hohe Chance, numerisch stabil zu sein. Der rechnende Physiker weiß das aus Erfahrung, dem Mathematiker gelingt manchmal der Beweis dafür.

Zur diskreten Imitation von Integralinvarianten sind in den letzten Jahren Prädiktor-Korrektor-Verfahren vorgeschlagen und in Testrechnungen auf ihre Brauchbarkeit untersucht worden (Sasaki [1], Isaacson [2], Navon [3]).

Ihre Grundidee ist, beim Übergang von einem Zeitniveau zum nächsten ein ohne Beachtung der Integralinvarianten entworfenes Differenzenschema zu verwenden und dann die für das neue Zeitniveau berechneten Werte so zu korrigieren, daß ein Summenanalogon der Integralinvariante erfüllt ist; ein Maß der Korrektur (vorzugsweise eine gewichtete Quadratsumme) soll hierbei möglichst klein ausfallen).

Im Falle von Diffusionsprozessen leuchtet die stabilisierende Wirkung der Nichtnegativitätserhaltung (und damit verbunden der Einebnungseigenschaften) unmittelbar ein; sie ist bei parabolischen Modellproblemen in vielen Lehrbüchern (z. B. [4]) fürs explizite Zweischichtschema auch bewiesen. Sogar notwendig ist sie beim expliziten Standardschema für die Gleichung $u_t = u_{xx}$, wenn man bei Gitterverfeinerung Konstanz von $\Delta t/(\Delta x)^2$ fordert. Beim Crank-Nicolson-Schema ist sie zwar für Stabilität nicht mehr notwendig, aber doch bei Vorliegen rauher Daten sehr nützlich, da bei ihrem Nichterfülltsein die Ausgleichseigenschaften des Schemas oft nicht gut genug sind. Beim voll-impliziten Schema ist die Nichtnegativitätserhaltung unter sehr allgemeinen Voraussetzungen automatisch erfüllt. Man vgl. etwa [5] und [6].

Wir beschäftigen uns in dieser Arbeit hauptsächlich mit der räumlich zweidimensionalen linearen Diffusionsgleichung $u_t = a(u_{xx} + u_{yy})$ in einem Rechteck mit reflektierendem Rand oder gegebenen

Sonderdruckanforderungen an Prof. Dr. R. Gorenflo, Freie Universität Berlin, Drittes Mathematisches Institut, Arnimallee 2–6, 1000 Berlin 33.

0340-4811 / 82 / 0800-0759 $ 01.30/0. — Please order a reprint rather than making your own copy.

Zuflüssen über den Rand und geben explizite und implizite Schemata der Standardklasse an, die sowohl *nichtnegativitätserhaltend* als auch *konservativ* sind. Ein Differenzenschema nennen wir „konservativ", wenn es die Erhaltungseigenschaft für Substanz (bzw. Energie) in einem Substanz- (bzw. Energie-)Diffusionsprozeß in Gestalt eines Summenanalogons exakt imitiert. Speziellen Wert legen wir hierbei auf Ersetzung der impliziten Verfahren durch lokal-eindimensionale Schemata (alternierende Richtungen) zur Verringerung des Rechenaufwandes. Wegen expliziter Schemata dieser Art für räumlich mehrdimensionale allgemeinere Diffusionsgleichungen vgl. man [7], explizite und implizite Schemata für räumlich eindimensionale Diffusionsprobleme werden in [6] analysiert.

Die Nützlichkeit solcher Schemata für Diffusionsprobleme ohne Quellen und Zufluß wird ausführlich diskutiert in den sehr anregenden Arbeiten [8] und [9] von Pert, der gründlich die Einebnung von Extremwerten analysiert und auch auf nichtkartesische Koordinatensysteme eingeht. Glashoff und Kreth untersuchen in [10] für eindimensionale Diffusionsprobleme die Frage der Variationsverminderung im Sinne der zeitlichen Abnahme der Anzahl der räumlichen Nullstellen der Lösung.

Wir wollen hier die Herleitung und Wirkungsweise unserer Schemata darstellen, aber an manchen Stellen auf strenge Beweise verzichten. Für diese verweisen wir auf [6]; dort wird zwar nur der räumlich eindimensionale Fall behandelt, die Beweise sind aber verallgemeinerbar. Um nicht die wesentlichen Ideen hinter mathematischer Abstraktheit zu verstecken, wollen wir hier nicht alles so allgemein formulieren, wie es möglich wäre. Wir stellen in § 2 die Analyse eines allgemeinen linearen Modellproblems vor und in § 3 unsere Diskretisierungsprinzipien, die wir in § 4 auf die zweidimensionale Wärmeleitungsgleichung anwenden und in § 5 modifizieren für die Methode der alternierenden Richtungen. In § 6 geben wir Ergebnisse einer numerischen Fallstudie.

§ 2. Das Modellproblem allgemein

Sei $G \subset \mathbb{R}^p$ ($= p$-dimensionaler euklidischer Raum) ein Gebiet mit genügend gutartigem Rand ∂G (mindestens stückweise glatt: der Gaußsche Integralsatz soll anwendbar sein). Für eine gesuchte reelle Funktion $u(x,t)$ mit $x = (x_1, x_2, \cdots, x_p) \in G \cup \partial G$, $t \geqq 0$ betrachten wir die *Anfangs-Randwert-Aufgabe* ((2.1), (2.2), (2.3)):

$$\gamma(x)\,\partial u/\partial t = \operatorname{div}(Bu) + f(x,t) \quad \text{für} \quad x \in G \cup \partial G, \tag{2.1}$$

$$u(x,0) = g(x) \quad \text{für} \quad x \in G \quad \text{(Anfangsbedingung)}, \tag{2.2}$$

$$Bu(x,t) \cdot \boldsymbol{\nu} = \psi(x,t) \quad \text{für} \quad x \in \partial G, \quad t \geqq 0 \quad \text{(Rand-Zufluß-Bedingung)}. \tag{2.3}$$

Es sei $\gamma(x) > 0$, und alle auftretenden Funktionen seien so gutartig (mindestens stetig) wie jeweils benötigt.

Mit einer symmetrischen ($a_{jk} = a_{kj}$) positiv-definiten Matrix

$$A(x) = (a_{jk}(x) \mid j = 1, \ldots, p;\ k = 1, \ldots, p)$$

und einem Spaltenvektor

$$\boldsymbol{b}(x) = (b_1(x), \ldots, b_p(x))^{\mathrm{T}}$$

ist hierbei der auf u wirkende Operator B *alternativ* gegeben als

$$Bu = (\operatorname{div}(\boldsymbol{a}_1(x)\,u), \ldots, \operatorname{div}(\boldsymbol{a}_p(x)\,u))^{\mathrm{T}} + \boldsymbol{b}(x)\,u \tag{2.4'}$$

mit

$$\boldsymbol{a}_j(x) = (a_{j1}(x),\ a_{j2}(x), \ldots, a_{jp}(x))^{\mathrm{T}}$$

oder

$$Bu = A(x)\,\operatorname{grad} u\,. \tag{2.4''}$$

Hierbei sind div und grad die üblichen bezüglich der Ortsvariablen x wirkenden Operationen der Vektoranalysis, und $\boldsymbol{\nu} = \boldsymbol{\nu}(x)$ bezeichnet den äußeren Normaleinheitsvektor für $x \in \partial G$.

Die Differentialgleichung (2.1) lautet dann *alternativ*

$$\gamma(x)\frac{\partial u}{\partial t} = \sum_{j=1}^{p}\sum_{k=1}^{p}\frac{\partial^2}{\partial x_j\,\partial x_k}(a_{jk}(x)\,u) + \sum_{j=1}^{p}\frac{\partial}{\partial x_j}(b_j(x)\,u) + f(x,t) \tag{2.1'}$$

oder (2.1'')

$$\gamma(x)\frac{\partial u}{\partial t} = \sum_{j=1}^{p}\sum_{k=1}^{p}\frac{\partial}{\partial x_j}\left(a_{jk}(x)\frac{\partial u}{\partial x_k}\right) + f(x,t)\,.$$

Diese Gleichungen dienen zur Beschreibung von Prozessen der Diffusion von Energie oder Substanz; etwa mit $u(x,t)$ als Temperatur, $\gamma(x)$ als spezifischer Wärme, $w(x,t) = \gamma(x)\,u(x,t)$ als Energiedichte oder im Sonderfall $\gamma(x) \equiv 1$ mit $u(x,t)$ auch als Sub-

stanzdichte. Die Matrix $A(x)$ mißt die Stärke der (i.allg. anisotropen und ortsabhängigen) *Diffusion*, und $-\boldsymbol{b}(x)$ gibt die Stärke der *Konvektion* („Drift"), $f(x,t)$ ist ein *Quellterm*, $(Bu)\cdot\boldsymbol{\nu}$ die *Zuflußrate* über den Rand ∂G.

Die gesamte *Energie oder im Sonderfall* $\gamma\equiv 1$ *auch Substanz* innerhalb G zur Zeit t ist

$$E(t)=\int_G \gamma(x)\,u(x,t)\,\mathrm{d}x\,. \tag{2.5}$$

Differentiation nach t (rechts hinterm Integralzeichen) bei Beachtung von (2.1) mit dem Gaußschen Integralsatz gibt

$$\frac{\mathrm{d}E(t)}{\mathrm{d}t}=\int_{\partial G} Bu(x,t)\cdot\boldsymbol{\nu}\,\mathrm{d}\sigma+\int_G f(x,t)\,\mathrm{d}x \tag{2.6}$$

mit $\boldsymbol{\nu}=\boldsymbol{\nu}(x)$ als nach außen gerichteter Einheitsnormale auf ∂G und $\mathrm{d}\sigma$ als Oberflächenelement. Es folgt

$$E(t)=E(0)+\int_0^t\left(\int_G f(x,t')\,\mathrm{d}x+\int_{\partial G}\psi(x,t')\,\mathrm{d}\sigma\right)\mathrm{d}t' \tag{2.7}$$

mit

$$E(0)=\int_G \gamma(x)\,g(x)\,\mathrm{d}x\,.$$

Im Sonderfall $f(x,t)\equiv 0$ und $\psi(x,t)\equiv 0$ („reflektierende Wand" ∂G) ist

$$E(t)=E(0)\quad\text{für alle}\quad t\geqq 0\,, \tag{2.8}$$

die Gesamtenergie (bzw. Gesamtsubstanz). (2.5) ist dann eine *Integralinvariante*. Im allgemeinen inhomogenen Fall ist $E(t)$ nach (2.7) explizit angebbar ohne vorherige Bestimmung der Lösung $u(x,t)$ der Anfangsrandwertaufgabe.

Eine weitere wesentliche Eigenschaft des dargestellten Diffusionsprozesses ist die *Nagumo-Westphal-Eigenschaft* der *Nichtnegativitäts-Erhaltung* (die Theorie hierzu findet man in sehr allgemeiner Form bei Walter [11]): *Ist überall und immer* $g\geqq 0$, $f\geqq 0$, $\psi\geqq 0$, *so auch überall und immer* $u\geqq 0$. Das leuchtet physikalisch ein, nirgends und nimmer entsteht von selber eine negative Dichte. Diese Eigenschaft hat weitere qualitative Eigenschaften zur Folge wie etwa die bekannten Maximum- und Minimumprinzipien und die Einebnungseigenschaften, die wir hier nicht eingehend auseinandersetzen wollen (man vgl. etwa Pert [9]).

§ 3. Diskretisierungsprinzipien

Um allzu häufige verbale Fallunterscheidungen zu vermeiden, sprechen wir im folgenden von *Energie-Diffusion* mit dem stillschweigenden Verständnis, daß es sich im Sonderfall $\gamma(x)\equiv 1$ auch um Substanzdiffusion handeln kann. Wir teilen das Gebiet G in *Zellen* Z_j ein, wobei der *Index* j über eine *endliche Indexmenge* J variiert. Ist $h>0$ die typische lineare Abmessung einer Zelle (Kantenlänge im Falle einer Würfeleinteilung), so ist $v_j\sim h^p$, wenn wir mit v_j das Volumen der Zelle Z_j bezeichnen. Die Zeitvariable t diskretisieren wir mit $t_n=n\tau$ mit dem Schritt $\tau>0$ und

$$n\in\{0\}\cup\mathbb{N}=\{0,1,2,\ldots\}\,.$$

Die gesamte Energie zum Zeitpunkt t_n denken wir uns in *Klumpen* $e_{j,n}=v_j w_{j,n}$ aufgeteilt, hierbei sitze der Klumpen $e_{j,n}$ im Zeitpunkt t_n in der Zelle Z_j. Da im kontinuierlichen Fall die *Energiedichte* $w(x,t)=\gamma(x)\,u(x,t)$ ist, lautet unser *Approximationswunsch* so:

$$e_{j,n}/v_j\approx w\,(x_{(j)},t_n)$$

oder auch

$$e_{j,n}/(v_j\gamma_j)\approx u\,(x_{(j)},t_n)$$

oder auch in Integralform $e_{j,n}\approx\int_{Z_j} w(x,t_n)\,\mathrm{d}x$.

Hierbei ist $\gamma_j=\gamma(x_{(j)})$ und $x_{(j)}$ ein jeweils ausgezeichneter Punkt (im Idealfall der Mittelpunkt) der Zelle Z_j. Sei $\varphi_{j,n}$ die in Zelle Z_j im Zeitintervall $t_n<t\leqq t_{n+1}$ durch die Quelldichte f entstehende und im Falle einer randbenachbarten Zelle zusätzlich gemäß der Zuflußrate ψ von außen zufließende Energie, also

$$\varphi_{j,n}=\int_{t_n}^{t_{n+1}}\int_{Z_j} f(x,t)\,\mathrm{d}x\,\mathrm{d}t+\varrho_{j,n} \tag{3.1}$$

mit

$$\varrho_{j,n}=0\,,\quad\text{falls}\quad \partial Z_j\cap\partial G\ \text{leer ist}\,,$$

andernfalls

$$\varrho_{j,n}=\int_{t_n}^{t_{n+1}}\int_{\partial Z_j\cap\partial G}\psi(x,t)\,\mathrm{d}\sigma\,\mathrm{d}t\,.$$

Ein *diskretes Diffusionsmodell* mit stets exakter Bilanz der Gesamtenergie hat dann die Gestalt

$$e_{j,0}=\int_{Z_j}\gamma(x)\,g(x)\,\mathrm{d}x\quad\text{für}\quad j\in J\,, \tag{3.2}$$

$$e_{j,n+1}=\sum_{k\in J}p_{jk}\,e_{k,n}+\sum_{k\in J}q_{jk}\,\varphi_{k,n} \tag{3.3}$$

$$\text{für}\quad j\in J\,,\quad n+1\in\mathbb{N}\,,$$

mit den Umverteilungsraten p_{jk}, q_{jk}, die den Bedingungen (K) der Konservativität und (NNE) der *Nichtnegativitätserhaltung* genügen:

$$\text{(K)} \quad \sum_{j\in J} p_{jk} = 1\,, \quad \sum_{j\in J} q_{jk} = 1 \quad \text{für alle} \quad k\in J\,,$$

$$\text{(NNE)} \quad \text{alle } p_{jk} \geqq 0\,, \quad \text{alle } q_{jk} \geqq 0\,.$$

Bei der Diskretisierung konkreter Anfangsrandwertaufgaben muß man natürlich zusätzlich auf das Erfülltsein üblicher Konsistenzbedingungen achten, so daß bei Verfeinerung der Zellenzerlegung die Diskretisierung allmählich in die Differentialgleichung mit Rand- und Anfangsbedingungen übergeht.

Es ist zweckmäßig, Spaltenvektoren $\tilde{e}_n$ und $\tilde{\varphi}_n$ mit den Komponenten $e_{j,n}$, $\varphi_{j,n}$, $j\in J$, in $\mathbb{R}^J$ einzuführen und die Umverteilungsraten $p_{j,k}$ und $q_{j,k}$, $j,k\in J$, zu Matrizen P und Q zusammenzufassen. Die Umverteilungsvorschrift (3.3) lautet dann

$$\tilde{e}_{n+1} = P\tilde{e}_n + Q\tilde{\varphi}_n \tag{3.3'}$$

und die Bedingungen (K) und (NNE) gehen über in

$$\text{(K')} \qquad \eta P = \eta\,, \quad \eta Q = \eta\,,$$

$$\text{(NNE')} \qquad P \geqq 0\,, \qquad Q \geqq 0\,.$$

Hierin ist $\eta = (1, 1, \ldots, 1)$ ein Zeilenvektor mit lauter Einsen, eine Eins für jedes Element der Indexmenge J, und die Zeichen „$\geqq$" in (NNE') sind (im Sinne von Varga [12]) gemäß (NNE) elementeweise zu verstehen. (K') und (NNE') bedeuten, daß die transponierten Matrizen P und Q *stochastisch* sein sollen.

Wir wollen statt (3.2) eine etwas flexiblere Anfangsbedingung

$$e_{j,0} = \int_{Z_j} \gamma(x)\, g(x)\, dx + \varepsilon_{j,0} \quad \text{für} \quad j\in J \tag{3.2*}$$

zulassen, da man manchmal lieber mit

$$e_{j,0} = \gamma_j v_j g_j = v_j\, \gamma(x_{(j)})\, g(x_{(j)})$$

rechnet (die Konsistenzuntersuchung gestaltet sich dann einfacher). Wenn g genügend glatt ist, so ist dann $e_{j,0}/(\gamma_j v_j) - g(x_{(j)}) = O(h^2)$. Im Falle (3.2) sind dann alle $\varepsilon_{j,0} = 0$. Die $\varepsilon_{j,0}$ fassen wir in einem Spaltenvektor $\tilde{\varepsilon}_0$ zusammen. Addition der Energieklumpen $\varepsilon_{j,n}$ zum Zeitpunkt t_n ergibt die diskretisierte Energie

$$E_n = \sum_{j\in J} e_{j,n} = \eta\tilde{e}_n \quad \text{für alle} \quad n\in\mathbb{N}_0\,, \tag{3.4}$$

und nach Erinnerung an die Definition der Energie $E(t)$ in (2.5) und ihre Darstellung (2.7) als Integral können wir einen Satz aussprechen.

Satz 3.1. Unter der Bedingung (K') *der Konservativität ist*

$$E_n = E(t_n) + \eta\tilde{\varepsilon}_0 \quad \textit{für alle} \quad n\in\mathbb{N}_0\,, \tag{3.5}$$

der eventuell vorhandene Energie-Anfangs-Fehler $\eta\tilde{\varepsilon}_0$ ist also persistent (von ihm abgesehen wird die Gesamtenergie stets exakt bilanziert).

Beweis. Aus (3.2*) folgt $E_0 = E(0) + \eta\tilde{\varepsilon}_0$, und für alle $n\in\mathbb{N}_0 = \{0\}\cup\mathbb{N}$ ist wegen (K'), (3.1) und (2.7)

$$\begin{aligned} E_{n+1} - E_n &= \eta\tilde{e}_{n+1} - \eta\tilde{e}_n = \eta(P\tilde{e}_n + Q\varphi_n) - \eta\tilde{e}_n \\ &= \eta\tilde{\varphi}_n = \int_{t_n}^{t_{n+1}} \left(\int_G f(x,t)\,dx + \int_{\partial G} \psi(x,t)\,d\sigma \right) dt \\ &= E(t_{n+1}) - E(t_n)\,, \end{aligned}$$

also $E_{n+1} - E_n = E(t_{n+1}) - E(t_n)$.

Durch (Teleskop-)Addition folgt hieraus

$$E_n - E_0 = E(t_n) - E(0) \quad \text{für} \quad n\in\mathbb{N}_0\,,$$

also

$$\begin{aligned} E_n &= E(t_n) + E_0 - E(0) \\ &= E(t_n) + \eta\tilde{\varepsilon}_0\,. \end{aligned}$$

□

Wir wollen jetzt die allgemeine *Theorie konservativer nichtnegativitätserhaltender Differenzenschemata* der Standardklasse für die *Anfangsrandwertaufgabe* [(2.1), (2.2), (2.3)] des § 2 darstellen.

Mit dem *Implizitätsparameter* θ erklären wir den in bezug auf den Zeitindex wirkenden linearen Interpolationsoperator:

$$I_\theta z_n = \theta z_{n+1} + \overline{\theta} z_n\,.$$

Hierbei ist $0 \leqq \theta \leqq 1$ und $\overline{\theta} = 1 - \theta$.

Mit $u_{j,n} = e_{j,n}/(v_j\gamma_j)$ bezeichnen wir die zu berechnenden Näherungswerte für $u(x_{(j)}, t_n)$, und mit $Lu_{j,n}$ eine konsistente Diskretisierung von $\operatorname{div}(Bu)$ mit geeigneter Berücksichtigung der Randbedingung für randbenachbarte Zellen (wie man das im konkreten Fall macht, wird in § 4 exemplarisch vorgeführt).

Mit der Anfangsbedingung

$$u_{j,0} = g(x_{(j)}) \quad \text{für} \quad j\in J \tag{3.6}$$

(dies entspricht (3.2*)) hat man dann in

$$(1/\tau)\,\gamma_j(u_{j,n+1} - u_{j,n}) = I_\theta L_{u_{j,n}} + \varphi_{j,n}/(\tau v_j)$$

ein Schema der Standardklasse (3.7)

($\theta = 0$: *explizites Schema*, $\theta = 1$: *vollimplizites Schema*,

$\theta = \frac{1}{2}$: *Crank-Nicolson*).

Wir müssen nun den örtlichen Differenzenoperator L so konstruieren, daß er möglichst gut konsistent zur Anfangsrandwertaufgabe ist und andererseits zu einem den Bedingungen (K′) und (NNE′) genügenden Umverteilungsschema (3.3′) führt. Zu diesem Zweck führen wir Diagonalmatrizen Γ und V ein gemäß

$$\Gamma = \operatorname{diag}(\gamma_j \mid j \in J), \quad V = \operatorname{diag}(v_j \mid j \in J).$$

Diese sind invertierbar (alle $\gamma_j > 0$, alle $v_j > 0$) und miteinander vertauschbar:

$$\Gamma V = V\Gamma, \quad (\Gamma V)^{-1} = V^{-1}\Gamma^{-1} = \Gamma^{-1}V^{-1}$$

(alle diese Matrizen sind diagonal). Mit der typischen linearen Zellenabmessung h sei

$$C = (1/h^p)\,V \tag{3.8}$$

die auf die Größenordnung 1 *skalierte Volumenmatrix*. Fassen wir die $u_{j,n}$ (für $j \in J$) zu einem Vektor $\tilde{u}_n$ zusammen, multiplizieren (3.7) mit v_j und ziehen aus dem Differenzenoperator L einen Faktor $1/h^2$ heraus ($\operatorname{div}(B\cdot)$ ist ja ein Differentialoperator der Ordnung 2), so bekommen wir mit der auf die Größenordnung 1 skalierten örtlichen Diskretisierungsmatrix M aus (3.7) das Schema

$$\frac{1}{\tau} V\Gamma(\tilde{u}_{n+1} - \tilde{u}_n) = \frac{1}{h^2} I_\theta V M \tilde{u}_n + \frac{1}{\tau}\tilde{\varphi}_n,$$

wegen

$$\tilde{u}_n = (V\Gamma)^{-1}\tilde{e}_n = V^{-1}\Gamma^{-1}\tilde{e}_n$$

mit dem Schrittweitenparameter

$$\mu = (\tau/h^2) > 0 \tag{3.9}$$

und der Skalierung (3.8) also für $n \in \mathbb{N}_0$

$$\begin{aligned}&\tilde{e}_{n+1} - \tilde{e}_n \\ &= \mu C M C^{-1}\Gamma^{-1}(\theta\tilde{e}_{n+1} + \overline{\theta}\tilde{e}_n) + \varphi_n. \end{aligned} \tag{3.10}$$

Formale Auflösung nach $\tilde{e}_{n+1}$ gibt (3.3′) mit

$$\begin{aligned} Q &= (I - \mu\theta C M C^{-1}\Gamma^{-1})^{-1}, \\ P &= Q(I + \mu\overline{\theta} C M C^{-1}\Gamma^{-1}), \end{aligned} \tag{3.11}$$

wobei I die Einheitsmatrix bedeutet.

Da die skalierte Volumenmatrix durch die Geometrie der Zellenzerlegung schon bestimmt ist, müssen wir in konkreten Fällen die räumliche Diskretisierungsmatrix so konstruieren, daß P und Q die gewünschten Eigenschaften haben.

Für die *Konservativität* geben wir eine *algebraische Bedingung* an.

Satz 3.2. Wenn die Matrix $I - \mu\theta C M C^{-1}\Gamma^{-1}$ invertierbar ist, so ist die Bedingung

$$\eta C M = \mathrm{o} \tag{3.12}$$

notwendig und hinreichend dafür, daß die Matrizen P und Q in (3.11) die Eigenschaft (K′) der Konservativität haben. Hierbei ist $\mathrm{o} = (0, 0, \ldots, 0)$ ein Null-Zeilenvektor, für jedes $j \in J$ eine Null.

Kommentar. (3.12) besagt, daß jede *Spalte* der Matrix CM die Summe 0 hat. (K′) besagt, daß jede Spalte der Matrix P und jede Spalte der Matrix Q die Summe 1 hat.

Beweis des Satzes.

(a) für „hinreichend“: Aus (3.12) folgt

$$\eta(I - \mu\theta C M C^{-1}\Gamma^{-1}) = \eta, \quad \text{also} \quad \eta = \eta Q,$$

und weiter

$$\eta P = \eta Q(I + \mu\overline{\theta} C M C^{-1}\Gamma^{-1}) = \eta.$$

(b) für „notwendig“: Es gelte (K′). Aus

$$\eta Q = \eta \quad \text{folgt} \quad \eta = \eta(I - \mu\theta C M C^{-1}\Gamma^{-1}),$$

also

(i) $\theta\eta C M = \mathrm{o}$,

und wegen $\eta Q = \eta$ folgt aus $\eta P = \eta$ dann

$$\eta(I + \mu\overline{\theta} C M C^{-1}\Gamma^{-1}) = \eta, \quad \text{also}$$

(ii) $\overline{\theta}\eta C M = \mathrm{o}$.

Da θ und $\overline{\theta}$ nicht beide gleichzeitig verschwinden können, hat man also $\eta C M = \mathrm{o}$. □

Für die Erfüllung der Bedingung (NNE′) ist Anwendung der Theorie der nichtnegativen Matrizen und der M-Matrizen zu empfehlen, die in den Büchern von Collatz [13] und Varga [12] dargestellt ist. In konkreten Fällen ergibt sich im Falle $\theta < 1$ eine Beschränkung für $\mu = \tau/h^2$, keine Beschränkung oder allenfalls $\mu = O(1/h)$ im Falle $\theta = 1$. Exemplarisch wird das in § 4 vorgeführt. Wir werden dafür folgenden Satz brauchen (man vgl. Collatz [13], Seite 297).

Satz 3.3. Hat die quadratische Matrix $A = (a_{jk})$ lauter nichtpositive Nichtdiagonalelemente (alle $a_{jk} \leqq 0$ für $j \neq k$) und sind alle ihre Zeilen Summen $\sum_k a_{ik} > 0$, so existiert ihre Inverse A^{-1} und ist $\geqq 0$ (alle Elemente der Inverse $\geqq 0$).

§ 4. Diskretisierung der Wärmeleitungsgleichung

Im allgemeinen Modellproblem [(2.1), (2.2), (2.3)] nehmen wir $p=2$ und bezeichnen die Ortsvariablen mit x und y statt mit x_1 und x_2. Wir nehmen

$$a \equiv \text{const} > 0\,, \quad b \equiv 0\,, \quad \gamma \equiv 1\,,$$

und für $G \cup \partial G$ ein Rechteck $[0, x^*] \times [0, y^*]$ mit $x^* > 0, y^* > 0$. Wir haben dann für $u = u(x, y, t)$ die Anfangsrandwertaufgabe (für $t > 0$)

$$\frac{\partial u}{\partial t} = a\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + f(x, y, t) \quad \text{für} \quad (x, y) \in G \cup \partial G\,, \tag{4.1}$$

$$u(x, y, 0) = g(x, y) \quad \text{für} \quad (x, y) \in G \cup \partial G\,, \tag{4.2}$$

$$\begin{aligned} -a\frac{\partial u}{\partial x} &= \psi(0, y, t) \quad \text{für} \quad x = 0\,, \quad 0 < y < y^*\,, \quad t > 0\,, \\ a\frac{\partial u}{\partial x} &= \psi(x^*, y, t) \quad \text{für} \quad x = x^*\,, \quad 0 < y < y^*\,, \quad t > 0\,, \\ -a\frac{\partial u}{\partial y} &= \psi(x, 0, t) \quad \text{für} \quad 0 < x < x^*\,, \quad y = 0\,, \quad t > 0\,, \\ a\frac{\partial u}{\partial y} &= \psi(x, y^*, t) \quad \text{für} \quad 0 < x < x^*\,, \quad y = y^*\,, \quad t > 0\,. \end{aligned} \tag{4.3}$$

Zur Anwendung der in § 3 entwickelten Theorie nehmen wir an, x^* und y^* seien ganzzahlige Multipla einer positiven Schrittweite h. *Alternativ* betrachten wir *zwei Typen* möglicher Zerlegungen des Rechtecks $[0, x^*] \times [0, y^*]$ in Zellen Z_j (gemäß § 3). Der *Doppelindex* $j = (r, s)$ durchlaufe die *Indexmenge*

$$J = \{\bar{r}, \bar{r}+1, \ldots, \bar{\bar{r}}-1, \bar{\bar{r}}\} \times \{\bar{s}, \bar{s}+1, \ldots, \bar{\bar{s}}\} \tag{4.4}$$

mit $\bar{\bar{r}} - \bar{r} \in \mathbb{N}$, $\bar{\bar{s}} - \bar{s} \in \mathbb{N}$, und es sei

$$\bar{r} = \bar{s} = \tfrac{1}{2}\,, \quad \bar{\bar{r}} = (x^*/h) - \tfrac{1}{2}\,, \quad \bar{\bar{s}} = (y^*/h) - \tfrac{1}{2} \quad \text{bei Typ 1}\,, \tag{4.5'}$$

$$\bar{r} = \bar{s} = 0\,, \quad \bar{\bar{r}} = (x^*/h)\,, \quad \bar{\bar{s}} = (y^*/h) \quad \text{bei Typ 2}\,. \tag{4.5''}$$

Die Punkte $(x, y)_j = jh = (rh, sh) = (x_r, y_s)$ nennen wir „*Gitterpunkte*". Um jeden Gitterpunkt $(x, y)_j$ als Zentrum legen wir ein achsenparalleles Quadrat mit der Seitenlänge h und nehmen als Z_j den Durchschnitt dieses Quadrats mit der Menge $G \cup \partial G$.

Für das skalierte Volumen $c_j = v_j/h^2$ (man beachte § 3) gilt: $c_j = 1$ für alle Zellen bei Typ 1 und für alle nicht randbenachbarten Zellen bei Typ 2.

Bei Typ 2 ist $c_j = 1/4$ für die vier eckenbenachbarten Zellen und $c_j = 1/2$ für alle anderen randbenachbarten Zellen. Wir haben dann die Matrix $C = \text{diag}(c_j \mid j \in J)$. Skizze 1 zeigt Typ 1, Skizze 2 Typ 2 der Zerlegung, beidemale mit $x^* = 4h$, $y^* = 3h$. Die Zahlen c_j sind in die Zellen eingetragen.

Ausgehend von der symmetrischen Diskretisierung $h^{-2}(u_{r+1,s} - 2u_{r,s} + u_{r-1,s})$ für $\partial^2 u/\partial x^2$ und analog für $\partial^2 u/\partial y^2$ (wir unterdrücken der besseren Übersichtlichkeit wegen hier den Zeitindex n) erhalten wir für eine nicht randbenachbarte Zelle Z_j die *Fünf-Punkt-Diskretisierung*

$$(CM\tilde{u})_j = a(u_{r,s-1} + u_{r-1,s} + u_{r+1,s} + u_{r,s+1} - 4u_{r,s})\,. \tag{4.6}$$

Zur Konstruktion der in (3.10) auftretenden Matrix CM gehen wir so vor. Die Reihenfolge der Zeilen (zu jedem Doppelindex $j \in J$ gehört eine Zeile) entspreche dem horizontalenweisen Durchlaufen der Zellen Z_j, links unten in G beginnend, rechts oben

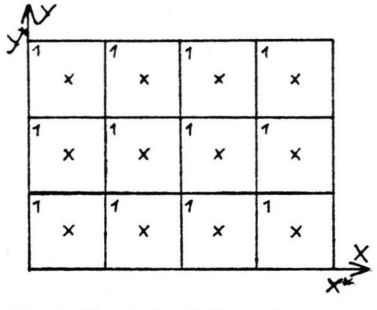

Fig. 1. Typ 1 der Zellenzerlegung.

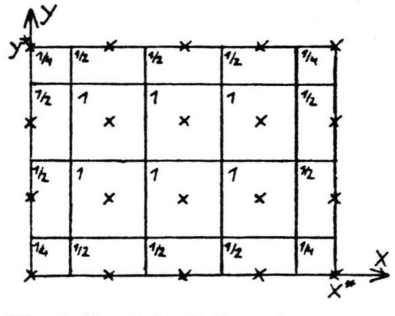

Fig. 2. Typ 2 der Zellenzerlegung.

in G aufhörend, also in der Reihenfolge

$$(\bar{r}, \bar{s}), (\bar{r}+1, \bar{s}), \ldots, (\bar{\bar{r}}, \bar{s}), (\bar{r}, \bar{s}+1),$$
$$(\bar{r}+1, \bar{s}+1), \ldots, (\bar{\bar{r}}, \bar{s}+1), \ldots (\bar{\bar{r}}-1, \bar{\bar{s}}), (\bar{\bar{r}}, \bar{\bar{s}}).$$

Zur Erzielung von *Konservativität* (K′) sollen alle Spaltensummen der Matrix CM verschwinden. Wir betrachten zunächst eine beliebige weit im Inneren liegende Zelle $Z_j = Z_{r,s}$ und dazu die Spalte mit dem Index j. In ihr stehen wegen (4.6) fünf Einträge, nämlich $-4a$ in der Zeile (r, s), und jeweils a in den Zeilen $(r, s-1)$, $(r-1, s)$, $(r+1, s)$, $(r, s+1)$. Wir haben wie gewünscht die Summe 0.

Sobald wir uns aber einer der vier Randstrecken des Rechtecks $G \cup \partial G$ nähern, fehlen Einträge, da dann nicht alle benötigten Fünfpunktsterne ganz im Rechteck liegen. Skizze 3 gibt einen Überblick, um welche Einträge es dabei geht. Exemplarisch betrachten wir die Spalten

$$j_1 = (\bar{r}, \bar{s}+2), \quad j_2 = (\bar{r}, \bar{s}+1), \quad j_3 = (\bar{r}, \bar{s}).$$

Für einige Zeilen der Matrix CM ist in der Skizze der jeweilige Doppelindex j angegeben, analog für die Spalten. Ein Kreischen „o“ in der Diagonale der Skizze bedeutet, daß im jeweiligen Diagonalelement von CM nur der zu $\partial^2 u/\partial x^2$ oder zu $\partial^2 u/\partial y^2$ gehörende Teil zu ergänzen ist, während ein „+“ angibt, daß das ganze Element zu bestimmen ist.

Es ist ansonsten üblich, sich durch Kombination der diskretisierten Randbedingung und der diskretisierten Differentialgleichung nach Elimination externer Werte u_j die fehlenden Einträge zu verschaffen. Tut man dies, so ist aber je nach Art der Diskretisierung die Konservativitätsbedingung eventuell nicht erfüllt. Nach Satz 3.2 können wir es uns leichter machen. Wir ergänzen einfach die fehlenden Einträge so, daß die entsprechenden Spaltensummen verschwinden.

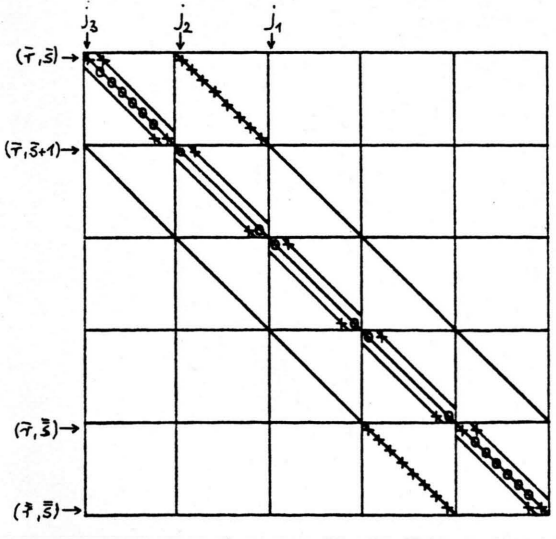

Fig. 3. Das Inzidenzschema der Matrix CM.

Wir beschreiben die Konstruktion exemplarisch für die Spalten j_1, j_2, j_3 und empfehlen dem Leser, einfachheitshalber zunächst nur an Zerlegungstyp 1 zu denken, da für diesen alle $c_j = 1$ sind. Die Spalte $j_1 = (\bar{r}, \bar{s}+2)$ hat vier Einträge, und zwar in den Zeilen

$$(\bar{r}, \bar{s}+1), \quad j_1 = (\bar{r}, \bar{s}+2), \quad (\bar{r}+1, \bar{s}+2),$$
$$(\bar{r}, \bar{s}+3).$$

Der Reihe nach ist dies a, $-2a+\alpha'$, a, a bei Typ 1 bzw. $a/2$, $-a+\alpha'$, a, $a/2$ bei Typ 2. Wir sehen, daß wir mit $\alpha' = -a$ am Ziel sind. Wir haben mit der Einführung der Größe α' bereits die zur Diskretisierung von $a(\partial^2 u/\partial x^2)$ gehörenden Anteile von denen getrennt, die zur Diskretisierung von $a(\partial^2 u/\partial y^2)$ gehören, da die zur y-Richtung gehörenden bereits die Summe 0 haben. Die Nützlichkeit dieser Trennung sehen wir bei der Spalte j_2.

In der Spalte $j_2 = (\bar{r}, \bar{s}+1)$ sind nämlich zwei Elemente zu ergänzen. Auf den ersten Blick scheint dies gar nicht eindeutig möglich zu sein, bei genauerem Hinsehen entdecken wir aber, daß im Diagonalelement der x-Anteil fehlt, während das andere Element zur Diskretisierung in y-Richtung gehört. In den Zeilen $(\bar{r}, \bar{s})$,

$$j_2 = (\bar{r}, \bar{s}+1), \quad (\bar{r}+1, \bar{s}+1), \quad (\bar{r}, \bar{s}+2)$$

haben wir der Reihe nach α_1, $-2a+\beta$, a, a bei Typ 1 bzw. α_2, $-a+\beta$, a, $a/2$ bei Typ 2. Wir ergänzen $\alpha_1 = a$ bei Typ 1, $\alpha_2 = a/2$ bei Typ 2 und $\beta = -a$ bei beiden Typen.

Schließlich betrachten wir die Spalte $j_3 = (\bar{r}, \bar{s})$, die zur linken unteren Zelle gehört. Hier ist das Diagonalelement vollständig zu ergänzen, da zwei Arme des Fünf-Punkte-Sterns aus dem Rechteck herausragen. Außer in der Zeile j_3 hat die Spalte j_3 noch Einträge in den Zeilen $(\bar{r}+1, \bar{s})$ und $(r, \bar{s}+1)$, in beiden a bei Typ 1, $a/2$ bei Typ 2. Das fehlende Diagonalelement an der Stelle (j_3, j_3) ist also $-2a$ bei Typ 1, $-a$ bei Typ 2.

Ganz analog ergänzt man alle anderen fehlenden Einträge der Matrix CM und bekommt so ein *konservatives Schema*.

Wir wollen nun noch eine Bedingung für *Nichtnegativitätserhaltung* (NNE′) des Schemas herleiten. Hinreichend für (NNE′) ist die Bedingung

$$Q \geqq 0, \quad I + \mu\bar{\theta} C M C^{-1} \Gamma^{-1} \geqq 0 \tag{4.7}$$

(man vgl. 3.11). Wegen $\gamma \equiv 1$ ist $\Gamma = I$, und als hinreichende Bedingung haben wir

$$I - \mu\theta C M C^{-1} \quad \text{ist } M\text{-Matrix und}$$
$$I + \mu\bar{\theta} C M C^{-1} \geqq 0. \tag{4.8}$$

Wir benutzen Satz 3.3. Alle Nichtdiagonalelemente der Matrix $I - \mu\theta C M C^{-1}$ sind 0 oder ein negatives Vielfaches von a, also $\leqq 0$, und genaue Inspektion der Matrix CM lehrt, daß sie symmetrisch ist. Also verschwinden auch alle ihre Zeilensummen, ebenso die von CMC^{-1}, und alle Zeilensummen der Matrix $I - \mu\theta CMC^{-1}$ haben den positiven Wert 1. Diese ist also eine M-Matrix.

Alle Nichtdiagonalelemente der Matrix

$$I + \mu\bar{\theta} C M C^{-1}$$

sind offensichtlich nichtnegativ. Ihre Diagonalelemente sind bei Typ 1 wegen $C = I$ für die inneren Zellen $1 - 4\mu\bar{\theta}a$, für die Eckzellen $1 - 2\mu\bar{\theta}a$, für die übrigen Randzellen $1 - 3\mu\bar{\theta}a$. Also ist

$$I + \mu\bar{\theta} C M C^{-1} \geqq 0,$$

wenn

$$4\mu\bar{\theta} a \leqq 1 \tag{4.9}$$

ist. Bei Typ 2 findet man, daß alle Diagonalelemente der Matrix CMC^{-1} den Wert $-4a$ haben, also ist (4.9) auch bei Typ 2 für Nichtnegativität der Matrix $I + \mu\bar{\theta} C M C^{-1}$ hinreichend.

Ergebnis: Unter der Bedingung (4.9) *hat das Schema die Eigenschaft* (NNE′) *der Nichtnegativitätserhaltung.*

Bemerkungen. (A) Besonders einfach zu behandeln ist der *Periodizitätsfall* des Diffusionsproblems (alles periodisch in x mit Periode x^*, in y mit Periode y^*). Dann hat man *Diffusion auf einem Torus*, somit *keine Randbedingungen*. Überall wird wie üblich mit dem Fünf-Punkte-Stern diskretisiert, die c_j haben alle den Wert 1, man braucht nicht zwischen Typ 1 und Typ 2 zu unterscheiden, die Matrizen sind dann allerdings *zirkulant*.

(B) Die Rechnung mit dem Schema (ein Zeitschritt nach dem anderen) kann man wahlweise mit (3.10) oder (3.7) durchführen, indem man bei jedem Zeitschritt jeweils den neuen Vektor $\tilde{e}_{n+1}$ oder $\tilde{u}_{n+1}$ ausrechnet.

(C) Das Schema hat die Genauigkeitsordnung $h^2 + \tau$ bei Typ 1 überall, bei Typ 2 außer in den Eckzellen, wo es die Ordnung $h + \tau$ hat. Im Sonderfall $\theta = 1/2$ kann man hier τ durch τ^2 ersetzen. Man findet dies durch (mühsame) Taylor-Entwicklung des Diskretisierungsfehlers. Numerisch scheint sich aber nach unseren Beispielrechnungen die geringere Konsistenzordnung außerhalb der Eckzellen nicht auszuwirken.

(D) Im Falle $\theta > 0$ (speziell im besonders schönen voll-impliziten Fall $\theta = 1$, bei dem (4.9) immer erfüllt ist) empfiehlt sich Auflösung des bei jedem Zeitschritt zu lösenden linearen Gleichungssystems nach dem in [4], Seite 58ff., beschriebenen Verfahren für block-tridiagonale Matrizen.

§ 5. Ein konservatives ADI-Verfahren für die Wärmeleitungsgleichung

Wir diskretisieren das Anfangsrandwertproblem [(4.1), (4.2), (4.3)] nach der von Peaceman und Rachford [14] dargestellten Methode der alternierenden Richtungen (Aufteilung des Zeitschritts τ in Halbschritte der Größe $\tau/2$, im ersten implizit in x-Richtung, im zweiten implizit in y-Richtung, so daß bei jedem Zeitschritt nur tridiagonale lineare Gleichungssysteme aufzulösen sind. Also

$$\begin{aligned} &\frac{1}{\tau/2}\,(u_{j,\,n+1/2} - u_{j,\,n}) \\ &\qquad = L_x u_{j,\,n+1/2} + L_y u_{j,\,n} + \frac{\bar{\varphi}_{j,\,n}}{v_j \tau/2}, \\ &\frac{1}{\tau/2}\,(u_{j,\,n+1} - u_{j,\,n}) \\ &\qquad = L_x u_{j,\,n+1/2} + L_y u_{j,\,n+1} + \frac{\bar{\varphi}_{j,\,n+1/2}}{v_j \tau/2} \end{aligned} \tag{5.1}$$

mit

$$\bar{\varphi}_{j,\,m} = \int_{t_m}^{t_{m+1/2}} \int_{Z_j} \int f(x, y, t)\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}t + \bar{\varrho}_{j,\,m}$$

mit

$$\bar{\varrho}_{j,\,m} = 0, \quad \text{falls} \quad \partial Z_j \cap \partial G = \text{leer ist},$$

andernfalls

$$\varrho_{j,\,m} = \int_{t_m}^{t_{m+1/2}} \int_{\partial Z_j \cap \partial G} \psi(x, y, t)\,\mathrm{d}\sigma\,\mathrm{d}t.$$

Hierbei ist $m=n$ oder $m=n+\frac{1}{2}$. Mit L_x bzw. L_y bezeichnen wir konsistente Diskretisierungen des Anteils des örtlichen Differentialoperators in x bzw. y-Richtung.

Für die Vektoren $\tilde{e}_m$ der Energieklumpen erhalten wir dann mit den zu § 3 analogen Bezeichnungen

$$\tilde{e}_{n+1/2}-\tilde{e}_n=\tfrac{1}{2}\mu CM_x C^{-1}\tilde{e}_{n+1/2}+\tfrac{1}{2}\mu CM_y C^{-1}\tilde{e}_n+\tilde{\tilde{\varphi}}_n,$$

$$\tilde{e}_{n+1}-\tilde{e}_{n+1/2}=\tfrac{1}{2}\mu CM_x C^{-1}\tilde{e}_{n+1/2}+\tfrac{1}{2}\mu CM_y C^{-1}\tilde{e}_{n+1}+\tilde{\tilde{\varphi}}_{n+1/2}.$$

Es gilt hierbei $M_x+M_y=M$. Durch Elimination von $\tilde{e}_{n+1/2}$ bekommen wir

$$\begin{aligned}&\left(I-\frac{\mu}{2}CM_x C^{-1}\right)\left(I-\frac{\mu}{2}CM_y C^{-1}\right)\tilde{e}_{n+1}\\&=\left(I+\frac{\mu}{2}CM_x C^{-1}\right)\left(I+\frac{\mu}{2}CM_y C^{-1}\right)\tilde{e}_n\\&\quad+\left(I+\frac{\mu}{2}CM_x C^{-1}\right)\tilde{\tilde{\varphi}}_n\\&\quad+\left(I-\frac{\mu}{2}CM_x C^{-1}\right)\tilde{\tilde{\varphi}}_{n+1/2}.\end{aligned}\tag{5.1'}$$

Ein noch etwas schnelleres Verfahren (nur einmalige Berechnung der Inhomogenität pro Zeitschritt) bekommen wir, indem wir rechts den Ausdruck

$$\left(I+\frac{\mu}{2}CM_x C^{-1}\right)\tilde{\tilde{\varphi}}_n+\left(I-\frac{\mu}{2}CM_x C^{-1}\right)\tilde{\tilde{\varphi}}_{n+1/2}$$

durch $\tilde{\tilde{\varphi}}_n+\tilde{\tilde{\varphi}}_{n+1/2}=\tilde{\varphi}_n$ ersetzen. Wir dürfen dies ohne Schaden für die Konsistenzordnung $h^2+\tau$ tun. Auflösung nach $\tilde{e}_{n+1}$ gibt jetzt ein Schema der Gestalt (3.3') mit

$$\begin{aligned}Q&=\left(I-\frac{\mu}{2}CM_y C^{-1}\right)^{-1}\cdot\left(I-\frac{\mu}{2}CM_x C^{-1}\right)^{-1},\\P&=Q\left(I+\frac{\mu}{2}CM_x C^{-1}\right)\cdot\left(I+\frac{\mu}{2}CM_y C^{-1}\right).\end{aligned}\tag{5.2}$$

Dieses Verfahren entspricht der Modifikation des ADI-Verfahrens nach Wirz [15], der allerdings Dirichletsche Randbedingungen behandelt.

Man kann nun analog zum Vorgehen in § 4 zeigen, daß die Matrizen

$$I-\frac{\mu}{2}CM_x C^{-1}\quad\text{und}\quad I-\frac{\mu}{2}CM_y C^{-1}$$

M-Matrizen sind, und analog zu den Überlegungen in § 3 folgt, daß die Bedingung

$$\eta CM_x=\mathrm{o}\quad\text{und}\quad\eta CM_y=\mathrm{o}\tag{5.3}$$

notwendig und hinreichend für *Konservativität* sowohl des Verfahrens (5.1') als auch des durch Zusammenziehung (in Analogie zu Wirz) der $\tilde{\tilde{\varphi}}$-Terme modifizierten Verfahrens ist. Als hinreichende Bedingung für die *Nichtnegativitätserhaltung* ergibt sich durch Inspektion der Matrizen

$$I+\frac{\mu}{2}CM_x C^{-1}\quad\text{und}\quad I+\frac{\mu}{2}CM_y C^{-1}$$

die Bedingung

$$\mu a\leqq 1.\tag{5.4}$$

Wir verzichten auf Ausführung der Beweise, vermerken lediglich, daß die Matrizen M_x und M_y durch Spaltung von M in x- und y-Anteil entstehen (man vergleiche hierzu [16]).

Bemerkung. Den allgemeineren Fall der Aufgabe [(2.1). (2.2), (2.3)] mit $p=2$ kann man analog behandeln, wenn der Ortsoperator keine gemischten Ableitungen enthält. An den entsprechenden Stellen der Schemata sind dann noch die Koeffizienten γ_j bzw. Matrizen Γ oder Γ^{-1} einzufügen.

§ 6. Numerische Testrechnungen

Zur Überprüfung und Illustration unserer Theorie haben wir umfangreiche Testrechnungen durchgeführt, die meisten allerdings für den allgemeineren Fall des § 2 mit $p=2$, aber ohne gemischte Ortsableitungen. Es ergab sich dabei bis auf kleine durch Rundung verursachte Fehler Erhaltung der Energie, außerdem Konvergenz in der Ordnung $h^2+\tau$ gegen die exakte Lösung u, außer in manchen Fällen bei Zerlegungstyp 2 in den Eckpunkten. Da wir die Einzelheiten der Konstruktion der Verfahren nur für die zweidimensionalen Wärmeleitungsgleichung dargestellt haben, soll hier nur über diese

ausführlicher berichtet werden, obwohl die Behandlung komplizierterer Probleme aussagekräftiger ist.

Bei jedem Zeitschritt τ ist im impliziten Fall $\theta > 0$ ein großes (schwach besetztes) lineares Gleichungssystem aufzulösen. Bei dem Verfahren des § 4 haben wir das mit dem Eliminationsalgorithmus für blocktridiagonale Matrizen ([4], Chapt. 2, Sect. 3.3) gemacht. Ist MN die Anzahl der Gitterpunkte in $[0, x^*] \times [0, y^*]$ (mit M für die x-Richtung, N für die y-Richtung), so braucht man $M^2(N-1)+2MN+3M$ Speicherplätze für die Links-Rechts-Zerlegung der Matrix $Q^{-1} = I - \mu\theta C M C^{-1} \Gamma^{-1}$ (siehe (3.11), in unserem speziellen Fall ist $\Gamma = \Gamma^{-1} = I$), was ziemliche Speicherprobleme aufwirft. Dies begründet teilweise den Vorteil der speicherplatzsparenden ADI-Verfahren des § 5, die außerdem pro Zeitschritt erheblich weniger Rechenzeit erfordern; für die entsprechende Matrix Q^{-1} braucht man dann nur $2 \cdot 5MN$ Speicherplätze. Die besten Eigenschaften bezüglich Nichtnegativitätserhaltung hat allerdings das voll-implizite Verfahren ($\theta = 1$) des § 4, da es keine Zeitschrittbeschränkung benötigt. Hingegen benötigt das ADI-Verfahren hierfür leider die Beschränkung $\tau = \mu h^2 \leqq h^2/a$ wegen (5.4), so daß der Vorteil der kürzeren Rechenzeit pro Zeitschritt nicht voll zur Geltung kommt. In der Praxis wendet man oft das ADI-Verfahren ohne die Beschränkung (5.4) an, im quadratischen Mittel hat man dann bekanntlich immer noch Konvergenz, bei rauhen Daten treten allerdings ziemlich große Ungenauigkeiten ein, da die Glättungseigenschaften der Wärmeleitung dann diskret nicht gut genug imitiert werden.

Wir haben einige Rechnungen durchgeführt auf einer Maschine CDC 170-835 des Rechenzentrums der FU Berlin, deren Wortlänge etwa 13 signifikanten Dezimalen entspricht, haben dabei aber nicht jede Anstrengung zur Minimierung der Rechenzeit unternommen, sondern legten mehr Wert auf Nachprüfung der Erhaltungseigenschaft. Wir nahmen das Problem [(4.1), (4.2), (4.3)] mit

$$a = 2, \quad f = \psi \equiv 0,$$
$$g(x, y) = e^{2x-y}/(\cosh 2 - 1)$$

im Rechteck $[0, 1] \times [0, 2]$.

Wir diskretisierten mit $h = 1/10$ und für $\theta = 1/2$ und $\theta = 1$ mit $\mu = 1/4$, für $\theta = 0$ mit $\mu = 1/8$, für das analog zu Wirz [17] in § 5 beschriebene ADI-Verfahren mit $\mu = 1/4$. Das führt auf jeweils 200 Gitterpunkte für Typ 1, auf 231 Gitterpunkte für Typ 2 der Zerlegung.

Die Wahl $f = \psi \equiv 0$ bedeutet „reflektierende Wände", für alle $t > 0$ ist $E(t) = E(0) =$ konstant, und wir haben die Anfangsfunktion g gerade so gewählt, daß $E(0) = 1$ ist. In die numerische Rechnung haben wir die Werte g als Integralmittelwerte über die einzelnen Zellen Z_j eingebracht, also gemäß (3.2) mit $\gamma = 1$.

Bei $t = 2$ ergaben sich für den Energiefehler

$$\delta := |\hat{E}_{2/\tau} - E(2)|$$

(wir schreiben $\hat{E}_{2/\tau}$ für den rundungsgestörten Wert, den die Rechenmaschine anstelle von $E_{2/\tau}$ ausgibt). Werte zwischen $9.45 \cdot 10^{-13}$ und $1.1 \cdot 10^{-11}$. Daß sie durch das Rundungsrauschen entstehen, sieht man auch an ihrer Kleinheit im Vergleich zum Abbruchfehler der Verfahren, der die Größenordnung $h^2 + \tau \sim h^2 = 10^{-2}$ hat. Die berechneten Näherungswerte für u waren bei $t = 2$ bei fast allen Rechnungen (auf 4 Dezimalen genau) auf konstant 0.5 abgeklungen.

[1] Y. Sasaki, J. Comput. Phys. **21**, 270 (1976).
[2] E. Isaacson, Advances in Computer Methods for Partial Differential Equations II (R. Vichnevetsky, ed.), 1977, pp. 251–255.
[3] I. M. Navon, Monthly Weather Rev. **109**, 16 (1981).
[4] E. Isaacson and H. B. Keller, Analysis of Numerical Methods, Wiley & Sons, New York 1966.
[5] R. Gorenflo und S. Kiesner, ISNM **58**, 73 (1982).
[6] R. Gorenflo und M. Niedack, Computing **25**, 299 (1980).
[7] R. Gorenflo, Numer. Math. **14**, 448 (1970).
[8] G. J. Pert, J. Comput. Phys. **39**, 251 (1981).
[9] G. J. Pert, J. Comput. Phys. **42**, 20 (1981).
[10] K. Glashoff und H. Kreth, Numer. Math. **35**, 343 (1980).
[11] W. Walter, Differential- und Integralungleichungen, Springer-Verlag, Berlin 1964.
[12] R. S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, New Jersey 1962.
[13] L. Collatz, Funktionalanalysis und numerische Mathematik. Springer-Verlag, Berlin 1964.
[14] D. W. Peaceman and H. H. Rachford, Jr., J. Soc. Ind. and Appl. Math. **3**, 28 (1955).
[15] H. J. Wirz, Z. Ang. Math. Mech. **52**, 329 (1972).
[16] N. N. Janenko, Die Zwischenschrittmethode zur Lösung mehrdimensionaler Probleme der mathematischen Physik, Lecture Notes in Mathematics **91**, Springer-Verlag, Berlin 1969.